Reprint of Tom Chittenden's Recent COVID-19 Article

by Kathy Kendrick, RFSPE

Last fall, when ISPE's Thousanders were participating in the election to promote Thomas Chittenden, PhD, to the rank of Diplomate, Tom himself was publishing yet another groundbreaking article regarding his work in artificial intelligence (AI) and machine learning (ML). In October 2021, in conjunction with over 60 additional authors, Tom published the paper titled, "Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort," in the journal, *Science Translational Medicine*.

To our *Telicom* readers, Tom reports, "Our COVID-19 paper was officially published in *Science Translational Medicine* on October 26, 2021. We are confident these findings will eventually lead to a novel, nonvaccine-based solution for the COVID-19 pandemic. We are now evaluating existing biotechnologies designed to disrupt virus–host interactions at the cell surface and thus block SARS-CoV-2 entry and replication in lung epithelium. As such, we are pursuing the industry's first potential AI/ML-based repositioning of an experimental Phase II drug for ovarian cancer."

Tom's impressive paper is reprinted below, followed by several detailed figures which have been enlarged for the convenience of our *Telicom* readers. Reference notations in the body of the paper are indicated as italicized numbers within parentheses. The paper may be accessed online at https://www.science.org/doi/10.1126/scitranslmed.abj7521, where the images can be further enlarged and the complete list of references, authors, affiliations, and other supplementary materials can be accessed.

Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort

by R. Carapito *et al*, including Thomas W. Chittenden via his affiliations with both Genuity AI Research Institute, Genuity Science (Boston, MA 02114, USA) and Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School (Boston, MA 02115, USA)

Corresponding author emails: tom.chittenden@genuitysci.com and siamak@unistra.fr.

ABSTRACT

The drivers of critical coronavirus disease 2019 (COVID-19) remain unknown. Given major confounding factors such as age and comorbidities, true mediators of this condition have remained elusive. We employed a multi-omics analysis combined with artificial intelligence in a young patient cohort where major comorbidities were excluded at the onset. The cohort included 47 "critical" (in the intensive care unit under mechanical ventilation) and 25 "non-critical" (in a non-critical care ward) patients with COVID-19 and 22 healthy individuals. The analyses included whole-genome sequencing, whole-blood RNA sequencing, plasma and blood mononuclear cells proteomics, cytokine profiling, and high-throughput immunophenotyping. An ensemble of machine learning, deep learning, quantum annealing, and structural causal modeling were employed. Patients with critical COVID-19 were characterized by exacerbated inflammation, perturbed lymphoid and myeloid compartments, increased coagulation, and viral cell biology. Among differentially expressed genes, we observed up-regulation of the metalloprotease *ADAM9*. This gene signature was validated in a second independent cohort of 81 critical and 73 recovered patients with COVID-19, and were further confirmed at the transcriptional and protein level as well as by proteolytic activity. Ex vivo ADAM9 inhibition decreased severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) uptake and replication in human lung epithelial cells. In conclusion, within a young, otherwise healthy, cohort of individuals with COVID-19, we provide the landscape of biological perturbations in vivo where a unique gene signature differentiated critical from non-critical patients. We further identified ADAM9 as a driver of disease severity and a candidate therapeutic target.

INTRODUCTION

Unlike many viral infections and most respiratory virus infections, coronavirus disease 2019 (COVID-19) is characterized by a complex and diversified spectrum of clinical manifestations (*1*). Indeed, upon infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), age-, sex-, and phenotype-matched individuals can be classified within four distinct groups: (1) asymptomatic individuals, (2) patients displaying influenza-like illnesses, (3) patients affected by respiratory dysfunction who eventually need an external oxygen supply, and (4) patients suffering from acute respiratory distress syndrome (ARDS) who need invasive mechanical ventilation in an intensive care unit (ICU). Even though the last group represents only a small fraction of COVID-19 patients, this group encompasses the most critical form of the disease and has an average case-fatality rate of approximately 25% (*2*). Despite intense investigation, the fundamental question of why the course of the disease shows such a marked difference in an otherwise, apparently indistinguishable set of individuals remains unanswered (*3–6*). To better understand this issue, high-resolution molecular analyses should be applied to well-defined cohorts of patients and controls where a maximum of confounding factors have been eliminated. These factors include older age as well as a number of comorbidities, such as cerebrovascular disease, types 1 and 2 diabetes, chronic kidney disease, chronic obstructive pulmonary disease, or heart conditions (*7*).

Several studies have used single, or a restricted number of, omics technologies to uncover molecular processes associated with disease severity, usually in unfiltered critical COVID-19 patients. Systemic inflammation with high concentrations of acute-phase proteins (C reactive protein; CRP, serum amyloid A; SAA, calprotectin) (*8*) and inflammatory cytokines, particularly interleukin (IL)-6 and IL-1β (*9–11*) has been found to be a hallmark of disease severity. In contrast, following an initial burst shortly after infection, the type I interferon (IFN) response is impaired at the RNA (*12*) and protein (*13*) level. Severity was also correlated with profound immune dysregulation, including modifications in the myeloid compartment with increases in neutrophils (*14, 15*), decreases in nonclassical monocytes (8) and dysregulation of macrophages (*10, 16*). The lymphoid compartment is also modified by both B cell activation (*17*) and an impaired T cell response, characterized by skewing toward a Th17 phenotype (*18, 19*). Moreover, coagulation defects have been identified in critically ill patients who are prone to thrombotic complications (*20–22*). Nevertheless, the full spectrum of omics technologies has not been applied to a highly curated cohort of patients with COVID-19 and controls that was established by eliminating a number of key confounding factors that affect severity and death, such as older age and comorbidities, at onset.

In this cross-sectional study, we aimed to analyze the SARS-CoV-2-induced molecular changes that are characteristic of critical patients and differentiate them from non-critical patients. We hypothesized that certain host driver genes might be responsible for the development of critical illness and that those genes might represent therapeutic targets. To test these hypotheses, we performed an ensemble artificial intelligence (AI) and machine learning (ML)-based multi-omics study of 47 young (under 50 years of age) patients with COVID-19 without comorbidities admitted to the ICU and under mechanical ventilation ("critical" patients), versus matched patients with COVID-19 needing only hospitalization in a non-critical care ward (25 "non-critical" patients) and an age- and sex-matched control group of 22 healthy individuals not infected with SARS-CoV-2 ("healthy"). The multi-omics approach included whole-genome sequencing (WGS), whole-blood RNA sequencing (RNA-seq), quantitative plasma and peripheral blood mononuclear cell (PBMC) proteomics, multiplex plasma cytokine profiling, and high-throughput immune cell phenotyping. These analyses were complemented by the status of anti-SARS-CoV-2 neutralizing antibodies and multitarget IgG serology as well as the measurement of neutralizing anti-type I IFN auto-antibodies in the entire cohort.

RESULTS

A young, comorbidity-free patient cohort was analyzed by a multiomics approach

The present study focused on patients who were hospitalized for COVID-19 at a university hospital network in northeast France (Alsace) during the first French wave of the pandemic (March to April 2020), before the routine use of corticosteroids. A total of 72 patients under 50 years of age without comorbidities were enrolled. Fifty-three of these patients were men (74%), and the median age of the patients was 40 [IQR 33; 46] years. The patients were divided into two groups: (i) a "critical" group consisting of 47 (65%) patients hospitalized in the ICU due to moderate or severe ARDS according to the Berlin criteria (*23*) with 45 requiring invasive mechanical ventilation and 2 requiring high-flow nasal oxygen and noninvasive mechanical ventilation due to acute respiratory failure and (ii) "non-critical" group consisting of 25 patients (35%) who stayed at a non-critical care ward. In the latter group, nineteen (76%) needed low-flow supplemental oxygen. Patients who were transferred from the non-critical care ward to the ICU (n=19) were considered "critical" patients and for these the sampling was done upon ICU admission in the same conditions as patients directly admitted to the ICU. The median simplified acute physiology score (SAPS) II of the patients at the ICU was 38 [IQR 33; 47] points, and the median PaO_2/FiO_2 ratio of these patients was 123 [IQR 95; 168] mmHg upon admission. All the patients were discharged from the hospital or were deceased at the time of data analysis. The overall hospital- and day-28 mortality rate was 8.3% (6 patients, all in the critical group, for a mortality of 13% in this group). The characteristics of the patients in both groups are summarized in Table 1.

Based on these two patient groups and an additional group of 22 healthy (SARS-Cov-2 negative) sex- and aged- matched controls, we applied a global multi-omics analysis strategy to identify pathways and drivers of ARDS (Fig. 1). PBMCs were analyzed by mass-cytometry (CyTOF) and shotgun proteomics. Plasma samples were used for multiplex cytokine quantification and shotgun proteomics. Serum samples were used for multiplex IgG serology (*24*), detection of anti-SARS-CoV-2 neutralizing antibodies and anti-type I IFN neutralizing autoantibodies. Finally, RNA-seq and WGS were performed using whole-blood samples. Unless otherwise specified, all measures were obtained from samples that were collected at the time of hospital admission (whether at the ICU or the non-critical care ward). Validation of the identified driver genes was performed using an ex vivo model of SARS-CoV-2 infection. The top 600 genes found by classification of patient cohort 1 were evaluated in a second, independent cohort of 81 critical patients and 73 recovered critical patients (table S1).

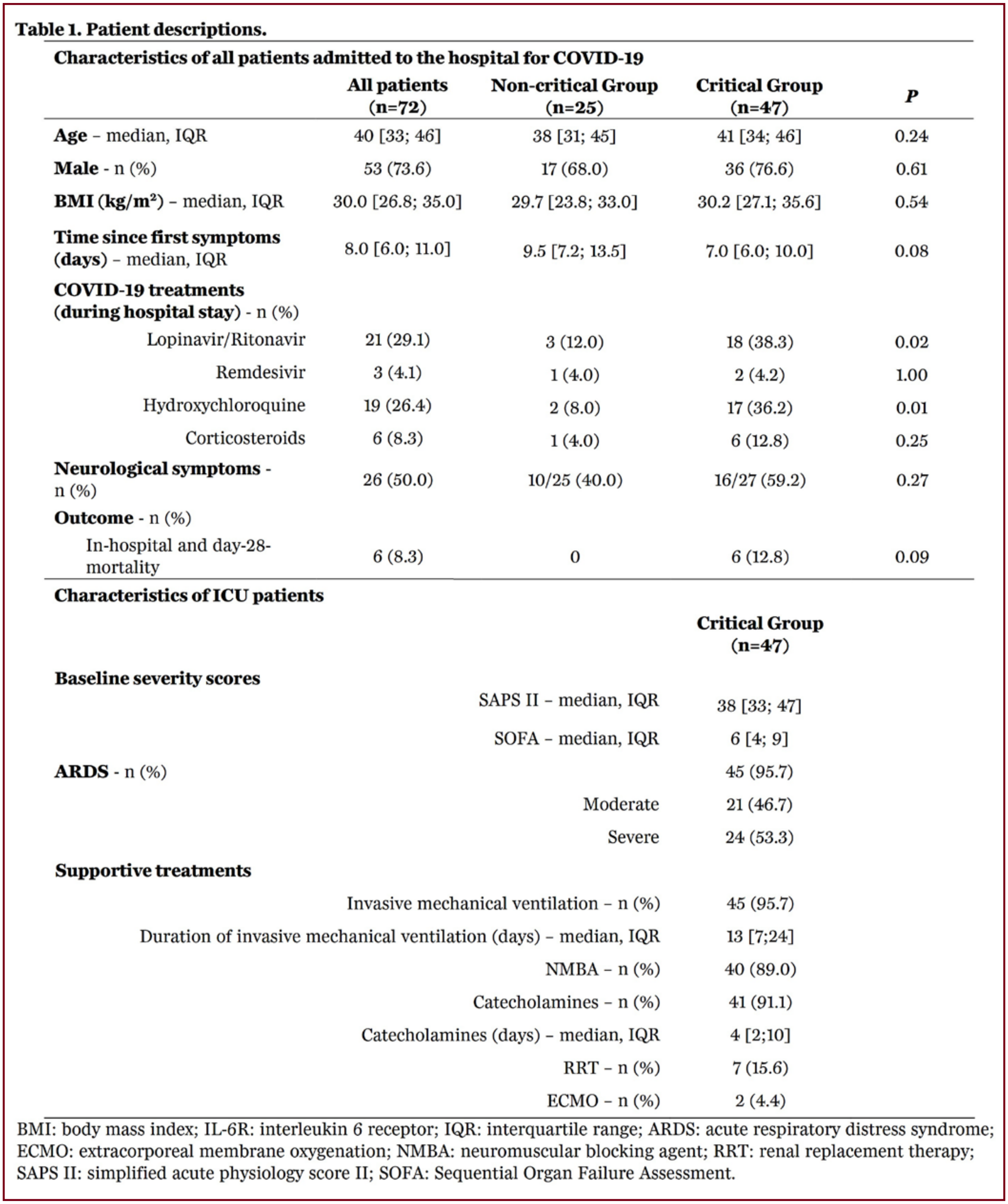


Table 1. Patient descriptions.

Characteristics of all patients admitted to the hospital for COVID-19	**All patients (n=72)**	**Non-critical Group (n=25)**	**Critical Group (n=47)**	***P***
Age – median, IQR	40 [33; 46]	38 [31; 45]	41 [34; 46]	0.24
Male - n (%)	53 (73.6)	17 (68.0)	36 (76.6)	0.61
BMI (kg/m^2) – median, IQR	30.0 [26.8; 35.0]	29.7 [23.8; 33.0]	30.2 [27.1; 35.6]	0.54
Time since first symptoms (days) – median, IQR	8.0 [6.0; 11.0]	9.5 [7.2; 13.5]	7.0 [6.0; 10.0]	0.08
COVID-19 treatments (during hospital stay) - n (%)				
Lopinavir/Ritonavir	21 (29.1)	3 (12.0)	18 (38.3)	0.02
Remdesivir	3 (4.1)	1 (4.0)	2 (4.2)	1.00
Hydroxychloroquine	19 (26.4)	2 (8.0)	17 (36.2)	0.01
Corticosteroids	6 (8.3)	1 (4.0)	6 (12.8)	0.25
Neurological symptoms - n (%)	26 (50.0)	10/25 (40.0)	16/27 (59.2)	0.27
Outcome - n (%)				
In-hospital and day-28-mortality	6 (8.3)	0	6 (12.8)	0.09

Characteristics of ICU patients	**Critical Group (n=47)**
Baseline severity scores	
SAPS II – median, IQR	38 [33; 47]
SOFA – median, IQR	6 [4; 9]
ARDS - n (%)	45 (95.7)
Moderate	21 (46.7)
Severe	24 (53.3)
Supportive treatments	
Invasive mechanical ventilation – n (%)	45 (95.7)
Duration of invasive mechanical ventilation (days) – median, IQR	13 [7;24]
NMBA – n (%)	40 (89.0)
Catecholamines – n (%)	41 (91.1)
Catecholamines (days) – median, IQR	4 [2;10]
RRT – n (%)	7 (15.6)
ECMO – n (%)	2 (4.4)

BMI: body mass index; IL-6R: interleukin 6 receptor; IQR: interquartile range; ARDS: acute respiratory distress syndrome; ECMO: extracorporeal membrane oxygenation; NMBA: neuromuscular blocking agent; RRT: renal replacement therapy; SAPS II: simplified acute physiology score II; SOFA: Sequential Organ Failure Assessment.

Table 1

Critical illness is characterized by a proinflammatory cytokine storm, changes in the T, B, dendritic and monocyte cell compartments and is independent of the extent of viral infection

The global proinflammatory cytokine profile showed significantly increased concentrations of IFN-γ (*P*=0.034), tumor necrosis factor (TNF)-α (*P*=0.022), IL-1β (*P*=0.0002), IL-4 (*P*=0.036), IL-6 (*P*<0.0001), IL-8 (*P*=0.0004), IL-10 (*P*=0.0002), and IL-12p70 (*P*=0.0221) in critical versus non-critical patients (Fig. 2A). This "cytokine storm" (*25*) was more pronounced in critical patients, as only IFN-γ, TNF-α and IL-10 were higher in non-critical patients as compared to healthy controls. Although the disease severity was initially associated with an RNA-seq based type I IFN signature, the absence of an increase in the plasma concentration of IFN-α in critical versus non-critical patients, the decrease in the IFN-α concentration during the ICU stay, and the reduction in the number of plasmacytoid dendritic cells, which are the main source of IFN-α, suggest that the IFN response is indeed impaired in critical patients (fig. S1) (*12*).

At a systemic level, lymphopenia is correlated with disease severity (*25–27*) (Fig. 2B). To further characterize the immune cells, we analyzed PBMCs by mass cytometry using an immune profiling assay covering 37 cell populations. Visualization of stochastic neighbor embedding (viSNE) showed a cell population density distribution pattern that was specific to the critical group (Fig. 2C). This pattern could be partly linked to the known immunosuppression phenomenon in critical patients (*12, 28, 29*), which was characterized by marked differences in the T cell compartments, where memory CD4 and CD8 T cells and Th17 cells were negatively correlated with disease severity (Fig. 2D). The latter observation is in line with the absence of a clear association between the plasma concentration of IL-17 and disease severity (Fig. 2A). In contrast, the B cell compartments of critical patients contained more naïve B cells and plasmablasts and fewer memory B cells than those of healthy controls (Fig. 2E). In accordance with previous reports (*17*), the number of plasmablasts tended to be higher in critical versus non-critical patients. Moreover, non-critical and critical patients were also characterized by lower numbers of dendritic cells and nonclassical monocytes (Fig. 2F and G). The remaining cell populations are presented in fig. 2. Altogether, the results indicate that critical illness was characterized by a proinflammatory cytokine storm and notable changes in the T, B, dendritic and monocyte cell compartments. These specific changes were independent from the extent of viral infection, as both the global anti-SARS-CoV-2 antibody concentrations and their neutralizing activity were not different in critical versus non-critical patients (fig. S3A and B).

To complete the immunologic profile, based on findings suggesting that at least 10% of critical patients have preexisting anti-type I IFN autoantibodies (*30, 31*), we measured anti-IFN-α2 and anti-IFN-ω neutralizing autoantibodies in patients and controls. Autoantibodies against type I IFNs were identified in two critical patients (fig. S3C) but none of the non-critical patients nor the healthy controls. Interestingly, in these two patients, the presence of autoantibodies was associated with an absence of SARS-CoV-2 neutralizing antibody titers (fig. S3D).

Quantitative plasma and PBMC proteomics high-light signatures of acute inflammation, myeloid activation, and dysregulated blood coagulation

Quantitative nanoLC-MS/MS analysis of whole unfractionated plasma samples identified a total of 336 proteins. Differential analysis was performed on an average of 178 ± 7, 189 ± 11 and 195 ± 8 proteins in healthy individuals, non-critical and critical patients, respectively (Fig. 3A). These experiments were conducted on crude liquid digested plasma samples without any fractionation or depletion of high abundant proteins to favor repeatability and robustness of quantification and differential analysis, at the cost of a lower proteome coverage. After validating the homogeneous distribution of the three groups using a multidimensional scaling plot, we performed a differential protein expression analysis to identify protein signatures that were specific to critical patients (Fig. 3B and C). In line with previous studies (*8, 32*), the antimicrobial calprotectin (heterodimer of S100A8 and S100A9) was among the top differentially expressed proteins (DEPs) in critical versus non-critical patients, which confirms that calprotectin is a robust marker for disease severity (Fig. 3D). Our data also showed dysregulation of multiple apolipoproteins including APOA1, APOA2, APOA4, APOM, APOD, APOC1 and APOL1 (Fig. 3C and E). Most of these proteins were associated with macrophage functions and were down-regulated in critical patients. Acute-phase proteins (CRP, CPN1, CPN2, C6, CFB, ORM1, ORM2, SERPINA3, and SAA1) were strongly up-regulated in critical patients (Fig. 3C and E). These findings are consistent with previous studies showing that acute inflammation and excessive immune cell infiltration are associated with disease severity (*26, 33, 34*).

Whole-cell lysates of PBMCs from the same groups of patients and controls were also subjected to quantitative nanoLC-MS/MS analysis, which led to the identification and quantification of a total of 2196 proteins. Differential analysis was performed on an average of 801 ± 213, 1050 ± 309 and 1052 ± 286 proteins in healthy individuals, non-critical and critical patients respectively (Fig. 3F). Although the human proteome coverage was relatively low after exclusion of contaminating fetal calf serum peptides and the distribution of the three groups in the multidimensional scaling plot was less clear than that found for plasma proteins, the differential expression analysis between non-critical and critical patients showed dysregulation of blood coagulation and myeloid cell differentiation (Fig. 3G to I). The latter observation involving the CA2, AHSP, SLC4A1, TFRC, DMTN, FASN, and PRTN3 proteins was in line with the plasma proteomics results evidencing dysregulation of macrophages and with other reports showing that severe COVID-19 is marked by a dysregulated myeloid cell compartment (*15*). The profile of the blood coagulation proteins HBB, HBD, HBE1, SLC4A1, PRDX2, SRI, ARF4, MANF, ITGA2, ORM1, and SERPINA1 confirmed that severity is also associated with coagulation-associated complications that can involve either bleeding or thrombosis (*35*).

Combined transcriptomics and proteomics analysis supports inflammatory pathways associated with critical disease

Consistent with the proteomics data, differential gene expression and gene set enrichment analysis of RNA-seq data from whole blood samples collected from the patients showed that regulation of the inflammatory response, myeloid cell activation and neutrophil degranulation were the main enriched pathways in critical patients with normalized enrichment scores of 2.33, 2.65 and 2.66, respectively (Fig. 4A and B). To identify enriched pathways that were supported by different omics layers, we performed nested GOSeq (nGOseq) (*36*) functional enrichment of the differentially expressed genes or proteins identified from the RNA-seq, plasma and PBMC proteomics data (Fig. 4C). In line with the cytokine profiling results (Fig. 2A), inflammatory signaling and the response to proinflammatory cytokine release (IL-1, IL-8 and IL-12) were supported by multiple omics datasets. As suggested by the results from immune cell profiling (Fig. 2C and D) and previous studies, the B cell response was activated, whereas the T cell response was impaired (*17, 37*). As previously observed (*8, 14, 15, 38*), the activation of neutrophils and monocytes was confirmed by the enrichment of nine different nGOseq terms (Fig. 4C). The nGOseq enrichment analysis also indicated that dysfunction of blood coagulation involves a fibrinolytic response; however, this observation could also be linked to the anticoagulant therapy administered to most critical patients. Moreover, nGOseq terms related to viral entry and even viral transcription were strongly enriched for patients with critical disease across the three omics datasets. This result was consistent with the identification of viral gene transcripts in the RNA-seq data of eight critical patients, but not in those from non-critical patients (table S2).

Integrated AI, ML, and probabilistic programming reveals a robust gene expression signature and identifies driver genes that differentiate critical from non-critical patients.

To robustly identify a set of genes that might differentiate between non-critical and critical COVID-19 patients and could thus be related to the progression to ARDS, we partitioned the 69 patient blood RNA-seq data (46 critical and 23 non-critical patients) 100 times to account for sampling variation, using 80% for training and 20% for testing, and evaluated the performance of seven distinct AI and ML algorithms, including a quantum support vector machine (qSVM), to differentiate between patients with non-critical and critical COVID-19. We have previously shown that quantum annealing is a more robust classifier for relatively small patient training sets (*39*). The receiver operating characteristic curves (ROCs) for the 100 partitions of the patient data as well as other classification performance metrics are shown in Fig. 5A and table S3. The classification performance on the test set provided a high degree of confidence that the signals learned by the various AI and ML algorithms are generalizable.

After successfully classifying non-critical versus critical patients based on whole-transcriptome RNA-seq profiling, we assessed feature scores across the six distinct ML algorithms and all partitions to determine an ensemble feature ranking, ignoring features from the partitions of patient data where the test Area Under the Receiver Operating Characteristic (AUROC) was less than 0.7. Aggregating the best performing features across both the algorithm and data partitions afforded a more robust and stable set of generalizable features.

This signature represented hundreds of genes that are differentially expressed and, by itself, did not distinguish between driver genes of critical COVID-19 and genes that react to the disease. Therefore, we then selected the top 600 most informative genes and used them as input for structural causal modeling (SCM) to identify likely drivers of critical COVID-19. We confirmed that these 600 genes are biologically relevant for distinguishing between critical and non-critical patients by retraining an ensemble ML classifier using only those 600 genes (table S4). Previous work has shown that SCM of RNA-seq data produces causal dependency structures, which are indicative of the signal transduction cascades that occur within cells and drive phenotypic and pathophenotypic development (*40*). However, this approach works best if the gene sets are stable and consistent across six different algorithms, as shown here. The resultant SCM output is presented as a directed acyclic graph (DAG) (Fig. 5B), a gene network representing the putative flow of causal information, with genes on the left predicted to have the greatest degree of influence on the entire state of the network. Perturbing these genes is the most likely to be disruptive to the state of the network (fig. S4) and is expected to exert the greatest effect on the expression of downstream genes. The top five genes associated with the greatest degree of putative causal dependency were *ADAM9*, *RAB10*, *MCEMP1*, *MS4A4A* and *GCLM*, and all five of these genes were significantly up-regulated in critical patients with false discovery rates (FDR) of $1.6x10^{-11}$, $3.1x10^{-12}$, $1.6x10^{-11}$, $1.0x10^{-9}$ and $5.3x10^{-13}$, respectively (Fig. 5C).

To further assess the informativeness of this COVID-19 gene expression signature, we employed a second independent patient cohort consisting of critical COVID-19 patients sampled at the time of entry into the ICU and recovered critical patients sampled at three months after discharge from the ICU. Patients in this second cohort were from a more typical COVID-19 ICU population as no exclusion criteria based on age or absence of comorbidities were applied (table S1). Although non-critical COVID-19 patients cannot be assumed to be the same as recovered critical COVID-19 patients, and thus the ML models from the first patient cohort cannot be directly applied to the second, the second patient cohort was used to provide additional evidence of the overall importance of the gene expression signature related to critical forms of COVID-19. The driver genes followed the same trend in the second patient cohort; namely that all five of these genes showed increased expression in the critical COVID-19 patient groups (fig. S5A). Moreover, an ensemble of ML classifiers trained on the second cohort using the 600 genes identified in the first group of patients was well able to differentiate between critical and recovered patients (fig. S5B and C); classification performance when training on the differentially expressed genes between critical and recovered patients was nearly the same as the first patient cohort (table S5), which further suggests a substantial degree of biological relevance of this gene signature.

ADAM9 is a driver of ARDS in critical COVID-19 patients.

Among the five driver genes identified by structural causal modeling, we primarily focused on experimentally determining the role of *ADAM9* (a disintegrin and a metalloprotease 9) in COVID-19 etiology because (i) it was the gene with the greatest degree of causal influence in the SCM DAG, (ii) it was the only driver gene that was previously shown to interact with SARS-CoV-2 through a global interactomics approach (*41, 42*) and (iii) it is an entry factor for another RNA virus, the encephalomyocarditis virus (*43*). ADAM9 is a metalloprotease with various functions that are mediated either by its disintegrin domain for adhesion or by its metalloprotease domain for the shedding of a large range of cell surface proteins (*44*). The *ADAM9* gene encodes two isoforms which are translated into either membrane-bound or secreted protein. Although neither isoform could be detected using our proteomics approach, *ADAM9* was up-regulated at the RNA level, and the secreted form was found at a higher concentration in the serum of critical versus non-critical patients (Fig. 6A and B). The transcriptional up-regulation of *ADAM9* was also found to be associated with disease severity in a previously published bulk RNA-seq dataset (fig. 6) (*45*). To assess the potential for increased metalloprotease activity in the critical cohort, we quantified the soluble form of the MICA protein (*46*), which is known to be cleaved by ADAM9 (*47*) by ELISA. The concentration of soluble MICA was indeed significantly higher in the plasma of critical patients as compared to non-critical patients (*P*=0.016) and healthy controls (*P*=0.0001; Fig. 6C). A global expression quantitative trait loci (eQTL) analysis using WGS and RNA-seq data identified eight SNPs associated with three of the top five putative driver genes with genome-wide significance (P<0.0001 for all SNPs, table S6). Among these SNPs, rs7840270 is localized just 0.3 kb upstream of the *ADAM9* gene and an eQTL for blood expression was reported in the Genotype-Tissue Expression database (GTEx). In the present cohort, including all 3 groups together, the C allele was associated with a higher abundance of ADAM9 transcripts (Fig. 6D) as it is in the GTEx dataset. The higher expressing allele C was indeed more frequent in critical than in non-critical patients (71.3% versus 50%, OR=2.48, 95% CI: [1.14-5.36], *P*=0.017). This was not due to any difference in ethnicities between critical and non-critical groups (fig. S7) (*48*). *ADAM9* RNA expression was significantly higher in the CC compared to CA (*P*=0.049) and AA (*P*=0.0046) genotypes only in the critical group, suggesting that the CC genotype may contribute to higher ADAM9 RNA expression in critical patients (fig. S8).

To assess the role of ADAM9 in viral infection, we set up an ex vivo assay in which *ADAM9* was silenced by siRNA in Vero 76 or A549-ACE2 (*49*) cells and subsequently infected the cells with SARS-CoV-2. Viral replication was monitored by flow cytometry quantification of the intracellular nucleocapsid protein and by quantitative viral real time polymerase chain reaction (qRT-PCR) of the culture supernatant (Fig. 6E). The average silencing efficiency reached 66% in Vero 76 cells and 93% in A549-ACE2 cells (fig. S9). In both cell lines, the amount of intracellular virus and the quantity of released virus were lower when *ADAM9* was silenced as compared to the control condition that was treated with a control siRNA (Fig. 6F and G). Our results collectively demonstrate that *ADAM9* is an in vivo up-regulated driver in critical patients. We also show a higher global proteolytic activity in serum samples of critical patients and demonstrate that a higher amount of ADAM9 facilitates viral infection and replication in an ex vivo cellular model.

DISCUSSION

A number of studies have detailed the molecular and cellular modifications associated with COVID-19 disease severity (*8, 11, 12, 15, 16, 34, 45, 50–54*), yet very few studies have targeted a young population with no comorbidities to reduce confounders that may also drive severity and mortality, and these confounders were limited to epidemiology or standard clinical parameters such as CRP, D-dimers or SOFA scores (*55–57*). A comprehensive understanding of the immune responses to SARS-CoV-2 infection is fundamental to develop an explanation as to why some young patients without comorbidities progress to critical illness whereas others do not, a phenomenon that has been exacerbated with new viral variants in current epidemic waves across the globe (*58, 59*). In particular, knowledge of the molecular drivers of critical COVID-19 is urgently needed to identify predictive biomarkers and more efficient therapeutic targets that function through drivers of critical COVID-19 rather than to downstream or secondary events (*60–62*).

Here, we used a multi-omics strategy associated with integrated AI, ML, and probabilistic programming methods to identify pathways and signatures that can differentiate critical from non-critical patients in a population of patients younger than 50 years without comorbidities. This in silico strategy provided a detailed view of the systemic immune response that was globally in accordance with previously published data. The thrust of our work, however, was to define a consistent transcriptomic signature that can robustly differentiate critical from non-critical patients, as shown by the classification performance metrics assessed in this study. Moreover, one can infer the biological relevance of the COVID-19 gene expression signature found in patient cohort 1 as the same classification performance was achieved in the second, independent patient cohort composed of 81 critically ill and 73 recovered critical patients.

Using the top 600 gene expression features of the signature as the input for structural causal modeling, we derived a causal network that uncovered five putative driver genes: *RAB10*, *MCEMP1*, *MS4A4A*, *GCLM* and *ADAM9*. RAB10 (Ras-related protein Rab-10) is a small GTPase that regulates macropinocytosis in phagocytes (*63*), which is a mechanism that has been suggested to be involved in the entry of SARS-CoV-2 into respiratory epithelial cells (*64*). MCEMP1 (mast cell expressed membrane protein 1) is a membrane protein specifically associated with lung mast cells, and decreasing the expression of this protein has been shown to reduce inflammation in septic mice (*65, 66*). MS4A4A (a member of the membrane-spanning, four domain family, subfamily A) is a surface marker for M2 macrophages that mediates immune responses in pathogen clearance (*67*) and regulates arginase 1 induction during macrophage polarization and lung inflammation in mice (*68*). GCLM (glutamate-cysteine ligase modifier subunit) is the first rate-limiting enzyme of glutathione synthesis and has been linked to severe COVID-19 (*68*). Although these four genes are all good candidates that can at least partially explain the severity of the disease, we focused our functional validations on ADAM9, which represented, from an in silico standpoint, the most promising driver gene. The confirmed up-regulation of ADAM9 at the RNA and protein levels in critical patients, which might partly be linked to pre-stored ADAM9 release by neutrophils (*69*), the increased metalloprotease activity in these same patients, and our ex vivo validation of its effect on viral uptake and replication are strong arguments supporting the targeting of this protein as a potential therapeutic strategy for the treatment or prevention of critical COVID-19. In vitro, we found that ADAM9 dramatically affects viral uptake or replication. The inhibition of this presumed mechanism of action of ADAM9 might represent a target for the treatment SARS-CoV-2 or other viral infections. Moreover, therapies that block viral uptake rather than host receptor binding are more likely to be variant-independent, a known virological behavior which might, at least partially, compromise current vaccination efforts (*70*).

Due to its implication in tumor progression and metastasis, ADAM9 is currently being tested as a target of antibody-drug-conjugate therapy for solid tumors (*71*). A repurposing strategy using ADAM9-blocking antibodies for the treatment of critical COVID-19 patients could therefore be envisioned. Alternatively, other therapeutic agents to reduce the ADAM9 concentration or activity could be pursued.

Our study has several limitations. Based on the present experimental results, we cannot conclude yet as to the molecular mechanism linking ADAM9 and viral uptake or replication. The predictive performance of ADAM9 as diagnostic marker for disease severity, as well as therapeutic target has to be evaluated in further studies. In addition, due to the differences in the first and second patient cohorts, we were unable to fully assess the generalizability of the RNA-seq gene signature found in the first patient cohort to the second patient cohort. Finally, we did not test the silencing of ADAM9 on various SARS-CoV-2 variants.

In conclusion, this study presents a detailed multi-omics investigation of a well-characterized cohort of young, previously healthy, critical COVID-19 patient series compared with non-critical patients and healthy controls. In addition to uncovering a landscape of molecular changes in the blood of critical patients, we applied a data-driven ensemble AI/ML strategy, which was independent of prior biological knowledge and thus minimized possible annotation biases, to gain insights into COVID-19 pathogenesis and to provide potential candidate diagnostic, prognostic and especially much needed therapeutic targets that might be helpful in combating the COVID-19 pandemic.

MATERIALS AND METHODS

Study design

In March and April 2020, patients aged less than 50 years, who had no comorbidities (of note, obesity alone was not considered an exclusion criterion) and were admitted for COVID-19 to the infectious disease unit (hereafter designated non-critical care ward) or to the designated ICUs at the university hospital network in northeast France (Alsace) were investigated within the framework of the present study. Follow-up was performed until hospital discharge. SARS-CoV-2 infection was confirmed in all the patients by qRT-PCR tests for COVID-19 nucleic acid of nasopharyngeal swabs (*72*). The ethics committee of Strasbourg University Hospitals approved the study (COVID-HUS, reference CE: 2020-34). Written informed consent was obtained from all the patients. The demographic characteristics, medical history, and symptoms were reported. Three groups were considered: (1) the “critical group” which included 47 patients admitted to the ICU, (2) the “non-critical group”, which was composed of 25 hospitalized patients at the non-critical care ward, and (3) the "healthy control group", which included 22 healthy age- and sex-matched blood donors aged less than 50 years. A second, independent cohort composed of 81 critical patients and 73 recovered critical patients from one of the ICU departments of Strasbourg University hospitals was used to further evaluated our molecular classification findings. No sample size calculations, randomization, or blinding was performed.

Sampling

Venipunctures were performed within the first hours after admission to the ICU or medical ward within the framework of routine diagnostic procedures. A subset of ICU patients (73%) were sampled every 4 to 8 days posthospitalization until discharge or death. Patient blood was collected into BD Vacutainer tubes with heparin (for plasma and PBMCs), EDTA (for DNA) or without additive (for serum) and into PAXgene Blood RNA tubes (Becton, Dickinson and Company). Blood from healthy donors was sampled in BD Vacutainer tubes with heparin, with EDTA or without additive. Plasma and serum fractions were collected after centrifugation at 900 × g at room temperature for 10 min, aliquoted, and stored at -80°C until use. PBMCs were prepared within 24 hours by Ficoll density gradient centrifugation. Aliquots of 1×10^6 dry cell pellets were frozen at -80°C until use for proteomics. Aliquots of a minimum of 5×10^6 cells were frozen at -80°C in 90% fetal calf serum (FCS)/10% dimethyl sulfoxide (DMSO). The EDTA and PAXgene tubes were stored at -80°C until use for DNA and RNA extraction, respectively.

Cytokine profiling

The plasma samples were analyzed using the V-PLEX Pro-inflammatory Panel 1 Human Kit (IL-6, IL-8, IL-10, TNF-α, IL-12p70, IL-1β, IL-2, IL-4 and IFN-γ) and the S-PLEX Human IFN-α2a Kit following the manufacturer's instructions (Mesoscale Discovery). Undiluted plasma was used for the S-PLEX Human IFN-α2a Kit and plasma was diluted two-fold for use with the V-PLEX Proinflammatory Panel 1. The MSD plates were analyzed with an MS2400 imager (Mesoscale Discovery). Soluble IL-17 in undiluted serum was quantified by Quantikine HS ELISA (Human IL-17 Immunoassay) following the manufacturer's instructions (R&D Systems). All standards and samples were measured in duplicate.

Immune phenotyping by mass cytometry

PBMCs were thawed rapidly, washed twice with 10 volumes of RPMI-1640 (Roswell Park Memorial Institute) medium (Thermo Fisher Scientific) and centrifuged for 7 min at 300 × g at room temperature between each washing step. Cells were then treated with 250 U of DNase (Thermo Fisher Scientific) in 10 volumes of RPMI-1640 medium for 30 min at 37°C in the presence of 5% CO_2. During this step, the viability and the number of the cells were determined with Trypan Blue (Thermo Fisher Scientific) and Türk's solution (Merck Millipore), respectively. After the elimination of DNase by centrifugation for 7 min at 300 × g at room temperature, a total of 3×10^6 cells were used for immunostaining with the Maxpar Direct Immune Profiling Assay kit (Fluidigm), following the manufacturer's instructions, except that we used 32% paraformaldehyde (Electron Microscopy Sciences). A red blood cell lysis step was included after the immunostaining following the manufacturer's instructions. The prepared cells were stored at -80°C until use for acquisition with a Helios mass cytometer system (Fluidigm). An average of 600000 events were acquired per sample. The mass cytometry standard files produced with the Helios instrument were analyzed using Maxpar Pathsetter software v.2.0.45 that was modified for live/dead parameters: the tallest peak was selected instead of the closest peak for the identification and quantification of the cell populations. The FCS files from each group (healthy, critical, non-critical) were then concatenated using CyTOF software v.7.0.8493.0 for viSNE analysis (Cytobank Inc). A total of 300000 events were used for the viSNE map that was generated with the following parameters: iterations (1000), perplexity (30) and theta (0.5). viSNE maps are presented as the means of all samples in each group.

Plasma proteomics analysis

Two microliters of plasma were prepared using the Pre-Omics iST Kit (PreOmics GmbH) according to the manufacturer's protocol prior to nanoLC-MS/MS analysis on a nanoAcquity Ultra-Performance LC (UPLC) device (Waters Corporation) coupled to a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific), as detailed in the supplementary materials. A sample pool comprising equal amounts of all protein extracts was constituted and regularly injected during the course of the experiment as an additional quality control.

The raw data obtained from each sample (45 critical patients, 23 non-critical patients, and 22 healthy controls) were processed using MaxQuant (version 1.6.14). Peaks were assigned using the Andromeda search engine with trypsin/P specificity. A database containing all human entries was extracted from the UniProtKB-SwissProt database (May 11 2020, 20410 entries). The minimal peptide length required was seven amino acids, and a maximum of one missed cleavage was allowed. Methionine oxidation and acetylation of the proteins' N-termini were set as variable modifications, and acetylated and modified methionine-containing peptides, as well as their unmodified counterparts, were excluded from the protein quantification step. Cysteine carbamidomethylation was set as a fixed modification. The "match between runs" option was enabled. The maximum false discovery rate was set to 1% at the peptide and protein levels with the use of a decoy strategy. The normalized label-free quantification (LFQ) intensities were extracted from the ProteinGroups.txt file after the removal of nonhuman and keratin contaminants, as well as reverse and proteins only identified by site. This resulted in 336 quantified proteins. Complete datasets have been deposited in the ProteomeXchange Consortium database with the identifier PXD025265 (*73*).

The LFQ values from MaxQuant were used for differential protein expression analysis. For each pairwise comparison, the proteins expressed in at least 80% of the samples in either group were retained. Variance stabilization normalization (Vsn) was performed using the justvsn function from the vsn R package (*74*). Missing values were imputed using the random forest approach (*75*). This process resulted in 161 proteins. Differential protein expression analysis was performed using the limma bioconductor package in R (*76*). Significant differentially expressed proteins were determined based on an adjusted *P*-value cutoff of 0.05 using the Benjamini-Hochberg method.

PBMC proteomics analysis

PBMC pellets were prepared using the PreOmics' iST Kit (PreOmics GmbH) according to the manufacturer's protocol prior to nanoLC-MS/MS analysis on a nanoAcquity UPLC device (Waters Corporation) coupled to a Q-Exactive HF-X mass spectrometer (Thermo Fisher Scientific, Waltham), as detailed in the supplementary materials.

The raw data obtained from each sample (34 critical patients, 21 non-critical patients and 22 healthy controls) were processed using MaxQuant (version 1.6.14). Peaks were assigned using the Andromeda search engine with trypsin/P specificity. A combined human and bovine database (because of contamination with fetal calf serum in the samples) was extracted from UniProtKB-SwissProt (8 September 2020, 26413 entries). The minimal peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. Methionine oxidation and acetylation of the proteins' N-termini were set as variable modifications, and acetylated and modified methionine-containing peptides, as well as their unmodified counterparts, were excluded from protein quantification. Cysteine carbamidomethylation was set as a fixed modification. The "match between runs" option was enabled. The maximum false discovery rate was set to 1% at the peptide and protein levels with the use of a decoy strategy. Only peptides unique to human entries were retained and their LFQ intensities were summed to derive the protein intensities. This process resulted in 2196 quantified proteins. Complete datasets have been deposited in the ProteomeXchange Consortium database with the identifier PXD 025265 (*73*).

Summed peptides normalized label-free quantification (LFQ values from MaxQuant software) values were used for differential protein expression analysis. For each pairwise comparison, proteins expressed in at least 80% of the samples in either group were retained. Variance stabilization normalization (Vsn) was performed using the justvsn function from the vsn R package (*74*). Missing values were imputed using the random forest approach (*75*). This resulted in 732 proteins. Differential protein expression analysis was performed using the limma bioconductor package in R (*76*). Significant differentially expressed proteins were determined based on an adjusted P-value cutoff of 0.05 using the Benjamini-Hochberg method.

Whole-genome sequencing (WGS)

WGS data was generated from DNA isolated from whole blood. Novaseq 6000 (Illumina Inc.) machines were used for DNA sequencing to a mean 30X coverage. The raw sequencing reads from FASTQ files were aligned using Burrows-Wheeler Aligner (BWA) (*77*), and Genomic Variant Call Format (GVCF) files were generated using Sentieon version 201808.03 (*78*). Functional annotation of the variants was performed using Variant Effect Predictor from Ensembl (version 101). GATK version 4 (*79, 80*) was used for the joint genotyping process and variant quality score recalibration (VQSR). We removed one duplicate sample based on kinship (king cutoff of 0.3) and retained 24476739 SNPs that were given a 'PASS' filter status by VQSR. The analysis of the 72 samples from the critical and non-critical groups identified 15870076 variants with MAF < 5%. The first two principal components were generated using plink2 on LD-pruned variants with Hardy-Weinberg equilibrium in the controls with a P-value $\geq 1 \times 10^{-6}$ and MAF > 5% and were used as covariates to correct for population stratification.

Analysis of expression quantitative trait loci (eQTLs)

We performed local (cis-) eQTL analysis to test for associations between genetic variants and gene expression in 67 samples having both RNA-seq and SNP genotype data. Briefly, we used the MatrixEQTL R package (*81*) where we selected a linear model and a maximum distance for gene-SNP pairs of 1×10^6. The top two principal components identified from the genotype principal component analysis were used as covariates to control for population stratification. We selected 304044 significant eQTLs with FDR $\leq$ 0.05.

RNA sequencing (RNA-seq)

Whole-blood RNA was extracted from PAXgene tubes with the PAXgene Blood RNA Kit following the manufacturer's instructions (Qiagen). A total of 69 samples, including 46 critical and 23 non-critical patients were processed. The RNA quantity and quality were assessed using the Agilent 4200 TapeStation system (for the RIN) (Agilent Technologies) and RiboGreen (for the concentration) (Thermo Fisher Scientific). RNA sequencing libraries were generated using the TruSeq Stranded Total RNA with Ribo-Zero Globin kit (Illumina) and sequenced on the Illumina NovaSeq 6000 instrument with S4 flow cells and 151-bp pairedend reads. The raw sequencing data were aligned to a reference human genome build 38 (GRCh38) using the short reads aligner STAR (*82*). Quantification of gene expression was performed using RSEM (*83*) with GENCODE annotation v25 (http://www.gencodegenes.org). Raw and processed datasets have been deposited in GEO with identifier GSE172114.

Differential gene expression (DGE) analysis was performed for two different purposes: 1) for the combined omics analysis of differentially expressed genes and proteins, and 2) as step to determine feature selection for classification in the in silico computational intelligence approach. For the combined omics analysis, we first removed lowly expressed genes for the 69 samples by removing genes with less than 1 count per million in less than 10% of the samples. We then performed DGE analysis on all 69 samples using the trimmed mean of M-values method (TMM) from the edgeR R package (*84, 85*).

In our computational intelligence approach, we performed DGE analysis for each partition of the train data using a frozen TMM normalization to calculate normalization factors based only on the training data, in order to avoid data leakage. Briefly, we removed lowly expressed genes for the 69 samples with genes with 1 count per million in less than 10% of samples. For each partition of the training data, we calculated the normalization factors, and then selected the library that had a normalization factor closest to 1. We used this library as a reference library to normalize all the samples keeping the training normalization factors unchanged. Differentially expressed genes were identified using quasi-likelihood F-test (QLF)-adjusted P-values from the edgeR R package. Differentially expressed genes with FDRs less than 0.05 were used for further downstream analysis.

Identification of potential driver genes through structural causal modeling.

To identify potential biomarkers that might differentiate patients in the non-critical group from those in the critical group, we used classification as a feature selection approach and then used the most informative features as input for structural causal modeling to identify potential driver genes. More specifically, classification was performed using the RNA-seq data by repeatedly partitioning non-critical and critical patients into 100 unique training and independent test sets representing 80% and 20% of the total data, respectively, ensuring that the proportions of non-critical and critical patients were consistent in each partition of the data. One hundred partitions of the data were used to capture the biological variation and to obtain increased statistical confidence in the results. After classification, feature scores for each method were determined and combined across all 100 partitions of the data and six of the ML algorithms, not including the deep learning algorithm. In order to capture as much information as possible while still being able to finish the analysis in a reasonable amount of time, the 600 most informative features were retained for structural causal modeling (600 features is the maximum that the structural causal modeling can finish in a reasonable amount of time).
We used seven distinct ML approaches for our classification models: LASSO, Ridge, support vector machine (SVM), quantum support vector machine (qSVM), eXtreme Gradient Boosting (XGB), random forest (RF), and a deep artificial neural network (DANN). A description of the algorithms and their relevant hyperparameters are mentioned in their respective sections in the supplementary materials. Hyperparameters were selected by using 10-fold cross-validation of the training data, and the performance was evaluated using the held-out test data.

Ensemble feature ranking

To derive an ensemble ranking of the feature importance, we first calculated the feature importances for each algorithm. LASSO, Ridge, SVM, and qSVM are linear models, and thus the feature importance was determined based on the value of the weight assigned to each feature, with a larger score corresponding to greater importance. RF creates a forest of decision trees, and as part of the fitting process, it determines an estimate of the feature importance by randomly permuting the features one at a time and determining the change in the accuracy. XGB calculates the feature importance by averaging the gain across all the trees, where the gain is the difference in the Gini purity of the parent node and the two children nodes.

The top 1000 most informative features of each model and for each partition of the data were retained. Because there were 100 partitions of the data, six algorithms (LASSO, Ridge, SVM, qSVM, RF, and XGB; DANN was not included because it lacks a robust approach to determine the feature importance), and up to 1000 features were retained, a total of up to 600000 possible features were considered for each feature set, though it may be lower as the features may not be unique such that the 1000 features for one partition of the data might exhibit some overlap with the top 1000 features for another partition of the data). We discarded the feature scores from an algorithm on any partition with a test AUROC < 0.7 in an attempt to exclude scores that might not truly be informative. To aggregate the scores, we scaled the scores by the most informative feature for each algorithm on each partition such that the feature scores were all between 0 and 1; for the first partition of the data, we scaled the 1000 most informative features from LASSO, then proceeded to do the same for Ridge, SVM, RF, and then repeated the process for each partition of the data. The scores were then averaged across all the partitions of the data to obtain a feature ranking for each method. If a feature was determined to be important for one partition of the data but not for others, it was given a value of 0 for all partitions of the data in which it did not appear. To determine a final ensemble feature ranking, the grand mean across all training partitions and algorithms was taken, and the features were sorted by the average score.

Structural causal modeling

We generated Bayesian Belief Networks (BBNs) for the top 600 most informative genes as defined by ensemble feature ranking described above on the first patient cohort (the informativeness of those 600 genes was evaluated in the second patient cohort). 600 genes were chosen to capture as much information possible while still allowing the algorithm to finish in a reasonable amount of time. A BBN is a directed acyclic graph (DAG), where the directionality of the arcs represents conditional dependencies between the nodes. Training of BBNs was performed in R using the *bnlearn* package (*86*). See supplementary materials for more details.

Real-time reverse transcription quantitative PCR (qRT-PCR)

Total RNA was extracted from cells using the RNeasy Mini Kit (Qiagen), and the RNA quality was assessed using an Agilent 2100 BioAnalyzer before reverse transcription into cDNA with Maxima H Minus Mastermix and following the manufacturer’s instructions (Thermo Fisher Scientific). RT-qPCR was performed using QuantStudio3 (Thermo Fisher Scientific) according to the manufacturer’s protocol, and using PowerTrack SYBR Green Master Mix (Thermo Fisher Scientific, Waltham, MA, USA). The following primers were used: *ADAM9*, forward 5′-GGACTCAGAGGATTGCTGCATTTAG-3′, reverse 5′-CTTCGAAGTAGCTGAGTCATGCTGG-3′; and GAPDH (housekeeping gene), forward 5′-GGTGAAGGTCGGAGTCAACGGA-3′ and 5′-GAGGGATCTCGCTCCTGGAAGA-3′ (Integrated DNA Technologies). The qRT-PCR protocol consisted of 95°C for 2 min followed by 40 cycles of 95°C for 5 s and 60°C for 30 s. All reactions were performed in duplicate, and the relative amounts of transcripts were calculated with the comparative Ct method. Gene expression changes were calculated using the $2^{-\Delta\Delta Ct}$ values calculated from averages of technical duplicates relative to the negative control. Melting-curve analysis was performed to assess the specificity of the PCR products.

Enzyme-linked immunosorbent assays (ELISA)

The concentrations of soluble ADAM9 (sADAM9) and soluble MICA (sMICA) in the serum of critical and non-critical patients and healthy controls were quantified by ELISA. For soluble ADAM9, we used the Human sADAM9 DuoSet ELISA kit (R&D Systems) following the manufacturer's instructions. sMICA concentrations were measured with an in-house developed sandwich ELISA using two monoclonal mouse antibodies for capture (A13-C485B10 and A9-C255A9 at concentrations of 2 mg/ml and 0.2 mg/ml, respectively) and one biotinylated monoclonal mouse antibody for detection (A15-C199B9 at 60 pg/ml); all three antibodies were made in house and described in Carapito *et al*. (*87*). Coating of Max-iSorp ELISA plates (Thermo Fisher Scientific) was performed in phosphate-buffered saline (PBS) at 4°C overnight. After three washing steps with PBS, the wells were blocked with 200 **μl** of 10% bovine serum albumin (BSA) in PBS for 1 hour at room temperature. All the following steps were carried out at room temperature with PBS/0.05% Tween 20/10% BSA, which was used as a diluent for all the reagents and serum samples. The plates were washed three times with PBS/0.05% Tween 20 between incubation steps. After blocking, the plates were incubated with 100 **μl** of sera, standards and controls for 2 hours, followed by incubation with 100 **μl** of biotinylated detection antibody for 1 hour. The plates were subsequently incubated for 1 hour with 100 **μl** of a 5000-fold dilution of streptavidin poly-horseradish peroxidase (HRP, Thermo Fisher Scientific) per well. The reactions were finally revealed using 3,3′,5,5′-tetramethylbenzidin (TMB) Ultra (Thermo Fisher Scientific) at 100 **μl**/well for 15 min and stopped with 100 **μl** of 1 M HCl. The absorbance was measured at 450 nm on a Varioskan LUX (Thermo Fisher Scientific).

Silencing and cell transfection

Vero 76 cell lines (Vero C1008, Cat. Nr. CRL-1586, Clone E6) were grown at 37 °C under 5% CO_2 and maintained in DMEM (Thermo Fisher Scientific) containing 100 units/ml penicillin and supplemented with 10% fetal bovine serum (Pan Biotech). ACE2-expressing A549 cells (A549-ACE2; a gift from Olivier Schwartz, Institut Pasteur) were grown at 37 °C under 5% CO_2 and maintained in DMEM (Thermo Fisher Scientific) containing 10 **μg**/ml of blasticidin S (Invitrogen). The cells were transfected with predesigned Stealth siRNA directed against ADAM9 (HSS112867) or the control Stealth RNAi Negative Control Duplex medium GC (45-55%) (Thermo Fisher Scientific) using Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher Scientific). One day prior to transfection, the cells were seeded in a 24-well plate at 0.05 × 10^6 cells per well. First, 1.5 **μl** of Lipofectamine RNAiMAX Transfection Reagent was added to 25 μl of Opti-MEM medium, followed by addition of the mix containing 5 pmoles of siRNA in 25 **μl** of Opti-MEM medium (Thermo Fisher Scientific). The mixture was incubated at room temperature for 5 min and then added to the cells. The cells were collected or infected after 48 hours.

In vitro viral infections

Vero 76 and A549-ACE2 cell lines were infected with wild-type SARS-CoV-2 virus at Multiplicities Of Infections (MOIs) of 10 and 400, respectively. The percentage of infected cells was determined by staining with SARS-CoV-2 nucleocapsid (% of nucleocapsid positive cells), and virus released into the supernatant was analyzed by RT-PCR (copies/ml), after 2 and 3 days of infection for Vero 76 and A549-ACE2 cells, respectively. The cells were fixed for 20 min in 3.6% paraformaldehyde at 4°C, washed in 5% FCS in PBS and stained with anti-nucleocapsid antibody (GTX135357, Genetex) at a 1:200 dilution in Perm/Wash (Becton, Dickinson and Company) for 45 min at room temperature. Samples were then incubated with Alexa Fluor 647-labeled goat anti-rabbit monoclonal antibody (Ab150083, Abcam, Cambridge, UK) diluted 1:200 in 5% FCS in PBS for 45 min at room temperature. Samples were acquired with a MACSQuant flow cytometer (Miltenyi Biotec) and analyzed with the Kaluza software (Beckman Coulter).

RNA was extracted from the supernatant of infected cells using the NucleoSpin Dx Virus Kit (Macherey-Nagel GmbH & Co.KG). RT-qPCR was performed using TaqPath 1-Step RT-qPCR Master Mix (CG) on the Quanstudio3 instrument (Thermo Fisher Scientific). The primer/probe mix used for absolute quantification of the virus was N1 and N2 from the 2019-nCoV RUO Kit (Integrated DNA Technologies), and the positive control for the standard curve was 2019-nCoV N Positive Control (Integrated DNA Technologies). The reaction was performed in 20 **µl**, which included 5 **µl** of eluted RNA, 5 **µl** of TaqPath Master Mix and 1.5 **µl** of the primer/probe. The qRT-PCR protocol consisted of 25°C for 2 min, 50°C for 15 min, and 95°C for 2 min, and 40 cycles of 95°C for 3 s and 60°C for 30 s. All reactions were performed in duplicate, and absolute quantification was calculated with the standard curve of the positive control.

Statistical analysis

Statistical analysis was performed with GraphPad Prism (GraphPad Software) unless stated otherwise. A *P*-value below 0.05 was considered significant. For two groups comparisons, data were analyzed by unpaired, two-sided Mann-Whitney or student's *t* test. For three or more groups comparisons, data were analyzed by unpaired, two-sided Kruskal-Wallis test, followed by Dunn's post-test; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$. In figures and tables, "n" represents the number of biological replicates and "N" the number of times an experiment was independently performed.

(Figures begin on the next page.)

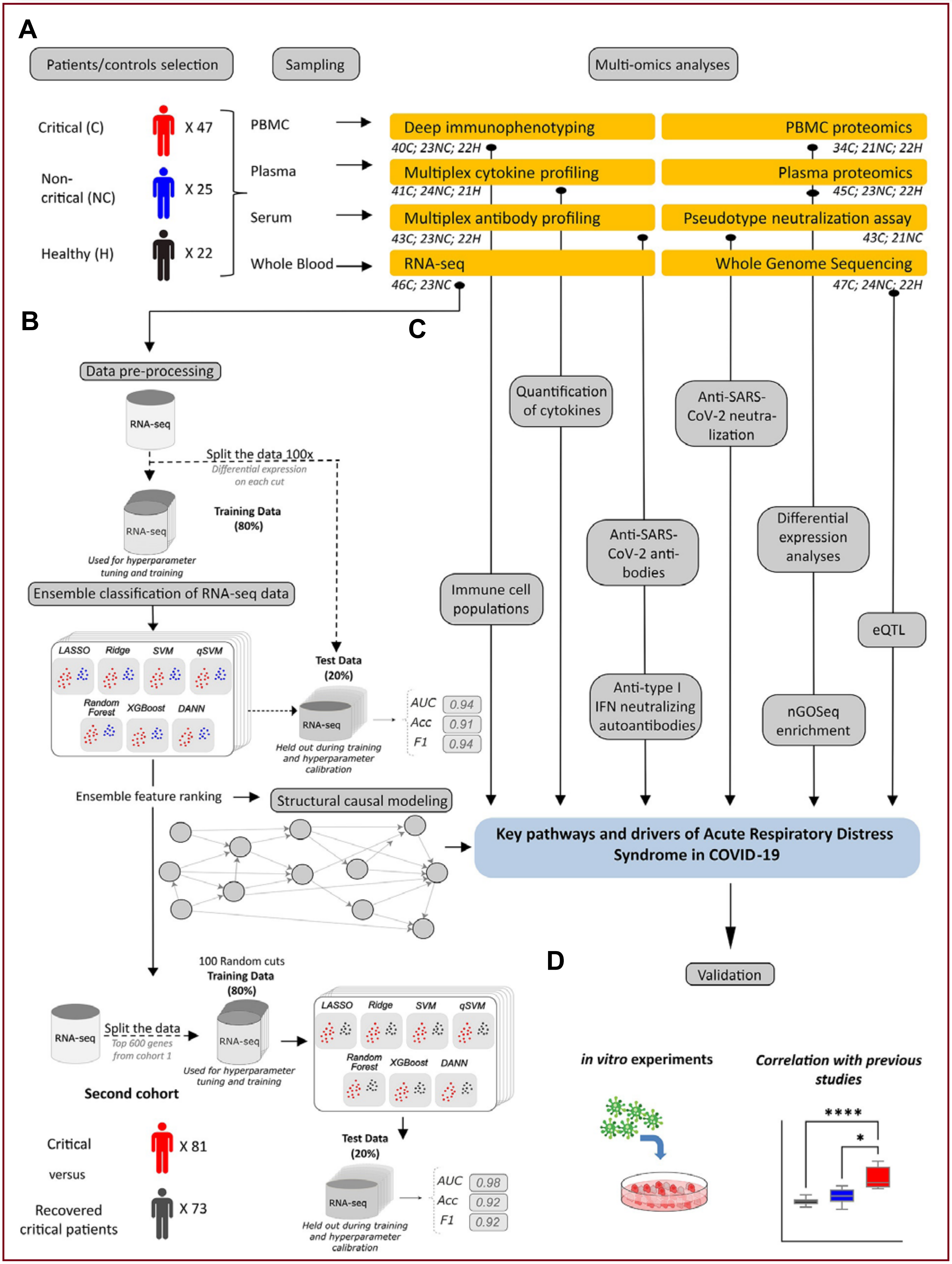

A
Patients/controls selection
Sampling
Multi-omics analyses
Critical (C)
X 47
Non-critical (NC)
X 25
Healthy (H)
X 22
PBMC
Plasma
Serum
Whole Blood
Deep immunophenotyping
40C; 23NC; 22H
PBMC proteomics
34C; 21NC; 22H
Multiplex cytokine profiling
41C; 24NC; 21H
Plasma proteomics
45C; 23NC; 22H
Multiplex antibody profiling
43C; 23NC; 22H
Pseudotype neutralization assay
43C; 21NC
RNA-seq
46C; 23NC
Whole Genome Sequencing
47C; 24NC; 22H
B
C
Data pre-processing
RNA-seq
Split the data 100x
Differential expression on each cut
Training Data (80%)
RNA-seq
Used for hyperparameter tuning and training
Ensemble classification of RNA-seq data
LASSO
Ridge
SVM
qSVM
Random Forest
XGBoost
DANN
Test Data (20%)
RNA-seq
Held out during training and hyperparameter calibration
AUC 0.94
Acc 0.91
F1 0.94
Quantification of cytokines
Anti-SARS-CoV-2 neutralization
Differential expression analyses
Anti-SARS-CoV-2 antibodies
Immune cell populations
eQTL
Anti-type I IFN neutralizing autoantibodies
nGOSeq enrichment
Ensemble feature ranking
Structural causal modeling
Key pathways and drivers of Acute Respiratory Distress Syndrome in COVID-19
Validation
D
100 Random cuts
Training Data (80%)
RNA-seq
Split the data
Top 600 genes from cohort 1
RNA-seq
Used for hyperparameter tuning and training
Second cohort
LASSO
Ridge
SVM
qSVM
Random Forest
XGBoost
DANN
Test Data (20%)
RNA-seq
Held out during training and hyperparameter calibration
AUC 0.98
Acc 0.92
F1 0.92
Critical
X 81
versus
Recovered critical patients
X 73
in vitro experiments
Correlation with previous studies
****
*

Figure 1

Fig. 1. A multi-omics analytical strategy identifies key pathways and drivers of Acute Respiratory Distress Syndrome in COVID-19.

(A) Forty-seven critical patients (C), 25 non-critical patients (NC) and 22 healthy controls (H) were enrolled in the study. PBMCs were isolated by density gradient and frozen until utilization for mass cytometry and whole proteomics. Plasma was used for cytokine profiling and whole proteomics. Serum was used to measure anti-type I IFN neutralizing antibodies, anti-SARS-CoV-2 neutralizing antibodies and multi-target antiviral serology. Whole blood was used for RNA-seq and whole-genome sequencing (WGS). The number of treated samples per group and per omics is indicated below each omics designation.

(B) The RNA-seq pipeline is shown based on the NC versus C comparison. To increase robustness of downstream analyses, an ensemble intelligence approach with seven algorithms was applied to multiple partitions of the RNA-seq data (see Methods) to classify NC versus C patients, performing differential analysis on each partition of the data. An ensemble ranking score across six of the seven algorithms and all partitions of the data was then determined, and the top 600 of those genes were used as the input for structural causal modeling to derive a putative causal network. To support the key findings from the first patient cohort, RNA-seq data from a second patient cohort consisting of 81 critical and 73 recovered critical patients were used. The data was partitioned analogously to the first patient cohort, but only the top 600 features from the first patient cohort were used to assess the informativeness of the gene signature.

(C) Cytokines and immune cells were quantified. WGS data were used for eQTL analysis together with the gene counts from the RNA-seq. Proteomics data were subjected to differential protein expression and nGOseq enrichment analyses.

(D) The key pathways and drivers resulting from the omics analyses in (B and C) were validated in a second cohort of 81 critical and 73 recovered critical patients. The differential expression of ADAM9, the main driver gene, was compared to publicly available bulk RNA-seq data. Finally, ex vivo infection experiments with SARS-CoV-2 were conducted to validate a driver gene candidate.

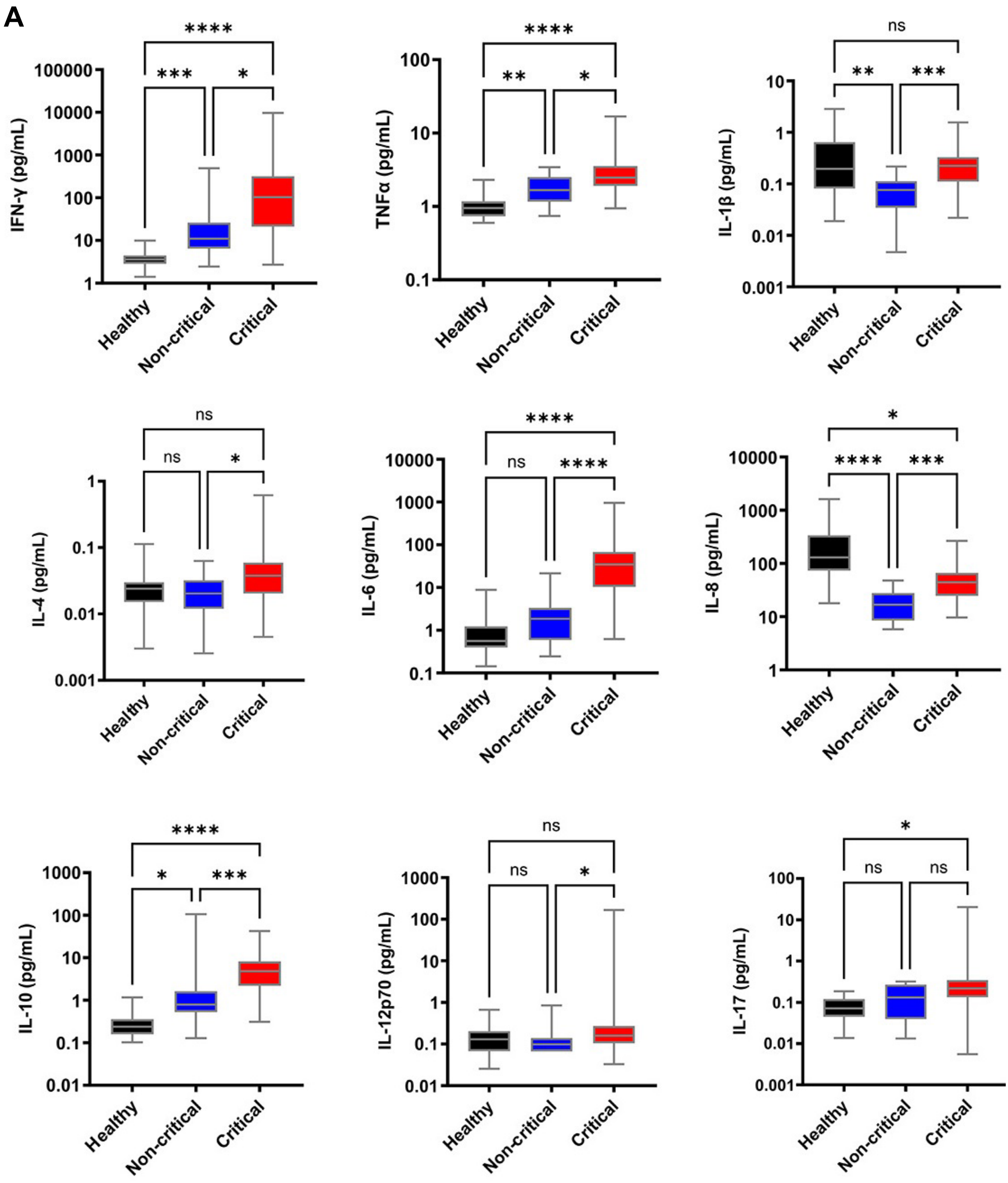

A
****
***
*
IFN-γ (pg/mL)
100000
10000
1000
100
10
1
Healthy
Non-critical
Critical
****
**
*
TNFα (pg/mL)
100
10
1
0.1
Healthy
Non-critical
Critical
ns
**
***
IL-1β (pg/mL)
10
1
0.1
0.01
0.001
Healthy
Non-critical
Critical
ns
ns
*
IL-4 (pg/mL)
1
0.1
0.01
0.001
Healthy
Non-critical
Critical
****
ns
****
IL-6 (pg/mL)
10000
1000
100
10
1
0.1
Healthy
Non-critical
Critical
*
****
***
IL-8 (pg/mL)
10000
1000
100
10
1
Healthy
Non-critical
Critical
****
*
***
IL-10 (pg/mL)
1000
100
10
1
0.1
0.01
Healthy
Non-critical
Critical
ns
ns
*
IL-12p70 (pg/mL)
1000
100
10
1
0.1
0.01
Healthy
Non-critical
Critical
*
ns
ns
IL-17 (pg/mL)
100
10
1
0.1
0.01
0.001
Healthy
Non-critical
Critical

Figure 2A

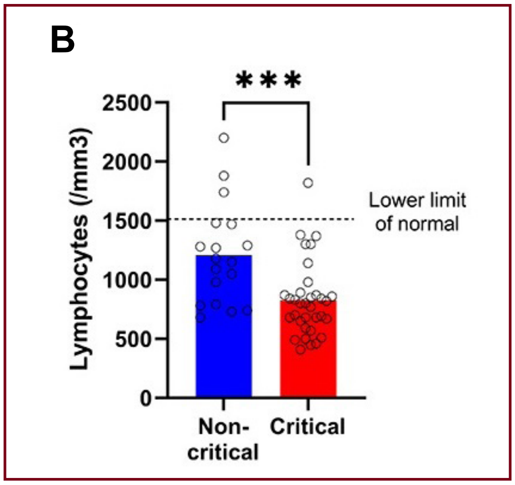


Figure 2B

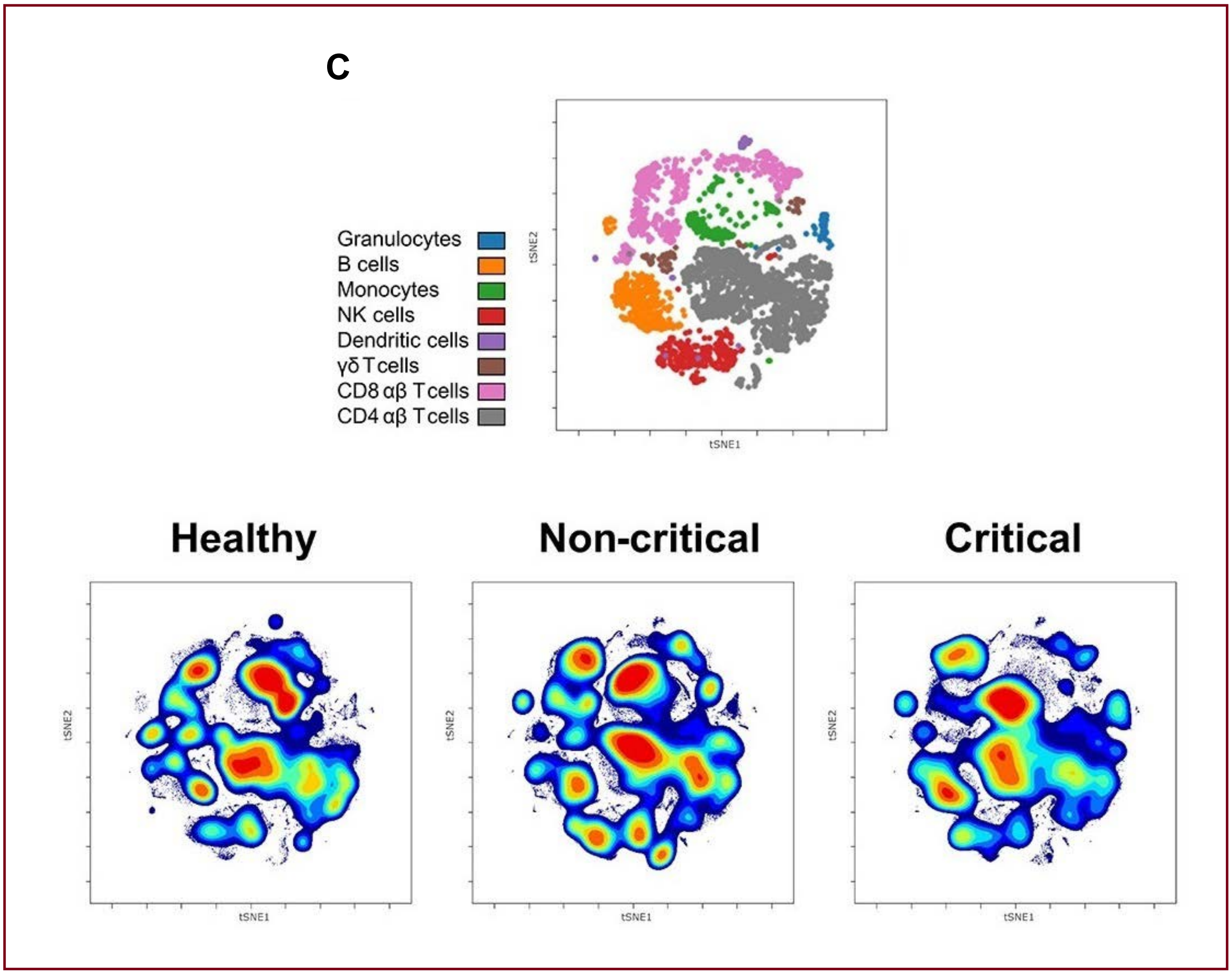


Figure 2C

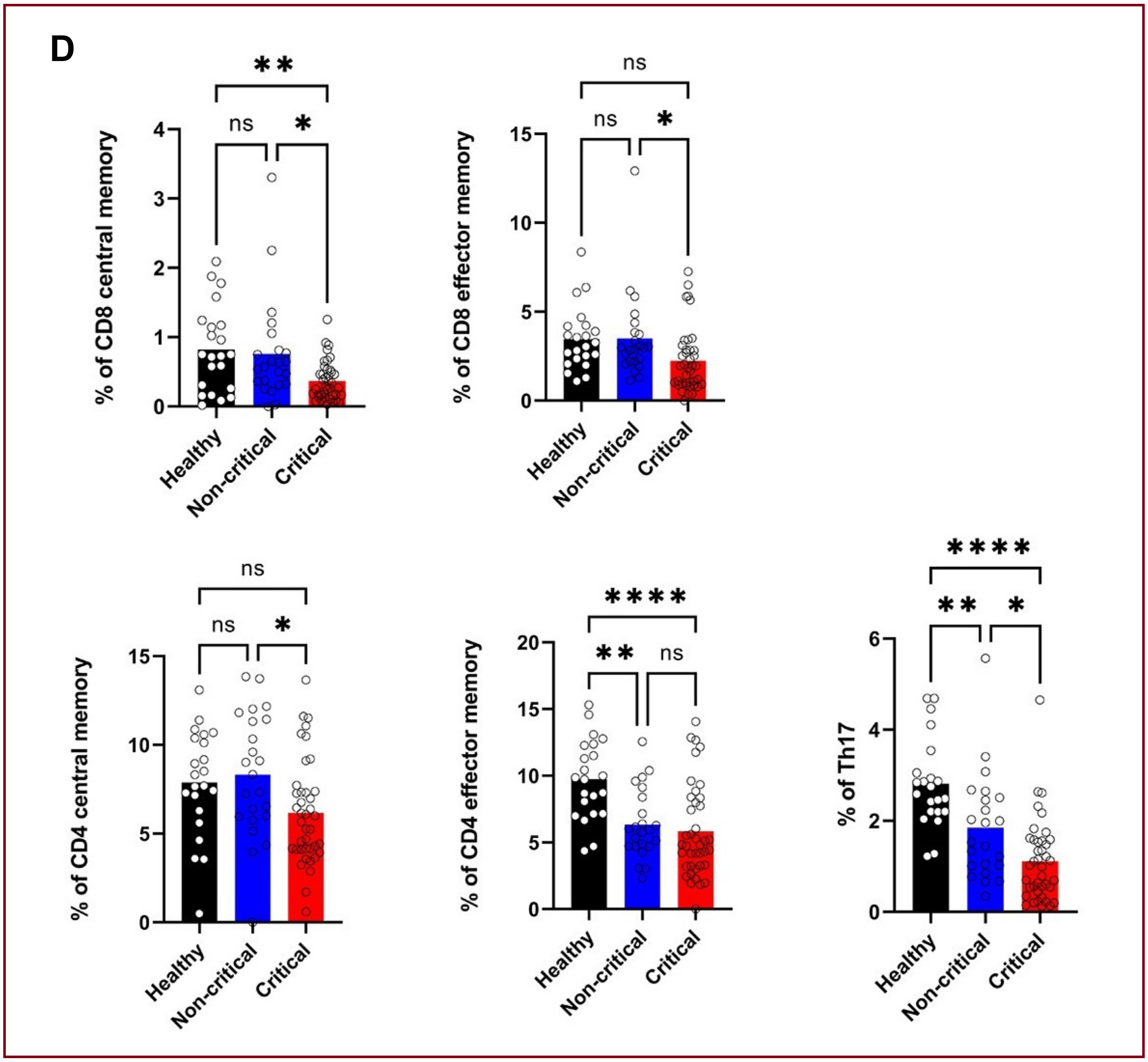


Figure 2D

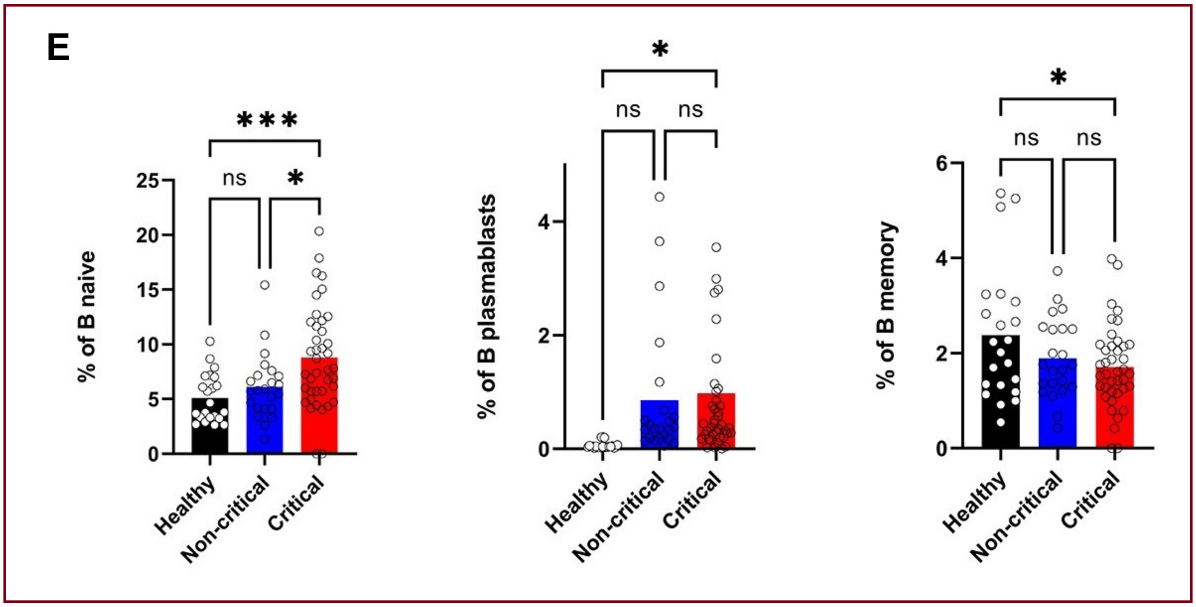


Figure 2E

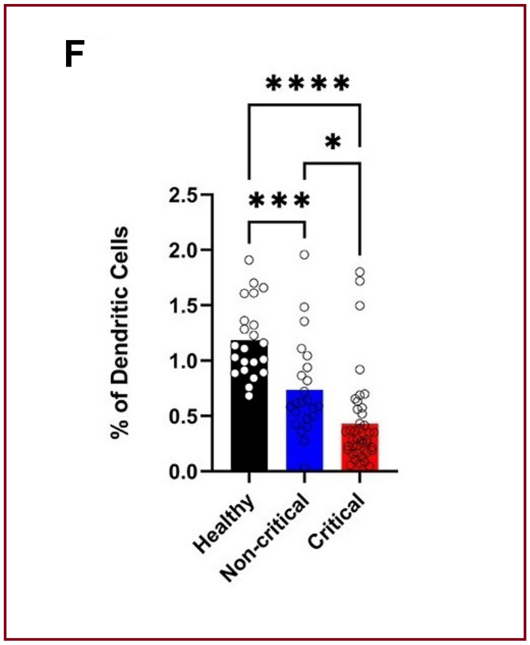


Figure 2F

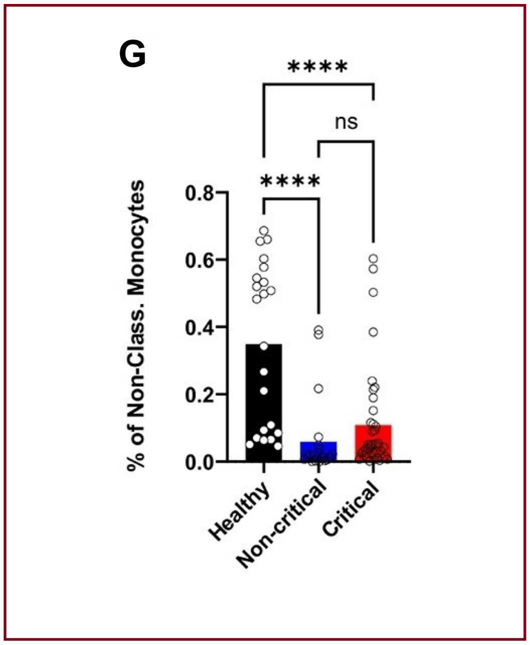


Figure 2G

Fig. 2. Immune profiling differentiates healthy individuals, non-critical patients with COVID-19 and critical patients with COVID-19.

(A) The concentrations of proinflammatory cytokines in plasma were quantified by cytokine profiling assays or ELISA.

(B) Absolute lymphocyte counts are shown. Each dot represents a single patient. The dashed horizontal line indicates the lower limit of normal lymphocyte concentrations.

(C) viSNE maps are shown colored according to the cell density across the three groups. Red indicates the highest density of cells. The plots are representative of 40 critical patients, 23 non-critical patients and 22 healthy controls.

(D to G) The proportions of modified lymphocyte subsets from patients with COVID-19 and healthy controls were determined by mass cytometry. Proportions of T cell subsets (**D**), B cell subsets (**E**), dendritic cells (**F**) and nonclassical monocytes (**G**) are shown. Each dot represents a single patient. In (A) and (D-G), the *P*-values were determined with the Kruskal-Wallis test followed by Dunn's posttest for multiple group comparisons; *$P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$; ns, not significant. In (B), the *P*-value was determined by a two-tailed unpaired *t* test; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$; ns, not significant. In (A), data are shown as box-and-whiskers plots with medians, 25th to 75th percentiles, maximal and minimal values, and include n=41 critical patients, n=24 non-critical patients and n=21 healthy controls. In (B), (D) and (E to G), all data points are shown and bars represent means with n=40 critical patients, n=23 non-critical patients and n=22 healthy controls.

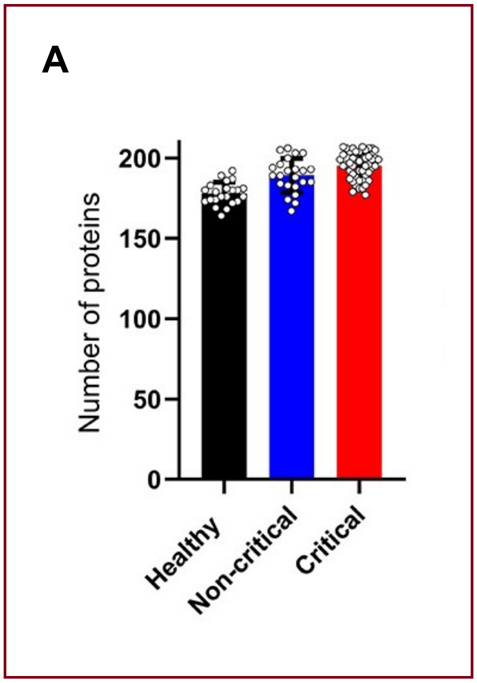


Figure 3A

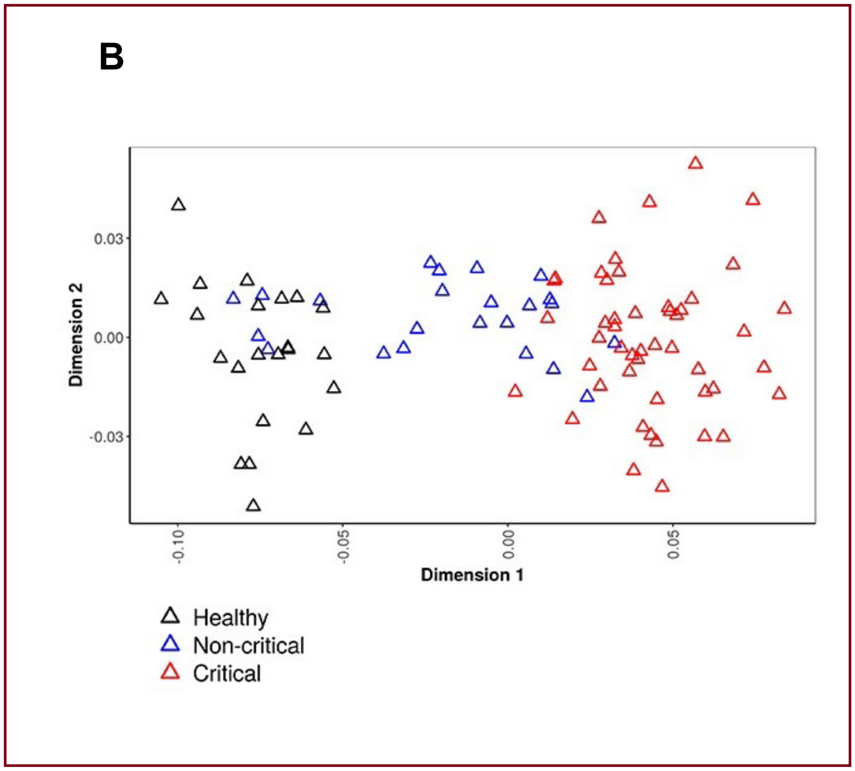


Figure 3B

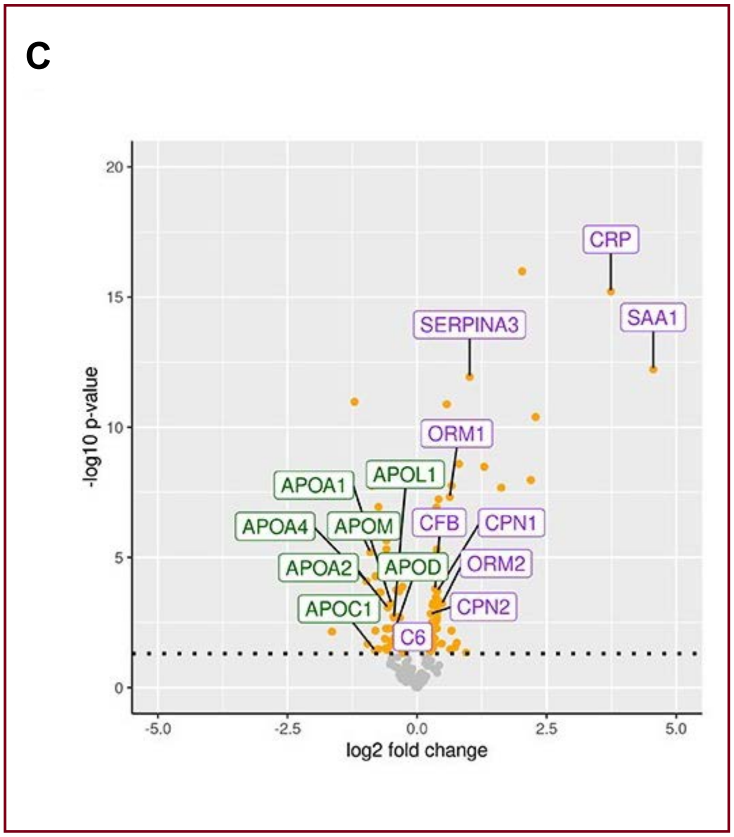


Figure 3C

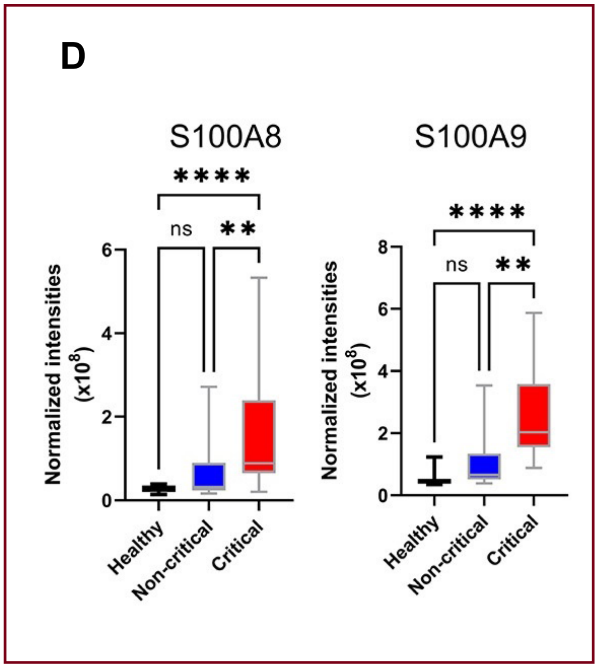


Figure 3D

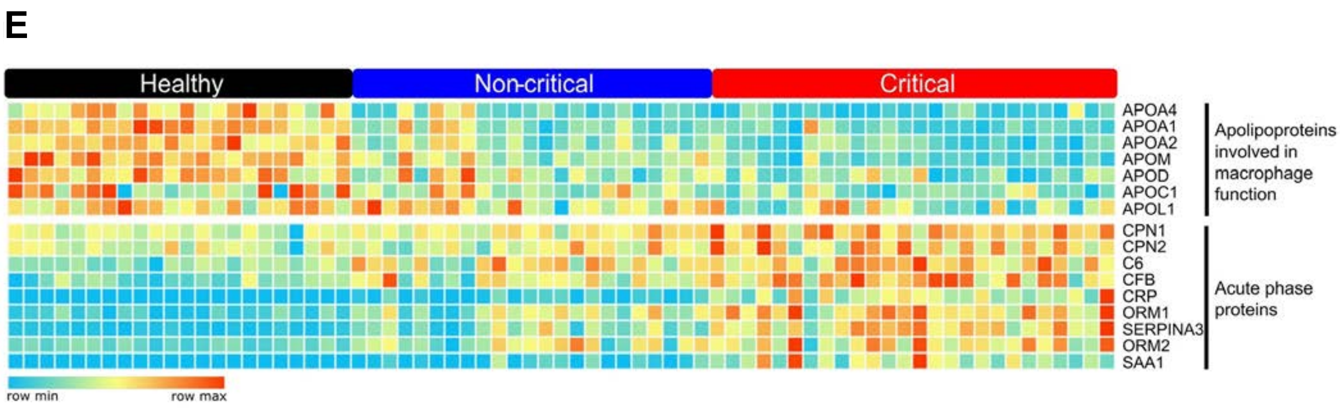


Figure 3E

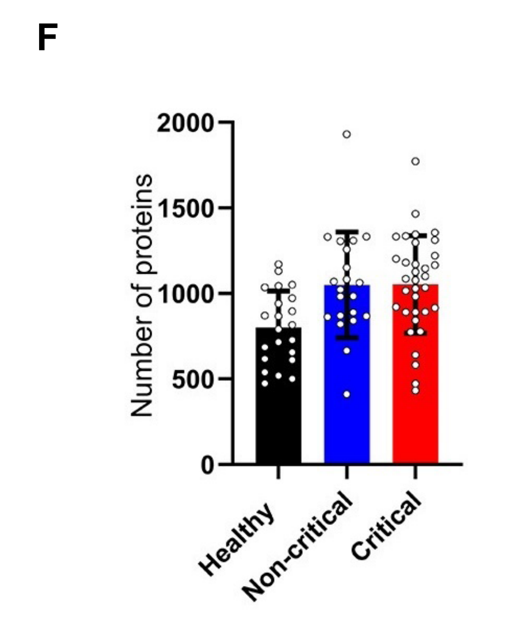


Figure 3F

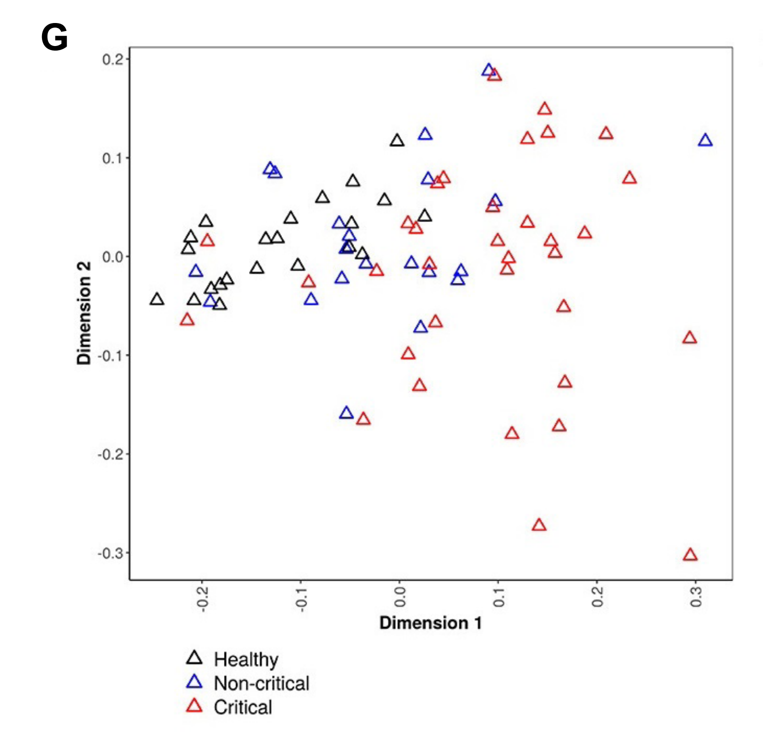


Figure 3G

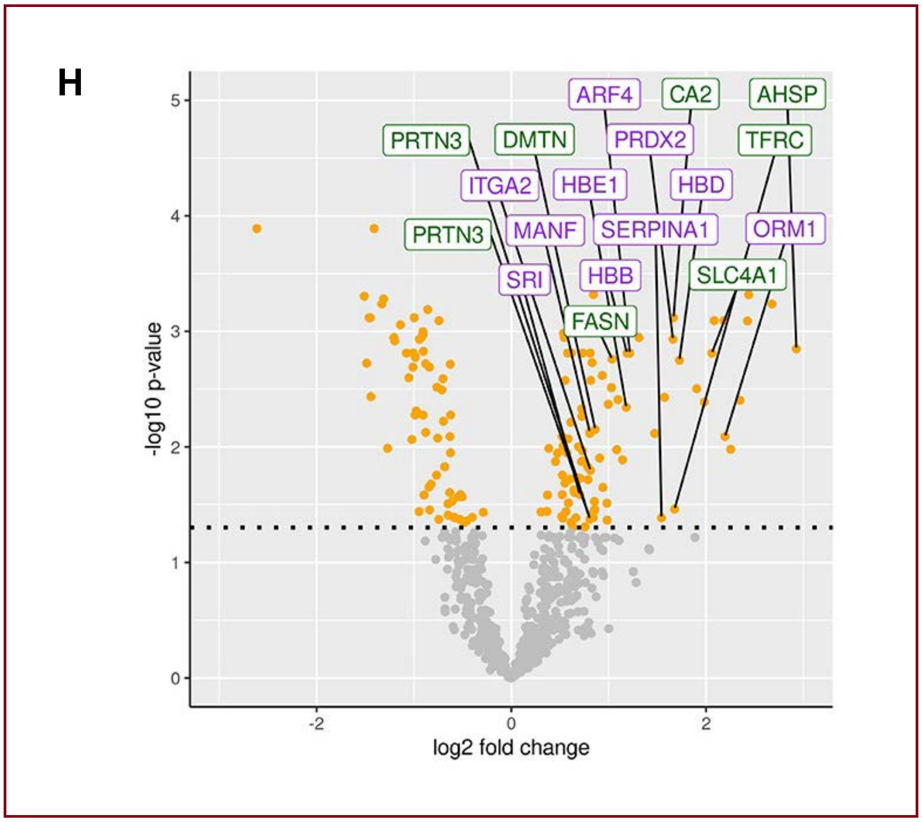


Figure 3H

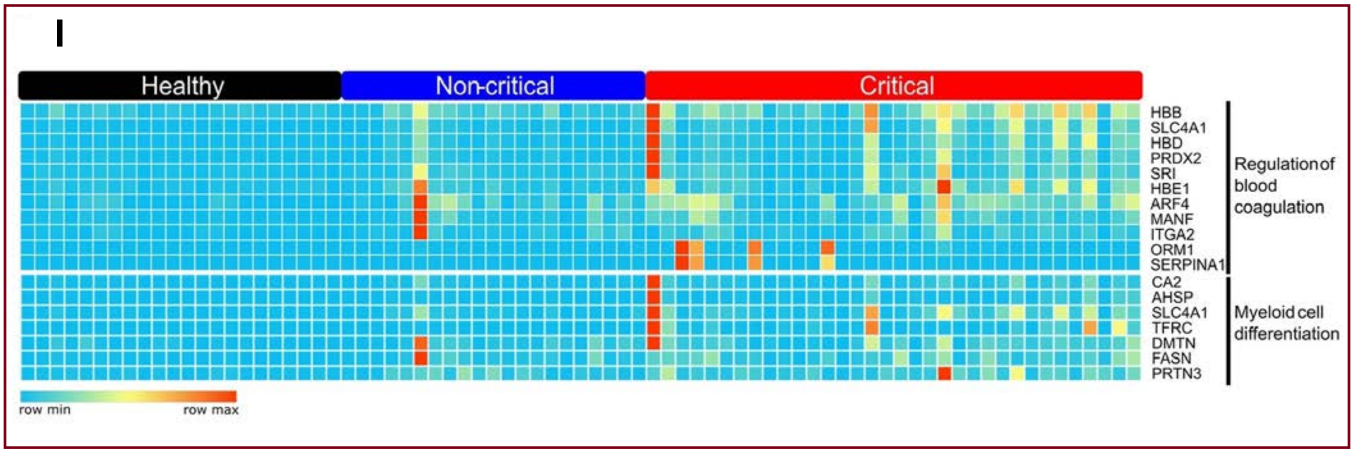


Figure 3I

Fig. 3. Plasma and PBMC proteomics distinguish healthy individuals, non-critical patients with COVID-19 and critical patients with COVID-19.

(A) The total number of proteins identified and used for quantification and differential analysis in the plasma of patients and healthy controls is shown. Each dot represents a patient. Bars represent means ± standard deviations.

(B) A multidimensional scaling plot of the normalized intensities of all individuals in the three groups is shown.

(C) A volcano plot representing the differentially expressed proteins (DEPs) in critical versus non-critical patients is shown. The orange dots represent the proteins that are differentially expressed with a corrected *P*-value < 0.05. Proteins labeled in green and purple represent down-regulated apolipoproteins and up-regulated acute phase proteins, respectively.

(D) Normalized intensities of the proteins S100A8 and S100A9 in the three groups are shown. Data are shown as box-and-whiskers plots with medians, 25th to 75th percentiles, maximal and minimal values, and include n=45 critical patients, n=23 non-critical patients and n=22 healthy controls. *P*-values were determined with the Kruskal-Wallis test, followed by Dunn's posttest for multiple group comparisons; **P* < 0.05, ** *P* < 0.01, *** *P* < 0.001, **** *P* < 0.0001; ns, not significant.

(E) The heatmap shows the expression of apolipoproteins involved in macrophage functions and acute phase proteins in the three groups. Up-regulated proteins are shown in red and down-regulated proteins are shown in light blue.

(F) The total number of proteins identified and used for quantification and differential analysis in PBMCs of patients and healthy controls is shown. Each dot represents a patient. Bars represent means ± standard deviations.

(G) A multidimensional scaling plot of the normalized intensities of all patients/individuals in the three groups is shown.

(H) A volcano plot representing the DEPs in critical versus non-critical patients is shown. The orange dots represent the proteins that are differentially expressed with a corrected *P*-value < 0.05. Proteins labeled in green and purple are up-regulated proteins involved in the regulation of blood coagulation and myeloid cell differentiation, respectively.

(I) The heatmap shows the expression of proteins involved in the regulation of blood coagulation and myeloid cell differentiation in the three groups. Up-regulated proteins are shown in red and down-regulated proteins are shown in light blue.

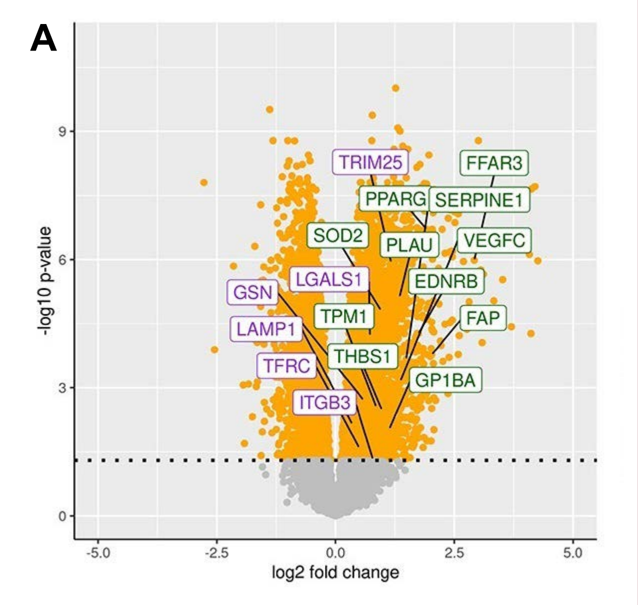
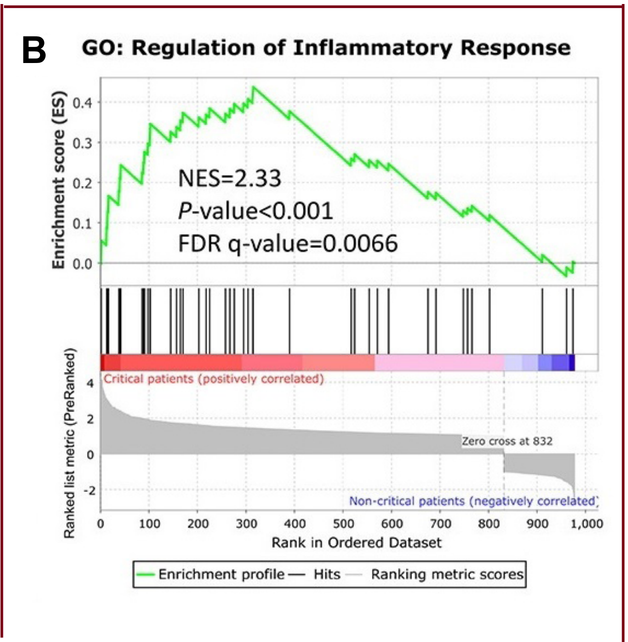


Figure 4A



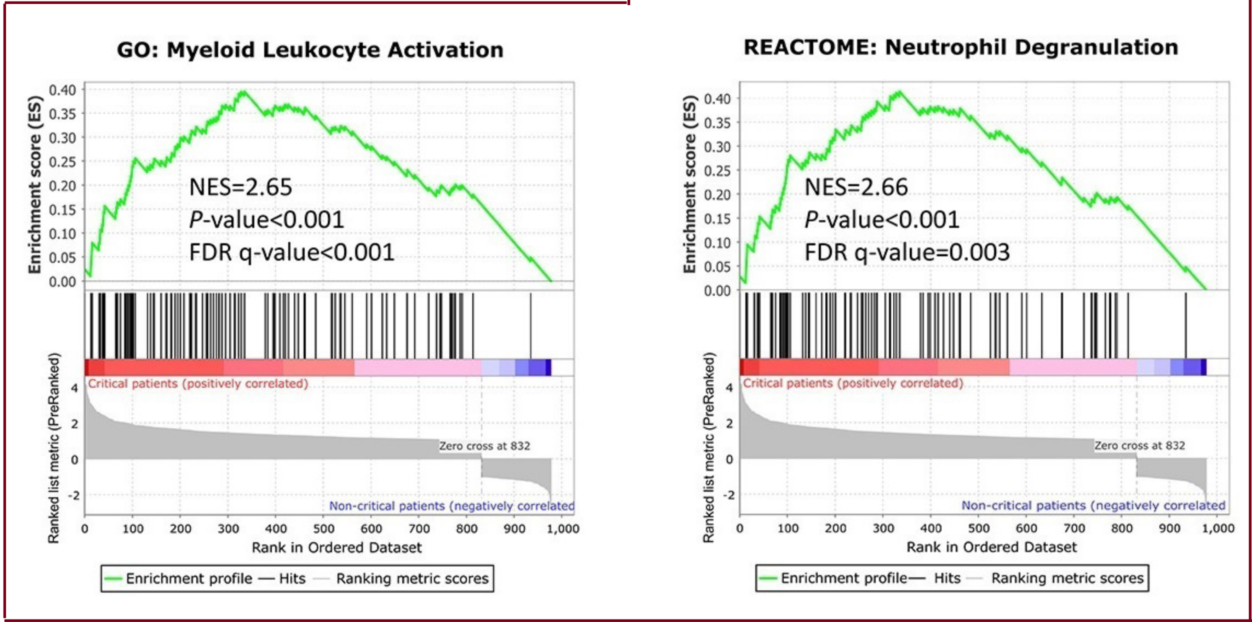


Figure 4B

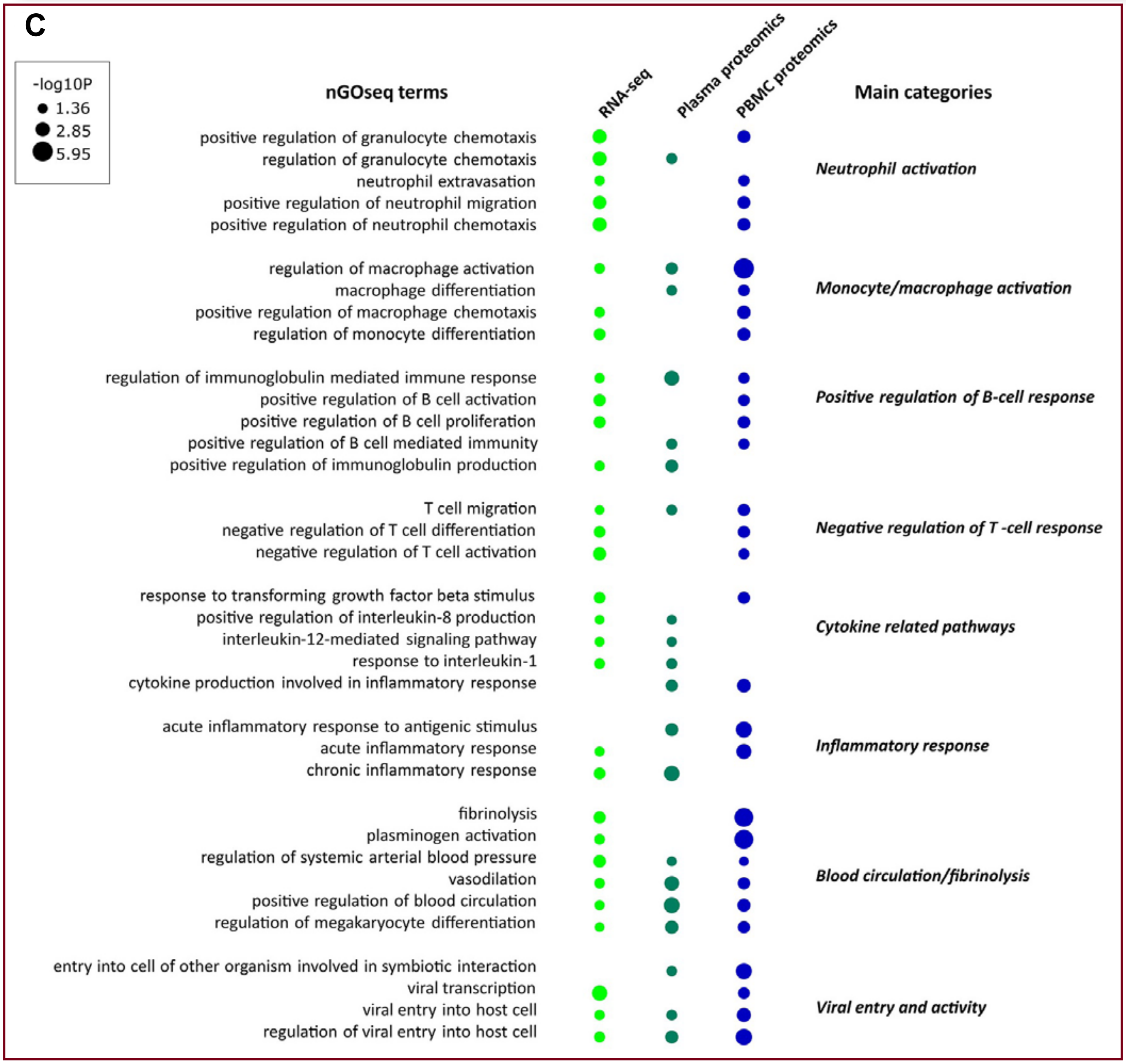


Figure 4C

Fig. 4. RNA-seq and combined omics analysis reveal critical patient-specific pathways.

(A) A volcano plot representing the differentially expressed genes in critical versus non-critical patients is shown. The orange dots represent the genes that are differentially expressed with a corrected *P*-value < 0.05. Proteins labeled in green and purple represent up-regulated genes involved in blood pressure regulation and viral entry, respectively.

(B) Gene set enrichment analysis plots show positive enrichment of inflammatory response, myeloid leukocyte activation and neutrophil degranulation pathways in samples from patients with critical COVID-19. NES, normalized enrichment score.

(C) Enriched nested gene ontology (nGO) categories are shown for critical versus non-critical patients using RNA-seq, plasma proteomics and PBMC proteomics data.

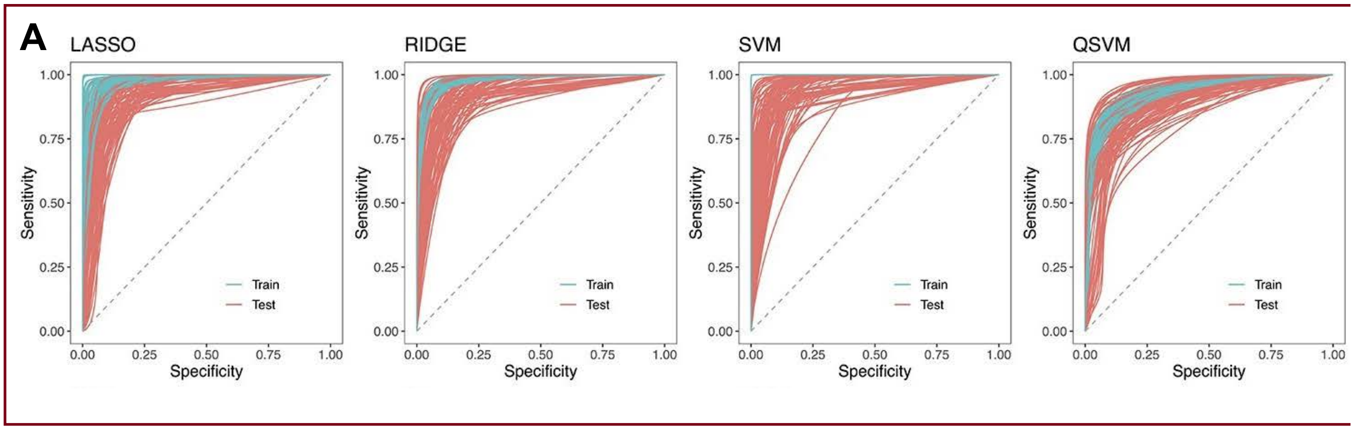


Figure 5A, left half

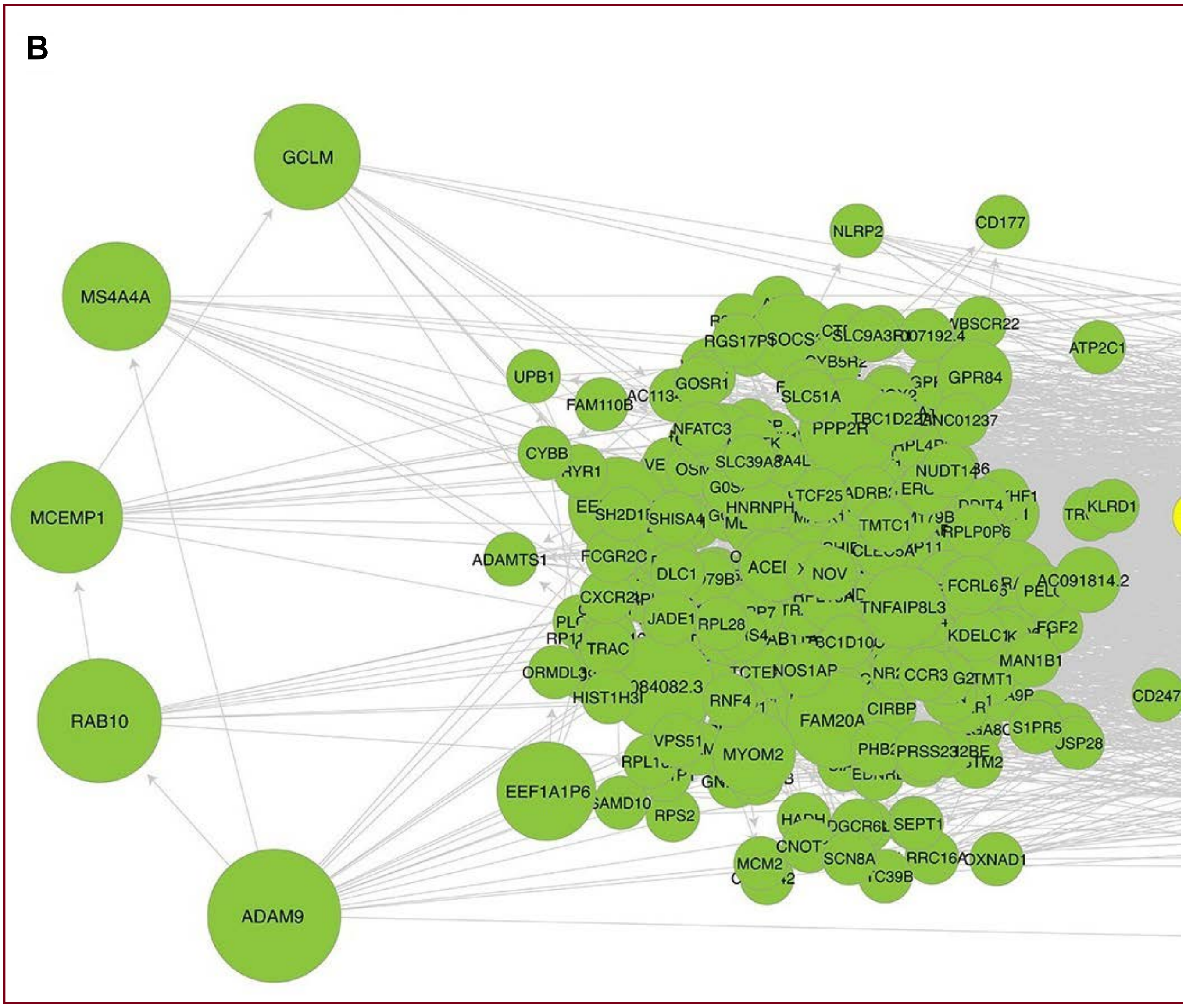


Figure 5B, left half

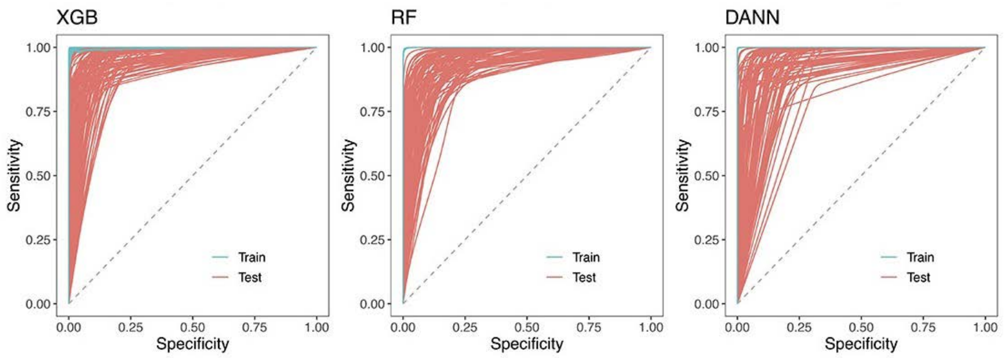


Figure 5A, right half

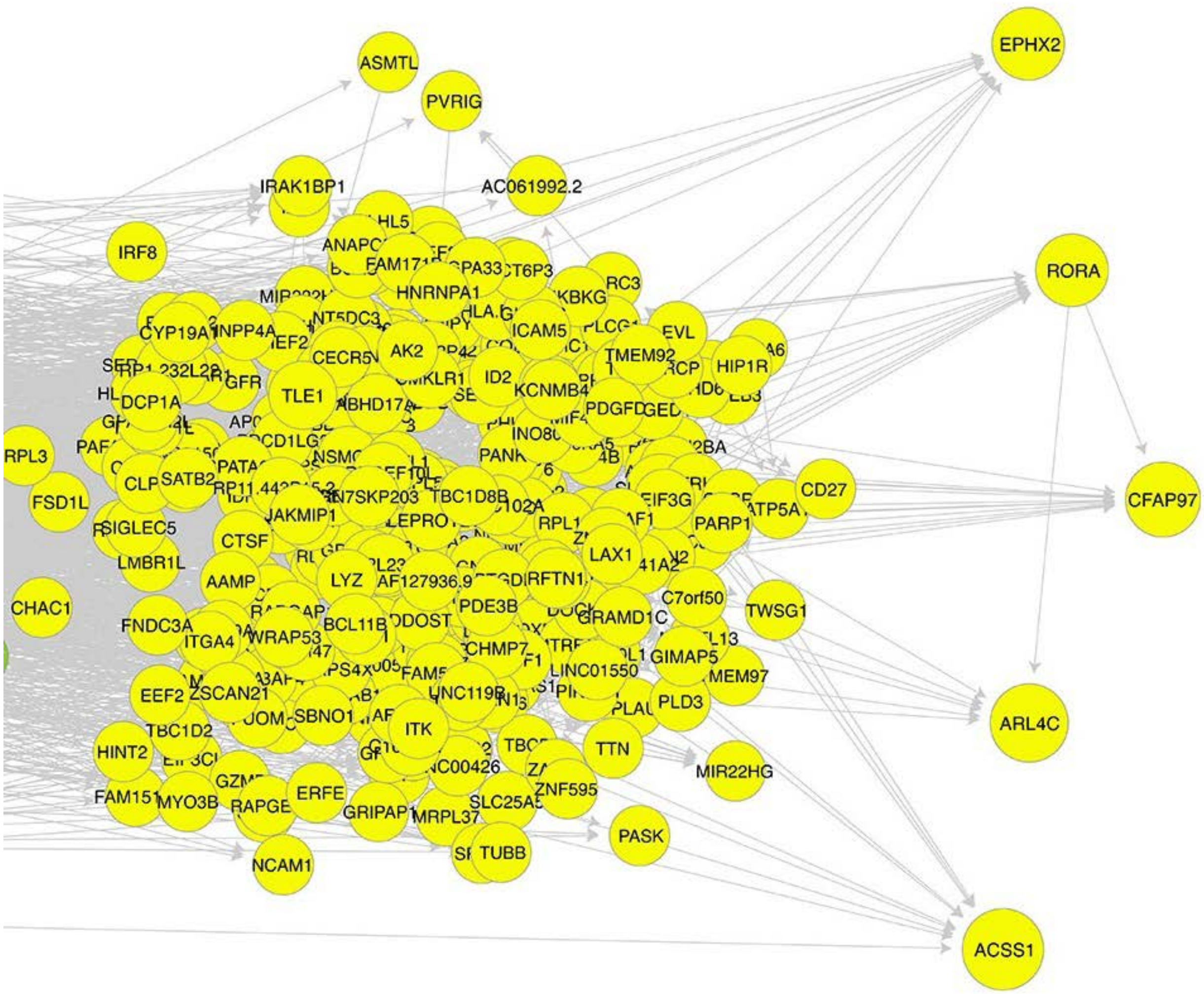


Figure 5B, right half

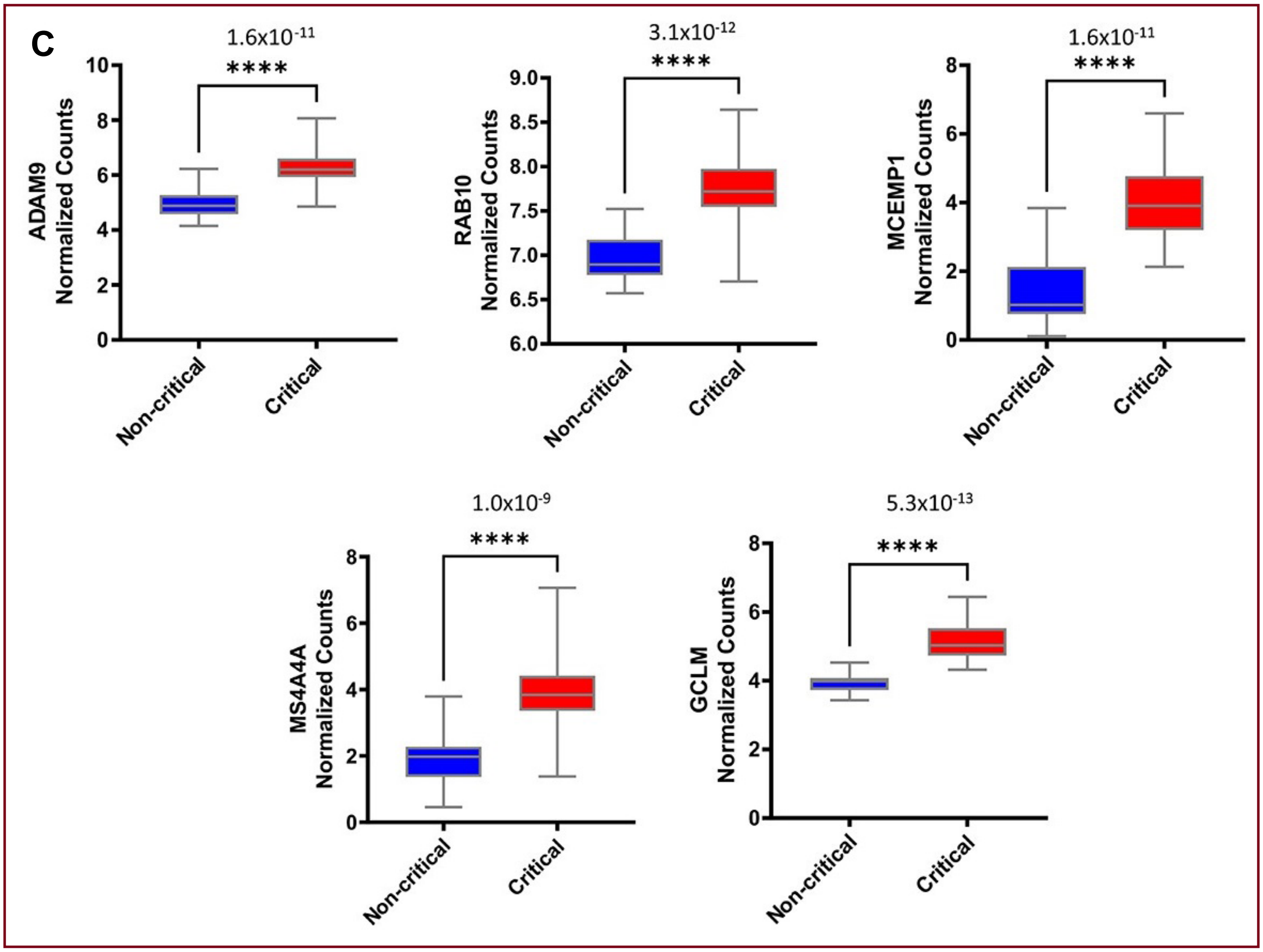


Figure 5C

Fig. 5. Integrated AI, ML, and probabilistic programming distinguishes non-critical and critical patients with COVID-19.

(A) ROCs of the train and test sets for critical versus non-critical comparisons are shown each of the seven modeling methods. All methods performed similarly. Other classification metrics are provided in table S3.

(B) A putative network shows the flow of causal information based on the top 600 most informative genes for classifying RNA-seq data of critical versus non-critical patients.

(C) Box plots show the normalized gene counts of the five driver genes identified that distinguish critical and non-critical patients. The indicated values correspond to the FDR. Data are shown as box-and-whiskers plots with medians, 25th to 75th percentiles, maximal and minimal values, and include n=46 critical and n=23 non-critical patients.

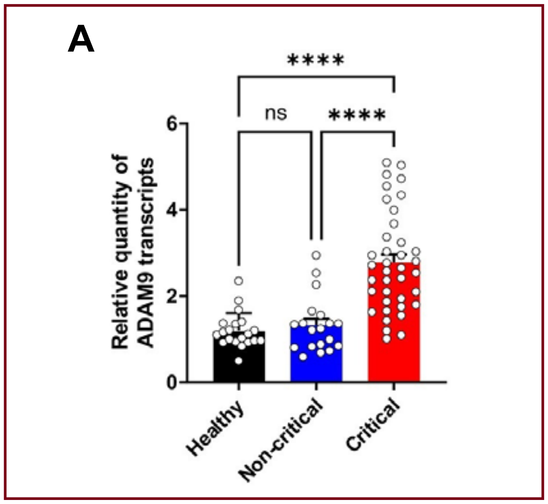


Figure 6A

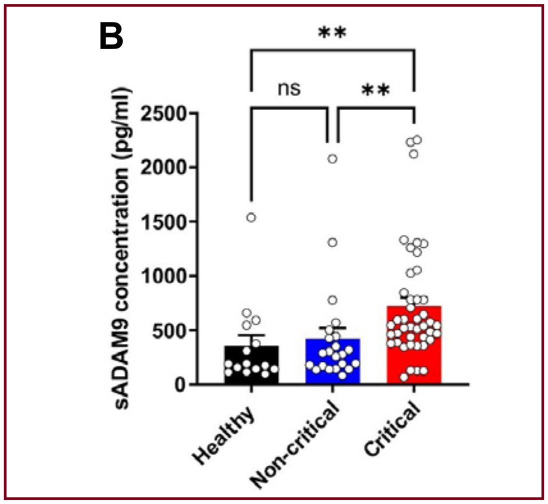


Figure 6B

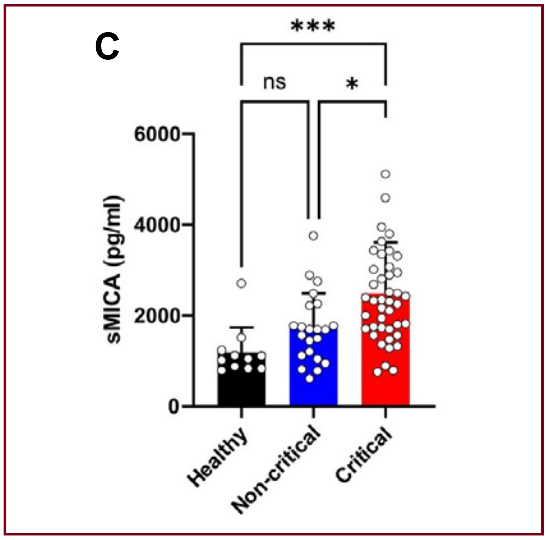


Figure 6C

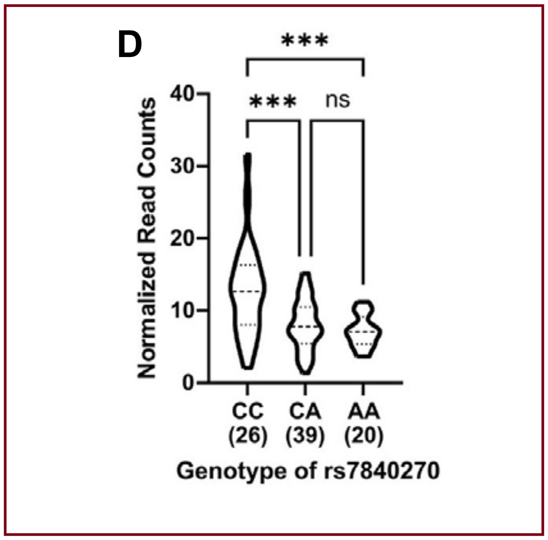


Figure 6D

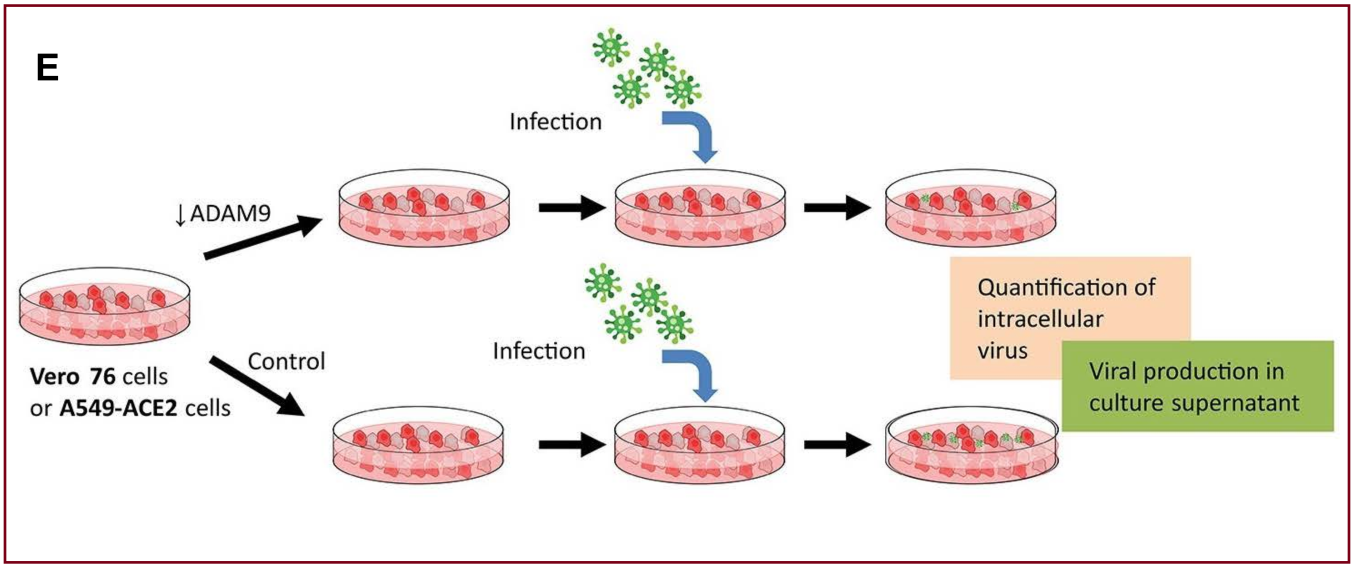

E
Infection
↓ADAM9
Vero 76 cells
or A549-ACE2 cells
Control
Infection
Quantification of intracellular virus
Viral production in culture supernatant

Figure 6E

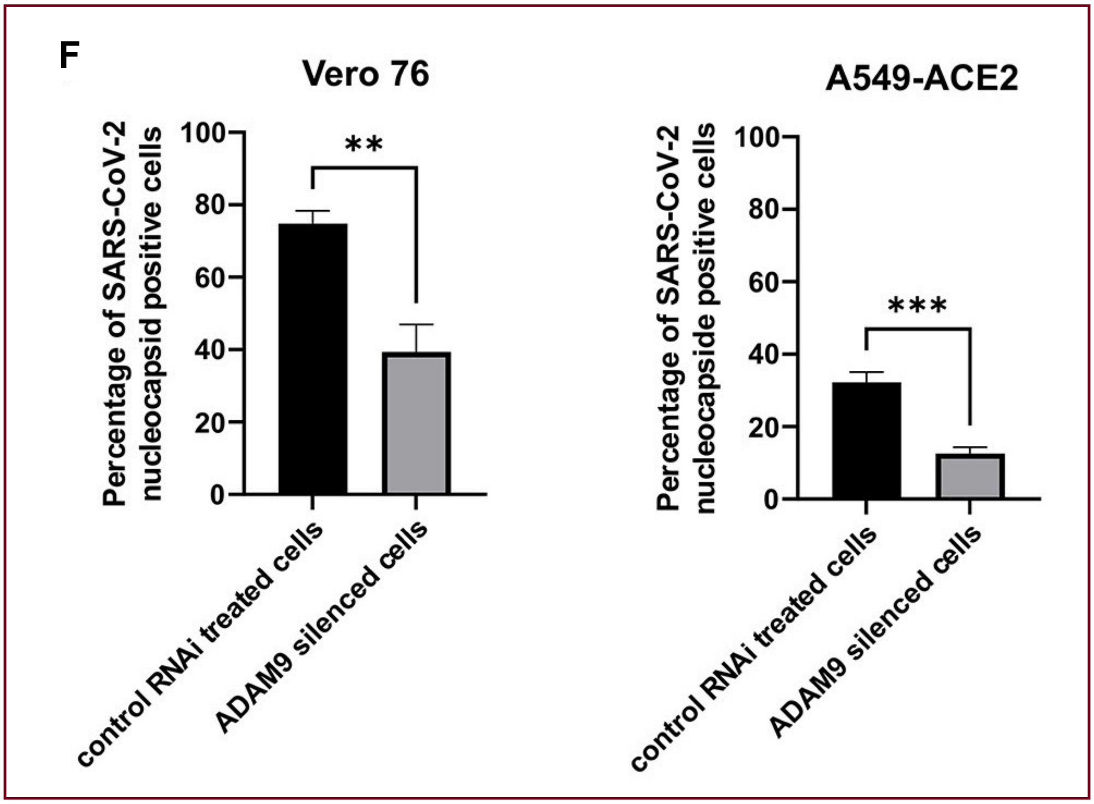


Figure 6F

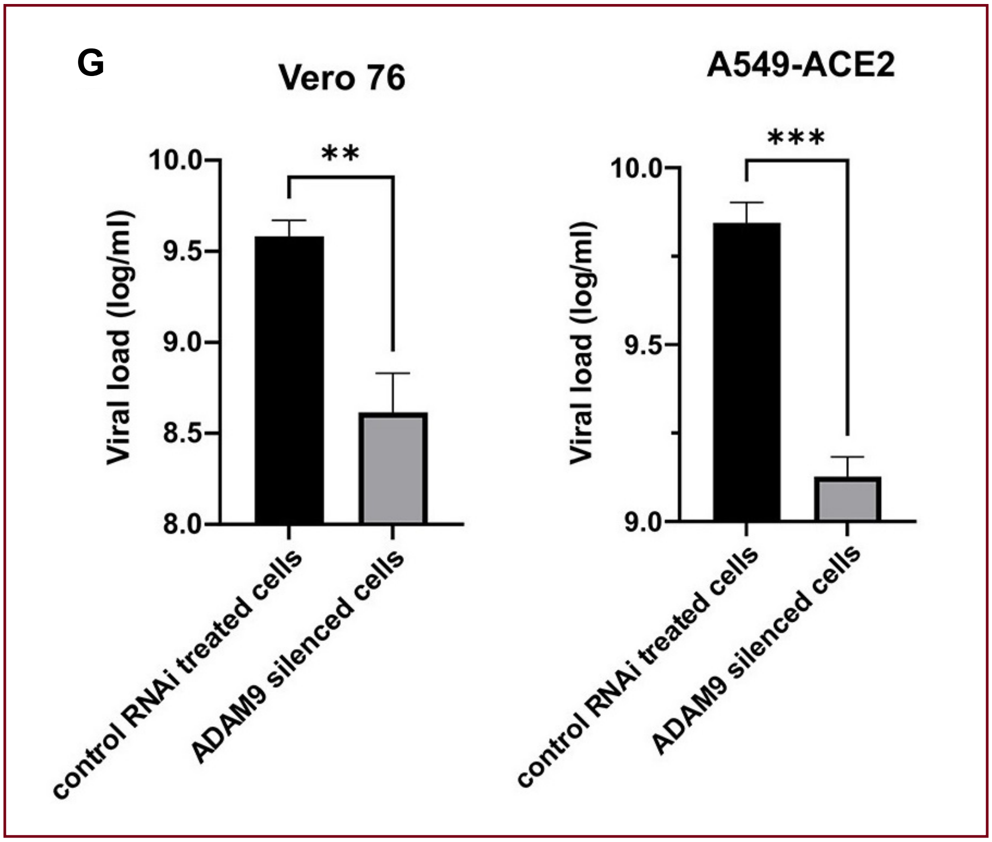


Figure 6G

Fig. 6. ADAM9 is a key driver of SARS-CoV-2 infection and replication in vitro.

(A) Quantitative RT-PCR confirmation of the differential expression of ADAM9 in non-critical (n=19) versus critical patients (n=38) and in healthy controls (n=20) is shown.

(B) Soluble ADAM9 (sADAM9) concentration in serum samples isolated from healthy controls (n=15), non-critical (n=22) and critical patients (n=43) was determined by ELISA.

(C) Soluble MICA concentration (sMICA) in serum samples isolated from healthy controls (n=11), non-critical (n=22) and critical patients (n=43) was determined by ELISA.

(D) Expression of ADAM9 according to the genotype of the eQTL rs7840270 is shown (n are indicated below the genotypes).

(E) The experimental approach for assessing viral uptake and viral replication in silenced Vero 76 or A549-ACE2 cells is shown.

(F) Flow cytometry-based intracellular nucleocapsid staining in control and ADAM9-silenced Vero 76 and A549-ACE2 cells was quantified. One representative experiment of N=3 independent experiments with n=3 in each group is shown.

(G) Quantitative RT-PCR for SARS-CoV-2 in culture supernatant after the silencing of *ADAM9* in Vero 76 or A549-ACE2 cells is shown. The results from probe N1 are shown. One representative experiment of N=3 independent experiments with n=3 in each group is shown. In (A to D), the *P*-values were determined with the Kruskal-Wallis test followed by Dunn's posttest for multiple group comparisons; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$; ns, not significant. In (F to G), the *P*-values were determined from a two-tailed unpaired *t* test; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$. In (A to C) and (F to G) bars represent means ± standard deviations.

The complete article and all of its supplementary materials can be accessed online at https://www.science.org/doi/10.1126/scitranslmed.abj7521. Thank you, Dr. Thomas Chittenden, for sharing this article with your fellow Thousanders. **Ω**