Unsupervised clustering of SARS-CoV-2 positive hospitalized patients identifies six endophenotypes of COVID-19 and points to FGFR and SHC4-signaling in acute respiratory distress syndrome

William Ma[1,*], Antoine Soulé[1,*], Katelyn Liu[2], Catherine Allard[3], Karine Tremblay[4, †], Simon Rousseau[2, †], and Amin Emad[1, 5, †]

[*] These authors contributed equally.

[1] Department of Electrical and Computer Engineering, McGill University, Montréal, QC, Canada

[2] The Meakins-Christie Laboratories at the Research Institute of the McGill University Heath Centre Research Institute, & Department of Medicine, Faculty of Medicine, McGill University, Montréal, QC, Canada

[3] Statistical department, Centre de recherche du Centre hospitalier universitaire de Sherbrooke (CRCHUS), Sherbrooke, Canada.

[4] Pharmacology-physiology Department, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Saguenay, QC, Canada; Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean, Saguenay, QC, Canada; CRCHUS, Sherbrooke, Canada.

[5] Mila, Quebec AI Institute, Montréal, QC, Canada

† Corresponding Authors:
Amin Emad,
755 McConnell Engineering Building, 3480 University Street, Montreal H3A 0E9, Canada
Email: amin.emad@mcgill.ca

Simon Rousseau,
RI-MUHC, E M3.2244, 1001 Décarie, Montréal H4A 3J1, Canada,
Email: simon.rousseau@mcgill.ca

Karine Tremblay,
Pavillon des Augustines, local AUG-5-01A, 225 St-Vallier street, Chicoutimi G7H 7P2, Canada
Email: karine.tremblay@usherbrooke.ca

Abstract

Defining the molecular mechanisms of novel emerging diseases like COVID-19 is crucial to identify treatable traits to improve patient care. To circumvent a priori bias and the lack of in-depth knowledge of a new disease, we opted for an unsupervised approach, using the detailed circulating proteome, as measured by 4985 aptamers (SOMAmers), of 731 SARS-CoV-2 PCR-positive hospitalized participants to *Biobanque québécoise de la COVID-19 (BQC19)*. The consensus clustering identified six endophenotypes (EPs) present in this cohort, with varying degrees of disease severity. One endophenotype, EP6, was associated with a greater proportion of ICU admission, mechanical ventilation, acute respiratory distress syndrome (ARDS) and death. Clinical features of this endophenotype, showed increased levels of C-reactive protein, D-dimers, elevated neutrophils, and depleted lymphocytes. Moreover, metabolomic analysis supported a role for immunothrombosis in severe COVID-19 ARDS. Furthermore, the approach enabled the identification of Fibroblast Growth Factor Receptor (FGFR) and SH2-containing transforming protein 4 (SHC4) signaling as features of the molecular pathways associated with severe COVID-19. Finally, this information was sufficient to train an accurate predictive model solely based on clinical laboratory measurements, suggesting the use of blood markers as surrogates for generalizing these EPs to new patients and automating identification of high-risk groups in the clinic.

Introduction

The coronavirus disease 2019 (COVID-19) is a new human disease caused by the coronavirus SARS-CoV-2 infection that has been assessed pandemic by the World Health Organization in March 2020. As SARS-CoV-2 infection spread, a breath of outcomes of the infection became apparent, from asymptomatic individuals to severely ill and dying, from complete recovery to long-lasting symptoms[1,2]. The emergence of novel diseases such as COVID-19 presents the medical and scientific community with numerous challenges. Among them, defining the molecular mechanisms of disease related to specific outcomes is important to identify treatable traits and improve the performance of healthcare systems facing the challenges brought by the pandemic.

Successfully reaching this precision medicine goal requires a more granular definition of the pathology. A symptom-based method to discover molecular mechanisms of the disease may result in a challenge emerging from the fact that the same higher-level phenomenon, such as COVID-19 severity, can be produced by several different molecular mechanisms, a phenomenon termed the «many-one» limitation[3]. Recent advances in computing strategies, such as machine learning, has enabled the development of methods that help overcome this limitation by starting from molecular profiles instead of symptoms to define endophenotypes, *i.e.* subgroups of individuals who are inapparent to traditional methods but share a common set of molecular factors that can lead to treatable traits[4]. Establishing successful treatment strategies requires a tailored approach to the underlying molecular mechanisms that can help predict and alter disease trajectories[5]. Endophenotypes can become apparent using extensive molecular

phenotyping combined with machine learning algorithms[6-8]. Current investigations of endophenotypes in COVID-19 have mainly relied on supervised approaches using fixed outcomes (such as disease severity) and integrating clinical variables at the onset[9]. We hypothesize that using an unsupervised approach exploiting a rich molecular dataset can provide novel mechanistic insights into the pathobiology of severe COVID-19 that can help physicians improve diagnosis and prognosis.

Therefore, this study aims to identify endophenotypes linked to diverse clinical trajectories of COVID-19 using the extensive molecular phenotyping of a cohort of 731 SARS-CoV-2 positive hospitalized patients from the *Biobanque québécoise de la COVID-19* (BQC19, www.quebeccovidbiobank.ca)[10], a prospective observational cohort of SARS-CoV-2 positive and negative participants recruited in the province of Québec, Canada, to improve our understanding of COVID-19 pathobiology and our capacity to alter disease outcomes.

In this manuscript, we report the identification of six endophenotypes in hospitalized SARS-CoV-2 positive participants to BQC19, associated with different clinical trajectories. The molecular information underpinning these endophenotypes were used to increase our understanding of pathobiology and predict the likelihood of patients admitted to the hospital to belong to each endophenotype using clinical blood workups.

Results

Unsupervised clustering of SARS-CoV-2-positive hospitalized BQC19 participants reveal endophenotypes associated with varying disease severity

In this study, we aimed to identify endophenotypes of COVID-19 based on the circulating proteome of patients in our cohort of SARS-CoV-2 positive hospitalized participants to BQC19 (n = 1,362, Table 1), using an unsupervised approach. Figure S1 shows the distribution of the time of hospital admission of the patients and the corresponding waves as defined by National Institute of Public Health of Quebec (https://www.inspq.qc.ca/covid-19). For this purpose, we performed consensus agglomerative clustering of the subset of patients (n = 731, Table S1) for whom data corresponding to circulating proteome measured by a multiplex SOMAmer affinity array (Somalogic, ~5,000 aptamers)[11] was available in BQC19. The remaining samples were kept aside for follow-up analysis. First, the optimal number of clusters (k = 6) was identified using two criteria, Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) (Figure 1A); then, consensus agglomerative clustering (Euclidean distance and Ward linkage)[12,13] using bootstrap subsampling was performed to obtain six robust clusters (see Methods for details) (Figure 1, Figures S2 and S3).

The clinical and pathological characteristics of patients in each endophenotype is provided in Table S1. To characterize the identified endophenotypes (EPs) with respect to disease severity, we performed two-sided Fisher's exact test to assess their enrichment (or depletion) in "severe" or "dead" outcome. EP6 was significantly enriched in the severe/dead outcomes (Benjamini-Hochberg false discovery rate (FDR) = 1.72E-18) with these outcomes observed in 66.1% of EP6

patients. Meanwhile, EP1 was significantly depleted in severe/dead outcomes (FDR = 8.23E-11) (Figure 2A, Table S1) with these outcomes only observed in 11.6% of EP1 patients. In addition, EP6 was enriched in participants 1) receiving oxygen therapy (FDR = 6.24E-12), 2) receiving ventilatory support (FDR = 6.63E-12), and 3) being admitted to intensive care unit (ICU) (FDR= 1.26E-22) (Figure 2A, Table S1). Kaplan Meier analysis[14] also confirmed that the identified EPs have a distinct temporal pattern of admission to ICU (multivariate logrank test P = 5.06E-36), with EP1 (EP6) having the highest (lowest) chance of not being admitted to ICU (or die prior to that) in a 40-day span since their admission to the hospital (Figure 2B). A similar pattern can be observed when patients that died before admission to ICU were excluded (Figure S4, multivariate logrank test P = 6.55E-36). A two-sided Mann-Whitney U (MWU) test showed that patients in EP5 were generally older than other EPs (FDR = 6.94E-5), while EP3 included younger patients (FDR = 1.82E-4). However, EP6 (which had the most severe patients) did not show enrichment in older patients or individuals with high BMI (two-sided MWU FDR>0.05) (Figure 2C, Table S1).

These analyses revealed that the unsupervised approach was able to identify endophenotypes with distinct disease characteristics and outcomes using the circulating proteome of the patients. We identified EP6 as a group of participants with an increase in many measures of COVID-19 disease severity.

EP6 is enriched with BQC19's participants having acute respiratory distress syndromes

In accordance with increase disease severity, EP6 was also enriched in COVID-19 medical complications (two-sided Fisher's exact test): acute respiratory distress syndrome (ARDS) (FDR =

1.43E-10), acute kidney injury (FDR = 6.37E-7), bacterial pneumonia (FDR=2.13E-6), liver dysfunction (FDR = 9.32E-3), and hyperglycemia (FDR = 1.56E-2) (Figure 3, Table S2). The frequency of ARDS was just below 8% in EP1 compared to greater than 44% in EP6, making this complication a key feature of this cluster (Figure 3, Table S2).

EP6 is enriched in blood metabolites associated with severe COVID-19.

To further characterize each EP and gain insight into mechanisms of disease, metabolomic profiling of plasma samples was done in parallel to the SOMAmer analysis. The results yielded data on 1,435 metabolites, of which 576 were found significantly altered in EP6 (two-sided MWU FDR < 0.01). Moreover, the metabolomic characterization of the plasma samples supported the distinction in blood composition at the levels of metabolomic sub-pathways and individual metabolites between the different EPs (Figure 4 and S3).

Pathway enrichment analysis identifies FGFR-signaling in severe COVID-19 acute respiratory distress syndrome

To gain insight into the molecular mechanisms underlying the severity in EP6, we performed pathway enrichment analysis using the Knowledge Engine for Genomics (KnowEnG)[6] for the aptamers associated with EP6. Since aptamers were used to identify the EPs, it is expected that many of them would be significantly associated with the EPs. As a result, we selected a very strict threshold of FDR < 10E-20 (two-sided MWU test) to identify top aptamers associated with EP6 (Table S4). We then used the gene set characterization pipeline of KnowEnG with Reactome [15] pathway collection. This analysis showed that while EP6 is characterized by pathways associated

with Interleukins and Cytokine Signaling in Immune system, multiple instances linked it with Fibroblast Growth Factor Receptor (FGFR) signaling, identifying this pathway as a potential driver of severe pathology that was not present in other EPs (Tables 2 and S4).

SHC4 genotype and protein expression levels are associated with higher odds of belonging to EP6

To further improve our understanding of the molecular mechanisms underlying EP6, we leveraged an additional dataset of Genome Wide association Study (GWAS) corresponding to these patients. We identified 25 single nucleotide variations (SNVs) distributed in 13 annotated genes, that were below a p-value threshold of 1E-4 differentiating EP6 versus the rest (Table 3). We then investigated each of the SNVs to which we could assign a gene and an aptamer, to assess whether their protein product in circulation was differentially regulated by the genotype (Table 4). We discovered two genes, *SHC4* (encoding SHC adaptor protein 4) and *CACNA2D3* (encoding calcium voltage-gated channel auxiliary subunit alpha2 delta3) for which there was a significant association between genotype and protein expression levels (p-values < 0.05). While *CACNA2D3* may have mild impact on EP6 membership (odd ratio = 0.61, Table 4), SHC4 was one of the top-enriched aptamers (position 32 out of 4,985), with odds ratio of 11.98 and 2.00 for the protein product and SNV, respectively of belonging to EP6 for the alternative allele. Therefore, the GWAS analysis revealed that the signaling adaptor protein SHC4 may play an important mechanistic role in contributing to severe disease pathology.

A predictive model based on blood markers predicts EPs in a separate validation cohort

To further characterize each EP, we assessed the clinical laboratory results obtained from blood draws and compared them between the groups. We focused on 21 markers that were measured in at least 50% of the patients used for consensus clustering (Figure 5A and Table S5) and used the summary value reported in the BQC19 database corresponding to the most extreme measurement among multiple blood draws (Table S5 provides this information for each blood marker). Figure 5A shows the elevation and depletion of these markers in the identified EPs. EP6 is characterized by abnormal values in markers of inflammation (lymphopenia, total white blood cells count, neutrophilia, C-reactive protein (CRP)), liver damage (alanine aminotransferase (ALT), albumin, lactate deshydrogenase (LDH)), blood clotting disorder (D-dimers, low hemoglobin, International Normalized Ratio (INR) and hyperglycemia (glucose).

To identify relationships that may shed light on factors influencing clinical laboratory results defining EP6, we also performed Spearman's rank correlation analyses between each blood test values and metabolites (Table S5).

Since EPs and particularly EP6, which we identified as the EP with worst outcome, showed a clear and distinct clinical laboratory result signature compared to other EPs, we sought to develop a predictive model based on these signatures. Due to the large number of missing values for these markers in our cohort, we developed a nearest-centroid classifier that is capable of dealing with missing values and can predict EPs based on blood markers (see Methods for details). To test the ability of this model on prediction of EPs on an independent yet similar dataset, we used data corresponding to 631 SARS-CoV-2 positive hospitalized BQC19's participants that did not have

circulating proteome data and hence were not used to identify the endophenotypes. The clinical and pathological characteristics of patients in each predicted endophenotype (PEP) is provided in Figure 5B-5E and Table S6.

Our predictive model identified 116 of these patients to belong to EP6. Fisher's exact test showed a significant enrichment of the predicted EP6 (PEP6) in severity/dead (FDR = 1.77E-22), while PEP1 and PEP2 were significantly depleted in these outcomes (FDR = 1.61E-4 and FDR = 1.30E-8, respectively), as shown in Figure 5B and Table S6. Similar to EP6, PEP6 was also significantly enriched in participants 1) receiving oxygen therapy (FDR = 2.38E-8), 2) receiving ventilatory support (FDR = 4.40E-8), and 3) being admitted to ICU (FDR = 1.23E-24) (Table S6). Kaplan Meier analysis also confirmed that these PEPs have a distinct temporal pattern of admission to ICU (multivariate logrank test P = 1.05E-34), with PEP6 having the lowest chance of not being admitted to ICU (or die prior to that) in a 40-day span since their admission to the hospital (Figure 5E).

These results suggest that our predictive model can use these 21 blood markers to generalize the definition of endophenotypes to patients for whom the proteome data is unavailable.

Discussion

The results presented herein came from a large cohort of deeply phenotyped SARS-CoV-2 positive hospitalized participants, combined with unsupervised clustering that provided both expected and novel findings into the molecular mechanisms regulating COVID-19 outcomes.

They led to the identification of an endophenotype (EP6) associated with worst clinical outcome of COVID-19 (enriched in acute respiratory distress syndrome) reflected by a greater proportion of ICU admission, mechanical ventilation, and severe/death outcomes (Figure 1). Clinical features of this endophenotype were consistent with published literature with increased levels of CRP, D-dimers, elevated neutrophils, and depleted lymphocytes (Figure 5A, Table S5). Our approach enabled the identification of interleukins, FGFR and SHC4 signaling as cardinal features of the molecular pathways associated with severe COVID-19. Importantly, this information was sufficient to train an accurate predictive model that could in the future support clinical care.

The approach: unsupervised clustering capacity at identifying clinically meaningful subpopulations

Our unsupervised clustering approach in conjunction with a rich molecular dataset enabled us to identify endophenotypes that could not be captured using traditional methods classifying the population in two bins solely based on severity. This is in part because of the «many-one» limitation: the same higher-level phenomenon (COVID-19 severity) can be produced by several different molecular mechanisms. Determining endophenotypes using an unsupervised method provides a higher granularity and increases the chance to identify distinct molecular mechanisms and pathways resulting in similar COVID-19 severity. Accordingly, we identified two endophenotypes with more favorable outcomes (EP1 and EP2), three endophenotypes with intermediate outcomes in terms of severity (EP3, EP4 and EP5) and one endophenotype which led to worst outcomes compared to all others (EP6).

The identification of endophenotypes was done systematically using robust consensus clustering of aptamer expression levels in which the optimum number of clusters was determined congruently using two well-established measures: AIC and BIC. The consensus clustering using bootstrap sampling (1000 times) ensured identification of robust clusters that are not sensitive to exclusion of some of the samples (20% randomly selected and excluded at each cycle). Moreover, identifying the best number of clusters using AIC/BIC (both of which agreed with each other) allowed us to reveal the patterns of the EPs directly from the data, instead of imposing a pattern onto it through human supervision. This is an important strength of the study that enabled us to identify distinct molecular patterns of patients that could have remained undetected using other traditional approaches.

Moreover, to improve the translational applicability of EPs, we developed a predictive model based only on laboratory measured blood markers to generalize the definition of these endophenotypes to unseen samples without measured aptamer expression levels. Characteristics of EPs predicted solely based on their blood markers were consistent with the original EPs, suggesting the use of blood markers as surrogates for generalizing these EPs to new patients and automating identification of high-risk groups in the clinic.

COVID-19 molecular pathology

The datasets used in this study carry rich molecular information on mechanisms of disease. SARS-CoV-2 infection has been shown to be associated with altered kynurenin levels associated with increase IL-6 and kidney injury[16]. Interestingly, we also observed Kynurenin to be enriched in EP6

but depleted in EP1, supporting the association of increase kynurenin with COVID-19 severity and the enrichment in acute kidney injury complication.

Spermidine/spermine N(1)-acetyltransferase (SSAT) contributes to polyamine synthesis. In addition, its extracellular metabolite, N1,N12-diacetylspermine, is one of the top 15 metabolites enriched in EP6 (Figure 4A), is positively correlated with Urea and creatinine, and is negatively correlated with lymphocyte numbers (Table S5). Deletion of SSAT in mice is protective against LPS-induced kidney injury[17]. SSAT activity is associated with white blood cell count in Acute Myeloid Leukemia and Chronic Myeloid Leukemia patients[18]. This suggest that the tryptophan and polyamine metabolisms are associated with acute kidney injury in COVID-19 and identifies potential pathways of disease progression.

COVID-19 ARDS

EP6 is characterized by its enrichment in ARDS (44% vs 8% in EP1, Figure 3). Low levels of Sphingosine 1-phosphate (S1P) are associated with ARDS and were shown to be associated with greater ICU admission and decrease survival in COVID-19[19]. Accordingly, we found that S1P levels are depleted in EP6 while they are enriched in EP1 (Figure 4A). Interestingly, the aptamers detecting neutral ceramidase, an enzyme converting ceramides into sphingosine, is enriched in EP1 (although changes in protein abundance as detected by aptamers may not necessarily reflect changes in enzymatic activity). Accordingly, dihydroceramides and ceramides are depleted in cluster EP1. Conversely, dihydroceramides and ceramides are significantly enriched in EP6, suggesting that there is a shunting of the pathway away from sphingosine towards more pro-

inflammatory ceramides in EP6 associated with metabolic disorders[20]. Moreover, one of the top aptamers found enriched in EP6, is the enzyme Serine palmitoyltransferase 2 (SPTLC2) (Table S4)[21,22]. These results suggest a counterbalance between ceramides and sphingosine, where the former is associated with poorer outcomes during critical illness, whereas higher levels of the latter has more favorable outcomes in ARDS.

Metabolomic profile of EP6 supports a role for immuno-thrombosis-mediated organ damage in COVID-19

To obtain a more comprehensive understanding of the alteration in metabolic profiles in EP6, we investigated which metabolic subpathways were significantly enriched in EP6 (Figure 4, Table S3). The two top pathways (FDR < 1E-2) are "Methionine, Cysteine, SAM and Taurine Metabolism" and "phosphatidinylethanolamine (PE)". Interestingly, these two pathways are known to interact[23], with PE methylation a major consumer of S-Adenosylmethionine (SAM) leading to the synthesis of S-Adenosylhomocysteine (SAH) and cystathionine[24], which itself has been found upstream of 2-hydroxybutyrate and 2-aminobutyrate in a model of hepatoxicity[25]. SAH, cysthathionine, 2-hydroxybutyrate and 2-aminobutyrate are significantly enriched in EP6. Congruently, EP6 is enriched (FDR < 1E-2) in liver dysfunction (Figure 3, Table S2), with markers of liver dysfunction all significantly altered in clinical blood works: ALT, albumin, bilirubin and LDH (Figure 5A, Table S5).

PEs become exposed at the surface of cell membranes upon exposure to stress, inflammation, and cell death[26,27]. In a Syrian hamster model, infection with SARS-CoV-2 had markedly increased

PE expression in the animals that were fed a high salt, high fat diet, demonstrating the interaction between infection and metabolic disorder with the abundance of circulating PE[28]. Phospholipids-containing microparticles from platelet activation contribute to Tissue Factor activation and pro-thrombinase activity[29]. Platelets-derived microparticles have a much greater procoagulant activity than activated platelets[30]. Exposure of glycerophospholipids in conjunction with phosphatidinylserine (PS) enhances factor X activation and increases pro-thrombinase activities[31,32]. Interestingly, red blood cells exposed to paclitaxel, PS were exposed to the surface by protein kinase C (PKC) zeta activation of scramblase[33]. The aptamer detecting PKC zeta is one of the top aptamers associated with EP6 (Table S4). PKC zeta can be activated by testosterone, Dehydroepiandrosterone(DHEA)[34] and dexamethasone[35], signals of relevance to EP6. The membranes of azurophilic granules, which contains Cathepsin G representing the most enriched aptamer in EP6, are enriched in PE relative to PC[36]. Accordingly, EP6 in addition to significantly enhanced PE abundance, showed decrease platelets count, increased D-Dimers, INR and activated partial thromboplastin time (APTT) (Figure 5A, Table S5), hallmarks of Disseminated Intravascular Coagulation (DIC), a serious and often lethal complication of sepsis[37]. Liver lesions are frequently observed in DIC[38], where liver damage can cause DIC, or exacerbate its manifestation due to its function in clearing activated products of the coagulation cascade. Taken together, the metabolomic profile of EP6 supports pro-coagulation activity in circulation that can be linked to organ damage.

Many early reports suggested a role for immunothrombosis involving neutrophil-mediated release of NETs contributing to endothelial dysfunction as a mechanism of microthrombosis in

COVID-19 associated ARDS[39-41]. These findings have been supported by several studies carried out in humans[42-50]. All of these studies were performed with 7 to 77 participants. Our study supports these findings in two important ways: 1) we used a much greater sample size (n = 731) and, 2) the identification of molecular factors associated with immuno-thrombosis emerged from an *unsupervised* analysis of deep phenotyping of the participant population. The strength of the extensive characterization performed in this study has enabled the finer definition of molecular mechanisms of disease by providing associations between the circulating proteome, metabolome and clinical laboratories results.

FGFR and SHC4 intracellular signaling in COVID-19 ARDS

Two of the outstanding novel molecular factors identified by our study associated with COVID-19 ARDS are FGFR and SHC4. Circulating levels of the pro-angiogenic FGF-2 has been associated with COVID-19 severity and creatine levels in a study of 208 SARS-CoV-2 positive participants[51]. It is noteworthy that the use of Nintedanib, an inhibitor of FGFR, Vascular Endothelial Growth Factor Receptor (VEGFR) and Platelet-Derived Growth Factor Receptor (PDGF-R) approved for use in interstitial lung disease, improved pulmonary inflammation and helped wean off mechanical ventilation of three middle-aged obese COVID-19 patients where lung function restoration has been challenging[52]. While the adaptor protein SHC4 has not been experimentally demonstrated to modulate FGFR signaling, a 12-gene biomarker signature associated with melanoma contains FGFR2, FGFR3 and SHC4[53]. In view of the limited knowledge of this understudied member of the SHC family, it is attractive to speculate that it may act downstream of FGFR or other associated growth factor receptors linked to angiogenesis, favoring

immunothrombosis associated with COVID-19 ARDS. Additional experimentation is required to establish a sound scientific basis for these hypotheses. Moreover, the identity of the cells expressing SHC4, leading to its presence in the circulation is not known, also the focus of ongoing investigations. Taken together, our identification of FGFR and SHC4 signaling pathways distinguishing EP6 from other endophenotypes, supports further investigation of antagonists of those pathways to treat severe lung manifestations of COVID-19 and their potential use as biomarkers of severe disease activity.

Limitations and considerations

The data presented in this study come from individuals participating to BQC19, a prospective observational cohort built to study COVID-19 in Québec (Canada) with its specific population profile as reported previously[10]. While the number of participants was sufficient to establish the endophenotypes using the extensive proteomic profile available in BQC19, it was insufficient for traditional genome-wide association study (GWAS) to identify relations between SNVs and the identified endophenotypes[54]. Instead, we exploited the top SNVs that distinguished EP6 from other EPs in a pQTL analysis. Because these results show a potential genetic functional causality, it gives us the confidence that these associations are likely not due to random chance; however, the robustness of this approach needs to be further tested in other studies. A chronological bias may also be present, as most of the participants used for endophenotyping in this study were recruited during the first two waves of the pandemic (Figure S1), prior to widespread vaccination in Québec and prior to the appearance of the Omicron variant and sub-variants. Therefore, some of the features of the identified endophenotypes may change over the course of the pandemic.

It will be essential to continue to assess the molecular profiles longitudinally to better understand the dynamic nature of host-pathogen interactions. It will also be interesting to compare the profiles of COVID-19 ARDS to other viral-induced ARDS, to identify commonalities as well as distinguishing features.

In this study, we used circulating proteome as determined by aptamers to identify endophenotypes, a task for which it is well suited as it can capture a dynamic landscape. However, several important considerations need to be mentioned. First, raw (unnormalized) expression values of the same aptamer can be used to compare different samples, but these values cannot be used to compare different aptamers against each other in the same sample, since they only show the relative abundance of expression and not absolute expression. As such, one needs to first normalize these values across samples (one aptamer at a time) and then subject them to follow-up analysis such as clustering, an approach that we adopted in this study. Second, since information only shows relative abundance, focusing on an individual aptamer and analyzing it will require additional measurements to establish absolute abundance. Moreover, if one wants to analyze aptamers individually (instead of the collective approach that we used in this study), they need to consider the effect of complexes and non-specific binding that may result in noisy data. As such, we suggest that the aptamer expression data be used collectively and after proper normalization, which enabled us to identify various EPs and important molecular mechanisms discussed in this study.

In this work, we developed a predictive model based on blood markers that enables us to generalize the definition of EPs to scenarios where data on circulating proteome is not available. This significantly increases the applicability of this approach and may enable automating identification of high-risk individuals in the clinic. However, time and follow-up studies are required to move this predictive model and the definitions of EPs from research to clinic and develop it as an acceptable clinical practice. Nonetheless, our approach can be used to study COVID-19 in different cohorts and identify characteristics that can guide the treatment of the disease.

Future directions

There is enormous amount of data within this study and BQC19 that can and should be exploited by the scientific and medical community to improve our understating of COVID-19 and novel emerging acute respiratory illnesses. Analyzing in more details the molecular profiles of the other EPs, in particular EP3-5, which lead to similar outcomes from distinct molecular pathways should further yield important insights into mechanisms of the disease. For example, EP5 is enriched in males and cardiovascular disease complications, while EP4 is enriched in female (like EP1) but with different distinct clinical trajectories pointing to sex-dependent and independent molecular mechanisms of disease.

Conclusion

The major strength of this study is its starting point: unsupervised analysis of a large and deeply phenotyped cohort. We showed that this approach can address both fundamental scientific

questions pertaining to mechanisms of disease and help the medical community improve patient outcomes through early identification of patient that may follow a severe clinical course during COVID-19.

Methods:

Datasets and preprocessing

The Biobanque québécoise de la COVID-19 (BQC19; www.quebeccovidbiobank.ca) is aimed at coordinating the collection of patients' data and samples for COVID-19 related research. Data and samples were collected from ten sites across the province of Québec (Canada)[10]. BQC19 organizes the collected data, including clinical information and multi-omics experimental data, before making it available in successive releases. For this study, we used release #5 of the clinical data published in December 2021, the circulating proteome determined using SOMAmers[54] and Metabolomics data[55]. BQC19 GWAS imputation data was generated by Tomoko Nakanishi at Brent Richards lab, Jewish General Hospital and McGill University. Detailed codes used for generating the data can be found in: https://github.com/richardslab/BQC19_genotype_pipeline

Our main corpus of analysis consisted of n = 1,362 hospitalized and SARS-CoV-2 positive patients (based on qRT-PCR) of BQC19. This included n = 731 patients for which both clinical and proteomic data were available as well as n = 631 patients for which proteomic data was not available, but their clinical data contained measurements for more than half (at least 11 out of 21) of the blood markers that we used as a validation set for the predictive model developed in this study.

We also obtained data (n = 731) corresponding to the circulating proteome measured between April 2, 2020 and April 20, 2021 by a multiplex SOMAmer affinity array (Somalogic, 4,985 aptamers) from BQC19 (release #3 Sep. 2021, associated patients' data updated in release #5 Dec. 2021). When measurements of the same patients but at different time points were available, we used the one corresponding to the first time point. SomaScan is a biotechnological protocol commercialized by the Somalogic company. It relies on a set of artificial aptamers linked to a fluorophore and each designed to bind a single protein. Once added to the sample, the activity of each aptamer is measured through fluorescence and used to approximate the expression level of the targeted protein. SomaScan protocol comprises several levels of calibration and normalization to correct technical biases[11]. Log2 and Z-score normalization were performed on each aptamer separately in addition to the manufacturer's provided normalized data (hybridization control normalization, intraplate median signal normalization, and median signal normalization). Since the data was analyzed by Somalogic in two separate batches, we applied the z-score transformation separately to each batch, to reduce batch effects. These additional transformations ensure that the measured values of different aptamers are comparable and can be used in cluster analysis.

We obtained metabolomic data (1,435 metabolites) from BQC19. We used the batch-normalized, missing values imputed and log-transformed version of the data.

Analyses of the GWAS dataset

For the GWAS analyses, annotation of SNVs were done using the biomaRt package[56] from R[57] and all analyses were done using R version 4.1.3. Quality control steps were derived in majority from a 2017 QC tutorial article[58]. At the beginning, we had 867,450 markers and 2,429 samples. We import Plink format data into R using the "read_plink" function from genio package[59] from R. We removed 103,592 non ACGT bi-allelic markers. We calculated the predicted sex by looking at the rate of homozygote markers on chromosome 23. We removed 3,588 markers with call rates < 98%, 448,932 monomorphic markers and markers with MAF <0.05, and 28,092 markers with Hardy Weinberg equilibrium < 1E-6 (calculated by the "HWE.exact" function from the genetics package[60] from R. For the EP6 cluster group analyses, we removed in addition 1,747 samples that were not in the cluster analysis, 8 samples with a sex discrepancy (based on predicted sex calculated earlier and reported sex), and 3 samples with a heterozygosity rate > 3 standard deviations. We finally removed a pair of samples who had approximately the same genome probably due to an error of manipulation. We couldn't know which one was the right sample, so we removed both of them. We also had 2 pairs of individuals who had a pi-hat of ~0.5 (meaning first degree relatives), we decided to keep one sample per pair, the one with the higher call rate. All other pairs of individuals had a pi-hat < 0.21 that is judge acceptable considering our population. Pi-hat were calculated with the "snpgdsIBDMoM" function from SNPRelate package[61] from R. At the end of quality controls, 283,246 markers on 655 samples have been used to perform association analyses.

To perform the principal component analyses (PCA), we took a subsample of independent markers (pruning) with a maximum sliding window of 500,000 base pairs and a linkage

458 disequilibrium (LD) threshold of 0.2 using the "snpgdsLDpruning" function from the SNPRelate
459 package from R. We ran the PCA with the "snpgdsPCA" function from SNPRelate package from
460 R. The first 2 principal components (PCs) were considered significant.
461
462 For the GWAS analyses of the 283,246 remaining markers between EP6 cluster compared to all
463 others, we modeled a logistic regression with the dichotomous variable indicating if the
464 participants belong to the EP6 cluster as the outcome variable. We used the additive model for
465 markers as an independent variable and we adjusted the models with the first two PCs. Odds
466 ratio and p-values were calculated on each model. QQ-plots have been performed as quality
467 control of the models; p-values were plotted using "qqplot.pvalues" function from gaston
468 package[62] from R (data not shown). We compared the aptamers' normalized level of expression
469 (based on normalization described earlier) between the three groups of genotypes for each
470 studied genes by performing standard ANOVA analyses followed by Tukey *post hoc* tests
471 (referred to in this study as pQTL analysis). Since aptamers tested are limited compared to the
472 SNVs, we have fixed significance p-values threshold below 1E-4 to report more SNVs instead of
473 the more common 1E-5 suggestive threshold. Finally, to identify if the cluster EP6 may be
474 explained by the aptamers and the SNVs, we performed multiple logistic regression analyses
475 models include aptamers expression values, SNV genotypes (additive model) and the two
476 principal components. OR are reported with 95% confidence intervals.
477
478 **Consensus agglomerative clustering**

Patients were clustered using agglomerative clustering, with Euclidean distance and Ward's linkage[12,13]. To identify number of clusters k, we used the elbow method based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). More specifically, we calculated the AIC and BIC for clustering using k = 2, 3, …, 20 and used the Kneedle algorithm[63] to identify the value of k corresponding to the "elbow", where increasing the value of k does not provide much better modeling of the data. Kneedle identified k = 6 as the number of clusters based on both AIC and BIC (Figure 1A).

Given the number of clusters in the data, we then used consensus clustering with sub-sampling to obtain robust endophenotypes. We randomly sampled 80% of the patients 1000 times. Each time, we used agglomerative clustering above with k = 6 to identify clusters. Given these 1000 clusterings, we calculated the frequency of two patients appearing in the same cluster, when both were present in the randomly formed dataset. We then performed one final agglomerative clustering of these frequency scores to identify the six endophenotypes (Figure S2A and Figure 1B). The distribution of Rand-Index, showing the concordance between each one of the 1000 clusterings and the final consensus clustering is provided in Figure S2B (mean Rand-Index = 0.823), reflecting a high degree of consistency.

Metabolomic pathway characterization of EP6

The 1435 metabolites measured were organized into 122 sub-pathways in the original dataset (denoted as "SUB_PATHWAYS"). We first identified metabolites whose values were significantly higher or lower in EP6 compared to other EPs (two-sided MWU test, FDR<0.01). Then, we used

these metabolites to perform pathway enrichment analysis (one-sided Fisher's exact test) based on 122 pathways. The resulting p-values (Table S3) were then corrected for multiple tests using Benjamini-Hochberg FDR.

Nearest-centroid predictor based on blood markers

In order to predict endophenotypes from blood tests, we developed a missing-value resilient nearest-centroid classifier. We used the set of patients that were used to form the original EPs (n = 731) as the training set and the set of patients that did not have proteome data as the validation set (n = 631). First, we z-score normalized each blood marker across all the patients in the training set, one marker at a time. We then formed a blood marker signature (a vector of length 21) for each EP. Each element of an EP's signature corresponds to the mean of the corresponding marker across all patients of that EP.

To predict the EP label of each patient in the test set, we first z-score normalized their blood marker measurements using the mean and standard deviation of the blood markers calculated from the training set. Then, we calculated the cosine distance between each test patient's blood marker profile and the centroids (excluding missing values) and identified the nearest EP as the predicted EP (PEP) label of the patient.

Acknowledgements

This work was made possible through open sharing of data and samples from the Biobanque québécoise de la COVID-19, funded by the Fonds de recherche du Québec - Santé, Génome

Québec, the Public Health Agency of Canada and, as of March 2022, the Ministère de la Santé et des Services Sociaux du Québec. We thank all participants to BQC19 for their contribution. This study was supported by the Fonds de recherche du Québec - Santé (FRQS)- Cardiometabolic Health, Diabetes and Obesity Research Network (CMDO)- Initiative. This work was also supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2019-04460 (AE).

Permissions

The study was approved by the Institutional Ethics Review Board of the "Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean" (CIUSSS-SLSJ) affiliated to Université de Sherbrooke [protocol #2021-369, 2021-014 CMDO – COVID19].

Figures

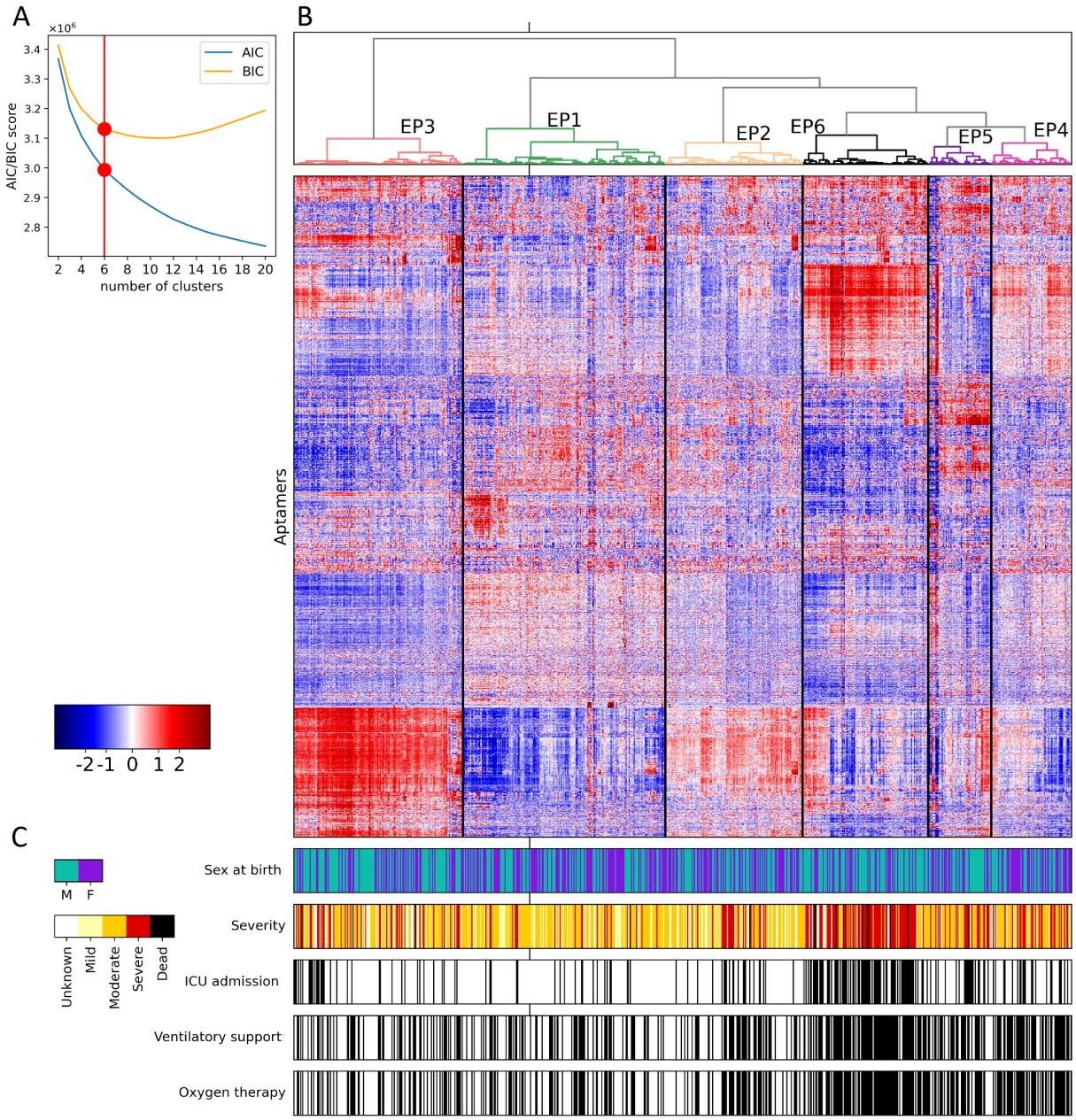


Figure 1: Unsupervised consensus clustering of SARS-CoV-2 positive patients. A) The elbow points (circles in red) of Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) curves versus number of clusters consistently corresponded to k=6 as the optimal number of clusters. B) The heatmap shows the expression of aptamers (rows) in each sample (columns). The dendogram shows the identified endophenotypes. C) Characterization of samples based on sex at birth, highest world health organization (WHO) severity level achieved, intensive care unit (ICU) admission, ventilatory support, and oxygen therapy. For the last three rows, a sample colored “black” reflects a label of “yes”.

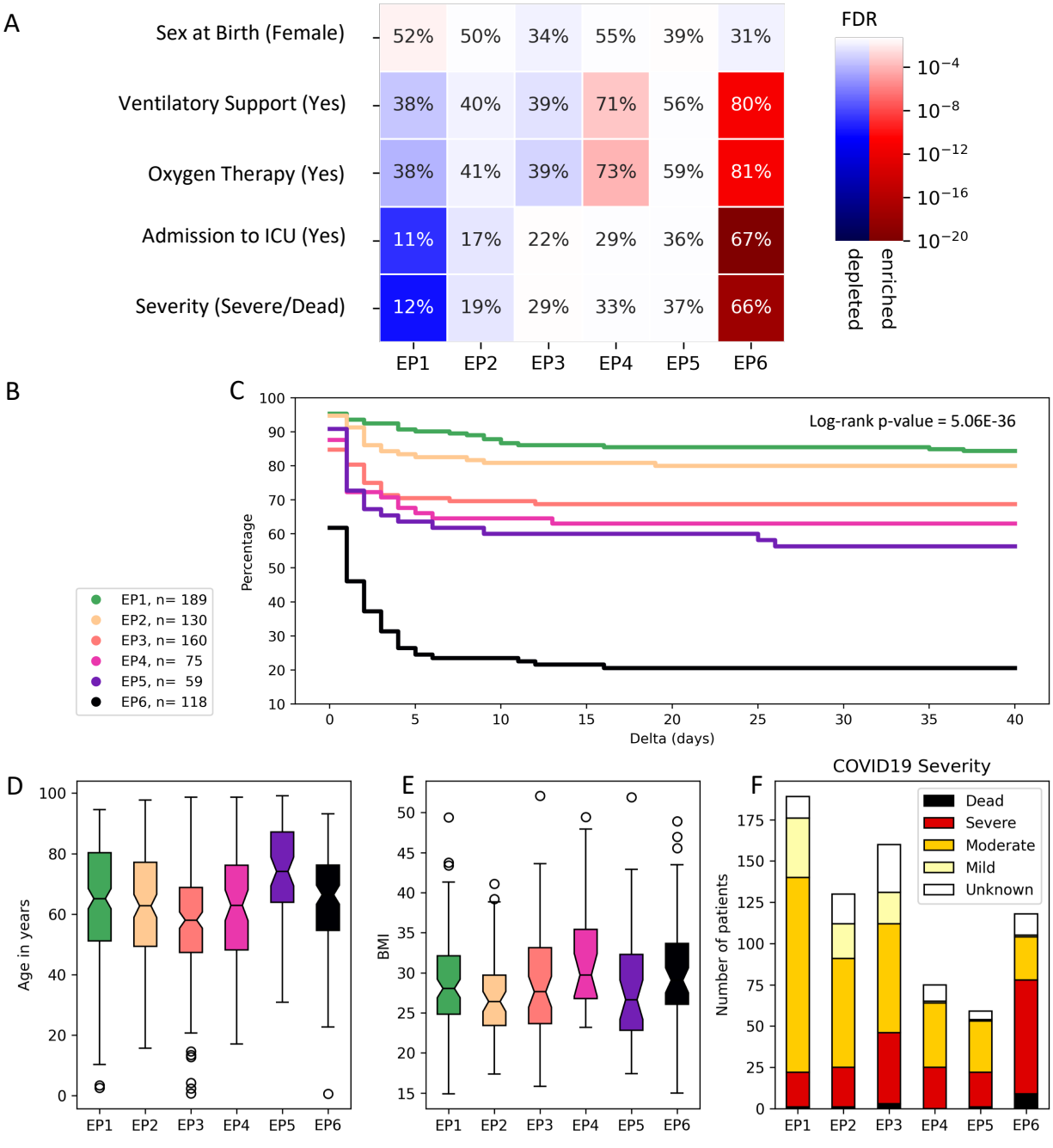


Figure 2: Characterization of endophenotypes (EPs).
A) Enrichment or depletion of each EP in clinical variables (one cluster versus rest). Two-sided Fisher's exact tests are used to calculate the p-values, which are corrected for multiple tests using Benjamini-Hochberg false discovery rate (FDR). Gradients of blue show depletion, while gradients of red show enrichment. FDR values above 0.05 are depicted as white. B) The number of patients in each EP and the colors used to represent them in panels C, D, and E. C) Kaplan-Meier analysis of the time between patients' admission to the hospital and their admission to intensive care unit (ICU) (or death if earlier) for each EP (Delta). D) Distribution of age in each EP. E) Distribution of BMI in each EP. F) COVID-19 severity in each EP.

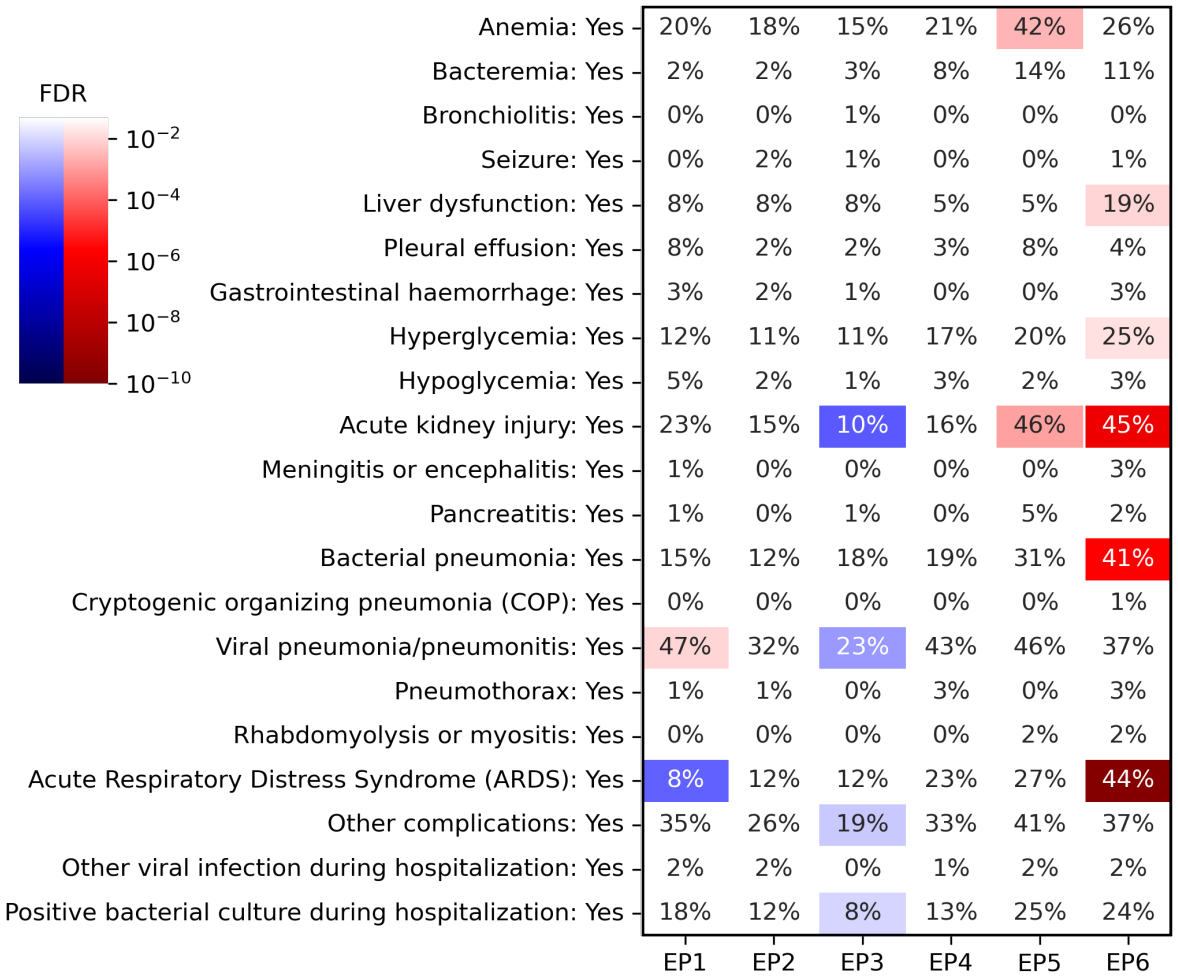


Figure 3: Frequency and significance of complications in different EPs.
The value in each cell shows the percentage of patients of that EP (column) that suffered from the complication (row). The colors represent two-sided Fisher's exact test false discovery rate (FDR, corrected for multiple tests). Red represents enrichment, while blue represents depletion. FDR values below 0.05 are shown as white.

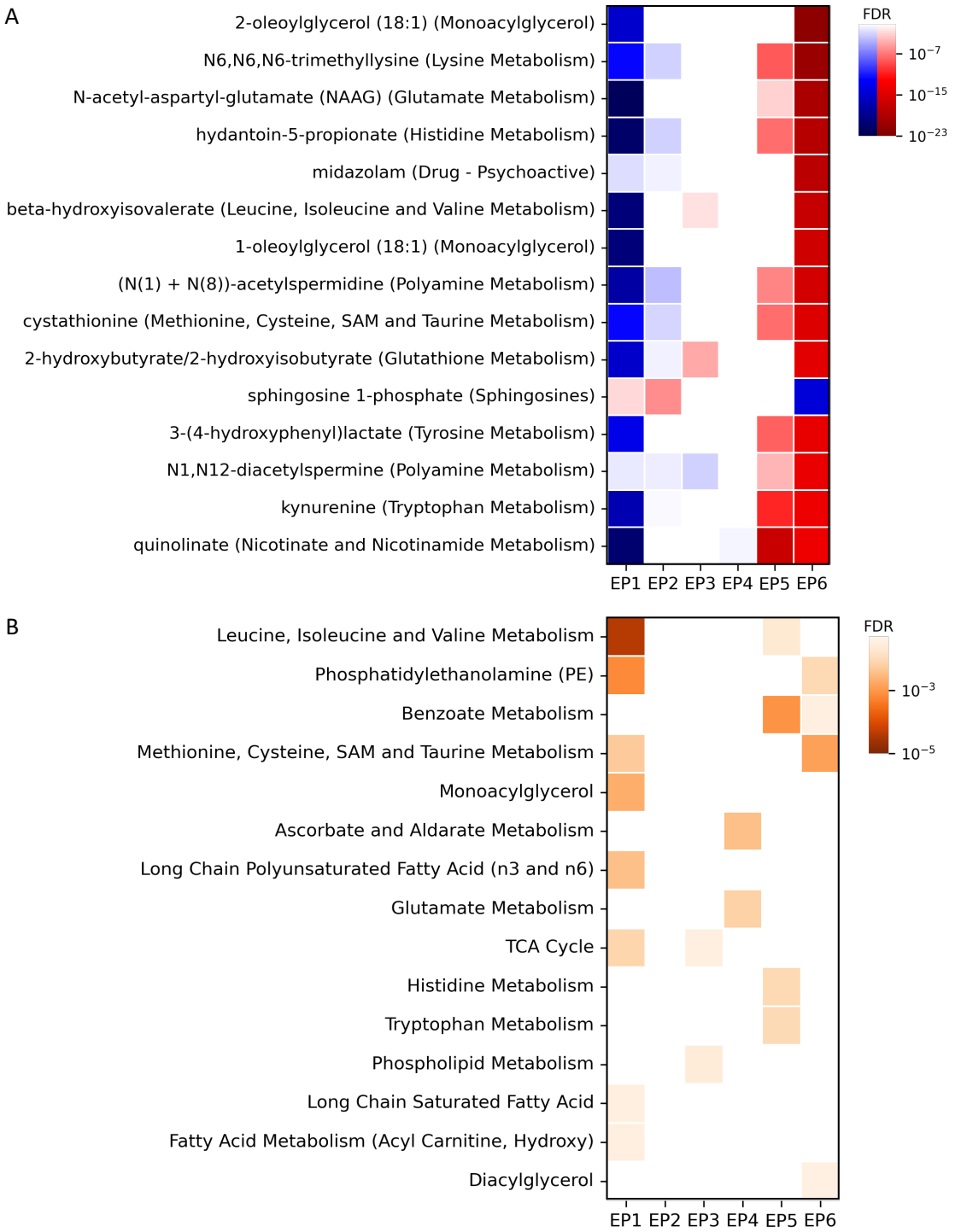
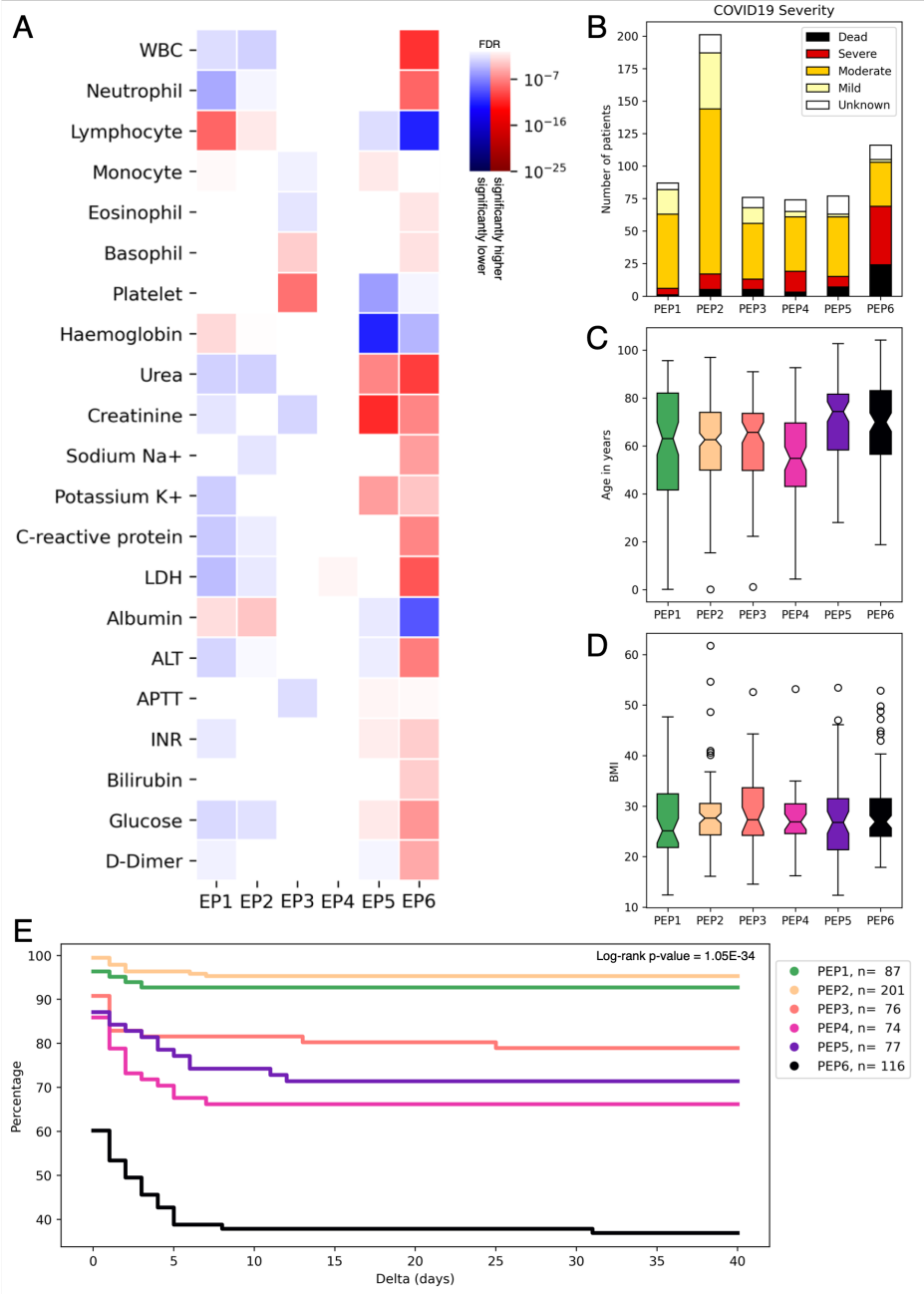


Figure 4: Metabolite characteristics of endophenotypes (EPs).
A) The heatmap shows the over-expression (red) and under-expression (blue) of metabolites in different EPs (two-sided Mann–Whitney U test). Row names show metabolites followed by the sub-pathway to which they belong in parentheses. Only top 15 metabolites (based on false discovery rate for EP6) for which a definite name and sub-pathway was available are shown. Full list is provided in Table S3. B) The heatmap shows the enrichment (one-sided Fisher's exact test) of EPs in different metabolite sub-pathways. See Table S3 for the full list.



568

569 **Figure 5:** Differential expression patterns of blood markers in the identified endophenotypes
570 (EPs) and characterization of predicted endophenotypes (PEPs) based on the predictive model.
571 A) Heatmaps of false discovery rates (FDR) values for two-sided one-vs-rest Mann–Whitney U
572 tests for 21 blood markers for each EP. Abbreviations used: WBC = white blood cells, LDH = lactate
573 deshydrogenase, ALT = alanine aminotransferase, APTT = activated partial thromboplastin time,
574 INR = International Normalized Ratio. FDR values below 0.05 are shown as white. B) World health
575 organization COVID-19 severity assessed in each PEP. C) Distribution of age in each PEP. D)
576 Distribution of BMI in each PEP. E) Kaplan-Meier analysis of the time between patients' admission
577 to the hospital and their admission to ICU (or death if earlier) for each PEP (Delta). The colormap

in panel E shows the number of patients in each PEP group and the colors used to represent them in panels C, D, and E. PEPs were predicted based on blood markers and were not used to originally identify the EPs.

581 **Tables**

582 **Table 1:** Clinical and pathological characteristics of the BQC19's participants used in this study.

		Cohort (n = 1,362) No. (%)
Age in years	<45	17.5
	45-65	33.6
	>65	48.7
	Unknown	0.3
Body mass index in kg/m²	<20	3.8
	20-25	13.0
	25-35	27.8
	>35	7.4
	Unknow	47.9
Sex at birth	Female	44.8
	Male	55.2
Severity	Dead	4.4
	Severe	21.8
	Moderate	51.0
	Mild	11.8
	Unknown	10.9
Oxygen therapy	Yes	48.2
	No	21.0
	Unknown	30.8
Ventilatory support	Yes	47.6
	No	14.0
	Unknown	38.5
Admission to intensive care unit	Yes	17.5
	No	33.6
	Unknown	48.7

583

584 **Table 2:** Reactome pathways associated with EP6, based on expression of aptamers. KnowEnG
585 analytical platform was used. The p-values were calculated using a one-sided Fisher's exact test
586 and were corrected for multiple tests using Benjamini-Hochberg method. Only signaling
587 pathways with FDR<5E-4 are shown in this table (see Table S4 for the full list).
588

Pathway	FDR
Signaling by Interleukins	2.75E-12
Cytokine Signaling in Immune system	6.25E-10
Immune System	2.28E-06
Signaling by FGFR	4.75E-06
Constitutive Signaling by Aberrant PI3K in Cancer	2.94E-05
FGFR2 ligand binding and activation	5.77E-05
Extracellular matrix organization	1.85E-04
FGFR3 ligand binding and activation	1.85E-04
FGFR3c ligand binding and activation	1.85E-04
FGFRL1 modulation of FGFR1 signaling	1.85E-04
TNFs bind their physiological receptors	2.58E-04
PI3K/AKT Signaling in Cancer	3.28E-04
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	4.14E-04
Interleukin-4 and 13 signaling	4.91E-04

589
590

591 **Table 3**: SNVs differentiating EP6 against all other endophenotype clusters.

SNV ID	Position	Gene Symbol	SNV [1]	MAF	HWE	OR [2]	p-value
rs1394671	chr5:7348567	-	G > A	0.240	0.095	2.28	2.41E-06
rs12186698	chr5:168880668	*SLIT3*	T > C	0.087	0.247	2.93	3.72E-06
rs11625406	chr14:50049474	-	C > A	0.438	0.137	0.47	1.03E-05
rs11862889	chr16:83828886	-	T > C	0.104	0.446	2.62	1.07E-05
rs2995918	chr4:37905775	*TBC1D1*	C > T	0.494	0.776	0.49	1.20E-05
rs16897810	chr5:68415296	-	G > A	0.093	0.001	2.67	1.43E-05
rs2376263	chr17:35436659	*SLFN13*	G > A	0.261	0.073	2.01	2.48E-05
rs4790712	chr17:1614620	*SLC43A2*	G > A	0.441	0.003	1.97	2.50E-05
rs2294566	chr20:41472939	*CHD6*	C > A	0.239	0.265	2.04	2.78E-05
rs7164451	chr15:48921859	*SHC4*	G > A	0.386	0.053	1.89	3.01E-05
rs57664621	chr8:22808443	*PEBP4*	G > C	0.174	0.016	2.16	3.07E-05
rs28482919	chr3:14903094	*FGD5*	T > C	0.187	0.053	0.32	3.20E-05
rs6559283	chr9:89712560	-	T > C	0.375	<0.001	1.96	3.27E-05
rs657075	chr5:132094425	-	A > G	0.104	0.328	2.49	4.28E-05
rs56235109	chr15:62424001	*TLN2*	A > G	0.235	0.366	0.38	4.67E-05
rs7620057	chr3:179473377	*GNB4*	T > C	0.099	1.000	2.51	4.67E-05
rs10466868	chr12:131455375	-	T > G	0.108	0.401	2.51	5.94E-05
rs2236798	chr1:18735127	*PAX7*	A > G	0.054	0.114	2.90	6.11E-05
rs3774814	chr4:5464702	*STK32B*	C > G	0.205	<0.001	0.36	6.75E-05
rs4497815	chr19:22903215	-	G > A	0.216	0.205	2.07	7.52E-05
rs6765694	chr3:54601810	*CACNA2D3*	G > A	0.404	0.833	0.51	7.66E-05
rs2797773	chr6:37559045	-	C > T	0.422	0.000	0.53	8.06E-05
rs17014760	chr4:129419725	-	A > A	0.316	0.000	1.89	9.33E-05
rs10948260	chr6:45835559	-	G > A	0.366	0.189	1.81	9.52E-05
rs12035677	chr1:232391209	-	A > G	0.063	0.000	3.02	9.95E-05

592 Abbreviations used: SNV = Single nucleotide variation, HWE = Hardy Weinberg Equilibrium, MAF
593 = Minor allele frequency, OR = Odd ratio.
594
595 [1] SNV are described following GWAS annotations: refence allele > alternative allele (e.g., G > A).
596 [2] Logistic regression analyses using additive model adjusted for the 2 principal components.
597

598 **Table 4:** Association between genotypes and aptamer expression levels

SNV	Gene Symbol	Nearest gene	Aptamers normalized expression [1]				Multiple logistic regression analyses [3]			
			HM_{ref}	HTZ	HM_{alt}	p-val [2]	Aptamers OR	p-val	SNV OR	p-val
rs6765694	*CACNA2D3*	-	-0.09 (±0.89)	0.01 (±1.10)	0.24 * (±0.99)	**0.011**	0.61 (0.46-0.79)	2.5E-4	0.52 (0.37-0.74)	2.4E-4
rs10948260	-	*CLIC5*	0.02 (±0.92)	0.02 (±1.09)	-0.02 (±0.97)	0.796	0.68 (0.55-0.83)	1.7E-4	1.83 (1.35-2.48)	8.8E-5
rs657075	-	*IL3*	-0.01 (±1.01)	0.05 (±0.99)	0.32 (±0.46)	0.328	1.37 (1.14-11.64)	8.2E-4	2.49 (1.60-3.88)	5.2E-5
rs7164451	*SHC4*	-	-0.07 (±0.90)	0.01 (±1.05)	0.23 * (±1.16)	**0.017**	11.98 (7.61-18.87)	8.4E-27	2.00 (1.30-3.09)	1.8E-3
rs12186698	*SLIT3*	-	0.00 (±1.02)	0.07 (±1.06)	0.15 (±0.89)	0.509	0.82 (0.64-1.06)	0.135	2.99 (1.89-4.73)	2.7E-6
rs56235109	*TLN2*	-	-0.03 (±0.98)	0.06 (±1.03)	0.04 (±1.27)	0.353	0.58 (0.45-0.75)	2.7E-5	0.38 (0.23-0.60)	4.7E-5

599 Abbreviations used: SNV = Single nucleotide variation, HM_{ref} = Homozygotes for the reference
600 allele, HM_{alt} = Homozygotes for the alternative allele, HTZ = Heterozygotes, OR = Odd ratio.
601
602 [1] Aptamers' normalized levels of expression are reported as mean (± standard deviation).
603 Normalization steps for aptamer expressions are described in Methods.
604 [2] p-value, standard ANOVA analyses followed by Tukey *post hoc* analyses. Asterisks (*) identify
605 difference between HM_{ref} and HM_{alt} genotypes.
606 [3] Multiple logistic regression analyses models include aptamers expression values, SNV
607 genotypes (additive model) and the two principal components. OR are reported with 95%
608 confidence intervals in parentheses.

References

1 Rubin, R. As their numbers grow, COVID-19 "long haulers" stump experts. *Jama* **324**, 1381-1383 (2020).

2 Marshall, J. C. *et al.* A minimal common outcome measure set for COVID-19 clinical research. *The Lancet Infectious Diseases* **20**, e192-e197 (2020).

3 Hull, D. L. in *PSA: Proceedings of the biennial meeting of the philosophy of science association.* 653-670 (Cambridge University Press).

4 Berrettini, W. H. Genetic bases for endophenotypes in psychiatric disorders. *Dialogues Clin Neurosci* **7**, 95-101 (2005). https://doi.org/10.31887/DCNS.2005.7.2/wberrettini

5 Russell, C. D. & Baillie, J. K. Treatable traits and therapeutic targets: goals for systems biology in infectious disease. *Current opinion in systems biology* **2**, 140-146 (2017).

6 Blatti III, C. *et al.* Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform. *PLoS biology* **18**, e3000583 (2020). https://doi.org/10.1371/journal.pbio.3000583

7 Emad, A. *et al.* Superior breast cancer metastasis risk stratification using an epithelial-mesenchymal-amoeboid transition gene signature. *Breast Cancer Research* **22**, 74 (2020). https://doi.org/10.1186/s13058-020-01304-8

8 Te Pas, M. F. W., Madsen, O., Calus, M. P. L. & Smits, M. A. The Importance of Endophenotypes to Evaluate the Relationship between Genotype and External Phenotype. *International Journal of Molecular Sciences* **18**, 472 (2017).

9 Al-Hadrawi, D. S., Al-Rubaye, H. T., Almulla, A. F., Al-Hakeim, H. K. & Maes, M. Lowered oxygen saturation and increased body temperature in acute COVID-19 largely predict chronic fatigue syndrome and affective symptoms due to Long COVID: A precision nomothetic approach. *Acta Neuropsychiatr*, 1-12 (2022). https://doi.org/10.1017/neu.2022.21

10 Tremblay, K. *et al.* The Biobanque québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. *PloS one* **16**, e0245031 (2021). https://doi.org/10.1371/journal.pone.0245031

11 Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *Nature Precedings*, 1-1 (2010).

12 Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 236-244 (1963).

13 Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* **31**, 274-295 (2014). https://doi.org/10.1007/s00357-014-9161-z

14 Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**, 457-481 (1958). https://doi.org/10.1080/01621459.1958.10501452

15 Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**, D687-D692 (2021). https://doi.org/10.1093/nar/gkab1028

16 Thomas, T. *et al.* COVID-19 infection alters kynurenine and fatty acid metabolism, correlating with IL-6 levels and renal status. *JCI insight* **5** (2020).

17 Zahedi, K. *et al.* The role of spermidine/spermine N 1-acetyltransferase in endotoxin-induced acute kidney injury. *American Journal of Physiology-Cell Physiology* **299**, C164-C174 (2010).

18 Pirnes-Karhu, S. *et al.* Spermidine/spermine N1-acetyltransferase activity associates with white blood cell count in myeloid leukemias. *Experimental Hematology* **42**, 574-580 (2014).

19 Marfia, G. *et al.* Decreased serum level of sphingosine-1-phosphate: a novel predictor of clinical severity in COVID-19. *EMBO molecular medicine* **13**, e13424 (2021).

20 Lachkar, F., Ferré, P., Foufelle, F. & Papaioannou, A. Dihydroceramides: their emerging physiological roles and functions in cancer and metabolic diseases. *American Journal of Physiology-Endocrinology and Metabolism* **320**, E122-E130 (2021).

21 Lone, M. A. *et al.* Subunit composition of the mammalian serine-palmitoyltransferase defines the spectrum of straight and methyl-branched long-chain bases. *Proceedings of the National Academy of Sciences* **117**, 15591-15598 (2020).

22 Han, G. *et al.* Identification of small subunits of mammalian serine palmitoyltransferase that confer distinct acyl-CoA substrate specificities. *Proceedings of the National Academy of Sciences* **106**, 8186-8191 (2009).

23 Blachier, F., Andriamihaja, M. & Blais, A. Sulfur-Containing Amino Acids and Lipid Metabolism. *The Journal of Nutrition* **150**, 2524S-2531S (2020). https://doi.org/10.1093/jn/nxaa243

24 Ye, C., Sutter, B. M., Wang, Y., Kuang, Z. & Tu, B. P. A metabolic function for phospholipid and histone methylation. *Molecular cell* **66**, 180-193. e188 (2017).

25 Parman, T. *et al.* Toxicogenomics and metabolomics of pentamethylchromanol (PMCol)-induced hepatotoxicity. *Toxicological Sciences* **124**, 487-501 (2011).

26 Stafford, J. H. & Thorpe, P. E. Increased exposure of phosphatidylethanolamine on the surface of tumor vascular endothelium. *Neoplasia* **13**, 299-IN292 (2011).

27 Ran, S., Downes, A. & Thorpe, P. E. Increased exposure of anionic phospholipids on the surface of tumor blood vessels. *Cancer research* **62**, 6132-6140 (2002).

28 Port, J. R. *et al.* High-fat high-sugar diet-induced changes in the lipid metabolism are associated with mildly increased COVID-19 severity and delayed recovery in the Syrian hamster. *Viruses* **13**, 2506 (2021).

29 Ataga, K. I. Hypercoagulability and thrombotic complications in hemolytic anemias. *Haematologica* **94**, 1481-1484 (2009). https://doi.org/10.3324/haematol.2009.013672

30 Sinauridze, E. I. *et al.* Platelet microparticle membranes have 50-to 100-fold higher specific procoagulant activity than activated platelets. *Thrombosis and haemostasis* **97**, 425-434 (2007).

31 Tavoosi, N. *et al.* Molecular Determinants of Phospholipid Synergy in Blood Clotting *. *Journal of Biological Chemistry* **286**, 23247-23253 (2011). https://doi.org/10.1074/jbc.M111.251769

32 Majumder, R., Liang, X., Quinn-Allen, M. A., Kane, W. H. & Lentz, B. R. Modulation of prothrombinase assembly and activity by phosphatidylethanolamine. *Journal of Biological Chemistry* **286**, 35535-35542 (2011).

33 Kim, K. *et al.* Pro-coagulant and pro-thrombotic effects of paclitaxel mediated by red blood cells. *Thrombosis and Haemostasis* **118**, 1765-1775 (2018).

34 Sato, K., Iemitsu, M., Aizawa, K. & Ajisaka, R. Testosterone and DHEA activate the glucose metabolism-related signaling pathway in skeletal muscle. *American Journal of Physiology-Endocrinology and Metabolism* **294**, E961-E968 (2008).

35 Kajita, K. *et al.* Glucocorticoid-induced insulin resistance associates with activation of protein kinase C isoforms. *Cellular signalling* **13**, 169-175 (2001).

36 MacDonald, J. & Sprecher, H. Distribution of arachidonic acid in choline-and ethanolamine-containing phosphoglycerides in subfractionated human neutrophils. *Journal of Biological Chemistry* **264**, 17718-17726 (1989).

37 Papageorgiou, C. *et al.* Disseminated intravascular coagulation: an update on pathogenesis, diagnosis, and therapeutic strategies. *Clinical and Applied Thrombosis/Hemostasis* **24**, 8S-28S (2018).

38 Esaki, Y., Hirokawa, K., Fukazawa, T. & Matsuda, T. Immunohistochemical study on the liver in autopsy cases with disseminated intravascular coagulation (DIC) with reference to clinicopathological analysis. *Virchows Archiv A* **404**, 229-241 (1984).

39 Barnes, B. J. *et al.* Targeting potential drivers of COVID-19: Neutrophil extracellular traps. *Journal of Experimental Medicine* **217** (2020).

40 Bonaventura, A. *et al.* Endothelial dysfunction and immunothrombosis as key pathogenic mechanisms in COVID-19. *Nature Reviews Immunology* **21**, 319-329 (2021).

41 Ding, J. *et al.* A network-informed analysis of SARS-CoV-2 and hemophagocytic lymphohistiocytosis genes' interactions points to Neutrophil extracellular traps as mediators of thrombosis in COVID-19. *PLoS Computational Biology* **17**, e1008810 (2021). https://doi.org/10.1371/journal.pcbi.1008810

42 Blasco, A. *et al.* Assessment of neutrophil extracellular traps in coronary thrombus of a case series of patients with COVID-19 and myocardial infarction. *JAMA cardiology* **6**, 469-474 (2021).

43 Desilles, J. P. *et al.* Impact of COVID-19 on thrombus composition and response to thrombolysis: Insights from a monocentric cohort population of COVID-19 patients with acute ischemic stroke. *Journal of Thrombosis and Haemostasis* **20**, 919-928 (2022).

44 Englert, H. *et al.* Defective NET clearance contributes to sustained FXII activation in COVID-19-associated pulmonary thrombo-inflammation. *EBioMedicine* **67**, 103382 (2021).

45 Leppkes, M. *et al.* Vascular occlusion by neutrophil extracellular traps in COVID-19. *EBioMedicine* **58**, 102925 (2020).

46 Middleton, E. A. *et al.* Neutrophil extracellular traps contribute to immunothrombosis in COVID-19 acute respiratory distress syndrome. *Blood* **136**, 1169-1179 (2020).

47 Obermayer, A. *et al.* Neutrophil extracellular traps in fatal COVID-19-associated lung injury. *Disease markers* **2021** (2021).

48 Ouwendijk, W. J. *et al.* High levels of neutrophil extracellular traps persist in the lower respiratory tract of critically ill patients with coronavirus disease 2019. *The Journal of infectious diseases* **223**, 1512-1521 (2021).

49 Petito, E. *et al.* Association of neutrophil activation, more than platelet activation, with thrombotic complications in coronavirus disease 2019. *The Journal of infectious diseases* **223**, 933-944 (2021).

50 Skendros, P. *et al.* Complement and tissue factor–enriched neutrophil extracellular traps are key drivers in COVID-19 immunothrombosis. *The Journal of clinical investigation* **130**, 6151-6157 (2020).

51 Smadja, D. M. *et al.* Placental growth factor level in plasma predicts COVID-19 severity and in-hospital mortality. *Journal of Thrombosis and Haemostasis* **19**, 1823-1830 (2021).

52 Bussolari, C. *et al.* Case Report: Nintedaninb May Accelerate Lung Recovery in Critical Coronavirus Disease 2019. *Frontiers in Medicine* **8**, 766486 (2021).

53 Liu, W., Peng, Y. & Tobin, D. J. A new 12-gene diagnostic biomarker signature of melanoma revealed by integrated microarray analysis. *PeerJ* **1**, e49 (2013). https://doi.org/10.7717/peerj.49

54 Zhou, S. *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nature Medicine* **27**, 659-667 (2021). https://doi.org/10.1038/s41591-021-01281-1

55 Ford, L. *et al.* Precision of a Clinical Metabolomics Profiling Platform for Use in the Identification of Inborn Errors of Metabolism. *The Journal of Applied Laboratory Medicine* **5**, 342-356 (2020). https://doi.org/10.1093/jalm/jfz026

56 Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4**, 1184-1191 (2009). https://doi.org/10.1038/nprot.2009.97

57 R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/. (R Foundation for Statistical Computing, Vienna, Austria, 2022).

58 Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res* **27**, e1608 (2018). https://doi.org/10.1002/mpr.1608

59 Ochoa, A. genio: Genetics Input/Output Functions. R package version 1.1.1. https://CRAN.R-project.org/package=genio (2022).

60 Warnes, G., Gorjanc, G., Leisch, F. & Man, M. genetics: Population Genetics. R package version 1.3.8.1.3. https://CRAN.R-project.org/package=genetics (2021).

61 Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012). https://doi.org/10.1093/bioinformatics/bts606

62 Perdry, H. & Dandine-Roulland, C. gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models. R package version 1.5.7. https://CRAN.R-project.org/package=gaston (2020).

63 Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. in *2011 31st international conference on distributed computing systems workshops.* 166-171 (IEEE).