nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19996-z OPEN

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

Maik Pietzner[1], Eleanor Wheeler[1], Julia Carrasco-Zanini[1], Johannes Raffler[2], Nicola D. Kerrison[1], Erin Oerton[1], Victoria P. W. Auyeung[1], Jian'an Luan[1], Chris Finan[3,4], Juan P. Casas[5,6], Rachel Ostroff[7], Steve A. Williams[7], Gabi Kastenmüller[2], Markus Ralser[8,9], Eric R. Gamazon[1,10], Nicholas J. Wareham[1,11], Aroon D. Hingorani[3,4,12]✉ & Claudia Langenberg[1,8,11,13]✉

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/).

[1] MRC Epidemiology Unit, University of Cambridge, Cambridge, UK. [2] Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany. [3] Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London WC1E 6BT, UK. [4] UCL BHF Research Accelerator centre, London, UK. [5] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [6] Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, Massachusetts, USA. [7] SomaLogic, Inc., Boulder, CO, USA. [8] The Molecular Biology of Metabolism Laboratory, The Francis Crick Institute, London, UK. [9] Department of Biochemistry, Charité University Medicine, Berlin, Germany. [10] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. [11] Health Data Research UK, Wellcome Genome Campus and University of Cambridge, Cambridge, UK. [12] Health Data Research UK, Institute of Health Informatics, University College London, London, UK. [13] Computational Medicine, Berlin Institute of Health (BIH), Charité University Medicine, Berlin, Germany. ✉email: a.hingorani@ucl.ac.uk; claudia.langenberg@mrc-epid.cam.ac.uk

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 biological class of drug targets and advances in high-throughput proteomic technologies have enabled systematic analysis of the "human druggable proteome" and genetic target validation to rapidly accelerate the prioritization (or de-prioritization) of therapeutic targets for new drug development or repurposing trials.

Identification and in-depth genetic characterization of proteins utilized by SARS-CoV-2 for entry and replication as well as those proteins involved in the maladaptive host response will help to understand the systemic consequences of COVID-19. For example, if confirmed, the reported protective effect of blood group O on COVID-19-induced respiratory failure[10] might well be mediated by the effect of genetically reduced activity of an ubiquitously expressed glycosyltransferase on a diverse range of proteins.

In this study, we integrate large-scale genomic and aptamer-based plasma proteomic data from a population-based study of 10,708 individuals prior to any SARS-CoV-2 infection or COVID-19 to characterize the genetic architecture of 179 host proteins relevant to COVID-19. We identify genetic variants that regulate host proteins interacting with SARS-CoV-2, or which may contribute to the maladaptive host response. We deeply characterize protein quantitative trait loci (pQTLs) in close proximity to protein-encoding genes (±500 kb window around the gene body), *cis*-pQTLs, and use genetic score analysis and phenome-wide scans to interrogate potential consequences for targeting those proteins by drugs. Our results enable the use of genetic variants as instruments for drug target validation in emerging genome-wide association studies (GWAS) of SARS-CoV-2 infection and COVID-19[10,11].

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©), 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©) for 32 out of the 179 candidate proteins in a subset of 485 Fenland

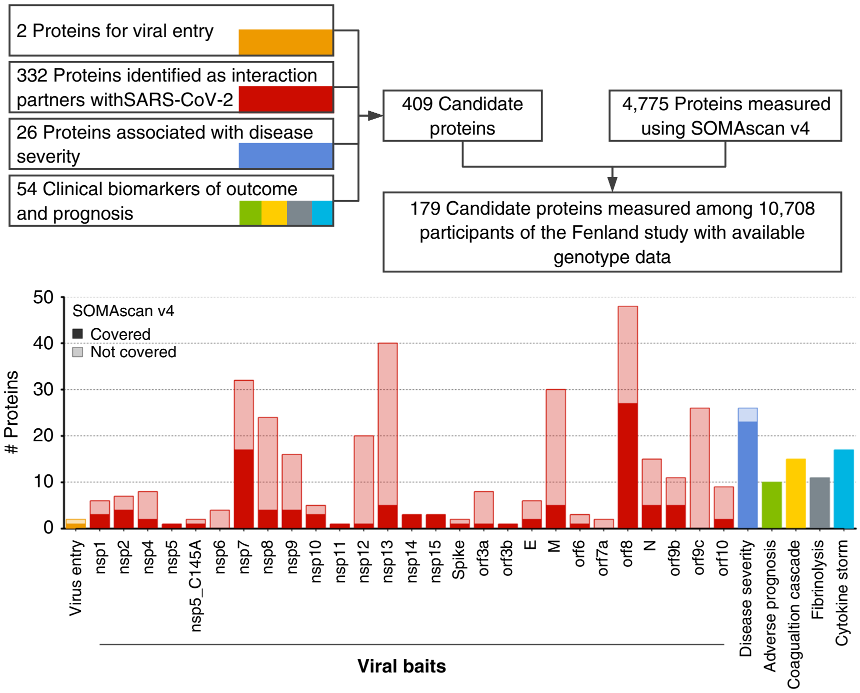


Fig. 1 Flowchart of the identification of candidate proteins and coverage by the SomaScan v4 platform within the Fenland cohort. More details for each protein targeted are given in Supplementary Data 1.

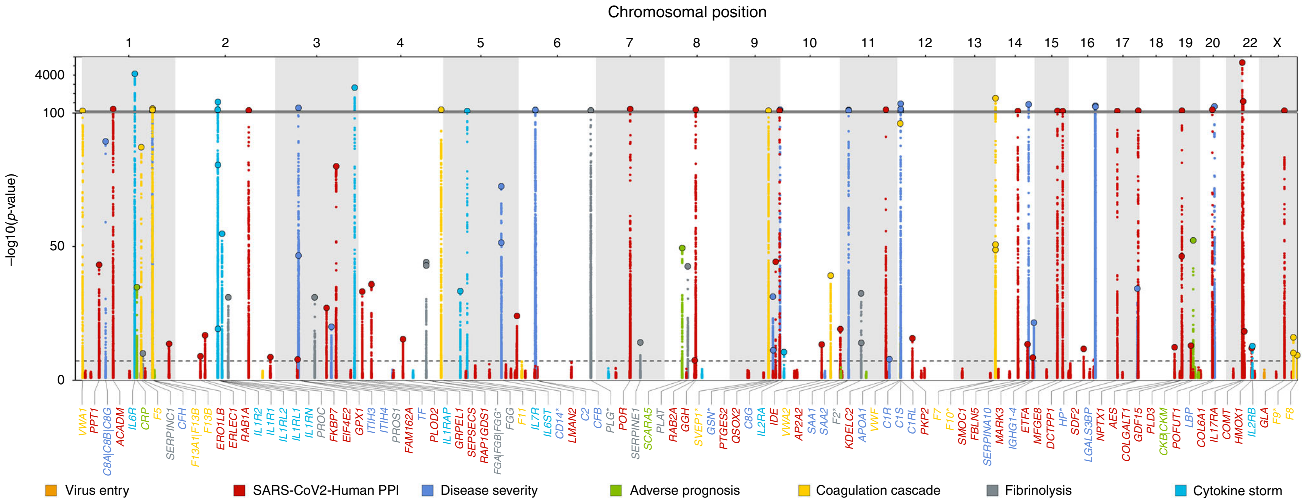


Fig. 2 Manhattan plot of *cis*-associations statistics (encoding gene ±500 kb) for 179 proteins. The most significant regional sentinel protein quantitative trait loci (pQTL) acting in *cis* are annotated by larger dots for 104 unique protein targets (dashed line; $p < 5 \times 10^{-8}$). Starred genes indicate those targeted by multiple aptamers ($n = 9$ genes).

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[14], 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-oligomer 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 in high linkage disequilibrium (LD) $r^2 > 0.8$) between aptamer pairs or triplets for seven proteins. Minor allele frequencies ranged from 0.01% to 49.9%, and the variance explained ranged from 0.3% to 70.1% for all *cis*-acting sentinel variants and 0.3% to 70.9% for *cis*-acting variants including 2–9 identified secondary signals at 57 targets, similar to what was observed considering all *cis*- and an additional 369 *trans*-acting variants identified for 98 aptamers (0.4–70.9%). Among the 97 proteins, 38 are targets of existing drugs, including 15 proteins (PLOD2, COMT, DCTPP1, GLA, ERO1LB, SDF2, MARK3, ERLEC1, FKBP7, PTGES2, EIF4E2, MFGE8, IL17RA, COL6A1, and PLAT) (eight with no known pQTL) that were previously identified[3] as interacting with structural or non-structural proteins encoded in the SARS-CoV-2 genome and 16 proteins (CD14, F2, F5, F8, F9, F10, FGB, IL1R1, IL2RA, IL2RB, IL6R, IL6ST, PLG, SERPINC1, SERPINE1, and VWF) (seven with no known pQTL) that encode biomarkers related to COVID-19 severity[7], prognosis, or outcome.

Drug targets are predicted to act specifically within a clearly defined pathway, i.e., being at most part of a known protein cascade. To classify such "vertical" pleiotropy and distinguish from "horizontal" pleiotropy, i.e., pQTLs associated with proteins across distinct pathways, we investigated associations of identified lead *cis*-pQTLs with all measured aptamers while mapping those to specific pathways using GO-terms ($N = 4776$ unique protein targets, see Methods). For 38 *cis*-pQTLs mapping to druggable 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, KDEL2, GDP/GTP exchange protein, PPT1, GT251, or PKP2) and five proteins related to cytokine storm (IL-1Rrp2 and IL-1Ra), fibrinolysis (PAI-1), coagulation (coagulation factor X(a)), 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). 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, FKBP7, 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 fibulin 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, $N = 4979$,

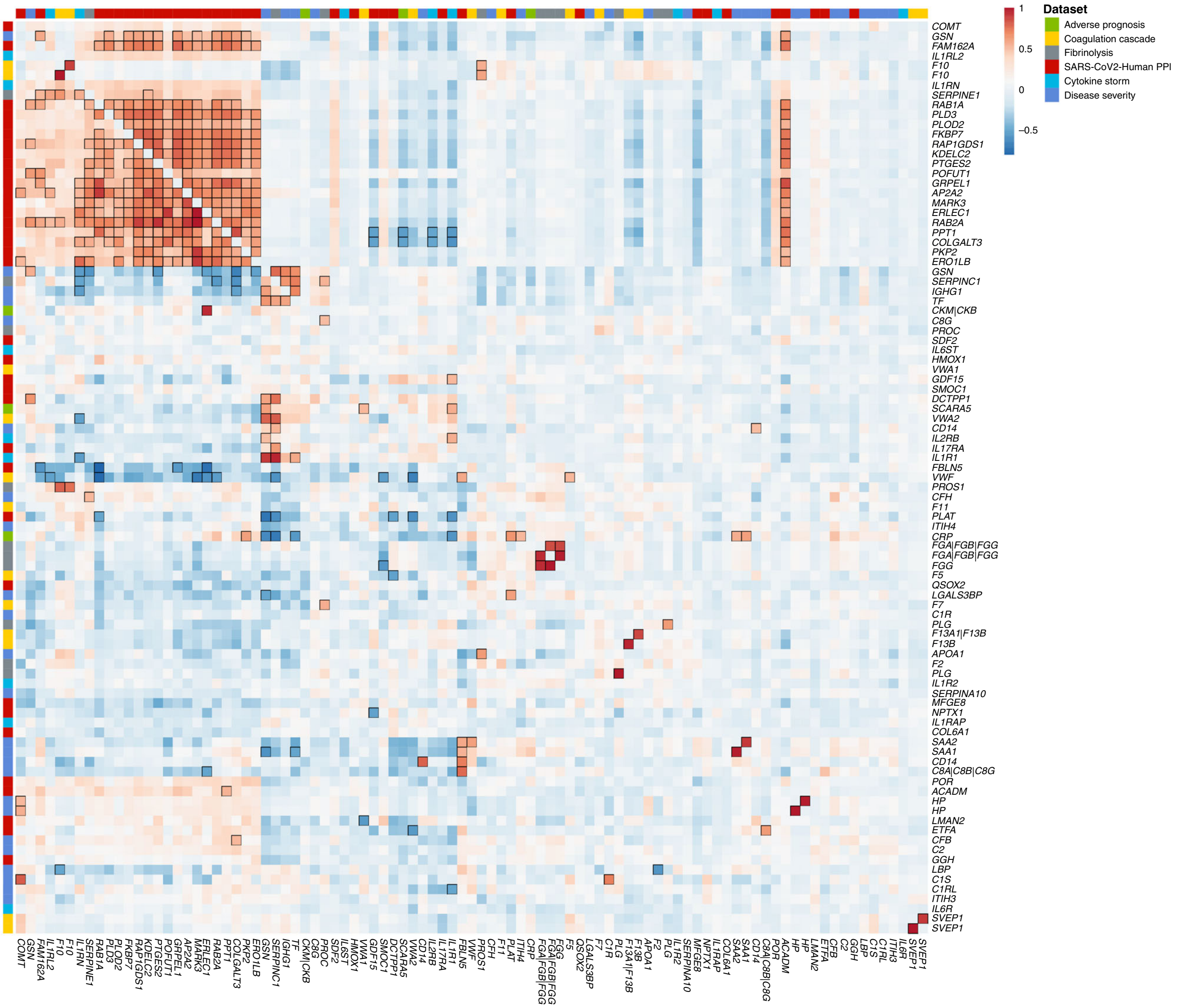


Fig. 3 Genetic (lower triangle) and observational (upper triangle) correlation matrix of 86 unique proteins targeted by 93 aptamers with reliable heritability estimates (see Methods). Aptamers were clustered based on absolute genetic correlations to take activation as well repression into account and protein-encoding genes were used as labels. The column on the far left indicates relevance to SARS-CoV-2 infection. Strong correlations (|*r*| > 0.5) are indicated by black frames.

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 pleiotropic *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 DNA-binding affinity of complement factor H[15], which may introduce unspecific binding of complement factor H to a variety of aptamers, being small DNA-fragments, and may therefore interfere with the method of measurement more generally, rather than presenting a biological effect on these proteins. A similar example is the *trans*-pQTL rs71674639 (MAF = 0.21) associated with 789 aptamers and in high LD (r^2=0.99) with a missense variant in *BCHE* (rs1803274).

Sample handling is an important contributor to the identification of non-specific *trans*-pQTL associations. Blood cells secrete a wide variety of biomolecules, including proteins, following activation or release such as consequence of stress-induced apoptosis or lysis. Interindividual genetic differences in blood cell composition can hence result in genetic differences in protein profiles depending on sample handling or delays in time-to-spin. A prominent example seen in our results and reported in a previous study[16] is variant rs1354034 in *ARHGEF3*, associated with over 1000 aptamers (on the full SomaScan platform). *ARHGEF3* is a known locus associated with platelet counts[17] and a master regulator of megakaryopoiesis[18], either genetically determined higher 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*[16] are not significant in our study.

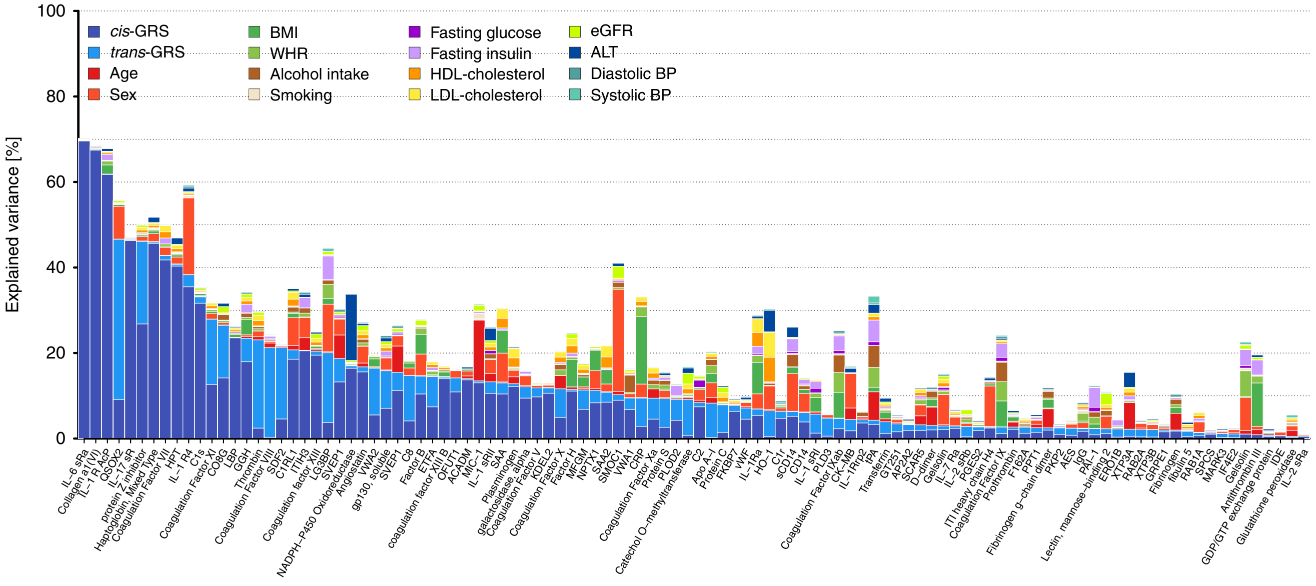


Fig. 4 Stacked bar chart showing the results from variance decomposition of plasma abundances of 106 aptamers targeting candidate proteins. For each candidate protein a model was fitted to decompose the variance in plasma levels including all 16 factors noted in the legend. These analyses were done based on 8006 participants with complete data. *cis*/*trans*-GRS = weighted genetic risk score based on all single nucleotide polymorphisms associated with the aptamer of interest acting in *cis* and *trans*, respectively. BMI (body mass index), WHR (waist-to-hip ratio), HDL (high-density lipoprotein), LDL (low-density lipoprotein), eGFR (estimated glomerular filtration rate), ALT (alanine amino transaminase), and BP (blood pressure).

We identified a few variants with evidence for vertical pleiotropic effects, including rs2289252 a *cis*-pQTL for coagulation factor XI, which was specifically associated ($p < 5 \times 10^{-8}$) 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 from different non-genetic factors for proteins such as LG3BP, SAA, IL-1Ra, or HO-1 implicating multiple, in part modifiable, factors with independent contributions to plasma levels of those proteins.

Patients with multiple chronic conditions are at higher risk of getting severe COVID-19 disease[2,19,20] 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 ($r^2 < 0.4$) and genetic colocalization of QSOX2 plasma levels and lung function[21] 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 IL1RL1. 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-19[7,22]. 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[23]. The higher plasma levels of VIP36, suggesting an enhanced immune response, among individuals with genetically higher WHR and lower kidney function appear contradictive as a higher WHR, indicating abdominal adiposity, and lower kidney function are considered as risk factors for COVID-19.

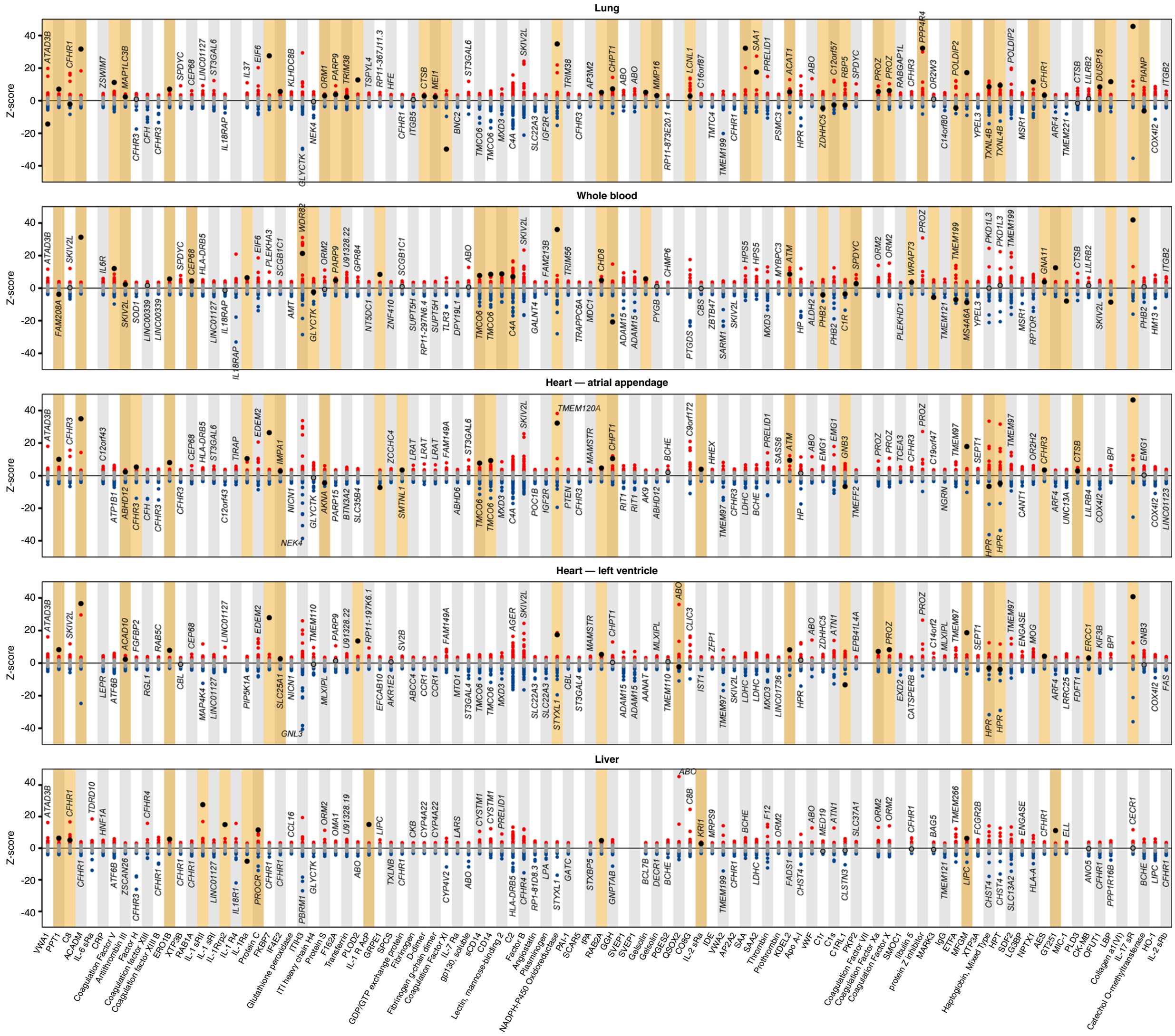


Fig. 5 Results of predicted gene expression in each of five tissues and plasma abundances of 102 aptamers with at least one *cis*-pQTL on one of the autosomes using PrediXcan. Each panel displays results for a tissue. Each column contains results across successful gene expression models for the association with the aptamer listed on the *x*-axis. Red indicates nominally significant ($p < 0.05$) positive *z*-scores (*y*-axis) and blue, nominally significant inverse *z*-scores for associated aptamers. Protein-encoding genes are highlighted by larger black circles. Orange background indicates all examples of significant associations between the protein-encoding gene and protein abundance in plasma regardless if this was the most significant one. Top genes were annotated if those differed from the protein-encoding gene.

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 druggable targets), *POR, RAB2A, KDELC2, C1RL, 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 *SERPINA10*.

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) = 2.5×10^{-5}), “glycoproteins” (FDR = 1.7×10^{-4}), or “disulfide bonds” (FDR = 2.8×10^{-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[29] (rs1227731, beta=0.59, $p < 6.5 \times 10^{-16}$). However, rs1227731 was a secondary signal for the SomaScan assay (beta=0.29, $p < 5.8 \times 10^{-66}$) 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[14]. 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 outlook 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 (aleplasinin), 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 migalastat (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 *F2* (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: I26) as well as phlebitis and thrombophlebitis (ICD-10 code: I80) (Supplementary Data 6).

To maximize power for disease outcomes, include clinically relevant risk factors, and allow for variant-specific effects, we complemented the phenome-wide strategy with a comprehensive look-up for genome-wide significant associations in the MR-Base platform[34].

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

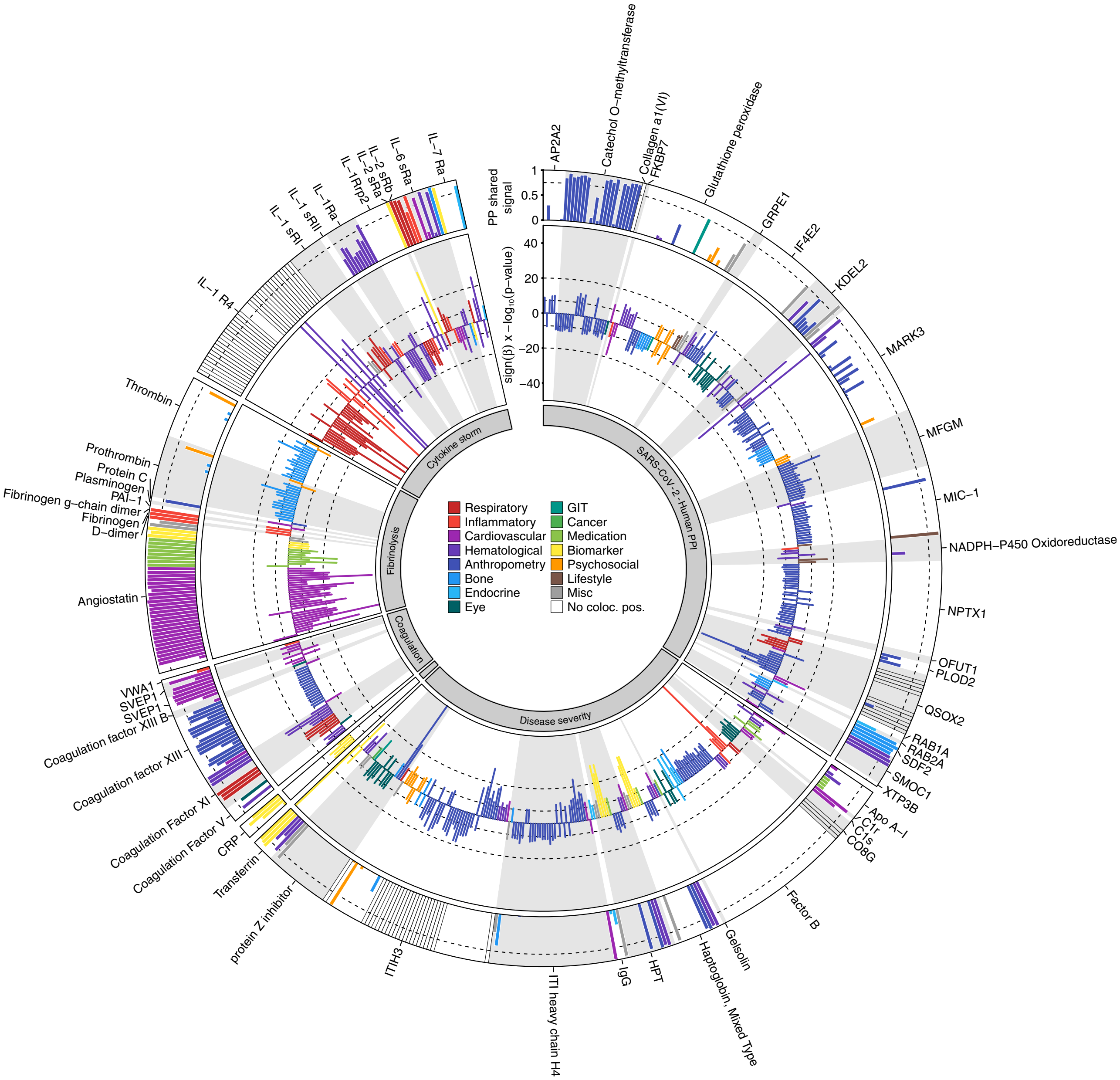


Fig. 6 Circos plot summarizing genome-wide significant associations between 74 *cis*-pQTLs and 239 traits[34] in the inner ring and results from statistical colocalization in the outer ring. The dashed line in the outer ring indicates a posterior probability of 75% of shared genetic signal between the 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.

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 colocalizing with *cis*-pQTLs, i.e., traits were closely related and effect estimates were consistent with phenotypic presentations (Supplementary Data 7 and Fig. 6). For instance, rs165656, a lead *cis*-pQTL increasing catechol o-methyltransferase plasma abundances, is a regional lead variant for BMI[35] and specifically colocalised with adiposity related traits, i.e., inversely associated with overall measures of body size such as BMI, weight, and fat-free mass. In general, phenotypic characterization of potential genetic instruments to simulate targeting abundances or activities of proteins can help to distinguish those with narrow and well-defined or target-specific from those with undesirable or broad phenotypic effects. Notable exceptions included the IL-6 receptor variant rs2228145, for which the protein increasing C allele was 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[10]. 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, van 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, 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[36].

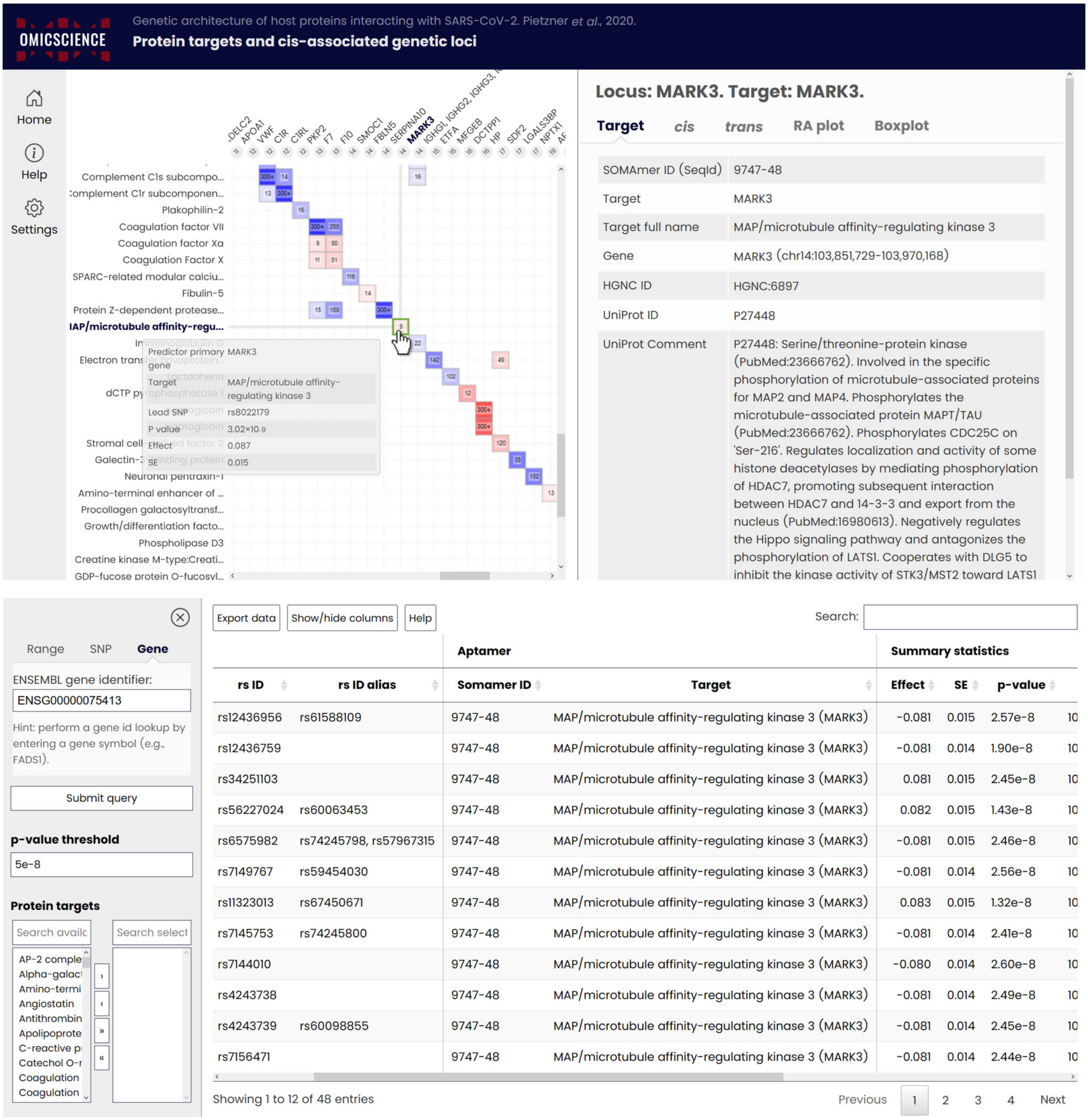
We exemplify the value of the data resource generated by linking a putative genomic risk variant for poor prognosis among COVID-19 patients, i.e., respiratory failure, at the *ABO* locus[10] to proteins related to the maladaptive response of the host, namely hypercoagulation, as well as two putative 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[37], 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.



Genetic architecture of host proteins interacting with SARS-CoV-2. Pietzner *et al.*, 2020.

Protein targets and cis-associated genetic loci

Predictor primary gene	MARK3
Target	MAP/microtubule affinity-regulating kinase 3
Lead SNP	rs8022179
P value	3.02×10-9
Effect	0.087
SE	0.015

Locus: MARK3. Target: MARK3.

SOMAmer ID (SeqId)	9747-48
Target	MARK3
Target full name	MAP/microtubule affinity-regulating kinase 3
Gene	MARK3 (chr14:103,851,729-103,970,168)
HGNC ID	HGNC:6897
UniProt ID	P27448
UniProt Comment	P27448: Serine/threonine-protein kinase (PubMed:23666762). Involved in the specific phosphorylation of microtubule-associated proteins for MAP2 and MAP4. Phosphorylates the microtubule-associated protein MAPT/TAU (PubMed:23666762). Phosphorylates CDC25C on 'Ser-216'. Regulates localization and activity of some histone deacetylases by mediating phosphorylation of HDAC7, promoting subsequent interaction between HDAC7 and 14-3-3 and export from the nucleus (PubMed:16980613). Negatively regulates the Hippo signaling pathway and antagonizes the phosphorylation of LATS1. Cooperates with DLG5 to inhibit the kinase activity of STK3/MST2 toward LATS1

ENSEMBL gene identifier: ENSG00000075413

Hint: perform a gene id lookup by entering a gene symbol (e.g., FADS1).

p-value threshold 5e-8

Protein targets

Search availa Search select

AP-2 comple
Alpha-galact
Amino-termi
Angiostatin
Antithrombin
Apolipoprote
C-reactive pı
Catechol O-r
Coagulation
Coagulation

		Aptamer		Summary statistics			
rs ID	rs ID alias	Somamer ID	Target	Effect	SE	p-value	
rs12436956	rs61588109	9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.081	0.015	2.57e-8	10
rs12436759		9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.081	0.014	1.90e-8	10
rs34251103		9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	0.081	0.015	2.45e-8	10
rs56227024	rs60063453	9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	0.082	0.015	1.43e-8	10
rs6575982	rs74245798, rs57967315	9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.081	0.015	2.46e-8	10
rs7149767	rs59454030	9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.081	0.014	2.56e-8	10
rs11323013	rs67450671	9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	0.083	0.015	1.32e-8	10
rs7145753	rs74245800	9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.081	0.014	2.41e-8	10
rs7144010		9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.080	0.014	2.60e-8	10
rs4243738		9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.081	0.014	2.49e-8	10
rs4243739	rs60098855	9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.081	0.014	2.45e-8	10
rs7156471		9747-48	MAP/microtubule affinity-regulating kinase 3 (MARK3)	-0.081	0.014	2.44e-8	10

Showing 1 to 12 of 48 entries

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.

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[42] 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[43]. 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 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[44]. 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 ("NPX") and subsequently $\log_2$-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 Infinium Core Exome 24v1 (Core-Exome, $N = 1060$), and Affymetrix SNP5.0 (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) unusually high number of singleton genotypes, or (6) impossible identity-by-descent values. Single nucleotide polymorphisms (SNPs) were removed if: (1) call rate <95%, (2) clusters failed Affymetrix SNPolisher 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 probeset according to Affymetrix SNPolisher), (5) Hardy-Weinberg equilibrium $p < 10^{-6}$, (6) did not match the reference, or (7) MAF = 0.

Autosomes for the OMICS and GWAS subsets were imputed to the HRC (r1) panel using IMPUTE4[45], and the Core-Exome subset and the X-chromosome (for all subsets) were imputed to HRC.r1.1 using the Sanger imputation server[46]. All three arrays subsets were also imputed to the UK10K + 1000Gphase3[47] 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, aptamer abundances were transformed to follow a 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 in STATA v14, and the residuals used as input for the genetic association analyses. Test site was omitted for protein abundances measured by Olink as those were all selected from the same test site. Genome-wide association was performed under an additive model using BGENIE (v1.3)[45]. Results for the three genotyping arrays were combined in a fixed-effects meta-analysis in METAL[48]. Following the meta-analysis, 17,652,797 genetic variants, also present in the largest subset of the Fenland data (Fenland-OMICS), were taken forward for further analysis.

Definition of genomic regions (including *cis/trans*). For each aptamer, we used a genome-wide significance threshold of 5×10^{-8} and defined non-overlapping regions by merging overlapping or adjoining 1 Mb intervals around all genome-wide significant variants (500 kb either side), considering the extended MHC region (chr6:25.5–34.0 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 shared across aptamers if regional sentinels of overlapping regions were in strong LD ($r^2 > 0.8$). We classified pQTLs as *cis*-acting instruments if the variant was less than 500 kb away from the gene body of the protein-encoding gene.

Conditional analysis. We performed conditional analysis as implemented in the GCTA software using the *slct* option for each genomic region–aptamer pair identified. We used a collinear cut-off of 0.1 and a *p*-value below 5×10^{-8} to identify secondary signals in a given region. As a quality control step, we fitted a final model including all identified variants for a given genomic region using individual level data in the largest available data set ("Fenland-OMICs") and discarded all variants no longer meeting genome-wide significance.

We performed a forward stepwise selection procedure to identify secondary signals at each locus on the X-chromosome using SNPTEST v.2.5.2 to compute conditional GWAS based on individual level data in the largest subset. Briefly, we defined conditionally independent signals as those emerging after conditioning on all previously selected signals in the locus until no signal was genome-wide significant.

Explained variance. To compute the explained variance for plasma abundancies of protein targets we fitted linear regression models with residual protein abundancies (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 R^2 from those models as explained variance.

Annotation of pQTLs. For each identified pQTL we first obtained all SNPs in at least moderate LD ($r^2 > 0.1$) using PLINK (version 2.0) and queried comprehensive annotations using the variant effect predictor software[49] (version 98.3) using the *pick* option. For each *cis*-pQTL we checked whether either the variant itself or a proxy in the encoding gene ($r^2 > 0.1$) 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 Finan at al.[14] based on common gene entries. We further added protein–drug combinations as recommended by Gordon et al.[3].

Identification of relevant GWAS traits. To enable linkage to reported GWAS-variants we downloaded all SNPs reported in the GWAS Catalog[50] (19/12/2019) and pruned the list of variant–outcome associations manually to omit previous protein-wide GWAS. For each SNP identified in the present study ($N = 671$) we tested whether the variant or a proxy in LD ($r^2 > 0.8$) has been reported to be associated with other outcomes previously.

Definition of novel pQTLs. To test whether any of the identified regional sentinel pQTLs has been reported previously, we obtained a list of published pQTLs[16,27,29,51,52] and defined novel pQTLs as those not in LD ($r^2 < 0.1$) with any previously identified variant. We note that this approach is rather conservative, since it only asks whether or not any of the reported SNPs has ever been reported to be associated with any protein measured with multiplex methods.

Assessment of pleiotropy. To evaluate possible protein-specific pleiotropy of pQTLs we computed association statistics for each of the 671 unique SNPs across 4979 aptamers ($N = 4775$ unique protein targets) with the same adjustment set as in the GWAS. This resulted in a protein profile for each variant defined as all aptamers significantly associated ($p < 5 \times 10^{-8}$). For all aptamers we retrieved all GO-terms referring to biological processes from the UniProt database using all possible UniProt-IDs as a query. GO-term annotation within the UniProt database has the advantage of being manually curated while aiming to omit unspecific parent terms. We tested for each pQTL if the associated aptamers fall into one of the following criteria: (1) solely associated with a specific protein, (2) all 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 protein-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 v.1.90[53] from all variants with MAF > 1% to calculate SNP-based heritability as implemented by biMM[54]. 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 biMM. 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 correlations with this protein could be computed probably due to its large contribution to heritability.

Variance decomposition. 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 used participants' age, sex, BMI, WHR, systolic and diastolic blood pressure, reported alcohol intake, smoking consumption, fasting plasma levels of glucose, insulin, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, alanine aminotransaminase, 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[55], type 2 diabetes[56], WHR, and BMI[57]), respiratory (FEV1, FVC[21], and asthma[58]), and cardiovascular traits (eGFR[59], systolic blood pressure[60], diastolic blood pressure[60], and CAD[33]) for Fenland-OMICs participants ($N =$ 8,350) to evaluate their 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. Associations 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 at least one *cis*-pQTL located on the autosomes and available gene expression models trained in GTEx v8[61], 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 pGWAS ($p < 4.5 \times 10^{-11}$) effort 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 efforts ($n = 6$).

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 grs.summary 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 "ieugwasr" 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 proxies. We restricted the search in the ieugwar tool to the batches "ebi-a", "ieu-a", and "ukb-b" to minimize redundant phenotypes.

Colocalization analysis. We used statistical colocalization[62] 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 trait. *PPs* above 75% were considered highly likely. In case the *cis*-pQTL was a secondary signal we computed conditional association statistics using the *cond* option 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-trait colocalization (HyPrColoc)[63] at the *ABO* locus (±200 kb) (1) 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 (2) 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 apatamers may share multiple causal variants at the same locus and provides some robustness against violation of the single causal variant assumption. We note that several protein targets had multiple 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. Publicly available 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 monocytes 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**, 11–24.e4 (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 mechanisms 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 (2020).
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. Silencing of RhoA nucleotide exchange factor, ARHGEF3, reveals its unexpected role in iron uptake. *Blood* **118**, 4967–4976 (2011).
19. WHO. Coronavirus disease. *World Health Organ.* **2019**, 2633 (2020).
20. Mehra, M. R., Desai, S. S., Kuy, S., Henry, T. D. & Patel, A. N. Cardiovascular disease, drug therapy, and mortality in Covid-19. *N. Engl. J. Med.* 382, e102 (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. *Cell Syst*. 1–18 https://doi.org/10.1016/j.cels.2020.10.003 (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 264.7 cells. *J. Biol. Chem.* **286**, 43154–43163 (2011).
24. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue 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. Suhre, 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 drug discovery. *Nature* **506**, 376–381 (2014).
33. Nikpay, M. et al. A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
34. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome.*Elife* **7**, e34408 (2018).
35. Kichaev, G. et al. Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
36. Zheng, J. et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet*. **52**, 1122–1131 (2020).
37. Tang, N. et al. Anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy. *J. Thromb. Haemost.* **18**, 1094–1099 (2020).
38. Zhao, J. et al. Relationship between the ABO blood group. and the COVID-19 susceptibility. Preprint at https://doi.org/10.1101/2020.03.11.20031096 (2020).
39. Li, J. et al. Association between ABO blood groups and risk of SARS-CoV-2 pneumonia. *Br. J. Haematol.* **190**, 24–27 (2020).
40. Pourali, F. et al. Relationship between blood group and risk of infection and death in COVID-19: a live meta-analysis. *New Microbes New Infect*. **37**, 100743 (2020).
41. Boudin, L., Janvier, F., Bylicki, O. & Dutasta, F. ABO blood groups are not associated with risk of acquiring the SARS-CoV-2 infection in young adults. Haematologica 105, haematol.2020.265066 (2020).
42. Bateman, A. et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
43. Williams, S. A. et al. Plasma protein patterns as comprehensive indicators of health. *Nat. Med.* **25**, 1851–1857 (2019).
44. Assarsson, E. et al. Homogenous 96-Plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS ONE* **9**, (2014).
45. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
46. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
47. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 1–9 (2015).
48. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
49. Mclaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 1–14 (2016).
50. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, 1005–1012 (2018).
51. Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).
52. Enroth, S. B. S., Johansson, Å., Enroth, S. B. S. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat. Commun.* **5**, 4684 (2014).
53. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
54. Pirinen, M. et al. BiMM: efficient estimation of genetic variances and covariances for cohorts with high-dimensional phenotype measurements. *Bioinformatics* **33**, 2405–2407 (2017).
55. Scott, R. A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
56. Trompet, S. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
57. Lotta, L. A. et al. Association of genetic variants related to gluteofemoral vs abdominal fat distribution with type 2 diabetes, coronary disease, and cardiovascular risk factors. *J. Am. Med. Assoc.* **320**, 2553–2563 (2018).
58. 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).
59. Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
60. Ehret, G. B. et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat. Genet.* **48**, 1171–1184 (2016).
61. Barbeira, 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).
62. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
63. Foley, C. N. et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. 44, 1–47 (2019).

Acknowledgements

The Fenland Study (10.22025/2017.10.101.00001) is funded by the Medical Research Council (MC_UU_12015/1). 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_12015/1). N.J.W. and A.D.H. are an NIHR 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). E.R.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 e:Med research and funding concept (grant no. 01ZX1912D). This work was supported by the UCL British Heart Foundation Research Accelerator Award (AA/18/6/34223), 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 NIHR-BHF Cardiovascular Partnership.

Author contributions

M.P., A.D.H., and C.L. designed the analysis and drafted the manuscript. M.P., E.W., J.C.S.Z., 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.