Genetic architecture of host proteins involved in SARS-CoV-2 infection

Maik Pietzner1, Eleanor Wheeler1, Julia Carrasco-Zanini1, Johannes Raffler2, Nicola D. Kerrison1, Erin Oerton1, Victoria P. W. Auyeung1, Jian’an Luan1, Chris Finan3,4, Juan P. Casas5,6, Rachel Ostroff7, Steve A. Williams7, Gabi Kastenmüller2, Markus Ralser8,9, Eric R. Gamazon1,10, Nicholas J. Wareham1,11, Aroon D. Hingorani3,4,12 & Claudia Langenberg1,8,11,13

Understanding the genetic architecture of host proteins interacting with SARS-CoV-2 or mediating the maladaptive host response to COVID-19 can help to identify new or repurpose existing drugs targeting those proteins. We present a genetic discovery study of 179 such host proteins among 10,708 individuals using an aptamer-based technique. We identify 220 host DNA sequence variants acting in cis (MAF 0.01-49.9%) and explaining 0.3-70.9% of the variance of 97 of these proteins, including 45 with no previously known protein quantitative trait loci (pQTL) and 38 encoding current drug targets. Systematic characterization of pQTLs across the phenome identified protein-drug-disease links and evidence that putative viral interaction partners such as MARK3 affect immune response. Our results accelerate the evaluation and prioritization of new drug development programmes and repurposing of trials to prevent, treat or reduce adverse outcomes. Rapid sharing and detailed interrogation of results is facilitated through an interactive webserver (https://omicscience.org/apps/covidpgwas/).
The pandemic of the novel coronavirus SARS-CoV-2 infection, the cause of COVID-19, is causing severe global disruption and excess mortality\(^1,^2\). While ultimately strategies are required that create vaccine-derived population immunity, in the medium term there is a need to develop new therapies or to repurpose existing drugs that are effective in treating patients with severe complications of COVID-19, and also to identify agents that might protect vulnerable individuals from becoming infected. The experimental characterization of 332 SARS-CoV-2-human protein–protein interactions and their mapping to 69 existing FDA-approved drugs, drugs in clinical trials, and/or preclinical compounds\(^3\) points to new therapeutic strategies, some of which are currently being tested. The measurement of circulating host proteins that associate with COVID-19 severity or mortality also provides insight into potentially targetable maladaptive host responses with current interest being focused on the innate immune response\(^4\), coagulation\(^5,^6\), and novel candidate proteins\(^7\).

Naturally occurring sequence variation in or near a human gene that is encoding a drug target and affecting its expression or activity can be used to provide direct support for drug mechanisms and safety in humans. This approach is now used by major pharmaceutical companies for drug target identification and validation for a wide range of non-communicable diseases, and to guide drug repurposing\(^8,^9\). Genetic evidence linking molecular targets to diseases relies on our understanding of the genetic architecture of drug targets. Proteins are the most common druggable targets in or near the protein-coding gene of a human gene that is encoding a drug target and affecting its expression. We identify genetic variants that are associated with COVID-19-related phenotype can be mediated by the effect of genetically reduced activity of an ubiquitously expressed enzyme.

Examples of evidence to be involved in the pathology of COVID-19: (1) two human proteins related to viral entry\(^12\), (2) 332 human proteins shown to interact with viral proteins\(^3\), (3) 26 proteomic markers of disease severity\(^7\), and (4) 54 protein biomarkers of adverse prognosis, complications, and disease deterioration\(^4,^6,^13\) (Fig. 1 and Supplementary Data 1). Of the 409 proteins prioritized, 179 were detectable by the currently most-comprehensive proteomic assay using an aptamer-based technology (SomaScan\(^6\)), including 28 recognized by more than one aptamer (i.e., 179 proteins recognized by 190 aptamers). We further included complementary data from proximity extensions assays (Olink\(^\circ\)) for 32 out of the 179 candidate proteins in a subset of 485 Fenland participants of the Fenland study with available genotype data.

**Results**

**Coverage of COVID-19-relevant proteins.** We identified COVID-19-relevant candidate proteins based on different layers of evidence to be involved in the pathology of COVID-19: (1) two human proteins related to viral entry\(^12\), (2) 332 human proteins shown to interact with viral proteins\(^3\), (3) 26 proteomic markers of disease severity\(^7\), and (4) 54 protein biomarkers of adverse prognosis, complications, and disease deterioration\(^4,^6,^13\) (Fig. 1 and Supplementary Data 1). Of the 409 proteins prioritized, 179 were detectable by the currently most-comprehensive proteomic assay using an aptamer-based technology (SomaScan\(^6\)), including 28 recognized by more than one aptamer (i.e., 179 proteins recognized by 190 aptamers). We further included complementary data from proximity extensions assays (Olink\(^\circ\)) for 32 out of the 179 candidate proteins in a subset of 485 Fenland participants of the Fenland study with available genotype data.

**Fig. 1 Flowchart of the identification of candidate proteins and coverage by the SomaScan v4 platform within the Fenland cohort.** More details for each protein targeted are given in Supplementary Data 1.
study individuals (Supplementary Data 1). Of these 179 proteins, 111 (Supplementary Data 1) were classified as druggable proteins, including 32 by existing or developmental drugs\(^1\), and another 22 highlighted by Gordon et al. as interacting with SARS-CoV-2 proteins\(^3\). To simplify the presentation of results we introduce the following terminology: we define a protein as a unique combination of UniProt entries, i.e., including single proteins and protein complexes. We further define a protein target as the gene product recognized by a specific aptamer, and, finally, an aptamer as a specific DNA-oligonucleotide designed to bind to a specific protein target.

Local genetic architecture of protein targets. We successfully identified 220 DNA sequence variants acting in cis for 97 proteins recognized by 106 aptamers (Fig. 2 and Supplementary Fig. 1 and Data 2). For 45 of these proteins, no pQTLs had previously been reported. Of the nine proteins recognized by more than one aptamer, sentinel sequence variants were concordant (identical or nearly identical) for 97 proteins (0.4–70.9\%). The cluster persisted in different sensitivity analyses, such as omitting highly pleiotropic genomic regions (associated with >20 aptamers) or lead cis-pQTLs (Supplementary Fig. 2). Manual curation highlighted protein modification and vesicle trafficking involving the endoplasmic reticulum as highly represented biological processes related to this cluster. Among these proteins, nine are the targets of known drugs (e.g., COMT, PGES2, PLOD2, ERO1B, XTP3B, FKB7, or MARK3). The high genetic correlation between these proteins indicates shared polygenic architecture acting in trans, which is unlikely to be driven by selected pleiotropic loci identified in the present study.

We further identified strong genetic correlations (|\(r| > 0.5\)) between smaller sets of proteins related to COVID-19 severity, and host proteins relevant to viral replication such as between IL-6-induced proteins (SAA1, SAA2, and CD14) and fibrinogen 5 (FBLN5).

A tiered system for trans-pQTLs. We created a pragmatic, tiered system to guide selection of trans-pQTLs for downstream analyses. We defined as (a) “specific” trans-pQTLs those solely associated with a single protein or protein targets creating a protein complex, (b) “vertically” pleiotropic trans-pQTLs those associated only with aptamers belonging to the same common biological process (GO-term), and (c) as “horizontally” pleiotropic trans-pQTLs all remaining ones, i.e., those associated with aptamers across diverse biological processes. We used the entire set of aptamers available on the SomaScan v4 platform, N = 4979, targets, we found evidence for (a) protein-specific effects for 23 regions, (b) possible vertical pleiotropy for six, and (c) horizontal pleiotropy for nine lead cis-pQTLs. A similar distribution across those categories was seen for the remaining cis-pQTLs (Fishers exact test \(p\)-value = 0.49).

To investigate dependencies between host proteins predicted to interact with SARS-CoV-2 and those related to the maladaptive host response we computed genetic correlations for all proteins with at least one cis-pQTL and reliable heritability estimates (see Methods). Among 86 considered proteins, we identified a highly connected subgroup of 24 proteins including 19 SARS-CoV-2-human protein interaction partners (e.g., RAB1A, RAB2A, AP2A2, PLD3, KDEI2, GDP/GTP exchange protein, PPT1, GT251, or PKP2) and five proteins related to cytokine storm (IL-1Rp2 and IL-1Ra), fibrinolysis (PAI-1), and severity of COVID-19 (GSN (gelsolin)) (Fig. 3). The cluster persisted in different sensitivity analyses, such as omitting highly pleiotropic genomic regions (associated with >20 aptamers) or lead cis-pQTLs (Supplementary Fig. 2).
to establish those tiers. Among 451 SNPs acting solely as trans-pQTLs, 114 (25.3%) were specific for a protein target, 29 (6.4%) showed evidence of vertical pleiotropy, and 308 (68.3%) evidence of horizontal pleiotropy, indicating that trans-pQTLs exert their effects on the circulating proteome through diverse mechanisms.

An apparent source of horizontal pleiotropy included possible artifacts of the measurement procedure. The most pleotropic trans-pQTL (rs4658046, minor allele frequency (MAF) = 0.39) showed associations with over 2000 aptamers and is in high LD ($r^2 = 0.99$) with a known missense variant at CFH (rs1061170). This missense variant was shown, among others, to increase platelet counts or higher susceptibility to platelet activation may result in the secretion of proteins into plasma during sample preparation. While we report such examples, the extremely standardized and well controlled sample handling of the contemporary and large Fenland cohort has minimized the effects of delayed sample handling on proteomic assessment, as compared to historical cohorts or convenience samples such as from blood donors, evidenced by the fact that previously reported and established sample handling related loci, such as rs62143194 in NLRP12, are not significant in our study.

A prominent example seen in our results and reported in a previous study is variant rs1354034 in ARHGEF3, associated with over 1000 aptamers (on the full SomaScan platform). ARHGEF3 is a known locus associated with platelet counts and a master regulator of megakaryopoiesis, either genetically determined higher platelet counts or higher susceptibility to platelet activation may result in the secretion of proteins into plasma during sample preparation.
We identified a few variants with evidence for vertical pleiotropic effects, including rs2289252 a cis-pQTL for coagulation factor XI, which was specifically associated (p < 5 × 10⁻⁸) with members of the coagulation cascade such as Kinigon 1, alpha-2-macroglobulin, kallikrein B, plasma (Fletcher factor) 1, and thrombin.

Finally, for 27 out of 98 aptamers with at least one cis- and trans-pQTL, we identified no or only very weak evidence for horizontal pleiotropy, i.e., associations in trans for no more than one aptamer, suggesting that those might be used as additional instruments to genetically predict protein levels in independent cohorts for causal assessment.

Host factors related to candidate proteins. We investigated host factors that may explain variance in the plasma abundances of aptamers targeting high-priority candidate proteins using a variance decomposition approach (see Methods). Genetic factors explained more variance compared to any other tested host factors for 63 out of 106 aptamers with IL-6 sRa, collagen a1(VI), or QSOX2 being the strongest genetically determined examples (Fig. 4). The composition of non-genetic host factors contributing most to the variance explained appeared to be protein-specific (Fig. 4). For SMOC1 and Interleukin-1 receptor-like 1, for example, sex explained 23.8% and 17.9% of their variance, respectively, indicating different distributions in men and women. Other examples for single factors with large contributions included plasma alanine aminotransferase (15.4% in the variance of NADPH-P450 oxidoreductase) or age (14.2%) in the variance of GDF-15/MIC-1. We observed a strong and diverse contribution of NADPH-P450 oxidoreductase) or age (14.2% in the variance of included plasma alanine aminotransferase (15.4% in the variance respectively, indicating different distributions in men and women.

Patients with multiple chronic conditions are at higher risk of getting severe COVID-19 disease²¹⁻²⁰ and we investigated the influence of disease susceptibility on protein targets of interest using genetic risk scores (GRS) for major metabolic (e.g., type 2 diabetes, body mass index (BMI), and waist-to-hip ratio (WHR)), respiratory (e.g., asthma), and cardiovascular (e.g., coronary artery disease (CAD)) phenotypes (Supplementary Fig. 3). We identified positive associations between the GRS for lung function and CAD with plasma abundances of the viral interaction partner QSOX2. However, as described below, these disease score to protein associations were likely driven by genetic confounding. Specifically, (cis) variants in proximity (±500 kb) to the protein-encoding gene (QSOX2) were genome-wide significant for forced expiratory volume (FEV1) and forced vital capacity (FVC), and exclusion of this region from the lung function genetic score abolished the score to QSOX2 association. None of the three lead cis-pQTLs were in strong LD with the lead lung function variant (r² < 0.4) and genetic colocalization of QSOX2 plasma levels and lung function²¹ showed strong evidence for distinct genetic signals (posterior probability of near 100%). The association with the CAD-GRS was attributed to the large contribution of the ABO locus to plasma levels of QSOX2, and exclusion of this locus from the CAD score led to the loss of association with QSOX2.

The GRSs for WHR (N = 11), estimated glomerular filtration rate (eGFR; N = 7), and CAD (N = 4) were associated with higher as well as lower abundance of different aptamers, and the asthma-GRS was specifically and positively associated with IL1R1L1. Individuals with genetic susceptibility to a higher WHR had higher abundances of four putative viral interaction partners (LMAN2, ETFA, TBCA, and SELENOS), and lower levels of albumin, GSN, and ITIH3. Lower plasma abundances of GSN have been repeatedly associated with severity of COVID-19²⁻². The association with plasma abundances of LMAN2 (or VIP36) was shared with the eGFR-GRS but in opposing direction (inversely). VIP36 is shed from the plasma membrane upon inflammatory stimuli and has been shown to enhance phagocytosis by macrophages. The higher plasma levels of VIP36, suggesting an enhanced immune response, among individuals with genetically higher WHR and lower kidney function appear contradictory as a higher WHR, indicating abdominal adiposity, and lower kidney function are considered as risk factors for COVID-19.
Integration of gene expression data. Plasma abundances of proteins are influenced by multiple processes, including expression of the encoding gene in diverse tissues. To test whether pQTLs are also gene expression QTLs, we integrated predicted gene expression data across five tissues of direct or indirect relevance to SARS-CoV-2 infection and COVID-19 (lung, whole blood, heart - left ventricle, heart - atrial appendage, and liver) from the GTEx project\(^{24,25}\) (version 8) using PrediXcan\(^{26}\). For 100 out of 102 high-priority protein targets we could establish genetically anchored gene prediction models, the protein-encoding gene was the strongest model in 31 (31%) cases in at least one tissue. Nevertheless, at 65 loci predicted gene expression of the protein-encoding gene and protein abundance were significantly associated \((p < 0.05)\) with varying tissue specificity (Fig. 5), similar to previous reports\(^{16,27}\). Predicted gene expression of \(IL17RA\), \(EIF4E2\), \(FKBP7\), \(SERPINC1\), \(EROLB1\) (all drug-gable targets), \(POR\), \(RAB2A\), \(KDEL2\), \(C1R\), \(AES\), and \(ACADM\), for example, was consistently associated with corresponding protein levels in plasma across at least three tissues, whereas gene expression in lung-only was associated with plasma levels of \(SAA1\), \(SAA2\), and \(SERPIN1\). For the majority of protein targets PrediXcan selected genes other than the protein-encoding gene as most strongly associated with pQTL data. Testing for enriched biological terms\(^{28}\) across all significantly associated predicted gene expression models \((p < 10^{-6}\), to account for multiple testing) in lung highlighted “signal peptide” (false discovery rate \((FDR) = 2.5 \times 10^{-5}\), “glycoproteins” \((FDR = 1.7 \times 10^{-4}\), or “disulfide bonds” \((FDR = 2.8 \times 10^{-4}\) as relevant processes. These are involved in the transport and post-translational modification of proteins before secretion and highlight the complexity of plasma proteins beyond a linear dose–response relationship with tissue abundance of the corresponding mRNA.

Cross-platform comparison. Protein measurements can be affected by the presence of protein altering variants (PAVs) changing binding epitopes and we tested cross-platform...
consistency of identified pQTLs using data on 33 protein targets measured across 12 panels using Olink's proximity extension assays, which rely on polyclonal antibodies, among 485 participants of the Fenland study. PAVs can be expected to affect only a subset, if any, of the binding epitopes targeted by the different antibody populations.

We compared effect estimates for 29 cis- and 96 trans-pQTLs based on a reciprocal look-up across both platforms (see Methods and Supplementary Data 5). We observed a strong correlation of effect estimates among 29 cis-pQTLs \((r = 0.75\) Supplementary Fig. 4) and slightly lower correlation for trans-pQTLs \((r = 0.54\) indicating good agreement between platforms for this specific subset of proteins. In detail, 36 pQTLs (30%) discovered using the far larger SOMAscan-based effort were replicated \((p < 0.05\) and directionally consistent) in the smaller subset of participants with overlapping measurements.

We identified evidence for inconsistent lead cis-pQTLs for two of these 33 protein targets. The lead cis-pQTL for GDF-15 from SOMAscan \((rs75347775\) was not significantly associated with GDF-15 levels measured using the Olink assay despite a clear and established signal in cis for the Olink measure\(^a\) \((rs1227731, beta = 0.59, p < 6.5 \times 10^{-16}\)\. However, rs1227731 was a secondary signal for the SomaScan assay \((beta = 0.29, p < 5.8 \times 10^{-6}\) highlighting the value of conditional analyses to recover true signals for cases where these are “overshadowed” by potential false positive lead signals caused by epitope effects. Another protein, the poliovirus receptor (PVR), did not have a cis-pQTL in the SomaScan but in the Olink-based discovery \((rs10419829, beta = 0.84, p < 2.9 \times 10^{-33}\)\), which in the context of an observational correlation of \(r = 0.02\) suggests that the two technologies target different protein targets or isoforms. A similar example is ACE2, the entry receptor for SARS-CoV-2, with a correlation of \(r = 0.05\) between assays and for which we identified only trans-pQTLs with evidence for horizontal pleiotropy (Supplementary Data 3).

The SCALLOP consortium investigates genetic association data focused on Olink protein measures, and can be a useful and complementary resource for the subset of proteins of interest that are captured (https://www.olink.com/scallop/).

Drug target analysis. We identified pQTLs for 105 COVID-19-relevant proteins, including 75 with at least one cis-pQTL, already the target of existing drugs or known to be druggable\(^4\). These cis-pQTLs can be used to try and emulate targeting of those proteins as treatment for COVID-19, once large and robust GWAS results on COVID-19 related outcomes become available, e.g., to carefully test whether patients requiring hospitalization differ from mild cases by the frequency of protein-increasing/decreasing alleles. Since (risk) alleles are randomly allocated during meiosis and inherited independent of virus exposure, they represent the effects of life-long lower or higher plasma protein levels, which may confer protection from or higher susceptibility to severe COVID-19. For example, one target identified through analysis of host–virus protein interactions is prostaglandin E synthase 2 (PGES2) involved in prostaglandin biosynthesis. Non-steroidal anti-inflammatory drugs (NSAIDs) are also known to suppress synthesis of prostaglandins and although the evidence is currently weak, concerns have been raised that NSAIDs may worsen outcomes in patients with COVID-19\(^30\)\. The cis-pQTLs we identified for PGES2 might be useful to explore this further.

Among the 105 proteins, 18 are targets of licensed or clinical phase compounds in the ChEMBL database. Thirteen of these were targets of drugs affecting coagulation or fibrinolytic pathways and five were targets of drugs influencing the inflammatory response. Drugs mapping to targets in the coagulation system included inhibitors of factor 2 (e.g., dabigatran and bivalirudin), factor 5 (drotrecogin alfa), factor 10 (e.g., apixaban, rivaroxaban), von Willebrand factor (caplacizumab), plasminogen activator inhibitor 1 (aleplasin), and tissue plasminogen activator. Drugs mapping to inflammation targets included toclizumab and satralizumab (targeting the interleukin-6 receptor), brodalumab (targeting the soluble interleukin-17 receptor), and anakinra (targeting interleukin-1 receptor type 1). Two targets with pQTLs (catechol O-methyltransferase and alpha-galactosidase-A) were identified as potential virus–host interacting proteins. The former is the target for a drug for Parkinson’s disease (entacapone) and the latter is deficient in Fabry’s disease, a lysosomal disorder for which miglustat (a drug that stabilizes certain mutant forms of alpha-galactosidase-A) is a treatment.

Another 24 protein targets have no current licensed medicines but are deemed to be druggable including multiple additional targets related to the inflammatory response, prioritized by untargeted proteomics analysis of COVID-19 patient plasma samples. These included multiple components of the complement cascade (e.g., Complement C2, Complement component C8, Complement component C8 gamma chain, and Complement factor H). A number of inhibitors of the complement cascade are licensed (e.g., the C5 inhibitor eculizumab) or in development, although none target the specific complement components prioritized in the current analysis.

Linking cis-pQTLs to clinical outcomes. To systematically assess phenotypic consequences, including possible adverse effects, of the identified cis-pQTLs we used three different strategies.

We first tested whether any of the 220 cis-pQTLs or proxies in high LD \((r^2 > 0.8\) have been reported in the GWAS Catalog and identified links between genetically verified drug targets and corresponding indications for lead cis-pQTLs at \(F_2\) (rs1799963 associated with venous thrombosis\(^31\)), IL6R (rs2228145 with rheumatoid arthritis\(^32\)) and PLG (rs4252185 associated with CAD\(^33\)).

To systematically evaluate whether higher plasma levels of candidate proteins are associated with disease risk, we tested GRS \((cis-GRS)\) for all 106 aptamers for their associations with 633 ICD-10 coded outcomes in the UK Biobank. We identified nine significant associations (false discovery rate <10%), including the druggable example of a thrombin–cis-GRS (2 cis-pQTLs as instruments) and increased risk of pulmonary embolism (ICD-10 code: 126) as well as phlebitis and thrombophlebitis (ICD-10 code: 180) (Supplementary Data 6).

To maximize power for disease outcomes, include clinically relevant risk factors, and allow for variant-specific effects, we complemented the phenotype-wide strategy with a comprehensive look-up for genome-wide significant associations in the MR-Base platform\(^4\).

Out of the 220 variants queried, 74 showed at least one genome-wide significant association (Fig. 6), 20 of which were cis-pQTLs for established drug targets. We obtained high posterior probabilities \((PP > 75%)\) for a shared genetic signals between 25 cis-pQTLs and at least one phenotypic trait using statistical (conditional) colocalization (Fig. 6 and Supplementary Data 7). Among these was rs8022179, a novel cis-pQTL for microtubule affinity-regulating kinase 3 (MARK3), a regional lead signal for monocyte count and granulocyte percentage of myeloid white cells\(^17\)\. The variant showed associations with higher plasma levels of MARK3 and monocyte count and therefore suppression of MARK3 expression with protein kinase inhibitors such as midostaurin may affect the protein host response to the virus. The important role of monocytes and macrophages in the pathology of COVID-19 has been recognized\(^4\), and a range of
immunomodulatory agents are currently evaluated in clinical trials, with a particular focus on the blockade of IL-6 and IL-1β. Our findings indicate that proteins utilized by the virus itself, such as MARK3, SMOC1, or IL-6 receptor, may increase the number of innate immune cells circulating in the blood and thereby contribute to a hyperinflammatory or hypercoagulable state. Stratification of large COVID-19 patient populations by cis-pQTL genotypes that contribute to stimulation/repression of a specific immune signaling pathway is one potential application of our results. However, such investigations would need to be large, i.e., include thousands of patients, and results need to be interpreted with caution as targeting those proteins can have effects not anticipated by the genetic analysis, which cannot mimic short term and dose-dependent “drug” exposure.

We observed general consistency among phenotypic traits colocalising with cis-pQTLs, i.e., traits were closely related and effect estimates were aligned to the protein increasing allele, i.e., bars with a positive -log10 (p-values) indicate positive associations with a trait from the database and vice versa. Clinical traits are grouped by higher-level categories and colored accordingly. GIT = gastrointestinal tract, Misc = Miscellaneous, No coloc. pos. = colocalization for secondary signals was not possible.

**Fig. 6** Circos plot summarizing genome-wide significant associations between 74 cis-pQTLs and 239 traits\(^4\) in the inner ring and results from statistical colocalization in the outer ring. The dashed line in the outer ring indicates a posterior probability of 75% of shared genetic signal between the protein and a phenotypic trait. Protein targets are classified on the basis of their reported relation to SARS-CoV-2 and COVID-19. Each slice contains any cis-pQTLs associated with the target protein annotated and effect estimates were aligned to the protein increasing allele, i.e., bars with a positive -log10 (p-values) indicate positive associations with a trait from the database and vice versa. Clinical traits are grouped by higher-level categories and colored accordingly. GIT = gastrointestinal tract, Misc = Miscellaneous, No coloc. pos. = colocalization for secondary signals was not possible.

---

**ARTICLE NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-020-19996-z**

---
inversely associated with the risk of coronary heart disease and rheumatoid arthritis but positively with the risk for allergic disease, such as asthma.

**Coagulation factors and the ABO locus.** A recent GWAS identified two independent genomic loci to be associated with an increased risk of respiratory failure in COVID-19 patients compared to healthy blood donors\(^9\). We observed six proteins to be associated positively with the lead signal (rs657152) at the ABO locus (coagulation factor VIII, sulfhydryl oxidase 2 (QSOX2), von Willebrand factor, SVEP1, and heme oxygenase 1) and one inverse association (interleukin-6 receptor subunit beta), but did not observe significantly associated proteins with the lead variant (rs11385942) at 3p21.31. We identified a cluster of 10 aptamers (targeting SVEP1, coagulation factor VIII, ferritin, heme oxygenase 1, von Willebrand factor, plasminogen, PLOD2, and CD14) sharing a genetic signal (regional probability: 0.88; rs941137; Supplementary Fig. 5), which was in high LD \((r^2 = 0.85)\) with the lead ABO signal associated with a higher risk for respiratory failure among COVID-19 patients.

**Webserver.** To facilitate in-depth exploration of candidate proteins, i.e., those with at least one cis-pQTL, we created an online resource (https://omicscience.org/apps/covidpgwas/). The webserver provides an intuitive representation of genetic findings, including the opportunity of customized look-ups and downloads of the summary statistics for specific genomic regions and protein targets of interest (Fig. 7). We further provide detailed information for each protein target, including links to relevant databases, such as UniProt or Reactome, information on currently available drugs or those in development as well as characterization of associated SNPs. The webserver further enables the query of SNPs across proteins to assess specificity and to find co-associated protein targets.

**Discussion**

We present a systematic genetic investigation of host proteins reported to interact with SARS-CoV-2 proteins, to be related to virus entry, host hyperimmune or procoagulant responses, or be associated with the severity of COVID-19. The integration of large-scale genomic and aptamer-based plasma proteomic data from 10,708 individuals improves our understanding of the genetic architecture of 97 of 179 investigated host proteins by identifying 220 cis-acting variants that explain up to 70% of the variance in these proteins, including 45 with no previously known pQTL and 38 encoding current drug targets. Our findings, shared in an interactive webserver (https://omicscience.org/apps/covidpgwas/), enable rapid “in silico” follow-up of these variants and assessment of their causal relevance as molecular targets for new or repurposed drugs in human genetic studies of SARS-CoV-2 and COVID-19, such as the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/).

The contribution of identified genetic variants outweighed the variance explained by most of the tested host factors for the majority of protein targets. Protein expression in plasma was also frequently associated with expression of protein-encoding genes in relevant tissues. We demonstrate that a large number of genetic variants acting in trans are non-specific and show evidence of substantial horizontal pleiotropy. Findings for these variants should be treated with caution in follow-up studies focused on protein-specific genetic effects.

The successful identification of candidate druggable targets for COVID-19 provides an insight both on potential therapies and on medications that might worsen outlook, depending on the direction of the genetic effect, and whether any associated compound inhibits or activates the target. We also found genetic evidence that selected protein targets, such as for MARK3 and monocyte count, have potential for adverse effects on other health outcomes, but note that this was not a general characteristic of all tested “druggable” targets. Further, in-depth characterization of the targets identified will be required as a first step in gauging the likely success of any new or repurposed drugs identified via this analysis\(^9\).

We exemplify the value of the data resource generated by linking a putative genomic risk variant for poor prognosis among COVID-19 patients, i.e., respiratory failure, at the ABO locus\(^9\) to proteins related to the maladaptive response of the host, namely hypercoagulation, as well as two putative viral interaction partners (heme oxygenase 1 and PLOD2). The risk increasing A allele of rs657152 was consistently associated with higher plasma levels of coagulation factor VIII and von Willebrand factor. Anticoagulation is associated with a better outcome in patients with severe COVID-19\(^35\), and randomized controlled trials are underway to evaluate the benefit or harms of anticoagulant therapies. We note that while there is some evidence\(^38 - 40\) of an increased risk for individuals of blood group A (tagged by rs657152) to experience more severe COVID-19 based on observational studies, results are not entirely consistent\(^41\). Support for the reported genetic findings also warrants further investigation, since early data releases from the COVID-19 Host Genetics Initiative did not replicate the published GWAS results for the ABO locus (https://www.covid19hg.org/, release 3, June 2020). This might be explained by a biased control group, i.e., healthy blood donors, who tend to be blood group O more often compared to the general population. From a drug-discovery perspective, GWAS among COVID-19 patients testing whether or not those require hospitalization would be the most promising application of our results.

Affinity-based proteomics techniques rely on conserved binding epitopes. Changes in the 3D-conformational structure of target proteins introduced by PAVs might change the binding affinity to the target, and hence measurements, without affecting biological activity of the protein. We identified 52 cis-pQTLs which were in LD \((r^2 > 0.1)\) with a PAV. However, 27 of those cis-pQTLs or a proxy in high LD \((r^2 > 0.8)\) have been previously identified as genome-wide significant signals for at least one trait in the GWAS Catalog (excluding any entries of platforms used in the present study) and might therefore carry biologically meaningful information.

This study was designed to provide a rapid open access platform to help prioritize drug discovery and repurposing efforts for the current COVID-19 pandemic. However, important limitations apply. Firstly, protein abundances have been measured in plasma, which may differ from the intracellular role of proteins, and include purposefully secreted as well as leaked proteins. Secondly, while aptamer-based techniques provide the broadest coverage of the plasma proteome, specificity can be compromised for specific protein targets and evidence using complementary techniques such as Olink or mass spectrometry efforts is useful for validation of signals. Thirdly, in-depth phenotypic characterization of the high-priority cis-pQTLs requires appropriate formal and statistical follow-up, such as colocalization, which needs further methodological development to allow for partially shared causal variants (and preferably across multiple traits), and cis-GRS evaluation in independent and adequately powered studies for the trait of interest.

**Methods**

**Study participants.** The Fenland study is a population-based cohort of 12,435 participants of Caucasian-ancestry born between 1950 and 1975 who underwent detailed phenotyping at the baseline visit from 2005 to 2015. Participants were recruited from general practice surgeries in the Cambridgeshire region in the UK.
Exclusion criteria were: clinically diagnosed diabetes mellitus, inability to walk unaided, terminal illness, clinically diagnosed psychotic disorder, pregnancy, or lactation. The study was approved by the Cambridge Local Research Ethics Committee (NRES Committee - East of England, Cambridge Central, ref. 04/Q0108/19) and all participants provided written informed consent. Fenland participants were on average 48.6 years old (standard deviation: 7.5 years) and 53.4% were female.

Mapping of protein targets across platforms. We mapped each candidate protein to its UniProt-ID and used those to select mapping aptamers and Olink measures based on annotation files provided by the vendors.

Proteomic profiling. Proteomic profiling of fasted EDTA plasma samples from 12,084 Fenland Study participants collected at baseline was performed by SomaLogic Inc. (Boulder, CO, USA) using an aptamer-based technology (SomaScan proteomic assay). Relative protein abundances of 4775 human protein targets were evaluated by 4979 aptamers (SomaLogic V4), and a detailed description can be found elsewhere. Briefly, the SomaScan assay utilizes a library of short single-stranded DNA molecules that are chemically modified to specifically bind to protein targets, and the relative amount of aptamers binding to protein targets is determined using DNA microarrays. To account for variation in hybridization within runs, hybridization control probes are used to generate a hybridization scale factor for each sample. To control for total signal differences between samples due

Fig. 7 Screenshots from the webserver. Upper panel: The matrix shows p-values and beta estimates from linear regression models for independently associated single nucleotide polymorphisms (SNPs) at a given locus across protein targets. The matrix can be customized to include only targets of interest or based on statistical criteria. Hovering above filled boxes shows information about the SNP and the associated target. The right-hand side shows information about the target, including associated SNPs in cis and trans, and includes regional associations (RA) plots as well as boxplots for plasma levels of the protein target across each SNP. We chose MAP/microtubule affinity-regulating kinase 3 (MARK3) as an example. Lower panel: Query options for variant-, gene-, or region-based queries of SNP associations across all targets.
to variation in overall protein concentration or technical factors such as reagent concentration, pipetting, or assay timing; a ratio between each aptamer’s measured value and a reference value is computed, and the median of these ratios is computed for each of the three dilution sets (40%, 1%, and 0.005%) and applied to each dilution set. Samples were removed if they were deemed by SomaLogic to have failed or did not meet our acceptance criteria of 0.25–4 for all scaling factors. In addition to passing SomaLogic QC, only human protein targets were taken forward for subsequent analysis (4979 out of the 5284 aptamers). Aptamers target annotation and mapping to UniProt accession numbers as well as Entrez gene identifiers were provided by SomaLogic.

Plasma samples for a subset of 500 Fenland participants were additionally measured using 12 Olink 92-protein panels using proximity extension assays. Of the 1104 Olink proteins, 1069 were unique (n = 35 on >1 panel, average correlation coefficient 0.90). We imputed values below the detection limit of the assay using raw fluorescence values. Protein levels were normalized ("NXP") and subsequently log-transformed for statistical analysis. A total of 15 samples were excluded based on quality thresholds recommended by Olink, leaving 485 samples for analysis.

Genotyping and imputation. Fenland participants were genotyped using three genotyping arrays: the Affymetrix UK Biobank Axiom array (OMICS, N = 8994), Illumina HumanCoreExome BeadChip (OMICS, N = 1960), and an Illumina Omni1-Quad BeadChip (GWAS, N = 1402). Samples were excluded for the following reasons: (1) failed channel contrast (DishQC <0.82), (2) low call rate (<95%), (3) gender mismatch between reported and genetic sex, (4) heterozygosity outlier, (5) unusual high number of singleton genotypes, or (6) impossible identity-by-descent values. All SNPs with failed Affymetrix and Illumina polymorphisms (SNPfilter) were removed if: (1) call rate <95%, (2) clusters failed Affymetrix SNPfilter standard tests and thresholds, (3) MAF was significantly affected by plate, (4) SNP was a duplicate based on chromosome, position and alleles (selecting the best probe set according to Affymetrix SNPfilter), (3) Hardy-Weinberg equilibrium p < 10^-6, (6) did not match the reference condition (7) MAF < 0.1. Autosomes from the OMICS and GWAS subsets were imputed to the HRC (r2 panel using IMPUTE49, and the Core-Exome subset and the X-chromosome (for all subsets) were imputed to HRC.r1 using the Sanger imputation server46. All three arrays subsets were also imputed to the UK10K + 1000G phase3 panel using the Sanger imputation server in order to obtain additional variants that do not exist in the HRC reference panel. Variants with MAF < 0.001, imputation quality (info) <0.4, or Hardy-Weinberg equilibrium p < 10^-7 in any of the genotyping subsets were excluded from further analyses.

GWAS and meta-analysis. After excluding ancestry outliers and related individuals, 10,708 Fenland participants had both phenotypes and genetic data for the GWAS (OMICS = 8350, Core-Exome =1026, and GWAS = 1332). Within each genotyping subset, any low abundance (t) was transformed to follow normal distribution using the rank-based inverse normal transformation. Transformed aptamer abundances were then adjusted for age, sex, sample collection site, and 10 principal components by merging overlapping or adjoining 1 Mb intervals around all genome-wide significant variants (500 kb either side), considering the extended MHC region (chrH:26.55–34.0 Mb) as one region. For each region we defined a regional sentinel variant as the most significant variant in the region. We defined genomic regions sharing at least 4, 2, or 1 regions with genomic regions as overlapping regions in strong LD (r² > 0.8). We classified pQTLs as cis-pQTLs if the majority (>50%) of associated aptamers belong to a single GO-term, (3) the majority (>50%) of associated aptamers but at least two belong to a single GO-term, and (4) no single GO-term covers more than 50% of the associated aptamers. We refer to category 1 as genome-specific association, categories 2 and 3 as vertical pleiotropy, and category 4 as horizontal pleiotropy.

Heritability estimates and genetic correlation. We used genome-wide genotype data from 8350 Fenland participants (Fenland-OMICS) to determine SNP-based heritability and genetic correlation estimates among the 102 protein targets with at least one cis-pQTL and excluding proteins encoded in the X-chromosome. We generated a genetic relationship matrix (GRM) using GCTA v1.1336 from all variants with MAF > 1% to calculate SNP-based heritability as implemented by bMM44. Genetic correlations were computed between all 4273 possible pairs among 93 protein targets with heritability estimates larger than 1.5 times its standard error, using the generated GRM by a bivariate linear mixed model as implemented by bMM. We further conducted two sensitivity analyses to evaluate whether the estimated genetic correlation could be largely attributable to the top cis-pQTL or to shared pleiotropic trans regions. To evaluate contribution of the top cis variant, each protein target was regressed against its sentinel cis variant in addition to age, sex, sample collection site, and 10 principal components, and the residuals were used as phenotypes to compute heritability and genetic correlation estimates. To assess the contribution of 29 pleiotropic trans regions, we excluded 2 Mb genomic regions around pleiotropic trans-pQTLs (associated with >20 aptamers) from the GRM to compute heritability and genetic correlation estimates. Genetic correlations could not be computed for pairs involving IL1RL1 in the main analysis and were therefore excluded. However, upon regressing out the sentinel cis variant, genetic correlation with this protein could be computed probably due to its large contribution to heritability.

Population substructure. We used linear mixed models as implemented in the R package variancePartition to decompose inverse rank-normal transformed plasma abundances of 106 aptamers with at least one cis-pQTL. To this end, we computed weighted genetic scores for each aptamer separating SNPs acting in cis (cis-GRS) and trans (trans-GRS). In addition to the GRS, we included BMI, WHR, systolic and diastolic blood pressure, reported alcohol intake, smoking consumption, fasting plasma levels of glucose, insulin, high-density lipoprotein

Explained variance. To compute the explained variance for plasma abundances of protein targets we fitted linear regression models with residual protein abundances (see GWAS section) as outcome and (1) only the lead cis-pQTL, (2) all cis-pQTLs, or (3) all identified pQTLs as exposure. We report the R² from those models as explained variance.

Annotation of pQTLs. For each identified pQTL we first obtained all SNPs in at least moderate LD (r² > 0.1) using PLINK (version 2.0) and queried comprehensive annotation data using the variant’s closest predictor. SNPs acting during the pick option. For each cis-pQTL we checked whether either the variant itself or a proxy in the encoding gene (r² > 0.1) is predicted to induce a change in the amino acid sequence of the associated protein, so-called protein altering variants (PAVs).

Mapping of cis-pQTLs to drug targets. To annotate druggable targets we merged the list of proteins targeted by the SomaScan V4 platform with the list of druggable genes from Fannin at all14 based on common gene entries. We further added protein–drug combinations as recommended by Groen et al.2.
cholesterol, low-density lipoprotein cholesterol, alanine aminotransferase, as well as a creatinine-based estimated glomerular filtration rate as explanatory factors. We implemented this analysis in the Fenland-OMICS data set leaving 8004 participants without any missing values in the factors considered.

**Genetic risk scores associations.** We computed weighted GRS for metabolic (Insulin resistance^35, type 2 diabetes^36, WHR, and BMI^37), respiratory (FEV1, FVC^24, and asthma^38), and cardiovascular traits (eGFR^39, systolic blood pressure^40, diastolic blood pressure^41, and CAD^42) for Fenland-OMICS participants (N = 8350). We labeled the association with plasma protein abundances. GRSs were computed from previously reported genome-wide significant variants and weighted by their reported beta coefficients for continuous outcomes or log(OR) for binary outcomes. Variants not available among Fenland genotypes, strand ambiguous or with low imputation quality (INFO < 0.6) were excluded from the GRSs. Asso- ciations between each scaled GRS and log_{10}-transformed and scaled protein levels were computed by linear regressions adjusted by age, sex, 10 genetic principal components, and sample collection site. We implemented this analysis for the 186 aptamers with at least one associated cis- or trans-pQTL. Associations with p-values < 0.05/186 were deemed significant according to Bonferroni correction for multiple comparisons.

**Incorporation of GTEx v8 data.** We leveraged gene expression data in five human tissues (lung, whole blood, heart - left ventricle, heart - atrial appendage, and liver), of relevance to COVID-19 and its potential adverse effects and complications, from the Genotype-Tissue Expression (GTEx) project^24,25. For the 102 SOMAmers with apparent in an in-house Olink-based pGWAS (NATURE COMMUNICATIONS| (2020) 11:6397 | https://doi.org/10.1038/s41467-020-19996-z | www.nature.com/naturecommunications), we performed summary-statistics-based PrediXcan^26 analysis to identify tissue-dependent genetically determined gene expression traits that significantly predict plasma protein levels. We used the standardized effect size (z-score) to investigate the tissue specificity or the consistency of the association across the tissues between the genetic component of the expression of the encoding gene and the corresponding protein. We performed DAVID functional enrichment analyses on all the genes significantly associated (Bonferroni-adjusted p < 0.05) with plasma levels of the proteins to identify biological processes (Benjamini-Hochberg adjusted p < 0.05) that may explain the associations found beyond the protein-encoding genes.

**Cross-platform comparison.** We selected 24 cis- and 101 trans-pQTLs mapping to 33 protein targets overlapping with Olink from the SomaScan-based discovery and obtained summary statistics from in-house genome-wide association studies (GWAS) based on corresponding Olink measures. To enable a more systematic reciprocal comparison, we further compared 13 pQTLs (for 11 proteins) only apparent in an in-house Olink-based GWAS (p < 4.5 x 10^{-11}) and obtained GWAS-summary statistics from corresponding aptamer measurements. We pruned the list for variants in high LD (r^2 > 0.8) and discarded SNPs not passing QC for both cohorts (n = 6).

**Phenome-wide scan among UK Biobank and look-up.** We obtained all ICD-10 codes-related genome-wide summary statistics from the most recent release of the Neale lab (http://www.nealelab.is/uk-biobank) with at least 100 cases resulting in 633 distinct ICD-10 codes. Among the 220 cis-pQTLs identified in the present study, 215 were included in the UK Biobank summary statistics (three aptamers had to be excluded due to unavailable lead cis-pQTLs or proxies in LD). We next aligned effect estimates between cis-pQTLs and UK Biobank statistics and used the gsum function from the ‘gtx’ R package to compute the effect of a weighted cis-GRS for an aptamer across all 633 ICD-codes. We applied a global testing correction across all cis-GRS – ICD-10 code combinations using the Benjamini-Hochberg procedure and declared a false discovery rate of 10% as a significance threshold.

We queried all 220 cis-pQTLs for genome-wide association results using the PheWAS function of the R package “pseudow” linked to the IEU GWAS database. We selected all variants in strong LD (r^2 > 0.8) with any of the cis-pQTLs to incorporate information on protein. We restricted the search in the search tool to the batches “eib-a”, “iej-a”, and “ukb-b” to minimize redundant phenotypes.

**Colocalization analysis.** We used statistical colocalization to test for a shared genetic signal between a protein target and a phenotype with evidence of a significant effect of the cis-pQTL (see above, Fig. 6). We obtained posterior probabilities (PP) of: H0, no signal; H1, signal unique to the protein target; H2, signal unique to the trait; H3, two distinct causal variants in the same locus, and; H4, presence of a shared causal variant between a protein target and a given phenotype with PP above 75% were considered highly likely. In case the cis-pQTL was a secondary signal we computed conditional association statistics using the cond function from GCTA-cojo to align with the identification of secondary signals. We conditioned on all other secondary signals in the locus. We note that conditioning on all other secondary variants in the locus failed to produce the desired conditional association statistics in a few cases probably due to moderate LD (r^2 > 0.1) between selected secondary variants and other putative secondary variants.

**Multi-trait colocalization at the ABO locus.** We used hypothesis prioritization in multi-variant colocalization (HyPrColoc) at the ABO locus (±200 kb) to identify protein targets sharing a common causal variant over and above what could be identified in the meta-analysis to increase statistical power, and to identify possible multiple causal variants with distinct associated protein clusters. Briefly, HyPrColoc aims to test the global hypothesis that multiple traits share a common genetic signal at a genomic location and further uses a clustering algorithm to partition possible clusters of traits with distinct causal variants within the same genomic region. HyPrColoc provides for each cluster three different types of output: (1) a posterior probability (PP) that all traits in the cluster share a common genetic signal, (2) a regional association probability, i.e., that all the aptamers share an association with one or more variants in the region, and (3) the proportion of the PP explained by the candidate variant. We considered a highly likely alignment of a genetic signal across various traits if the regional association probability > 80%. This criterion takes, to some extent, into account that aptamers may share multiple causal variants at the same locus and provides some robustness against vio- lation of the single causal variant assumption. We note that several protein targets had no independent signals at the ABO locus (Supplementary Data 4). We further filtered protein targets with no evidence of a likely genetic signal (p > 10^{-5}) in the region before performing HyPrColoc, which improved clustering across traits due to minimizing noise.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability.** All genome-wide summary statistics are made available through an interactive webserver (https://omicscience.org/apps/covidpgwas). Data from the Fenland cohort can be requested by bona fide researchers for specified scientific purposes via the study website (https://www.mrc-epid.cam.ac.uk/research/studies/fenland/information-for-researchers/). Data will either be shared through an institutional data sharing agreement or arrangements will be made for analyses to be conducted remotely without the necessity for data transfer. Publishable summary statistics for look-up and colocalization of pQTLs were obtained from https://gwas.mrcieu.ac.uk/, https://www.ebi. ac.uk/gwas/, and http://www.nealelab.is/uk-biobank.

Received: 19 June 2020; Accepted: 28 October 2020; Published online: 16 December 2020

**References**

1. Banerjee, A. et al. Articles estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. Lancet 6736, 1–11 (2020).
2. 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 (2020).
3. Gordon, D. E. et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583, 459–468 (2020).
4. Merad, M. & Martin, J. C. Pathological inflammation in patients with COVID-19: a key role for monocyes and macrophages. Nat. Rev. Immunol. 20, 355–362 (2020).
5. Zhang, L. et al. D-dimer levels on admission to predict in-hospital mortality in patients with Covid-19. J. Thromb. Haemost. 18, 1324–1329 (2020).
6. Violi, F., Pastori, D., Cangemi, R., Pignatelli, P. & Loffredo, L. Hypercoagulation and antithrombotic treatment in coronavirus 2019: a new challenge. Thromb Haemost 120, 949–956 (2020).
7. Messner, C. B. et al. Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 Infection. Cell Syst. 11, 314–34 (2020).
8. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. Nat. Genet. 47, 856–860 (2015).
9. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug developments on the probability of drug approval. PLOS Genet. 15, e1008489 (2019).
10. Ellinghaus, D. et al. Genomewide association study of severe COVID-19 with respiratory failure. N. Engl. J. Med. 383, 1522–1534 (2020).
11. Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in Covid-19. Preprint at http://medrxiv.org/content/10.1101/2020.09.24.20200048v2.
12. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181, 271–280.e8 (2020).
13. Jose, R. J. & Manuel, A. COVID-19 cytokine storm: the interplay between inflammation and coagulation. The Lancet Respiratory Medicine 8, e46–e47 (2020).
14. Finan, C. et al. The druggable genome and support for target identification and validation in drug development. Sci. Transl. Med. 9, (2017).

15. Sjöberg, A. P. et al. The factor H variant associated with age-related macular degeneration (His-384) and the non-disease-associated form bind differentially to C-reactive protein, fibromodulin, DNA, and necrotic cells. J. Biol. Chem. 282, 10894–10900 (2007).

16. Sun, B. B. et al. Genomic atlas of the human plasma proteome. Nature 558, 73–79 (2018).

17. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell 167, 1415–1429.e19 (2016).

18. Serbanovic-Canic, J. et al. SILENCE of RhoA nucleotide exchange factor, ARHGFE3, reveals its unexpected role in iron uptake. Blood 118, 4967–4976 (2011).

19. HCO3- Coronavirus disease. World Health Organ. 2019, 2633 (2020).

20. Grasselli G. et al. Risk factors associated with mortality among patients with COVID-19 in intensive care units in Lombardy, Italy. JAMA Intern. Med. 180, 1345–1355 (2020).

21. Shrine, N. et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat. Genet. 51, 481–493 (2019).

22. Overmyer, K. A. et al. Large-scale multi-omic analysis of COVID-19 severity. Nat. Commun. 11, 6397 (2020).

23. Shirakabe, K., Hattori, S., Seiki, M., Koyasu, S. & Okada, Y. VIP36 protein is a target of ectodomain shedding and regulates phagocytosis in macrophage raw cells. J. Biol. Chem. 286, 43154–43163 (2011).

24. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multisite tissue gene regulation in humans. Science 348, 648–660 (2015).

25. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat. Genet. 50, 956–967 (2018).

26. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. 47, 1091–1098 (2015).

27. Suher, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat. Commun. 8, 14357 (2017).

28. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57 (2009).

29. Folkersen, L. et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genet. 13, e1006706 (2017).

30. Little, P. Non-steroidal anti-inflammatory drugs and covid-19. BMJ 368, 1–2 (2020).

31. Klarin, D. et al. Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. Nat. Genet. 51, 1574–1579 (2019).

32. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and cardiovascular risk factors. J. Am. Med. Assoc. 320, 2553–2563 (2018).

33. Olafsdottir, T. A. et al. Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis. Nat. Commun. 11, 393 (2020).

34. Wuttke, M. et al. A catalog of genetic loci associated with kidney function and the impact of covariate cut-offs. PLoS Genet. 16, e1009307 (2020).

35. Barbéria, A. N., Bonazzola, R., Gamazon, E. R. & Liang, Y. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Preprint at http://biorxiv.org/10.1101/814350 (2020).

36. Foley, C. N. et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. 41, 1–47 (2019).

Acknowledgements
The Fenland Study (10.22057/2010.10.00001) is funded by the Medical Research Council (MC_UU_120015/1). We are grateful to all the volunteers and to the General Practitioners and practice staff for assistance with recruitment. We thank the Fenland Study Investigators, Fenland Study Co-ordination team and the Epidemiology Field, Data and Laboratory teams. We further acknowledge support for genomics from the Medical Research Council (MC_PC_13046). Proteomic measurements were supported and governed by a collaboration agreement between the University of Cambridge and Somalogic. J.C.Z. and V.P.W.A. are supported by a 4-year Wellcome Trust PhD studentship and the Cambridge Trust; C.L., E.W., and N.J.W. are funded by the Medical Research Council (MC_UU_120015/1). N.J.W. and A.D.H. are an NIH Senior Investigator. G.K. is supported by grants from the National Institute on Aging (NIA): R01 AG057452, RF1 AG058942, RF1 AG059093, U01 AG061359, and U19 AG063744. M.R. acknowledges funding from the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001134), the UK Medical Research Council (FC001134), and the Wellcome Trust (FC001134). U.K.G. is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Numbers R35HG010718 and R01HG011138. J.R. is supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the eMed research and funding concept (grant no. 01ZX1912D). This work was supported by the UCL British Heart Foundation Research Accelerator Award (AA18/6/34223), the National Institute for Health Research University College London Hospitals Biomedical Research Centre, and arises from one of the national “Covid-19 Cardiovascular Disease Flagship Projects” designated by the NIH-HRF Cardiovascular Partnership.

Author contributions
M.P., A.D.H., and C.L. designed the analysis and drafted the manuscript. M.P., E.W., I.C.S., V.P.W.A., and J.L. analyzed the data. N.K. and E.O. performed quality control of proteomic
measurements. J.R. and G.K. designed and implemented the webserver. R.O. and S.W. advised proteome measurements and assisted in quality control. E.G. did the gene expression analysis and interpretation of results. J.P.C. and M.R. provided critical review and intellectual contribution to the discussion of results. N.J.W. is PI of the Fenland cohort. All authors contributed to the interpretation of results and critically reviewed the manuscript.

Competing interests
S.W. and R.O. are employees of SomaLogic. The remaining authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19996-z.

Correspondence and requests for materials should be addressed to A.D.H. or C.L.

Peer review information Nature Communications thanks Eric Fauman, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020, corrected publication 2021