Identification of Immune complement function as a determinant of adverse SARS-CoV-2 infection outcome

Vijendra Ramlall
Columbia University
Phyllis M. Thangara
Columbia University
Nicholas P. Tatonetti (✉ nick.tatonetti@columbia.edu)
Columbia University
Sagi D. Shapira (✉ ss4197@columbia.edu)
Columbia University

Research Article

Keywords: SARS-CoV-2, complement activation disorders, coagulation disorders, Complement Factor H, Complement Component 4 Binding Protein Alpha

DOI: https://doi.org/10.21203/rs.3.rs-26451/v1

Abstract

Understanding the pathophysiology of SARS-CoV-2 infection is critical for therapeutics and public health intervention strategies. Viral-host interactions can guide discovery of regulators of disease outcomes, and protein structure function analysis points to several immune pathways, including complement and coagulation, as targets of the coronavirus proteome. To determine if conditions associated with dysregulation of the complement or coagulation systems impact adverse clinical outcomes associated with SARS-CoV-2 infection, we performed a retrospective observational study of 11,116 patients suspected of SARS-CoV-2 infection. We found that history of macular degeneration (a proxy for complement activation disorders) and history of coagulation disorders (thrombocytopenia, thrombosis, and hemorrhage) are risk factors for morbidity and mortality in SARS-CoV-2 infected patients – effects that could not be explained by age or sex. In addition, using data from the UK Biobank, we implemented a candidate driven approach to evaluate linkage between severe SARS-CoV-2 disease and genetic variation associated with complement and coagulation pathways. Among our findings, our scan identified an eQTL for CD55 (a negative regulator of complement activation) and SNPs in Complement Factor H (CFH) and Complement Component 4 Binding Protein Alpha (C4BPA), which play central roles in complement activation and innate immunity and were previously linked to Age Related Macular Degeneration (AMD) in a Genome-Wide Association Study (GWAS). In addition to providing evidence that complement function modulates SARS-CoV-2 infection outcome, the data point to several putative genetic markers of susceptibility. The results highlight the value of using a multi-modal analytical approach, combining molecular information from virus protein structure-function analysis with clinical informatics and genomics to reveal determinants and predictors of immunity, susceptibility, and clinical outcome associated with infection.

Introduction

The SARS-CoV-2 pandemic has had profound economic, social, and public health impact with over 3 million confirmed cases and over 210,000 deaths across the globe. The infection causes respiratory illness with symptoms ranging from cough and fever to difficulty breathing. While highly variable age-dependent mortality rates have been widely reported, the comorbidities that drive this dependence are not fully understood. Further, with some notable exceptions[1-3], molecular studies have largely focused on ACE-2, the receptor and determinant of cell entry and viral replication[3]. While ACE-2 expression is critical, viruses employ a wide range of molecular strategies to infect cells, avoid detection, and proliferate. In addition, viral replication and immune mediated pathology are the primary drivers of morbidity and mortality associated with SARS-CoV-2 infection[4,5]. Therefore, understanding how virus-host interactions manifest as SARS-CoV-2 risk factors will facilitate clinical management, choice of therapeutic interventions, and setting of appropriate social and public health measures.

Knowledge of the precise molecular interactions that control viral replicative cycles can delineate regulatory programs that mediate immune pathology associated with infection and provide valuable clues about disease determinants. For example, viruses, including SARS-CoV-2, deploy an array of genetically encoded strategies to co-opt host machinery. Among the strategies, viruses encode multifunctional proteins that harness or disrupt cellular functions, including nucleic acid metabolism and modulation of immune responses, through protein-protein interactions and molecular mimicry – structural similarity between viral and host proteins (for a full discussion please see accompanying paper). Recently, we employed protein structure modeling to systematically chart interactions across all human infecting viruses[6] and in an accompanying paper, performed a virome-wide scan for molecular mimics. This analysis points to broad diversification of strategies deployed by human infecting viruses and identifies biological processes that underlie human disease. Of particular interest, we mapped over 140 cellular proteins that are mimicked by coronaviruses (CoV). Among these, we identified components of the complement and coagulation pathways as targets of structural mimicry across all CoV strains (see companion paper).

Through activation of one of three cascades, (i) the classical pathway triggered by an antibody–antigen complex, (ii) the alternative pathway triggered by binding to a host cell or pathogen surface, and (iii) the lectin pathway triggered by polysaccharides on microbial surfaces, the complement system is a critical regulator of host defense against pathogens including viruses[7]. When dysregulated by age-related effects or excessive acute and chronic tissue damage, complement activation can contribute to pathologies mediated by inflammation[7,8]. Similarly, inflammation-induced coagulatory programs as well as crosstalk between pro-inflammatory cytokines and the coagulative and anticoagulant pathways play pivotal roles in controlling pathogenesis associated with infections. Therefore, while the age-related differences in susceptibility to SARS-CoV-2 are likely a consequence of multiple underlying variables, virally encoded structural mimics of complement and coagulation pathway components may contribute to CoV associated immune mediated pathology. Moreover, a corollary of these observations is that dysfunctions associated with complement and/or coagulation may impact clinical outcome of SARS-CoV-2 infection. For example, the companion study suggests that coagulation disorders, such as thrombocytopenia, thrombosis and hemorrhage, may represent risk factors for SARS-CoV-2 clinical outcome. Among complement-associated disorders, multiple genetic and experimental evidence (including animal models of disease, histological examination of affected tissue, and germline mutational analysis) point to dysregulation of the complement system as the major driver of both early-onset, and age-related macular degeneration (AMD)[9,10]. A hyperinflammatory phenotype mediated by complement leads to progressive immune-mediated deterioration of the central retina. While AMD, the leading cause of blindness in elderly individuals (affecting roughly 200 million people worldwide[11]), is likely the result of multiple pathological processes, dysregulation of complement activation has emerged as the most widely accepted cause of disease[11-13].

To determine if conditions associated with dysregulation of the complement or coagulation systems impact adverse clinical outcomes associated with SARS-CoV-2 infection, we conducted a retrospective observational study of 11,116 patients at New York-Presbyterian/Columbia University Irving Medical Center. In agreement with previous reports[14], survival analysis identified significant risk of mechanical respiration and mortality associated with age and sex, as well as history of hypertension, obesity, and type 2 diabetes (T2D), coronary artery disease (CAD). Moreover, we found that history of macular degeneration (a proxy for complement activation disorders) and coagulation disorders (thrombocytopenia, thrombosis, and hemorrhage) were at significantly increased risk of adverse clinical outcomes (including mechanical respiration and death) following SARS-CoV-2 infection. Importantly, these effects could not be explained by either age or sex. Conversely, albeit in a small number of individuals, we observed that no patients with complement deficiency disorders required mechanical respiration or succumbed to their illness. Finally, in an independent analysis of data from the UK Biobank that focused on variants associated with the complement and coagulation pathways, we found significant genetic markers in patients presenting with severe SARS-CoV-2 infection. In particular, we identified variants in CD55 (a negative regulator of complement activation[15]), CFH and C4BPA, which play central roles in complement activation and innate immunity, to be associated with adverse clinical outcome. In addition to providing evidence that complement function modulates SARS-CoV-2 infection, the data point to several putative genetic markers of susceptibility. The results highlight the value of using a multi-modal analytical approach, combining molecular information from virus protein structure-function analysis with clinical informatics and genomics to reveal determinants and predictors of immunity, susceptibility, and clinical outcome associated with infection.

Results

Comorbidity statistics and covariances in our retrospective observational clinical cohort

To explore if conditions associated with dysregulation of the complement or coagulation systems impact adverse clinical outcomes associated with SARS-CoV-2, we conducted a retrospective observational study of patients treated at New York-Presbyterian/Columbia University Irving Medical Center for suspected infection (Table 1). Electronic health records (EHR) were used to define sex and age as well as histories of macular degeneration, thrombocytopenia, thrombosis, and hemorrhage, hypertension, type 2 diabetes, coronary artery disease, and obesity (see Methods). As shown in Table 1, of the 11,116 patients that presented to the hospital between February 1, 2020 and April 25, 2020 with suspected SARS-CoV-2 infection, 6,398 tested positive for the virus. Among these, 88 were patients with history of macular degeneration, four patients with complement deficiency disorders, and 1,179 patients with disorders associated with the coagulatory system. In addition, hypertension, coronary artery disease, diabetes, obesity, and annotated cough were represented by 1,922, 1,566, 847, 791, and 727 patients, respectively (Table 1). While CAD, hypertension, T2D, obesity, and coagulation disorders represent a group with the highest covariance, we find lower co-occurrence between these conditions and macular degeneration in both SARS-CoV-2 positive and negative individuals (Figure S1). Finally, of patients who are put on mechanical ventilation, we observed a 35% mortality rate, and 31% of deceased patients had been on mechanical respiration.

Macular degeneration and coagulation disorders are associated with SARS-CoV-2 outcomes

We estimated the univariate and age- and sex-corrected risk associated with baseline clinical history of previously reported SARS-CoV-2 risk factors (including hypertension, obesity, type 2 diabetes, and coronary artery disease) as well as coagulation and complement disorders using survival analysis and Cox proportional hazards regression modeling. As shown in Figure 1 and Table 1, we identified significant risk of mechanical respiration and mortality associated with age and sex, as well as history of hypertension, obesity, and type 2 diabetes (T2D), coronary artery disease (CAD). Moreover, we found that history of macular degeneration (a proxy for complement activation disorders) and coagulation disorders (thrombocytopenia, thrombosis, and hemorrhage) were at significantly increased risk of adverse clinical outcomes (including mechanical respiration and death) following SARS-CoV-2 infection (Figure 1, Table 1). Specifically, we observed a mechanical respiration rate of 15.9% (95% CI: 8.3-23.6) and a mortality rate of 25% (95% CI: 16.0-34.0) among patients with a history of macular degeneration, and rates of 9.4% (95% CI: 7.7-11.1) and 14.7% (95% CI: 12.7-16.7) for mechanical respiration and mortality, respectively, among patients with coagulation disorders (Table 1). Moreover, as shown in Figure 1b, patients with a history of macular degeneration appear to succumb to disease more rapidly than others. Critically, the contribution of age and sex was not sufficient to explain the increased risks associated with history of macular degeneration (Age/Sex-Corrected mechanical respiration HR=1.8 95% CI: 1.1-3.2, *P*value = 0.024; Age/Sex-Corrected mortality HR=1.7 95% CI: 1.1-2.5, *P*value = 0.022). Conversely, albeit in a small number of individuals, we observed that among patients with complement deficiency disorders, who are normally at increased risk of complications associated with infections, none required mechanical respiration or succumbed to their illness (Figure 1a and 1b). While we cannot rule out comorbidities that may be associated with macular degeneration, as shown in Figure S1, the correlation between macular degeneration and established covariates included in this study is low (correlation coefficients between 0.09 and 0.15). Together, these data suggest that hyper-active complement and coagulative states predispose individuals to adverse outcomes associated with SARS-CoV-2 infection, and that deficiencies in complement components may be protective. Importantly, given the low incidence rate of deficiencies in either complement or coagulation pathways, further analysis with larger clinical cohorts is warranted.

Genetic variation in complement and coagulation pathway components is associated with adverse SARS-CoV-2 infection outcome

The data highlighted above provide evidence that macular degeneration and coagulation disorders play a role in SARS-CoV-2 infection outcome. Importantly, macular degeneration and coagulation disorders have established genetic markers associated with regulators of these functions. However, any genetic components that may underlie the clinical trends we observed remain hidden due to the retrospective nature of the study and the lack of available genetic data on these patients. On the other hand, the UK Biobank, a prospective cohort study with deep genetic, physical, and health data collected on ~500,000 individuals across the United Kingdom[16], allows for genetic and epidemiological associations to be made. Among UK Biobank participants, recently released data include SARS-CoV-2-related clinical information on 1,474 suspected cases, including 669 patients who tested positive and 572 who required hospitalization. In a candidate driven approach, we leveraged this resource to evaluate if SNPs associated with components of complement or coagulation pathways are associated with SARS-CoV-2 infection or hospitalization. Briefly, we focused our analysis on 337,147 (181,032 female) subjects of White British descent, excluding 3rd degree and above relatedness and without aneuploidy[16]. Applying these restrictions to the UK Biobank SARS-CoV-2 cohort resulted in 957 patients with suspected infection (388 positive, 332 positive and hospitalized; see Methods).

Of the 805,426 genetic variants profiled in the UK Biobank, 4,248 are associated with 67 genes with known roles in regulating complement or coagulation pathways (see Methods). As highlighted in Figure 2 and further delineated in Table 2, we identified 10 loci (*P*value = $3x10^{-6}$; see Methods) representing 7 genes with study-wide significance at a minor allele frequency of 0.005 with multiple-hypothesis adjusted p-values less than 0.05. Among these and proximal to coagulation factor III (F3) is variant rs72729504, which we find to be associated with increased risk of adverse clinical outcome associated with SARS-CoV-2 infection (OR: 1.93 95% CI 1.34-2.79). Fibrin fragment D-dimer, one of several peptides produced when cross-linked fibrin is degraded by plasmin, is the most widely used clinical marker of activated blood coagulation. Among the genetic loci that influence D-dimer levels, GWAS studies have identified mutations in F3 as having the strongest association[17]. Importantly, increased D-dimer levels were recently reported to correlate with poor clinical outcome in SARS-CoV-2 infected patients[14]. So, while the functional role of rs72729504 remains to be elucidated, our observations suggest that this locus may represent a genetic marker of SARS-CoV-2 susceptibility and outcomes.

In addition to the SNP highlighted above, we identified 4 variants (rs45574833, rs61821114, rs61821041, and rs12064775) previously identified as risk alleles for AMD in the UKBB dataset[18]. Moreover, we find that each of these variants predisposes carriers to adverse clinical outcome (i.e. hospitalization) following SARS-CoV-2 infection (OR: 2.13-2.65; see Table 3 for variant specific 95% CI). A fifth variant, rs2230199, which maps to complement C3, was shown to be linked to AMD in an independent GWAS, however, this variant has not been associated with increased AMD risk in the UK population. The three SNPs that map to C3 each appear to confer some protection associated with SARS-CoV-2 infection (OR: 0.66-0.68 see Table 3 for variant specific 95% CI). In addition, two of the identified variants (rs61821114 and rs61821041) map to expression quantitative trait loci (eQTL) associated with Complement Decay-Accelerating Factor (CD55). This protein negatively regulates complement activation by accelerating the decay of complement proteins, thereby disrupting the cascade and preventing immune-mediated damage[7]. As shown in Figure 2b, these eQTLs result in decreased expression of CD55, thereby relieving the restraining function of this protein. In agreement with the functional role of CD55, we observe that these variants are associated with increased risk of adverse clinical outcome associated with SARS-CoV-2 infection (OR: 2.34-2.4 see Table 3 for variant specific 95% CI). Together, our observations point to genetic variation in complement and coagulation components as a contributing factor in SARS-CoV-2 mediated disease.

Discussion

Zoonotic infections like the SARS-CoV-2 pandemic pose tremendous risk to public health and socioeconomic factors on a global scale. While the innate and adaptive arms of the immune system are exquisitely equipped to deal with noxious agents including viruses, interactions between emerging pathogens and their human hosts can manifest in unpredictable ways. In the case of SARS-CoV-2 infection a combination of viral replication and immune mediated pathology are the primary drivers of morbidity and mortality. While recent analysis of coronavirus patients in China, suggests that high serum levels of interleukin-6 (IL-6), a proinflammatory cytokine, is associated with poor prognosis[14], further delineation of the regulatory programs that mediate immune pathology associated with SARS-CoV-2 infection is necessary. As illustrated in the accompanying paper and by the results presented herein, knowledge of molecular interactions between virus and host can refine hypothesis-driven discovery of disease determinants.

Our scan for virus-encoded structural mimics across Earth's virome pointed to molecular mimicry as a pervasive strategy employed by viruses and indicated that the protein structure space used by a given virus is dictated by the host proteome (see accompanying paper). Moreover, observations about how coronaviruses exploit this strategy provided clues about the cellular processes driving pathogenesis. Together with knowledge that CoV infections, including the SARS-CoV outbreak in 2002-2003 and the current SARS-CoV-2 outbreak[14], result in hyper-coagulative phenotypes[19], our protein structure-function analysis led us to hypothesize that conditions associated with complement or coagulatory dysfunction may influence outcomes of SARS-CoV-2 infections. Of these, among the most common are AMD (which is associated with hyper-activation of the complement pathway) and hyper-coagulative disorders. Their relatively high incidence rates together with SARS-CoV-2 prevalence in and around New York City made them reasonable candidates for a retrospective clinical study.

As presented above, in addition to rediscovering previously identified risk factors including age, sex, hypertension, and CAD we found that history of macular degeneration or coagulatory dysfunctions predispose patients to poor clinical outcomes (including increased risk of mechanical ventilation and death) following SARS-CoV-2 infection. Complement deficiencies on the other hand, appear to be protective. Their low incidence rates, however, make for a small sample size and invite further investigation. Further, retrospective studies of observational data have notable limitations in their data completeness, selection biases, and methods of data capture. As a result, claims on causality cannot be made - nor can we definitively rule out other clinical factors as possible drivers. Recognizing these limitations and that AMD and coagulative dysfunctions can have acquired and congenital etiologies, we implemented a focused, candidate-driven analysis of UK Biobank data to evaluate linkage between severe SARS-CoV-2 disease and genetic variation associated with complement and coagulation pathways. Our analysis identified 10 complement and coagulation associated loci including 4 that have been associated with AMD and 2 eQTLs that negatively impact expression of CD55, a critical negative regulator of the complement cascade. Though interpretation of our results may be limited by sample size, site-specific biases in clinical care decisions, ancestral homogeneity in the biobank data, and socioeconomic status of affected populations, to our knowledge, this is the first study to identify complement and coagulation functions as an underlying risk-factors of SARS-CoV-2 disease outcome. In addition, given an existing menu of immune-modulatory therapies that target complement and coagulation pathways, the discovery provides a rationale to investigate these options for the treatment of SARS-CoV-2 associated pathology.

Our study highlights the value of combining molecular information from virus protein structure-function analysis with orthogonal clinical data analysis to reveal determinants and/or predictors of immunity, susceptibility, and clinical outcome associated with infection. Such a framework can help refine large-scale genomics efforts and help power genomics studies based on informed biological and clinical conjectures. While identification of CoV encoded structural mimics guided our retrospective clinical studies, a molecular and functional link between those observations and our discovery of complement and coagulation functions as risk factors for SARS-CoV-2 pathogenesis remains to be elucidated. Nevertheless, the findings advance our understanding how SARS-CoV-2 infection leads to disease and can help explain variability in clinical outcomes. Among the implications, the data warrant heightened public health awareness for individuals most vulnerable to developing adverse SARS-CoV-2 mediated pathology.

Methods

Retrospective Clinical Study

Cohort and Study Description

In this observational cohort study, we used a data warehouse derived from electronic health records (EHRs) from 11,116 patients treated at New York-Presbyterian/Columbia University Irving Medical Center for suspected cases of SARS-CoV-2 infection. For these patients we collected contemporary data from their current encounter (i.e. the encounter associated with their suspected SARS-CoV-2 infection) as well as historical data, if available, from their previous encounters. Contemporary data (data collected between February 1, 2020 and April 12, 2020) included insurance billing information, laboratory measurements, procedures, and SARS-CoV-2 diagnostic test results. These data were derived from the data warehouse tables in Epic. 6,927 patients have historical data (data collected prior to September 24, 2019) available from an OMOP v5 instance stored using MySQL, which included all of the standard tables for recording condition, procedure, medication, and measurement data (among others). Of these we used the insurance billing information from the condition occurrence table and demographics from the person table. See *Preparation of data for modeling* for further details on data preparation.

We used the contemporary data to define inclusion criteria and outcomes (requiring mechanical respiration and mortality) and used historical data to define patient comorbidities. We defined three hypothesized comorbidity covariates, macular degeneration, complement deficiency disorders, and disorders of coagulation. We used historical data to define these comorbidities, age, and sex. We did not include race and ethnicity data in the modeling as we have previously found issues with the data quality[20]. The race/ethnicity data we do have is included in the tables for reference. We also modeled other comorbidities previously associated with morbidity and mortality (Zhou et al and others), including history of cardiovascular disease, hypertension, obesity, and diabetes (Table 1, Table S1) – all derived from the historical data. Coded covariate definitions, as well as lists of which diagnosis codes are most common in each group, are available in the supplemental materials and methods. We used established institutional procedures and an institutional clinical data warehouse to extract all data from the EHR.

Defining patient outcomes

Outcome definitions were defined by data derived from the electronic health record between February 1, 2020 and April 12, 2020. Mortality is derived from a death note filed by a resident or primary provider that records the date and time of death. Intubation was used as an intermediary endpoint and is a proxy for a patient requiring mechanical respiration. We used note types that were developed for patients with SARS-CoV-2 infection to record that this procedure was completed. We validated outcome data derived from notes against the patient's medical record using manual review.

Ethics and Data Governance Approval

The study is approved by the Columbia University Irving Medical Center Institutional Review Board (IRB# AAAL0601) and the requirement for an informed consent was waived. A data request associated with this protocol was submitted to the Tri-Institutional Request Assessment Committee (TRAC) of New-York Presbyterian, Columbia, and Cornell and approved. The research on the UK Biobank data has been conducted using the UK Biobank Resource under Application Number 41039.

Preparation of data for modeling

We used MySQL and python libraries (pymysql, pandas) to extract and prepare the data for modeling. The code for data preparation is available in the github (https://github.com/tatonetti-lab/complementcovid) as a Jupyter Notebook titled Data Setup. We begin by creating a master list of suspected covid patients. These are patients that are either diagnosed with the disease, as indicated by a ICD10 code for SARS-CoV-2 infection, in their billing data or a patient that was tested for the presence of the virus using RT-PCR as indicated by a "lab" order for the test. We found 2,821 using the former method and 11,116 patients using the latter. We then extracted birthdates, death dates (if the patient had died or a null value otherwise), and sex codes (1 for female, 2 for male). Patients which had sex codes for non-binary genders were excluded from our analysis. We then define a "first diagnosis date" for each patient as either their first diagnosis date (by billing code) or the first date that they tested positive for SARS-CoV-2, whichever comes first. Next, we calculate each patient's age at the time of this "first diagnosis date." Each of the outcomes and covariates are extracted from their respective tables as detailed in the github. Whenever possible, we use the highest-level ancestor code (from the structured vocabulary in OMOP) that represents the concept we want to model. We then use the concept ancestor tables to grab all the descendant codes. Note that diabetic kidney disease was considered for inclusion and so is represented in the data preparation script, however, it was never modeled. Cough is included as a covariate as a reference symptom for comparison. The last step in the preparation process was to compute the censor dates. To do, we iterated through each patient in our master list and computed their time (in days) to intubation (if they required mechanical respiration) or death (if they died). If not, then the study end date (April 25, 2020) was used as the patient's censored time (in days). Finally, for any patients that were not SARS-CoV-2 positive, their time-to-event values were set to a null indicator to be dropped from the dataset later. Finally, the data are all combined in a pandas (version 1.0.3) dataframe and saved to disk as a pickle file for efficient loading.

Statistical Model

Our patient timelines may be censored since our study cohort included patients that were being treated at the time of analysis. We performed survival analysis on the intubation orders and death using a Cox proportional-hazards model and visualized the risk using Kaplan-Meier curves using the lifelines python package (version 0.24.4). Error estimates on the Kaplan-Meier curves are estimated using Greenwood's Exponential Formula[21]. We fit both univariate models and models fit on the covariate, age, and sex and used log-likelihood to assess significance. We reported Cox proportional hazards coefficients and their 95% confidence intervals (Table 1). We modeled whether or not a patient had macular degeneration, a complement deficiency disorder, or a coagulation disorder as binary variables (1=yes, 0=no). Code definitions provided in Table S1. We also included other significant comorbidities suggested by previous studies, CAD, hypertension, T2DM, or obesity as binary variables (1=yes, 0=no), sex as a binary variable (0=female, 1=male), age as quantitative variable, older age (65+), and outcome as a binary variable (1=yes, 0=no). The outcome of interest was coded as 0 until the day it occurred (the date of the first intubation order following admission or the death date) or the date of analysis, whichever occurred first. Survival curves are generated for the indicated variables by setting all other variables to their respected averages within the training data. Note that we dropped patients who experienced the outcome before their initial diagnosis. This is either due to patients being hospitalized prior to infection (in the case of intubation) or errors in the coded data. We dropped 121 patients for intubation prior to infection and 12 patients for prior death. We also restricted the study to 90 days from the start date. One patient was removed for having an event outside of this range.

Covariate Correlations

Using the data prepared as discussed above, we computed pairwise statistical correlations between age, sex as well as history of macular degeneration, complement deficiency disorders, coagulation disorders, HTN, T2DM, obesity, and CAD. We computed them using data from all suspected patients (tested both positive and negative) as well as only those patients who tested positive. We chose spearman rho as our measure of correlation.

Statistical Software

Models were generated each day that data was available beginning on March 23rd, 2019 with data from patients available through that day. We used Jupyter Notebooks (jupyter-client version 5.3.4 and jupytercore version 4.6.1) running Python 3.7 and all fit models using the python lifelines package (version 0.24.4).

UK Biobank Genetic Analysis

Data Source

UK Biobank subjects that were of White British descent, in the UK Biobank PCA calculations and therefore without 3rd degree and above relatedness and without aneuploidy, were used in this study, totaling 337,147 subjects (181,032 females and 156,115 males) (Bycroft 2018). Of the nearly 500,000 participants, approximately 50,000 subjects were genotyped on the UK BiLEVE Array by Affymetrix while the rest were genotyped using the Applied Biosystems UK Biobank Axiom Array, with over 800,000 markers using build GRCh37 (hg19). The arrays share 95% marker coverage. We extracted markers with a minor allele frequency greater than 0.005, INFO score greater than 0.3, and Hardy-Weinberg equilibrium test mid-p value greater than 10-10 using PLINK2[22]. UKBB version 3 Imputation combined the Haplotype Research Consortium with the UK10K haplotype resource using the software IMPUTE4 (UK Biobank White paper). Association analyses were performed using a logistic regression model with additive gene dosage and covariates including age at 2018, sex, first 10 principal components (provided by the UK Biobank), and the genotyping array the sample was carried out on. We adjusted for multiple testing with FDR-corrected p-values using the Benjamini-Hochberg method.

Genetic Association Studies

We performed three study-wide association analyses: (i) comparing variants for SARS-CoV-2 positive patients against the entire population of 337,147 subjects, (ii) comparing positive patients who required hospitalization against the entire population, and (iii) comparing patients who tested negative versus the entire population.

Targeted Gene Set Definition

We identified a set of 69 genes relating to the complement and coagulation cascades from the Kyoto Encyclopedia of Genes and Genomes (KEGG accession id: hsa04610). For each gene, we used the transcriptional start and stop site from the hg19 build of the human genome to define a catchment window of 60kbp. From the 805,426 variants profiled in the UK Biobank genotyping data after quality control, 4,248 variants within the transcribed region of the genes of interest or within 60kbp flanking the transcribed region. After applying additional QC filters using PLINK2 (see *Data Source* above), 2,097 SNPs remained for analysis. We calculated counts for each variant for each of our groups of interest listed in *Genetic Association Studies* above.

SNP Set Empirical Statistical Evaluation

To assess the probability of getting 10 study-wide significant hits (using BH corrected p-value < 0.05), we used empirical sampling to generate 100 sets of randomly chosen SNPs. In each sample, 69 genes were chosen at random from the genome and mapped to nearby SNPs (within a 60kbp flanking region), resulting in sets of SNPs sized 1712 to 2945 – similar to the 2097 that resulted from our complement and coagulation cascade set. We then repeated the association study and counted the number of significant hits (using BH corrected p-value < 0.05). We fit the empirical data to a Poisson distribution and used the derived lambda to compute p-values for our observations of 10, 4, and 1 hit (corresponding to the number of significant results from our severe analysis, positive analysis, and negative, respectively). We performed a chi-squared goodness-of-fit test to verify the data were consistent with a Poisson.

Software

We used PLINK v2.00a2LM 64-bit Intel (26 Aug 2019) to run the genetic association analysis.

Declarations

Acknowledgements

This work was funded by NIH grants 5R01GM109018 and 5U54CA209997 to SS, R35GM131905 to NPT, F30HL140946 to PT, and equipment grants S10OD012351 and S10OD021764 to the Columbia University Department of Systems Biology.

Declaration of interests

The authors declare no competing interests

References

1 Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. *Science* **368**, 409-412, doi:10.1126/science.abb3405 (2020).

2 Dai, W. *et al.* Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science*, doi:10.1126/science.abb4489 (2020).

3 Gordon, D. E. *et al.* A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. *bioRxiv*, 2020.2003.2022.002386, doi:10.1101/2020.03.22.002386 (2020).

4 Chen, G. *et al.* Clinical and immunological features of severe and moderate coronavirus disease 2019. *J Clin Invest*, doi:10.1172/JCI137244 (2020).

5 Moore, B. J. B. & June, C. H. Cytokine release syndrome in severe COVID-19. *Science*, doi:10.1126/science.abb8925 (2020).

6 Lasso, G. *et al.* A Structure-Informed Atlas of Human-Virus Interactions. *Cell* **178**, 1526-1541 e1516, doi:10.1016/j.cell.2019.08.005 (2019).

7 Merle, N. S., Church, S. E., Fremeaux-Bacchi, V. & Roumenina, L. T. Complement System Part I - Molecular Mechanisms of Activation and Regulation. *Front Immunol* **6**, 262, doi:10.3389/fimmu.2015.00262 (2015).

8 Holers, V. M. Complement and its receptors: new insights into human disease. *Annu Rev Immunol* **32**, 433-459, doi:10.1146/annurev-immunol-032713-120154 (2014).

9 Wu, J. & Sun, X. Complement system and age-related macular degeneration: drugs and challenges. *Drug Des Devel Ther* **13**, 2413-2425, doi:10.2147/DDDT.S206355 (2019).

10 Ambati, J., Atkinson, J. P. & Gelfand, B. D. Immunology of age-related macular degeneration. *Nat Rev Immunol* **13**, 438-451, doi:10.1038/nri3459 (2013).

11 Khandhadia, S., Cipriani, V., Yates, J. R. & Lotery, A. J. Age-related macular degeneration and the complement system. *Immunobiology* **217**, 127-146, doi:10.1016/j.imbio.2011.07.019 (2012).

12 Degn, S. E., Jensenius, J. C. & Thiel, S. Disease-causing mutations in genes of the complement system. *Am J Hum Genet* **88**, 689-705, doi:10.1016/j.ajhg.2011.05.011 (2011).

13 Liszewski, M. K., Java, A., Schramm, E. C. & Atkinson, J. P. Complement Dysregulation and Disease: Insights from Contemporary Genetics. *Annu Rev Pathol* **12**, 25-52, doi:10.1146/annurev-pathol-012615-044145 (2017).

14 Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054-1062, doi:10.1016/S0140-6736(20)30566-3 (2020).

15 Nicholson-Weller, A. & Wang, C. E. Structure and function of decay accelerating factor CD55. *J Lab Clin Med* **123**, 485-491 (1994).

16 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).

17 Smith, N. L. *et al.* Genetic predictors of fibrin D-dimer levels in healthy adults. *Circulation* **123**, 1864-1872, doi:10.1161/CIRCULATIONAHA.110.009480 (2011).

18 Han, X. *et al.* Genome-wide meta-analysis identifies novel loci associated with age-related macular degeneration. *J Hum Genet*, doi:10.1038/s10038-020-0750-x (2020).

19 Goeijenbier, M. *et al.* Review: Viral infections and mechanisms of thrombosis and bleeding. *J Med Virol* **84**, 1680-1696, doi:10.1002/jmv.23354 (2012).

20 Polubriaginof, F. C. G. *et al.* Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc* **26**, 730-736, doi:10.1093/jamia/ocz113 (2019).

21 Hosmer, D. W., Lemeshow, S. & May, S. *Applied survival analysis : regression modeling of time-to-event data*. 2nd edn, (Wiley-Interscience, 2008).

22 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

Supplemental Legends

Figure S1| Covariate correlation in clinical data. **a**, Spearman correlation between modeled covariates in patients were diagnosed or tested positive for SARS-CoV-2: age, sex, macular degeneration (macula), complement deficiency disorders (CD), coagulation disorders (coagulation), hypertension, Type 2 Diabetes, obesity, and coronary artery disease (CAD). **b**, Spearman correlations, as in (**a**), for all patients (includes patients who tested negative for SARS-CoV-2).

Figures

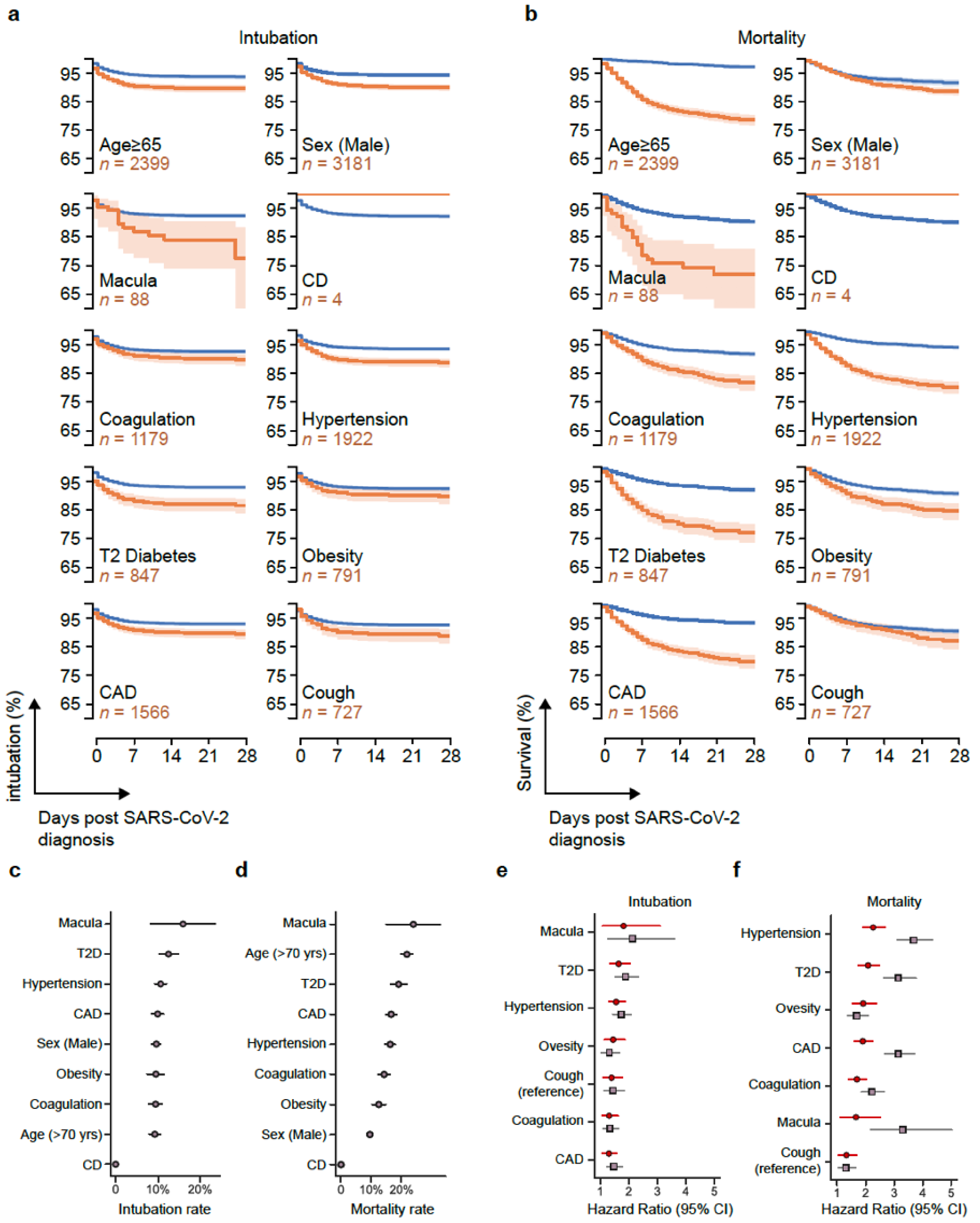
a
Intubation
Age≥65
n = 2399
Sex (Male)
n = 3181
Macula
n = 88
CD
n = 4
Coagulation
n = 1179
Hypertension
n = 1922
T2 Diabetes
n = 847
Obesity
n = 791
CAD
n = 1566
Cough
n = 727
95
85
75
65
0
7
14
21
28
intubation (%)
Days post SARS-CoV-2 diagnosis
b
Mortality
Survival (%)
Days post SARS-CoV-2 diagnosis
c
Macula
T2D
Hypertension
CAD
Sex (Male)
Obesity
Coagulation
Age (>70 yrs)
CD
0
10%
20%
Intubation rate
d
Macula
Age (>70 yrs)
T2D
CAD
Hypertension
Coagulation
Obesity
Sex (Male)
CD
0
10%
20%
Mortality rate
e
Intubation
Macula
T2D
Hypertension
Ovesity
Cough (reference)
Coagulation
CAD
1
2
3
4
5
Hazard Ratio (95% CI)
f
Mortality
Hypertension
T2D
Ovesity
CAD
Coagulation
Macula
Cough (reference)
1
2
3
4
5
Hazard Ratio (95% CI)

Figure 1

History of macular degeneration and coagulation disorders are associated with adverse outcomes after confirmed SARS-CoV-2 infection. a, Kaplan-Meier curves for 10 binary conditions: age over 70, male sex, macular degeneration (Macula), complement deficiency disorders (CD), coagulation, hypertension, type 2 diabetes (T2DM), obesity, coronary artery disease (CAD), and cough. The survival for the patients with the named condition are shown in orange. The shaded region indicates the 95% confidence interval. The blue survival line is for patients without the named condition. Note that none of the four patients with CD required mechanical ventilation. b, Kaplan-Meier curves for the same 10 conditions as in (a). All four patients with CD survived (not statistically significant). c, Intubation rates across the binary conditions. Mortality (N=88) was highest in patients with a history of macular degeneration, followed by Type 2 Diabetes and Hypertension. d, Mortality rates across the binary conditions. Patients with a history of macular degeneration saw the highest mortality rates, followed by Age ≥ 65 and Type 2 Diabetes. e, Hazard ratios, estimated using a Cox proportional hazards model, for risk if intubation (as a validated proxy for requiring mechanical respiration). f, Similarly, hazard ratios for mortality, estimated using a Cox proportional hazards model. Hazard ratios and statistical significances are shown in Table 1.

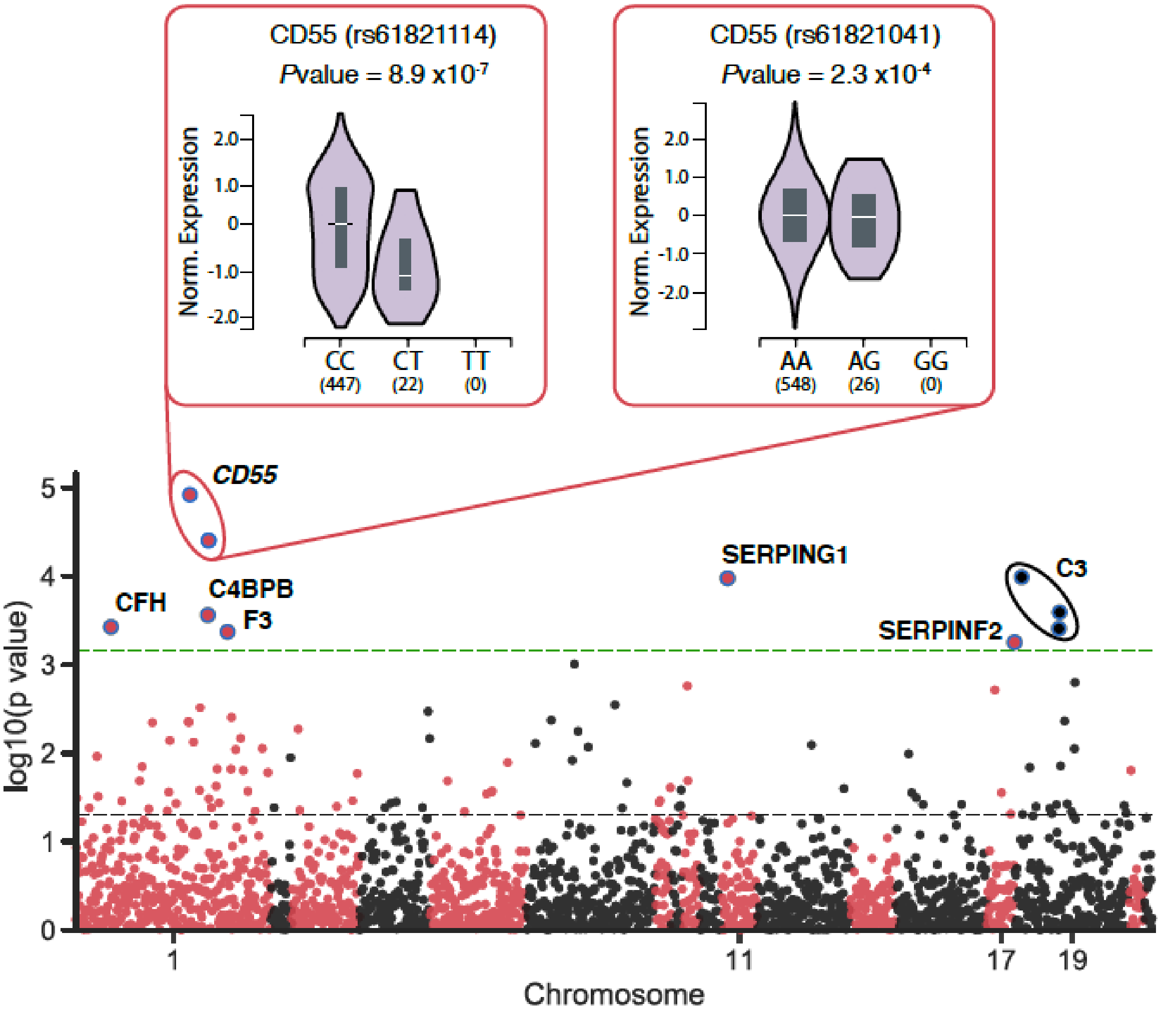

Figure 2

Genetic association study of 332 SARS-CoV-2 infected patients who required hospitalization in the UK Biobank. Shown is a Manhattan plot for 2,097 single nucleotide polymorphisms (SNPs) associated with the complement and coagulation pathway genes. Study-wide significance was determined using a Benjamini-Hochberg adjusted Pvalue < 0.05. The 10 study-wide significant loci are annotated with related complement genes, full details in Table 2. The panels above show expression quantitative trait relationships between two study-wide significant SNPs and CD55. The minor allele is associated with significantly reduced CD55 expression in omentum and thyroid.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 3729811dataset3422369q9lyls1.xlsx
- SuppTable1.pdf
- SuppFig1.pdf
- SuppTable2.pdf