Journal of Clinical Immunology (2023) 43:286–298
https://doi.org/10.1007/s10875-022-01386-3

ORIGINAL ARTICLE

Combination of WFDC2, CHI3L1, and KRT19 in Plasma Defines a Clinically Useful Molecular Phenotype Associated with Prognosis in Critically Ill COVID-19 Patients

Takeshi Ebihara[1] · Tsunehiro Matsubara[1] · Yuki Togami[1] · Hisatake Matsumoto[1] · Jotaro Tachino[1] · Hiroshi Matsuura[1,2] · Takashi Kojima[3] · Fuminori Sugihara[4] · Shigeto Seno[5] · Daisuke Okuzaki[6] · Haruhiko Hirata[7] · Hiroshi Ogura[1]

Received: 7 June 2022 / Accepted: 17 October 2022 / Published online: 4 November 2022

Abstract

Background COVID-19 is now a common disease, but its pathogenesis remains unknown. Blood circulating proteins reflect host defenses against COVID-19. We investigated whether evaluation of longitudinal blood proteomics for COVID-19 and merging with clinical information would allow elucidation of its pathogenesis and develop a useful clinical phenotype.

Methods To achieve the first goal (determining key proteins), we derived plasma proteins related to disease severity by using a first discovery cohort. We then assessed the association of the derived proteins with clinical outcome in a second discovery cohort. Finally, the candidates were validated by enzyme-linked immunosorbent assay in a validation cohort to determine key proteins. For the second goal (understanding the associations of the clinical phenotypes with 28-day mortality and clinical outcome), we assessed the associations between clinical phenotypes derived by latent cluster analysis with the key proteins and 28-day mortality and clinical outcome.

Results We identified four key proteins (WFDC2, GDF15, CHI3L1, and KRT19) involved in critical pathogenesis from the three different cohorts. These key proteins were related to the function of cell adhesion and not immune response. Considering the multicollinearity, three clinical phenotypes based on WFDC2, CHI3L1, and KRT19 were identified that were associated with mortality and clinical outcome.

Conclusion The use of these easily measured key proteins offered new insight into the pathogenesis of COVID-19 and could be useful in a potential clinical application.

Keywords Biomarker · Cluster analysis · Network analysis · Plasma proteomics

Takeshi Ebihara, Tsunehiro Matsubara, and Yuki Togami contributed equally to this work.

The first author was determined alphabetically.

✉ Hisatake Matsumoto
h-matsumoto@hp-emerg.med.osaka-u.ac.jp

[1] Department of Traumatology and Acute Critical Medicine, Osaka University Graduate School of Medicine, Suita, Osaka, Japan

[2] Osaka Prefectural Nakakawachi Emergency and Critical Care Center, Higashiosaka, Osaka, Japan

[3] Laboratory for Clinical Investigation, Osaka University Hospital, Suita, Osaka, Japan

[4] Core Instrumentation Facility, Immunology Frontier Research Center and Research Institute for Microbial Diseases, Osaka University, Osaka, Japan

[5] Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

[6] Genome Information Research Center, Research Institute for Microbial Diseases, Osaka University, Osaka, Japan

[7] Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine, Suita, Osaka, Japan

Background

Although the majority of individuals infected by COVID-19 exhibit no or mild-to-moderate symptoms, approximately 5–20% of subjects hospitalized required prolong treatment in an intensive care unit (ICU) with invasive mechanical ventilation (IMV) [1–4]. In these cases, excessive inflammation following a COVID-19 infection might lead to systemic inflammatory response syndrome and multiple organ failure [5].

From a clinical perspective, it is important to evaluate the blood circulating proteins that can reflect the systemic inflammation. The evaluation process is easy and rapid, and most of the biomarkers used in the ICU are based on the circulating proteins [6]. Previously, we showed that a key network of cytokine proteins based on a blood circulating cytokine profile and combined key cytokines score was related to prognosis and severity in critically ill patients including those with sepsis and burn [7–9]. Therefore, we hypothesized that a key protein network would also play an important role in critical COVID-19.

Recently, researchers have investigated new therapeutic targets by combining unsupervised clustering analysis and key biological indicators in various diseases to clarify potential sub-phenotypes [10–12]. Using key proteins to identify COVID-19 patient sub-phenotypes with poor outcomes may enable the discovery of new therapeutic strategies and target populations.

The present study approach involved several datasets and a statistical approach. To obtain globally versatile results, we used two discovery cohorts (i.e., American and Japanese cohorts) that had different patient characteristics (e.g., age, sex, and race). To achieve the primary goal, we derived plasma proteins related to COVID-19 pathogenesis from an innovative method, Olink proteomics, by using the two different cohorts. The candidates were validated by a classical method, enzyme-linked immunosorbent assay (ELISA), in a validation cohort to determine key proteins. To achieve the secondary goal, we derived the clinical phenotypes based on these key proteins and assessed their associations with mortality and clinical outcome.

Methods

Cohort Data and Measurement of Plasma Proteins

We used data from three observational cohorts. The first discovery cohort comprised publicly available data provided by the Massachusetts General Hospital Emergency Department COVID-19 Cohort [13] with Olink Proteomics (Olink® Explore 1536; https://www.olink.com/mgh-covid-study/) (Supplemental Methods, Statistical analysis, and Results), which was conducted from March 2020 to April 2020. In this study, we used proteomics data of days 1 and 8.

The second discovery cohort was composed of COVID-19 patients who were admitted to Osaka University Hospital from July 2020 to February 2021. Blood samples were obtained from the patients on days 1 (day of admission) or 2 and days 6–8 and once from healthy volunteers who were enrolled via public poster advertisements. Plasma proteomics were performed by using Olink® Explore 1536.

The validation cohort was composed of COVID-19 patients admitted to Osaka University Hospital or Osaka Prefectural Nakakawachi Emergency and Critical Care Center from December 2020 to January 2021 and April 2021, who were treated with IMV. Blood samples were obtained from the patients until hospital discharge or death on day 1 (day of admission) and days 6–8 and once from healthy volunteers who were enrolled via public poster advertisements. The plasma proteins were measured by ELISA. Details of the discovery and validation cohorts are shown in the Supplemental Methods, Statistical analysis, and Results.

Definition of Disease Severity: Critical and Non-critical

Acuity scores were based on the World Health Organization ordinal outcomes scale [14]: A1, dead; A2, intubated, survived; A3, hospitalized with oxygen; A4, hospitalized without oxygen; A5, discharged. Disease severity was classified according to the maximum acuity score (acuity max 1 or 2). We defined "critical" as acuity max 1 and 2 subjects and "non-critical" as acuity max 3, 4, and 5 subjects.

Definition of Timing of Sample Collection: Phase 1 and Phase 2

For easy clinical application, day 1 referred to the day of visiting the emergency department or of admission to the hospitals in this study, not to the day of disease onset or testing as PCR positive. We defined two different types of measurement timing: phase 1, days 1–2, and phase 2, days 6–10.

Definitions of Clinical Outcome: Early Recovery and Late Recovery

We defined the clinical outcome of patients who were treated with IMV for ≤ 12 days or not treated with IMV as early recovery and IMV > 12 days or 28-day non-survivors as late recovery as in our previous study [15]. We divided the COVID-19 patients of the second derivation cohort and the

validation cohort into two groups based on early recovery and late recovery and assessed them.

Statistical Analysis

In the first discovery cohort, patients were divided into the critical group and non-critical group, and differences in the expression of 1463 plasma proteins were evaluated. Instead of normalization to the total protein concentration, Olink proteomics data was normalized using three internal and three external controls that were used for quality control and data normalization. The proteins levels were expressed as values of normalized protein expression (NPX), which was an arbitrary unit on a log2 scale [16]. The difference in NPX was used to detect the difference of protein expressions as previously described [17]. Differential expression analysis was conducted using the Welch 2-sample *t*-test. The false discovery rate was calculated by the Benjamin-Hochberg method [18]. Proteins with false discovery rate < 0.01 and |NPX difference|> 1 were considered to be significantly expressed. The plasma proteins reaching significance in both phase 1 and phase 2 were extracted as candidates of the first discovery cohort. The phase 1 NPX values of the candidates were compared among acuity scores with the Kruskal–Wallis test.

In the second discovery cohort, the candidates from the first discovery cohort were evaluated. The patients were divided into those with early recovery and late recovery. The Dunnet test was used to evaluate the levels of each candidate of the first discovery cohort between the healthy volunteers and COVID-19 patients in each phase. The Welch 2-sample *t*-test was used to evaluate the differences in the levels of candidates of the first discovery cohort between the two groups in each phase. The correlation analysis between the number of days since onset and the candidates of the first discovery cohort was performed by Spearman's rank correlation. The trends of the two groups (early recovery and late recovery) are shown by linear regressions (solid lines) with 95% confidence intervals (gray areas).

We also evaluated the protein co-expression network only for critical COVID-19 patients in the second discovery cohort. Protein co-expression network analysis was performed with the R package "WGCNA" (weighted gene co-expression network analysis) as previously described [19] using the day 1 data of patients with acuity max scores 1 and 2. The module network is displayed graphically using Cytoscape® software (www.cytoscape.org) version 3.8.0 [20]. The biological functions of the proteins in each module were investigated by performing GO (Gene Ontology) pathway analysis [21] and KEGG (Kyoto Encyclopedia of Gene and Genomes) pathway analysis [22]. The details of WGCNA are shown in the Supplemental Methods, Statistical analysis, and Results.

In the validation cohort, the candidates from the second discovery cohort were validated. The levels of the candidates were transformed to common logarithmic values to normalize data distributions before analysis. The Dunnet test was used to evaluate the levels of each candidate between the healthy volunteers and COVID-19 patients in each phase. Then, the patients were divided into two groups, those with early recovery and late recovery. The candidates of the second discovery cohort were compared by Wilcoxon rank sum test between the two groups in each phase. The plasma proteins that were significantly increased (*P* values < 0.05) in the patients with late recovery compared with those in the patients with early recovery in both phases were extracted as key proteins. The association of these key proteins with 28-day and hospital mortality was also evaluated. The levels of key proteins in both phases were compared between the 28-day survivors and 28-day non-survivors or between hospital survivors and non-survivors by Wilcoxon rank sum test. The associations between phase 1 key proteins and body mass index (BMI), age, and comorbidities were also analyzed by Wilcoxon rank sum test or correlation analysis using Spearman's rank correlation. We also evaluated these key proteins in the patients who were not treated with IMV. Details of the characteristics of these patients and the method are described in the Supplemental Methods.

Latent class analysis (LCA) was performed using a combination of key proteins to identify the new clinical phenotypes. Phase 1 key proteins were transformed to common logarithmic values and scaled to become candidate variables for LCA. Because the high correlation of variables in LCA caused lower accuracy of model fit statistics with an overestimation of the true number of classes, the correlation matrix of the four key proteins was evaluated. One in any pair in which a strong correlation (Pearson correlation coefficient > 0.6) was observed was eliminated [23]. The optimal number of phenotypes was identified by evaluating the Bayesian information criterion (BIC), the appropriate size of each phenotype, and the misclassification rate of each phenotype [24, 25]. The optimal number of phenotypes was selected based on the largest BIC, considering the misclassification rate and interpretability [11]. The latent class analysis calculation was performed using the VarSelLCM package in R, in which the largest BIC is interpreted as optimal. Cumulative mortality is illustrated using Kaplan–Meier curves, and the phenotypes were compared by the log rank test.

A two-sided $P < 0.05$ was considered statistically significant. For all statistical analyses, a fully scripted data management pathway was created within the R environment for statistical computing, version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria). Categorical variables are reported as number and percentages, and significance was detected by χ^2 or

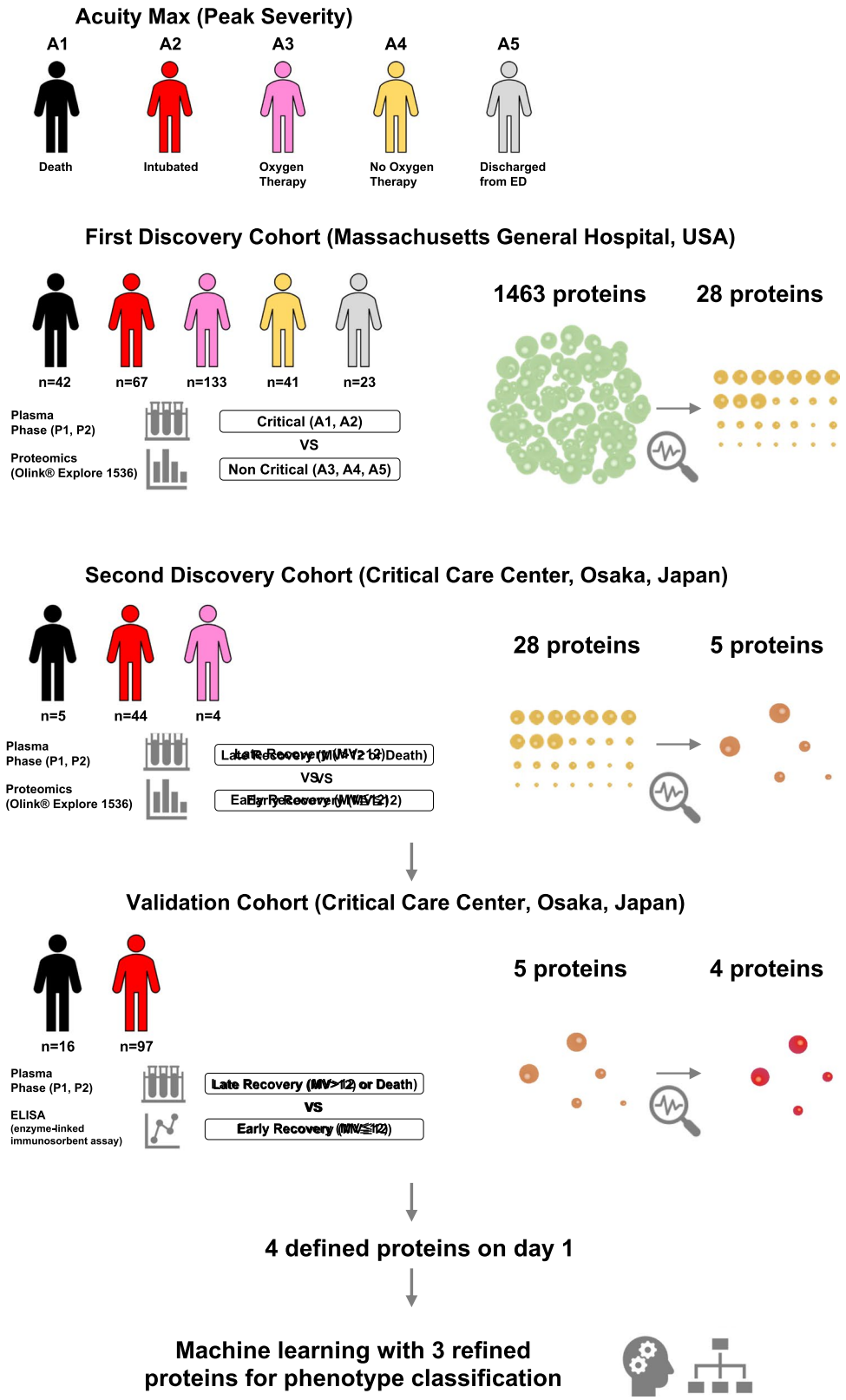


Fig. 1 Flow chart of participants. COVID-19 phenotypes were evaluated using three plasma proteins including WFDC2, CHI3L1, and KRT19. P1: phase 1 (days 1–2), P2: phase 2 (days 6–10)

Fisher's exact test. The continuous variables are described using mean and standard error or compared using the Mann–Whitney *U* test or Kruskal–Wallis rank sum test described using median and interquartile range (IQR) values. There were no missing data on plasma proteomics in the first and second discovery cohorts or on plasma proteins levels in the validation cohort. However, 5 of 113 patients (4%) in the validation cohort had missing values for BMI, but no imputation was made for this missing data.

Table 1 Clinical and demographic characteristics of COVID-19 patients in derivation and validation cohorts

	First discovery cohort, MGH, USA	Second discovery cohort, Osaka, Japan	Validation cohort, Osaka, Japan	*P* value
	(*n* = 306)	(*n* = 53)	(*n* = 113)	
Male sex, n (%)	162 (52.9)	37 (69.8)	80 (70.8)	0.001
Age, years, median (IQR)	58 (45–75)	73 (62–78)	65 (55–74)	
Age group, *n* (%)				0.001
20–34 years	32 (10.5)	0 (0)	1 (0.8)	
35–49 years	66 (21.6)	3 (5.7)	11 (9.7)	
50–64 years	89 (29.1)	13 (26.4)	44 (38.9)	
65–79 years	65 (21.1)	28 (49.1)	47 (41.6)	
Over 80 years	54 (17.6)	10 (18.9)	10 (8.9)	
Comorbidities, *n* (%)				
Heart disease	48 (16.5)	4 (7.7)	12 (10.6)	0.17
Lung disease	66 (21.1)	10 (19.2)	12 (10.6)	0.03
Kidney disease	41 (14.3)	8 (15.4)	12 (10.6)	0.64
Immunocompromised condition	25 (7.7)	5 (9.6)	4 (3.5)	0.21
Hypertension	146 (48.1)	24 (46.2)	47 (41.6)	0.53
Diabetes	111 (35.4)	25 (48.1)	41 (36.3)	0.25
BMI, kg/m^2, median (IQR)	29 (26–34)	23 (22–26)	25 (22–28)	
BMI, *n* (%)				< 0.001
0–24.9 kg/m^2	46 (15.1)	35 (66)	49 (43.4)	
25.0–39.9 kg/m^2	205 (66.9)	16 (30.2)	58 (51.3)	
≥ 40 kg/m^2	35 (11.4)	0 (0)	1 (0.8)	
Unknown	20 (6.5)	2 (3.8)	5 (4.4)	
Acuity max score, *n* (%)				
1, 28-day mortality	42 (13.7)	5 (9.6)	16 (14.2)	
2, intubated/Ventilated	67 (21.9)	44 (83)	97 (85.8)	
3, hospitalized, O_2 required	133 (43.5)	4 (7.5)	0 (0)	
4, hospitalized, no O_2 required	41 (13.4)	0 (0)	0 (0)	
5, discharged/not hospitalized	23 (7.5)	0 (0)	0 (0)	
Ratio of SARS-CoV-2 alpha variant, %	0 (0)	28 (52.8)	57 (50.4)	
SOFA score, median (IQR)	2 (1–7)	5 (3–6)	5 (3–6)	
Outcome				
28-day mortality, *n* (%)	42 (13.7)	1 (1.9)	12 (10.7)	
Hospital mortality, *n* (%)	42 (13.7)	5 (9.6)	16 (14.2)	0.69

Data are reported as number (percentage) or median (IQR, interquartile range) as appropriate *P* value: for the comparison between each cohort

BMI body mass index, *Heart disease* coronary artery disease, congestive heart failure, valvular disease, *Lung disease* asthma, COPD, requiring home O_2 and any chronic lung condition, *Kidney disease* chronic kidney disease, baseline creatinine > 1.5, *Immunocompromised condition* active cancer, chemotherapy, transplant and immunosuppressant agents, asplenic, *SOFA* Sequential Organ Failure Assessment

Results

The study approach involved several datasets and the statistical approach shown in Fig. 1.

Exploration of Candidate Plasma Proteins from First and Second Discovery Cohorts

Patient characteristics of the two discovery cohorts are shown in Table 1. In the publicly available first discovery cohort [13], one of the 306 COVID-19 patients was flagged as an outlier and removed from the final dataset, thus leaving 305 phase 1 samples and 139 phase 2 samples. The cohort comprised 109 critical patients and 197 non-critical patients. The distribution of patients by age group was statistically different between the critical and non-critical patients. Other details of the patients' characteristics are shown in Suppl. Table S1. Plasma proteins showing statistically significant changes in expression are indicated in red in the volcano plots for each phase (Fig. 2A, B), and those that showed statistically significant changes in expression in both phases 1 and 2 are indicated in Fig. 2C as candidates

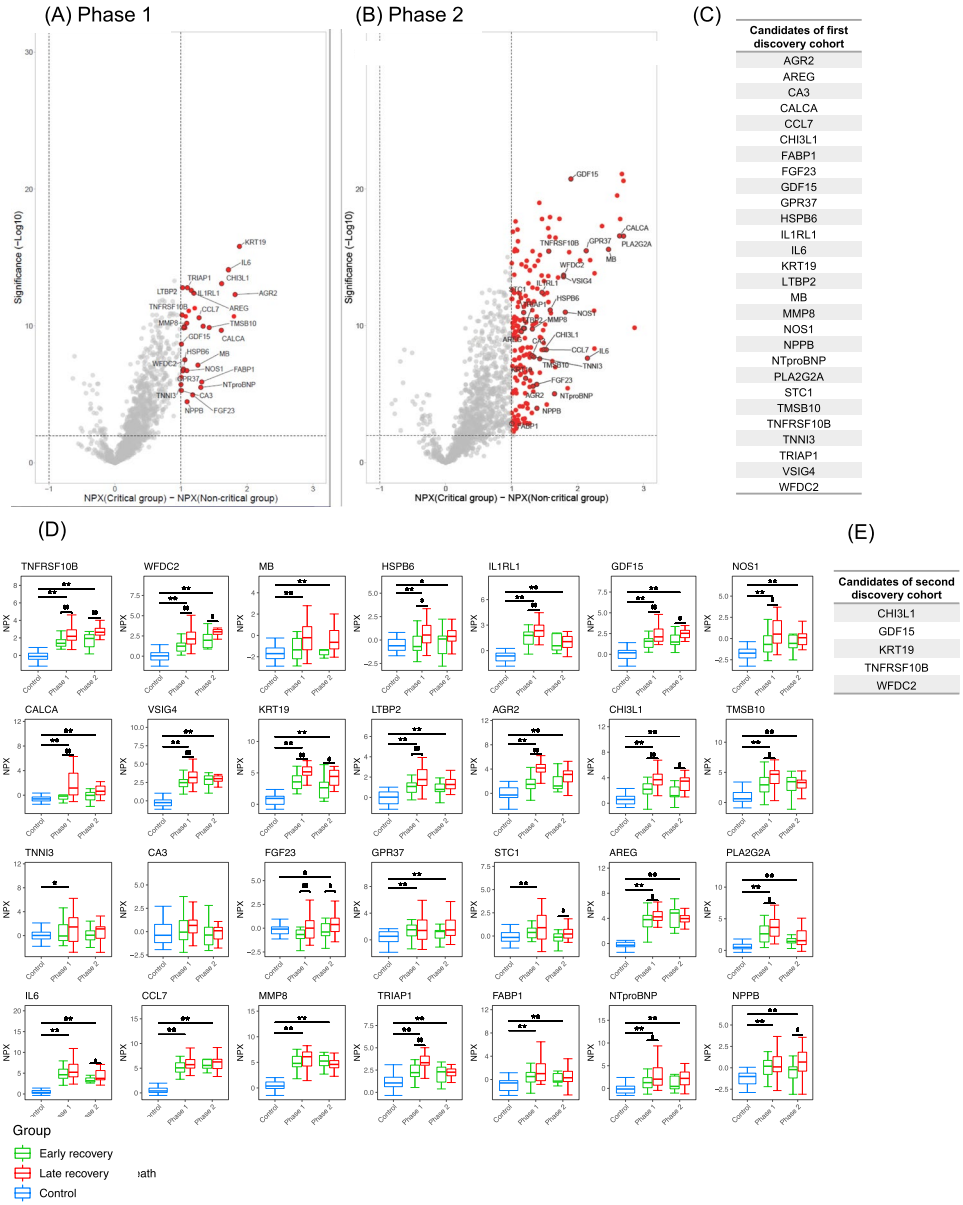


Fig. 2 Derivation of candidates' proteins from the first and second discovery cohorts. Volcano plots show the differentially expressed plasma proteins between critical (acuity max score: A1 or A2) and non-critical (A3, A4, or A5) groups in **A** phase 1 and **B** phase 2 in the first discovery cohort. The *X* axis represents the difference in normalized protein expression (NPX) between the critical group and non-critical group, and the *Y* axis represents log10 significance (adjusted *P* values). Significantly differentially expressed proteins were defined as proteins with adjusted *P* values < 0.01, |difference| > 1. **C** The 28 proteins labeled by gene names that were significantly increased in both phase 1 and phase 2 are shown as candidates of the first discovery cohort. **D** The 28 candidates of the first discovery cohort were evaluated in the second discovery cohort. The levels of the 28 candidates of the first discovery cohort were compared between healthy and COVID-19 individuals in phase 1 and phase 2. The difference between healthy volunteers and COVID-19 patients was measured by Dunnet test (*$P < 0.05$; **$P < 0.01$). The COVID-19 individuals were further classified into two groups, "early recovery" and "late recovery" in phase 1 and phase 2. The NPX values are plotted on the *Y* axes. In all box plots, the boxes show median, upper, and lower quartiles, and the whiskers show 5th to 95th percentiles. The difference between two groups was measured by Wilcoxon rank sum test ($^{\$}P < 0.05$; $^{\$\$}P < 0.01$). **E** The five candidates of the second discovery cohort are listed

of the first discovery cohort and are labeled in the volcano plots. We derived 28 plasma protein candidates from the first discovery cohort. Phase 1 NPX values of the 28 proteins were associated with acuity max (Suppl. Fig. S1).

These 28 candidates were then evaluated in the second discovery cohort that included 53 COVID-19 patients and 20 healthy volunteers (Suppl. Table S2), among whom 49 COVID-19 patients were critical and 4 were treated

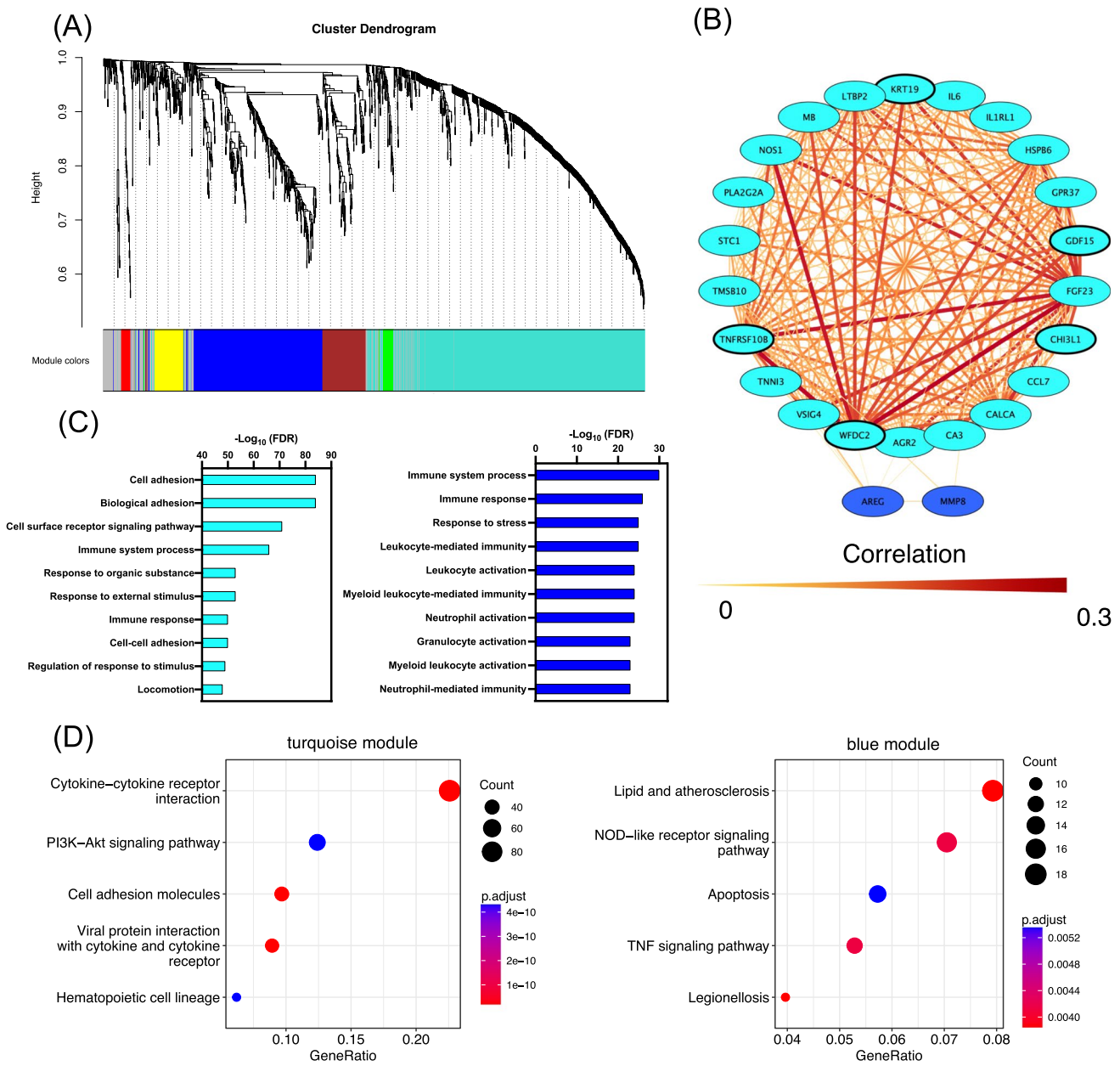
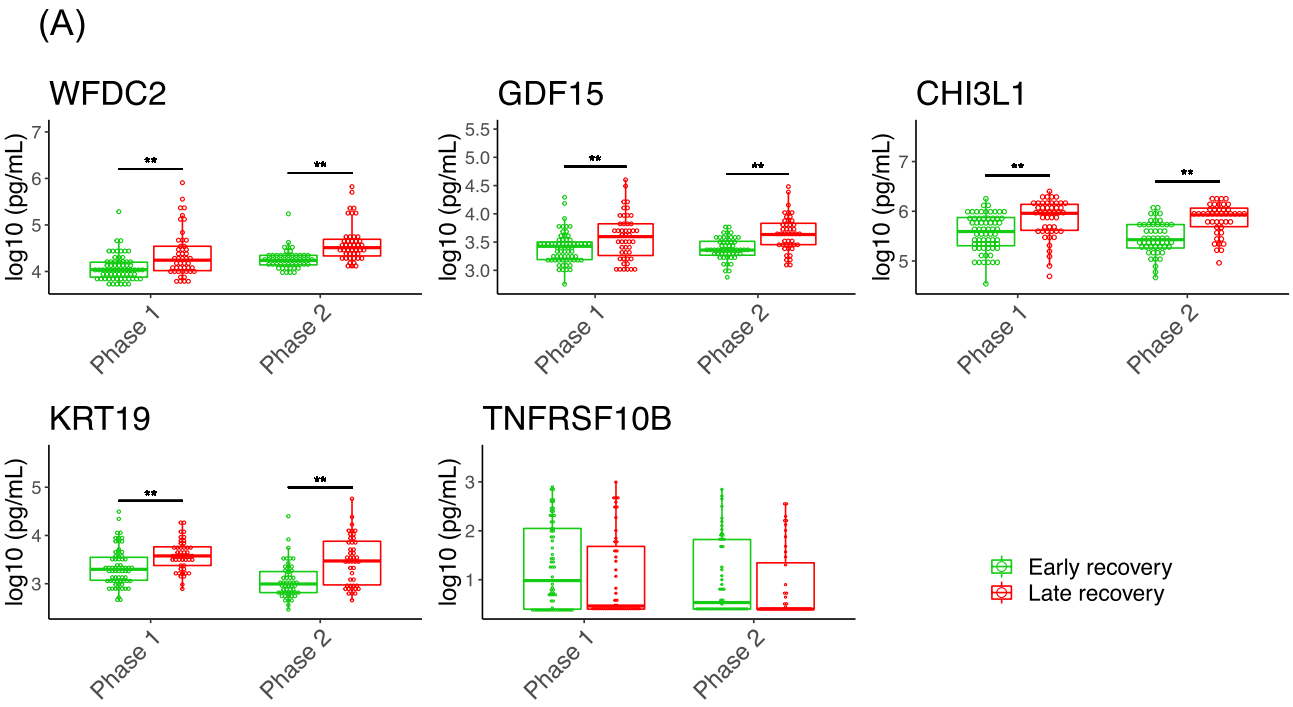
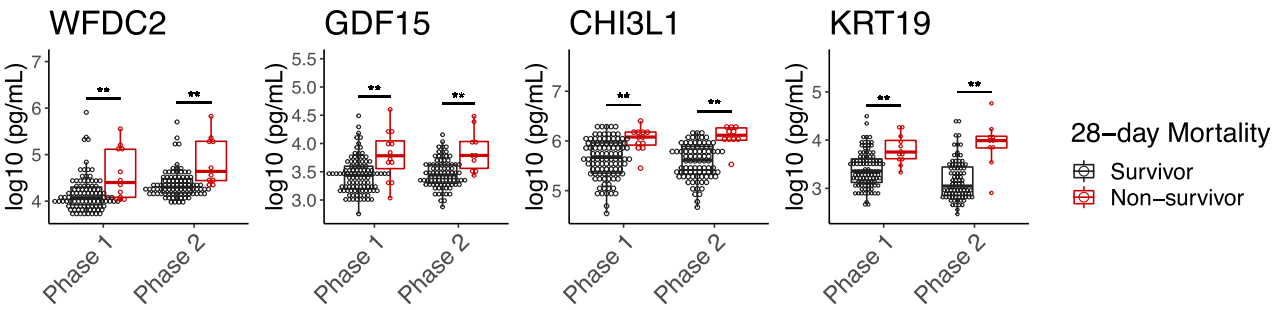
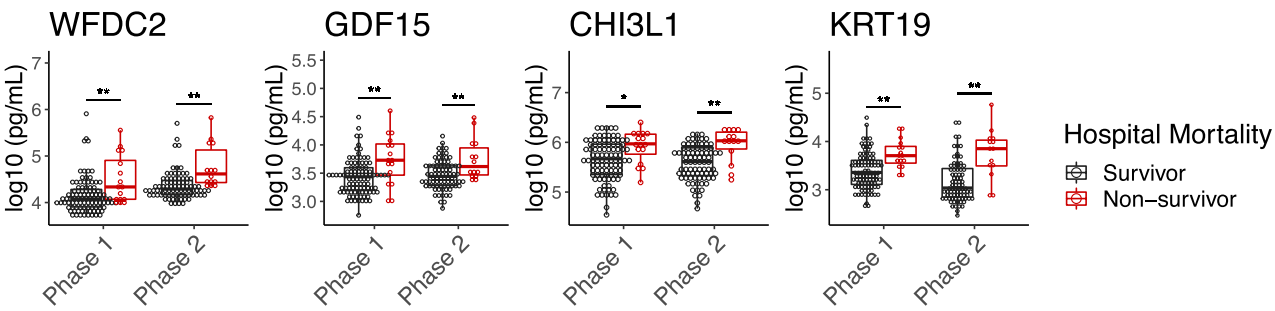


Fig. 3 Protein co-expression network in phase 1 of the second discovery cohort. **A** The hierarchical cluster tree of all proteins in the proteomic dataset on the basis of topological overlap. Modules correspond to branches of the tree. The branches and module proteins are colored, and gray indicates proteins outside the appropriate module, as can be seen from the color bands at the bottom of the tree. **B** Network depiction of protein co-expression modules of the 24 candidates of the first discovery cohort. Nodes represent proteins, and edges (lines) indicate connections between the nodes. The color of the nodes corresponds to the module, and the width and color of the edges correspond to the weight of the connected nodes. Bolded nodes indicate the five candidates of the second discovery cohort. **C** Gene ontology enrichment analysis of differentially expressed proteins of the turquoise and blue modules in **A**; the top 10 ontologies for each module are shown. Significantly enriched gene ontology terms are shown with Benjamin-Hochberg false discovery rate–corrected *P* values. **D** KEGG pathway analysis of differentially expressed proteins of the turquoise and blue modules in **A**

in the ICU without intubation (Table 1). The numbers of blood samples obtained in phases 1 and 2 were 49 and 34, respectively. We classified the patients into the early recovery group ($n = 23$) and late recovery group ($n = 30$). The age of the late recovery group was statistically higher than that of the early recovery group. The number of days since onset was not different between the two groups (Suppl. Table S3). The common plasma proteins that were higher in the COVID-19 patients than controls and that were higher in the late recovery group than the early


(A)
WFDC2
GDF15
CHI3L1
KRT19
TNFRSF10B
log10 (pg/mL)
Phase 1
Phase 2
**
Early recovery
Late recovery


(B)
WFDC2
GDF15
CHI3L1
KRT19
log10 (pg/mL)
Phase 1
Phase 2
**
28–day Mortality
Survivor
Non–survivor


(C)
WFDC2
GDF15
CHI3L1
KRT19
log10 (pg/mL)
Phase 1
Phase 2
*
**
Hospital Mortality
Survivor
Non–survivor

◂**Fig. 4** Discovery of four key proteins related to mortality and clinical outcome from the validation cohort. **A** The levels of WFDC2, GDF15, CHI3L1, KRT19, and TNFRSF10B associated with early recovery and late recovery or death in each phase in the COVID-19 groups. **B** The WFDC2, GDF15, CHI3L1, and KRT19 levels for 28-day survivors and non-survivors on each day in the COVID-19 groups. **C** The WFDC2, GDF15, CHI3L1, and KRT19 levels for hospital survivors and non-survivors on each day in the COVID-19 groups. The protein levels were transformed to common logarithmic values to normalize the data distribution. In all box plots, the boxes show median, upper, and lower quartiles, and the whiskers show 5th to 95th percentiles. Asterisks indicate a statistically significant difference (*$P < 0.05$, **$P < 0.01$) between two groups on each day by Wilcoxon rank sum test

recovery group for phases 1 and 2 were WFDC2, CHI3L1, GDF15, KRT19, and TNFRSF10B (Fig. 2D). Expression of these five proteins was higher than that in the control soon after onset, and a correlation between protein expression and the number of days since onset was not clear (Suppl. Fig. S2A). These five proteins in late recovery patients tended to remain high (Suppl. Fig. S2B). We derived these five proteins as candidates of the second discovery cohort (Fig. 2E).

Network Analysis of 1463 Plasma Proteins in Critical COVID-19 Patients in the Second Discovery Cohort

In total, six modules were identified (Fig. 3A). Twenty-five of the 28 candidates of the first discovery cohort were included in the turquoise module, as were all five candidates of the second discovery cohort. The 28 candidates of the first discovery cohort were reconstructed and visualized using cytoscape [18] (Fig. 3B). The top 10 GO results for the turquoise and blue modules are shown in Fig. 3C, and the top 5 KEGG results for the turquoise and blue modules are shown in Fig. 3D. The turquoise module is highly related to cell adhesion and biological adhesion. In this analysis, the five candidates of the second discovery cohort were associated with each other, and all had a function involving cell adhesion. The details are shown in the Supplemental Methods, Statistical analysis, and Results (Suppl. Fig. S3). The KEGG pathway of cell adhesion molecules is shown in Suppl. Fig. S4.

Validation of Five Candidate Plasma Proteins by ELISA

We assessed the candidate proteins in the validation cohort by ELISA. The validation cohort comprised 113 critical COVID-19 patients including 12 28-day non-survivors, and 16 healthy volunteers. The difference in age between the COVID-19 patients and healthy volunteers was not statistically significant. (Suppl. Table S4). The late recovery group was characterized by older patients, higher D-dimer, creatinine and LDH levels, and lower P/F ratio than those of the early recovery group. The number of days since onset was not different between the two groups (Suppl. Table S5). The numbers of blood samples collected in phases 1 and 2 were 113 and 110, respectively. The levels of WFDC2, GDF15, CHI3L1, and KRT19 were statistically significantly higher in both the COVID-19 patients compared to the controls (Suppl. Fig. S5) and in the late recovery group compared to the early recovery group (Fig. 4A) in both phases. The higher levels of WFDC2, GDF15, CHI3L1, and KRT19 were more frequently observed in the 28-day or hospital non-survivors than in the 28-day or hospital survivors, respectively, in both phases (Fig. 4B, C). There were no relationships between WFDC2, GDF15, CHI3L1, and KRT19, and sex and comorbidities for the control (Suppl. Fig. S6A, B). Only KRT19 of the control was associated with BMI (Suppl. Fig. S6C). GDF15, WFDC2, and CHI3L1 were associated with age in the control and COVID-19 patients (Suppl. Fig. S7). WFDC2, GDF15, CHI3L1, and KRT19 were elevated in the patients who were treated without IMV (Suppl. Table S6, Suppl. Fig. S8). We thus concluded that WFDC2, GDF15, CHI3L1, and KRT19 were four key proteins related to COVID-19 severity.

Identification of New Clinical Phenotypes Using Latent Cluster Analysis

There were no missing data for the levels of key proteins. A high correlation coefficient (> 0.6) was observed between GDF-15 and WFDC2 (Fig. 5A). WFDC2 can be measured as a tumor marker in Japan, and thus, GDF15 was excluded and WFDC2, CHI3L1, and KRT19 were used as the variables of classification. The BIC was highest for a three-class model and afterwards decreased in proportion to the number of added classes, suggesting that additional classes do not add substantial information to the model (Fig. 5B). The clinical phenotypes were named the α, β, and γ phenotypes. These phenotypes are visualized with principal component analysis in Fig. 5C. The relationships between the clinical phenotypes and the levels of WFDC2, CHI3L1, and KRT19 are visualized in Fig. 5D. The γ phenotype showed high levels of all three proteins. The log rank test indicated significant differences between the survival curves among the phenotypes (Fig. 5E). The associations of the clinical phenotypes with clinical data are shown in Suppl. Table S7. The γ phenotype was characterized by high creatinine and D-dimer levels and was associated with 28-day and hospital mortality (Suppl. Table S7).

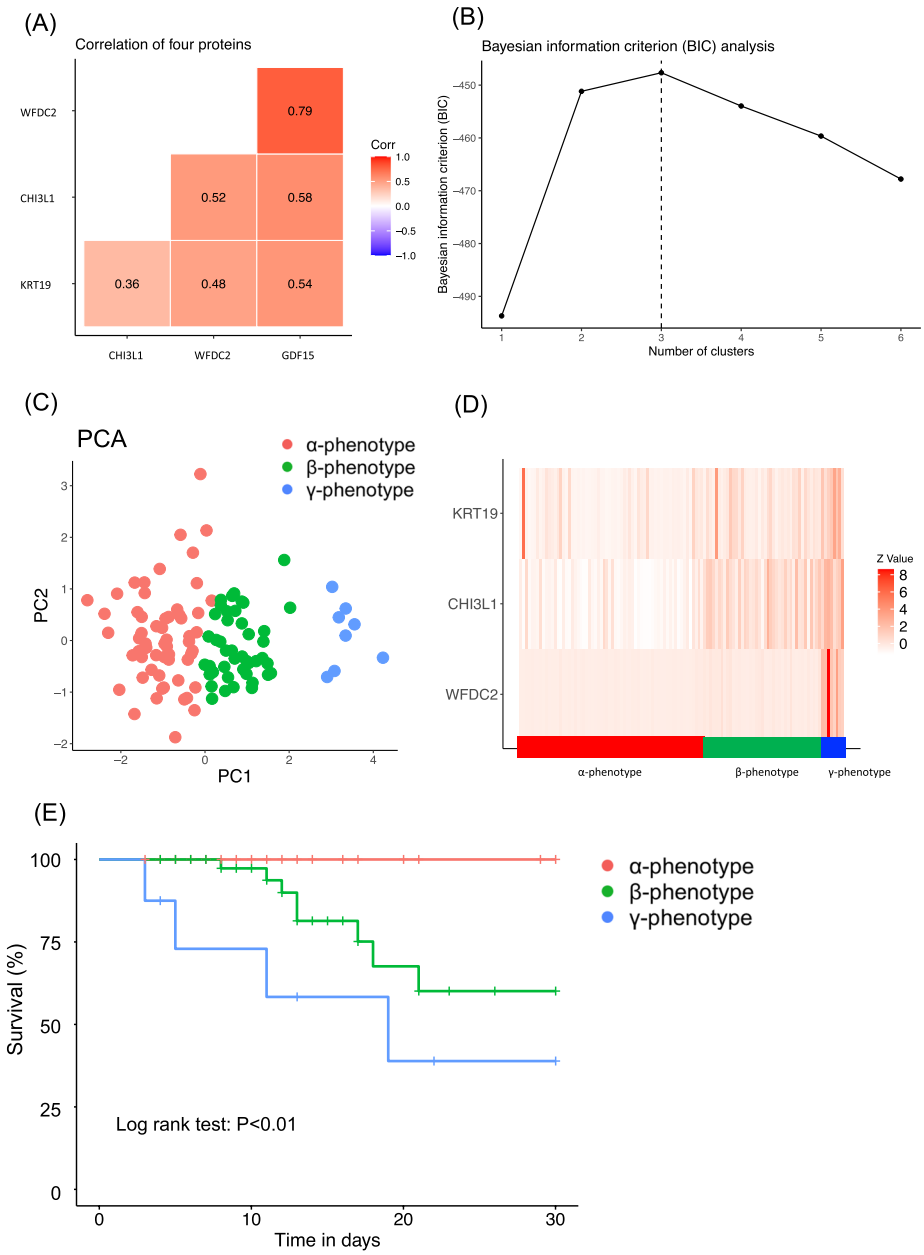


Fig. 5 Latent class analysis based on key proteins in the validation cohort. **A** The correlations of WFDC2, CHI3L1, and KRT19 are visualized by heat map. The numbers indicate the Pearson correlation. **B** BIC analysis with the number of clusters on the *X* axis. The BIC was highest for the three-class model. The latent class analysis calculation was performed using the VarSeILCM package in R, where the largest BIC is interpreted as optimal. **C** Visualization of phenotypes using principal component analysis in the validation data. **D** Heat map indicating the impact of the levels of the three proteins (WFDC2, CHI3L1, and KRT19) on the three phenotypes. White signifies the lowest and red the highest *Z*-score. The actual cytokine levels are transformed to *Z*-scores. **E** Kaplan–Meier curve of 28-day survival stratified by latent class analysis-derived phenotype. The log rank test showed significant differences between the three phenotypes ($P < 0.01$)

Discussion

Our study showed that four key proteins, WFDC2 (WAP four-disulfide core domain protein 2, also known as human epididymis protein 4 [HE4]) [26, 27], GDF-15 (growth differentiation factor 15) [28], CHI3L1 (chitinase-3 like-protein-1, also known as YKL-40) [29], and KRT19 (keratin, type I cytoskeletal 19) [30], were associated with the prognosis of COVID-19, and this is supported by the previous reports.

WFDC2 is highly expressed in ovarian cancer [31], systemic sclerosis–related interstitial lung disease [32] and lung adenocarcinoma [33]. It is also expressed in some epithelial cells of the upper airways, mucous cells, and ducts of the submucosal glands and is thought to be involved in innate immunity of the mucosal oral cavity and nasopharynx [26]. Previous reports have shown an association between the severity and prognosis of COVID-19 and WFDC2 [26, 27].

GDF-15 is a member of the transforming growth factor-β molecule superfamily [34] and is highly expressed in macrophages, airway epithelial cells, and vascular endothelial cells [35]. It has been reported to be an independent prognostic factor in cardiovascular disease, lung disease, and sepsis [36–38]. Several reports show the association between the levels of GDF-15 and disease severity in COVID-19 [28, 39]. As mentioned above, GDF-15 is well observed to be upregulated under stress conditions, but the mechanism for this is unclear. Further research is needed, including that into which tissues express GDF-15 in COVID-19.

CHI3L1 is a member of glycoside hydrolase family 18 and is synthesized and secreted by many cells, including macrophages, neutrophils, synoviocytes, smooth muscle cells, and tumor cells [40]. CHI3L1 has been reported to promote cancer growth, production of proinflammatory cytokines, and microglial activation [41]. It is strongly associated with diseases such as asthma, arthritis, sepsis, diabetes, liver fibrosis, and coronary artery disease [29]. In COVID-19, CHI3L1 is reported to be associated with severe disease, although it is not correlated with mortality. It is suggested that CHI3L1 is a major stimulator of ACE2, promotes binding and activation of SC2 S-protein-receptor, and enhances infection and the spread of COVID-19 [29, 42]. In the present study, GDF15, WFDC2, and CHI3LI were correlated with age in the COVID-19 patients. It has been reported that the levels of circulating CHI3L1 increase with aging in healthy controls, whereas levels of circulating CHI3L1 increase in patients with severe COVID19 compared to healthy controls regardless of age [29].

KRT19 is one of the most important cytokeratins expressed in epithelial and mesothelial tissues, and its overexpression has been reported in more than 30 malignant neoplasms, including lung and breast cancer [43]. Cyfra21-1 proteins, a fragment of KRT19, have been reported to be useful among lung cancers as a marker for non-small cell lung cancer (squamous cell carcinoma) [44]. In COVID-19, although Gisby et al. [45] reported the association of severity with upregulation of KRT19 using a proteomics approach, there are still few COVID-19-related reports on this cytokeratin.

Our previous studies showed that the key cytokine proteins formed a cytokine network in the patients including those with sepsis and burn [7–9]. In the present study, WGCNA revealed six protein network clusters. The GO enrichment and KEGG pathway analyses for each cluster indicated that the four key proteins are mainly involved in clusters related to cell adhesion pathways in addition to previously reported immune responses [46, 47]. This suggests that the four key proteins interact with each other in the cell adhesion pathways, which may play a key role in the pathogenesis of critical COVID-19.

The immune response associates with a complex interaction of factors involving comorbidities, age, weight, sex, ethnic background, pathogen types, and environment in the patients, thus resulting in a heterogeneous disease phenotype. The phenotype also varies between individuals over the time course of the disease. Mathew et al. showed that based on high-dimensional cytometry information, three immunotypes were associated with poor clinical trajectories in COVID-19 patients [47]. Also, Shu et al. distinguished different severity using LC–MS/MS on the basis of machine-learning models [48]. In our study, the critical COVID-19 patients were divided into three phenotypes (α, β, γ) using WFDC2, CHI3L1, and KRT19 on day 1 by latent class analysis. Patients with the β and γ phenotypes had lower survival rates and more prolonged ventilation times than those with the α phenotype, indicating that the β and γ phenotypes could be potential therapeutic targets for intervention in critical COVID-19.

This study has several limitations. First, age could have affected the plasma proteins levels. The difference in ages between the critical and non-critical patients in the first discovery cohort, between the control and COVID-19 patients, and between patients with early and late recovery in the second discovery cohort could have affected the process to derive the five candidate proteins. Second, the phase was defined as the time from visiting the emergency department or admission to the hospital, and thus, the time from onset was not considered. It was not clear what triggers the protein elevation, when the protein elevation occurs and how long the proteins elevation continued; therefore, the possibility of missing important proteins due to focusing on specific periods, phase 1 and phase 2, remains. However, in clinical practice, the time of infection varies as the time of admission to the emergency department or ICU, and this study may be more relevant to actual clinical practice. Third, we used three cohorts that included different variables. Therefore, information on unmeasured confounders and treatment details is lacking that may have biased the results. Fourth, basic treatment strategies of the participating facilities may have differed in their details. Such variation between the treatment centers could slightly influence the levels of proteins and the findings in this analysis. Finally, we did not perform a validation of the clinical phenotypes and the prediction model in another cohort.

Conclusion

The use of a new plasma proteomics approach revealed four key proteins in the blood validated by ELISA that were associated with COVID-19 pathogenesis. The clinical phenotypes based on WFDC2, CHI3L1, and KRT19 were significantly associated with patient prognosis.

Abbreviations *BIC*: Bayesian information criterion; *CHI3L1*: Chitinase-3 like-protein-1; *COVID-19*: Coronavirus disease 2019; *ELISA*: Enzyme-linked immunosorbent assay; *GDF15*: Growth differentiation factor 15; *ICU*: Intensive care unit; *IMV*: Invasive mechanical ventilation; *KRT19*: Keratin, type I cytoskeletal 19; *NPX*: Normalized protein expression; *TNFRSF10B*: TNF Receptor Superfamily Member 10b; *WFDC2*: WAP four-disulfide core domain protein 2; *WGCNA*: Weighted gene co-expression network analysis

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10875-022-01386-3.

Acknowledgements We greatly appreciate the patients, families, and healthy volunteers involved in this study. We also thank all of the medical staff who cooperated with this study.

Author Contribution TE, TM, and YT equally conceived and designed this study, acquired, and analyzed the data and wrote the manuscript. Hisatake M. helped with study design and data interpretation and conducted the literature review. JT analyzed the data. Hiroshi M., TK, and HH contributed to data acquisition. FS, SS, and DO helped analyze the data. HO conducted the literature review.

Funding This study was supported by the Japan Agency for Medical Research and Development Grant Number [20fk0108404h0001] and JSPS KAKENHI Grant Number [JP19H03760].

Data Availability Original Olink proteomics data have been deposited to Mendeley Data: https://doi.org/10.17632/2cbxgsn7vx.1.

Declarations

Ethics Approval and Consent to Participate This study was performed according to the principles of the Declaration of Helsinki and received approval from the institutional review board of Osaka University Hospital (Permit Numbers: 885 [Osaka University Critical Care Consortium Novel Omix Project; Occonomix Project]). Informed consent was obtained from all patients or their relatives, and the healthy volunteers gave their informed consent for the collection of their blood samples.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

1. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395:507–13. https://doi.org/10.1016/S0140-6736(20)30211-7.
2. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020;382:1708–20. https://doi.org/10.1056/NEJMoa2002032.
3. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. JAMA. 2020;323:2052–9. https://doi.org/10.1001/jama.2020.6775.
4. Hur K, Price CPE, Gray EL, Gulati RK, Maksimoski M, Racette SD, et al. Factors associated with intubation and prolonged intubation in hospitalized patients with COVID-19. Otolaryngol Head Neck Surg. 2020;163(1):170–8. https://doi.org/10.1177/0194599820929640.
5. Bonaventura A, Vecchié A, Dagna L, Martinod K, Dixon DL, Van Tassell BW, et al. Endothelial dysfunction and immunothrombosis as key pathogenic mechanisms in COVID-19. Nat Rev Immunol. 2021;21:319–29. https://doi.org/10.1038/s41577-021-00536-9.
6. Kermali M, Khalsa RK, Pillai K, Ismail Z, Harky A. The role of biomarkers in diagnosis of COVID-19 - a systematic review. Life Sci. 2020;254:117788. https://doi.org/10.1016/j.lfs.2020.117788.
7. Matsumoto H, Ogura H, Shimizu K, Ikeda M, Hirose T, Matsuura H, et al. The clinical importance of a cytokine network in the acute phase of sepsis. Sci Rep. 2018;8:13995. https://doi.org/10.1038/s41598-018-32275-8.
8. Matsuura H, Matsumoto H, Osuka A, Ogura H, Shimizu K, Kang S, et al. Clinical importance of a cytokine network in major burns. Shock. 2019;51:185–93. https://doi.org/10.1097/SHK.0000000000001152.
9. Ebihara T, Matsumoto H, Matsubara T, Matsuura H, Hirose T, Shimizu K, et al. Adipocytokine profile reveals resistin forming a prognostic-related cytokine network in the acute phase of sepsis. Shock. 2021;56:718–26. https://doi.org/10.1097/SHK.0000000000001756.
10. Sinha P, Delucchi KL, McAuley DF, O'Kane CM, Matthay MA, Calfee CS. Development and validation of parsimonious algorithms to classify ARDS phenotypes: secondary analyses of randomised controlled trials. Lancet Respir Med. 2020;8:247–57. https://doi.org/10.1016/S2213-2600(19)30369-8.
11. Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019;321:2003–17. https://doi.org/10.1001/jama.2019.5791.
12. Scicluna BP, van Vught LA, Zwinderman AH, Wiewel MA, Davenport EE, Burnham KL, et al. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. Lancet Respir Med. 2017;5:816–26. https://doi.org/10.1016/S2213-2600(17)30294-1.
13. Filbin MR, Mehta A, Schneider AM, Kays KR, Guess JR, Gentili M, et al. Longitudinal proteomic analysis of severe COVID-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions. Cell Rep Med. 2021;2:100287. https://doi.org/10.1016/j.xcrm.2021.100287.
14. COVID-19 therapeutic trial synopsis. https://www.who.int/publications-detail-redirect/covid-19-therapeutic-trial-synopsis. Accessed 11 June 2021.
15. Ebihara T, Matsumoto H, Matsubara T, Togami Y, Nakao S, Matsuura H, et al. Cytokine elevation in severe COVID-19 from longitudinal proteomics analysis: comparison with sepsis. Front Immunol. 2022;12:798338. https://doi.org/10.3389/fimmu.2021.798338.
16. Wik L, Nordberg N, Broberg J, Björkesten J, Assarsson E, Henriksson S, et al. Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. Mol Cell Proteomics. 2021;20:100168. https://doi.org/10.1016/j.mcpro.2021.100168.
17. Li Y, Schneider AM, Mehta A, Sade-Feldman M, Kays KR, Gentili M, et al. SARS-CoV-2 viremia is associated with distinct proteomic pathways and predicts COVID-19 outcomes. J Clin Invest. 2021;131:e148635. https://doi.org/10.1172/JCI148635.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.
20. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of

biomolecular interaction networks. Genome Res. 2003;13:2498–504. https://doi.org/10.1101/gr.1239303.
21. Fröhlich H, Speer N, Poustka A, Beißbarth T. GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. BMC Bioinformatics. 2007;8:166. https://doi.org/10.1186/1471-2105-8-166.
22. KEGG: Kyoto encyclopedia of genes and genomes [Internet]. https://www.kegg.jp/. Accessed 18 Sep 2022.
23. Sinha P, Calfee CS, Delucchi KL. Practitioner's guide to latent class analysis: methodological considerations and common pitfalls. Crit Care Med. 2021;49:e63-79. https://doi.org/10.1097/CCM.0000000000004710.
24. Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct Equ Model Multidiscip J. 2007;14:535–69. https://doi.org/10.1080/10705510701575396.
25. Nagin DS. Group-based modeling of development. Cambridge, MA: Harvard University Press; 2005.
26. Schirinzi A, Cazzolla AP, Lovero R, Lo Muzio L, Testa NF, Ciavarella D, et al. New insights in laboratory testing for COVID-19 patients: looking for the role and predictive value of Human epididymis secretory protein 4 (HE4) and the innate immunity of the oral cavity and respiratory tract. Microorganisms. 2020;8:1718. https://doi.org/10.3390/microorganisms8111718.
27. Wei X, Su J, Yang K, Wei J, Wan H, Cao X, et al. Elevations of serum cancer biomarkers correlate with severity of COVID-19. J Med Virol. 2020;92:2036–41. https://doi.org/10.1002/jmv.25957.
28. Myhre PL, Prebensen C, Strand H, Røysland R, Jonassen CM, Rangberg A, et al. Growth Differentiation Factor 15 provides prognostic information superior to established cardiovascular and inflammatory biomarkers in unselected patients hospitalized with COVID-19. Circulation. 2020;142:2128–37. https://doi.org/10.1161/CIRCULATIONAHA.120.050360.
29. Kamle S, Ma B, He CH, Akosman B, Zhou Y, Lee CM, et al. Chitinase 3-like-1 is a therapeutic target that mediates the effects of aging in COVID-19. JCI Insight. 2021;6:e148749. https://doi.org/10.1172/jci.insight.148749.
30. Gisby J, Clarke CL, Medjeral-Thomas N, Malik TH, Papadaki A, Mortimer PM, et al. Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death. eLife. 2021;10:e64827. https://doi.org/10.7554/eLife.64827.
31. Dochez V, Caillon H, Vaucel E, Dimet J, Winer N, Ducarme G. Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review. J Ovarian Res. 2019;12:28. https://doi.org/10.1186/s13048-019-0503-7.
32 Zhang M, Zhang L, Linning E, Xu K, Wang XF, Zhang B, et al. Increased levels of HE4 (WFDC2) in systemic sclerosis: a novel biomarker reflecting interstitial lung disease severity? Ther Adv Chronic Dis. 2020;11:2040622320956420. https://doi.org/10.1177/2040622320956420.
33. Bingle L, Cross SS, High AS, Wallace WA, Rassl D, Yuan G, et al. WFDC2 (HE4): a potential role in the innate immunity of the oral cavity and respiratory tract and the development of adenocarcinomas of the lung. Respir Res. 2006;7:61. https://doi.org/10.1186/1465-9921-7-61.
34. Bootcov MR, Bauskin AR, Valenzuela SM, Moore AG, Bansal M, He XY, et al. MIC-1, a novel macrophage inhibitory cytokine, is a divergent member of the TGF-beta superfamily. Proc Natl Acad Sci U S A. 1997;94:11514–9. https://doi.org/10.1073/pnas.94.21.11514.
35. Verhamme FM, Freeman CM, Brusselle GG, Bracke KR, Curtis JL. GDF-15 in pulmonary and critical care medicine. Am J Respir Cell Mol Biol. 2019;60:621–8. https://doi.org/10.1165/rcmb.2018-0379TR.
36. Buendgens L, Yagmur E, Bruensing J, Herbers U, Baeck C, Trautwein C, et al. Growth differentiation factor-15 is a predictor of mortality in critically ill patients with sepsis. Dis Markers. 2017;2017:5271203. https://doi.org/10.1155/2017/5271203.
37. Husebø GR, Grønseth R, Lerner L, Gyuris J, Hardie JA, Bakke PS, et al. Growth differentiation factor-15 is a predictor of important disease outcomes in patients with COPD. Eur Respir J. 2017;49:1601298. https://doi.org/10.1183/13993003.01298-2016.
38. Baek SJ, Eling T. Growth differentiation factor 15 (GDF15): a survival protein with therapeutic potential in metabolic diseases. Pharmacol Ther. 2019;198:46–58. https://doi.org/10.1016/j.pharmthera.2019.02.008.
39. Teng X, Zhang J, Shi Y, Liu Y, Yang Y, He J, et al. Comprehensive profiling of inflammatory factors revealed that growth differentiation factor-15 is an indicator of disease severity in COVID-19 patients. Front Immunol. 2021;12:662465. https://doi.org/10.3389/fimmu.2021.662465.
40. Zhao T, Su Z, Li Y, Zhang X, You Q. Chitinase-3 like-protein-1 function and its role in diseases. Sig Transduct Target Ther. 2020;5:1–20. https://doi.org/10.1038/s41392-020-00303-7.
41. Yeo IJ, Lee C-K, Han S-B, Yun J, Hong JT. Roles of chitinase 3-like 1 in the development of cancer, neurodegenerative diseases, and inflammatory diseases. Pharmacol Ther. 2019;203:107394. https://doi.org/10.1016/j.pharmthera.2019.107394.
42. Schoneveld L, Ladang A, Henket M, Frix AN, Cavalier E, Guiot J. YKL-40 as a new promising prognostic marker of severity in COVID infection. Crit Care. 2021;25:66. https://doi.org/10.1186/s13054-020-03383-7.
43. Hamesch K, Guldiken N, Aly M, Hüser N, Hartmann D, Rufat P, et al. Serum keratin 19 (CYFRA21-1) links ductular reaction with portal hypertension and outcome of various advanced liver diseases. BMC Med. 2020;18:336. https://doi.org/10.1186/s12916-020-01784-7.
44. Reinmuth N, Brandt B, Semik M, Kunze WP, Achatzy R, Scheld HH, et al. Prognostic impact of Cyfra21-1 and other serum markers in completely resected non-small cell lung cancer. Lung Cancer. 2002;36:265–70. https://doi.org/10.1016/s0169-5002(02)00009-0.
45. Gisby J, Clarke CL, Medjeral-Thomas N, Malik TH, Papadaki A, Mortimer PM, et al. Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death. Elife. 2021;10:e64827. https://doi.org/10.7554/eLife.64827.
46. Lucas C, Wong P, Klein J, Castro TBR, Silva J, Sundaram M, et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature. 2020;584:463–9. https://doi.org/10.1038/s41586-020-2588-y.
47 Mathew D, Giles JR, Baxter AE, Oldridge DA, Greenplate AR, Wu JE, et al. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science. 2020;369(6508):eabc8511. https://doi.org/10.1126/science.abc8511.
48. Shu T, Ning W, Wu D, Xu J, Han Q, Huang M, et al. Plasma proteomics identify biomarkers and pathogenesis of COVID-19. Immunity. 2020;53:1108-22.e5. https://doi.org/10.1016/j.immuni.2020.10.008.