

1 Genetic architecture of host proteins interacting with SARS-CoV-2

2
3 Maik Pietzner¹, Eleanor Wheeler¹, Julia Carrasco-Zanini¹, Johannes Raffler², Nicola D. Kerrison¹,
4 Erin Oerton¹, Victoria P.W. Auyeung¹, Jian'an Luan¹, Chris Finan^{3,4}, Juan P. Casas^{5,6}, Rachel
5 Ostroff⁷, Steve A. Williams⁷, Gabi Kastenmüller², Markus Ralser^{8,9}, Eric R. Gamazon^{1,10}, Nicholas
6 J. Wareham^{1,11}, Aroon D. Hingorani^{3,4,12*}, Claudia Langenberg^{1,8,11*}

7 8 Affiliations

9 ¹MRC Epidemiology Unit, University of Cambridge, Cambridge, UK

10 ²Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for
11 Environmental Health, Neuherberg, Germany

12 ³Institute of Cardiovascular Science, Faculty of Population Health, University College London, London
13 WC1E 6BT, UK

14 ⁴UCL BHF Research Accelerator centre

15 ⁵Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

16 ⁶Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston
17 Healthcare System, Boston, Massachusetts, USA

18 ⁷SomaLogic, Inc., Boulder, CO, USA

19 ⁸The Molecular Biology of Metabolism Laboratory, The Francis Crick Institute, London, UK

20 ⁹Department of Biochemistry, Charité University Medicine, Berlin, Germany

21 ¹⁰Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

22 ¹¹Health Data Research UK, Wellcome Genome Campus and University of Cambridge, UK

23 ¹²Health Data Research UK, Institute of Health Informatics, University College London, UK

24
25 *Correspondence to Dr Claudia Langenberg (claudia.langenberg@mrc-epid.cam.ac.uk) and Prof Aroon
26 Hingorani (a.hingorani@ucl.ac.uk)
27

28 **ABSTRACT**

29 Strategies to develop therapeutics for SARS-CoV-2 infection may be informed by experimental
30 identification of viral-host protein interactions in cellular assays and measurement of host
31 response proteins in COVID-19 patients. Identification of genetic variants that influence the
32 level or activity of these proteins in the host could enable rapid 'in silico' assessment in human
33 genetic studies of their causal relevance as molecular targets for new or repurposed drugs to
34 treat COVID-19. We integrated large-scale genomic and aptamer-based plasma proteomic data
35 from 10,708 individuals to characterize the genetic architecture of 179 host proteins reported
36 to interact with SARS-CoV-2 proteins or to participate in the host response to COVID-19. We
37 identified 220 host DNA sequence variants acting in *cis* (MAF 0.01-49.9%) and explaining 0.3-
38 70.9% of the variance of 97 of these proteins, including 45 with no previously known protein
39 quantitative trait loci (pQTL) and 38 encoding current drug targets. Systematic characterization
40 of pQTLs across the phenome identified protein-drug-disease links, evidence that putative viral
41 interaction partners such as MARK3 affect immune response, and establish the first link
42 between a recently reported variant for respiratory failure of COVID-19 patients at the *ABO*
43 locus and hypercoagulation, i.e. maladaptive host response. Our results accelerate the
44 evaluation and prioritization of new drug development programmes and repurposing of trials to
45 prevent, treat or reduce adverse outcomes. Rapid sharing and dynamic and detailed
46 interrogation of results is facilitated through an interactive webserver
47 (<https://omicscience.org/apps/covidpgwas/>).

48 INTRODUCTION

49 The pandemic of the novel coronavirus SARS-CoV-2 infection, the cause of COVID-19, is causing
50 severe global disruption and excess mortality^{1,2}. Whilst ultimately strategies are required that
51 create vaccine-derived herd immunity, in the medium term there is a need to develop new
52 therapies or to repurpose existing drugs that are effective in treating patients with severe
53 complications of COVID-19, and also to identify agents that might protect vulnerable individuals
54 from becoming infected. The experimental characterization of 332 SARS-CoV-2-human protein-
55 protein interactions and their mapping to 69 existing FDA-approved drugs, drugs in clinical trials
56 and/or preclinical compounds³ points to new therapeutic strategies, some of which are
57 currently being tested. The measurement of circulating host proteins that associate with
58 COVID-19 severity or mortality also provides insight into potentially targetable maladaptive
59 host responses with current interest being focused on the innate immune response⁴,
60 coagulation^{5,6}, and novel candidate proteins⁷.

61 Naturally-occurring sequence variation in or near a human gene encoding a drug target and
62 affecting its expression or activity can be used to provide direct support for drug mechanisms
63 and safety in humans. This approach is now used by major pharmaceutical companies for drug
64 target identification and validation for a wide range of non-communicable diseases, and to
65 guide drug repurposing^{8,9}. Genetic evidence linking molecular targets to diseases relies on our
66 understanding of the genetic architecture of drug targets. Proteins are the most common
67 biological class of drug targets and advances in high-throughput proteomic technologies have
68 enabled systematic analysis of the “human druggable proteome” and genetic target validation
69 to rapidly accelerate the prioritization (or de-prioritisation) of therapeutic targets for new drug
70 development or repurposing trials.

71 Identification and in-depth genetic characterization of proteins utilized by SARS-CoV-2 for entry
72 and replication as well as those proteins involved in the maladaptive host response will help to
73 understand the systemic consequences of COVID-19. For example, if confirmed, the reported
74 protective effect of blood group O on COVID-19-induced respiratory failure¹⁰ might well be
75 mediated by the effect of genetically reduced activity of an ubiquitously expressed
76 glycosyltransferase on a diverse range of proteins.

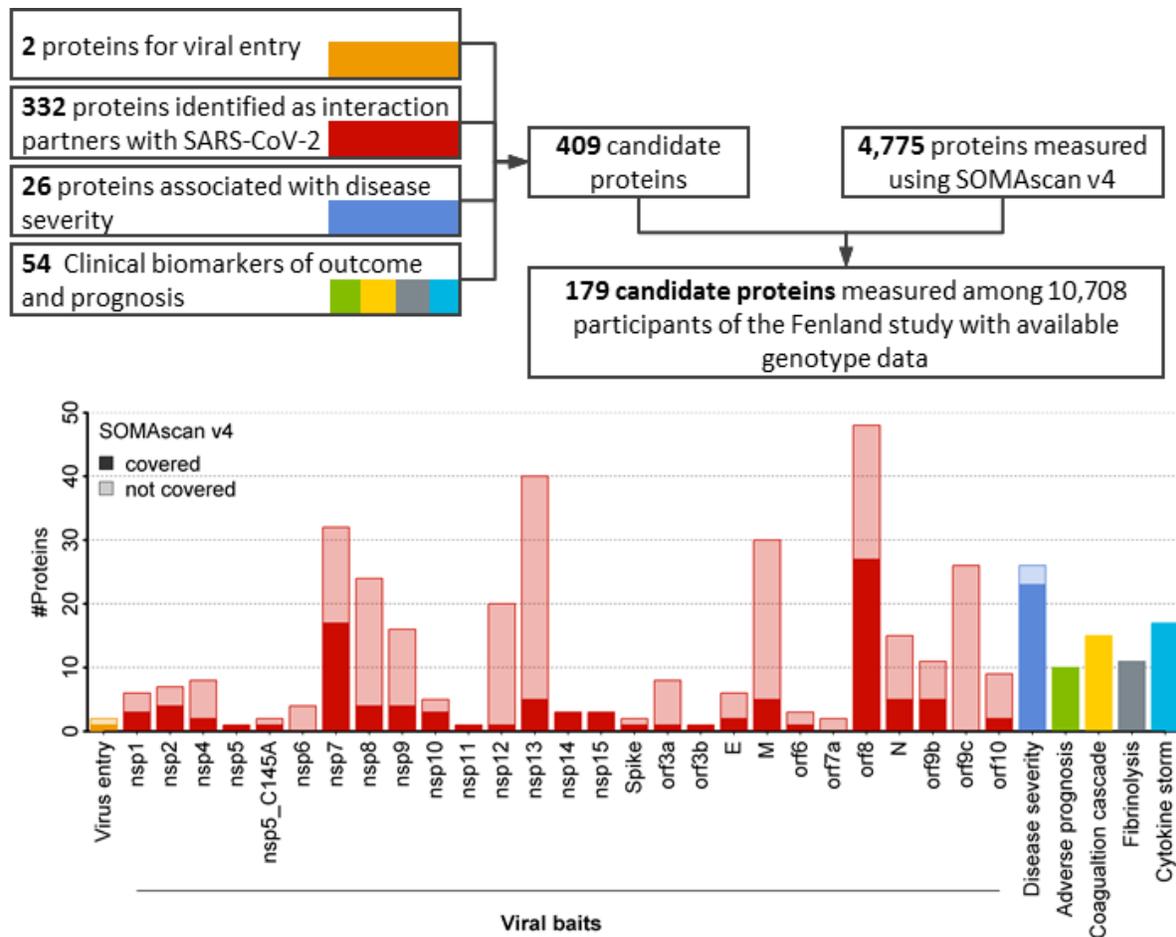
77 In this study we integrated large-scale genomic and aptamer-based plasma proteomic data
78 from a population-based study of 10,708 individuals to characterize the genetic architecture of
79 179 host proteins relevant to COVID-19. We identified genetic variants that regulate host
80 proteins that interact with SARS-CoV-2, or which may contribute to the maladaptive host
81 response. We deeply characterized protein quantitative trait loci (pQTLs) in close proximity to
82 protein encoding genes, *cis*-pQTLs, and used genetic score analysis and phenome-wide scans to
83 interrogate potential consequences for targeting those proteins by drugs. Our results enable
84 the use of genetic variants as instruments for drug target validation in emerging genome-wide
85 associations studies (GWAS) of SARS-CoV-2 infection and COVID-19.

86 **RESULTS**

87 *Coverage of COVID-19-relevant proteins*

88 We identified candidate proteins based on different layers of evidence to be involved in the
89 pathology of COVID-19: 1) two human proteins related to viral entry¹¹, 2) 332 human proteins
90 shown to interact with viral proteins³, 3) 26 proteomic markers of disease severity⁷, and 4) 54
91 protein biomarkers of adverse prognosis, complications, and disease deterioration^{4-6,12} (**Fig. 1**).
92 Of 409 proteins prioritised, 179 were detectable by an aptamer-based technology (SomaScan[®]),
93 including 28 recognised by more than 1 aptamer (i.e. 179 proteins recognised by 190 aptamers)
94 and 32 also measured using the Olink[®] proximity extension assay in a subset of 485 Fenland
95 study individuals (**Supplemental Tab. S1**). Of these 179 proteins, 111 (**Supplemental Tab. S1**)
96 were classified as druggable proteins, including 32 by existing or developmental drugs¹³, and 22
97 highlighted by Gordon et al. as interacting with SARS-CoV-2 proteins³. To simplify the
98 presentation of results we introduce the following terminology: we define a protein as a unique
99 combination of UniProt entries, i.e. including single proteins and protein complexes. We further
100 define a protein target as the gene product recognised by a specific aptamer, and, finally, an
101 aptamer as a specific DNA-oligomer designed to bind to a specific protein target.

102



103

104 **Figure 1** Flowchart of the identification of candidate proteins and coverage by the SomaScan v4
 105 platform within the Fenland cohort. More details for each protein targeted are given in
 106 Supplemental Table S1.

107

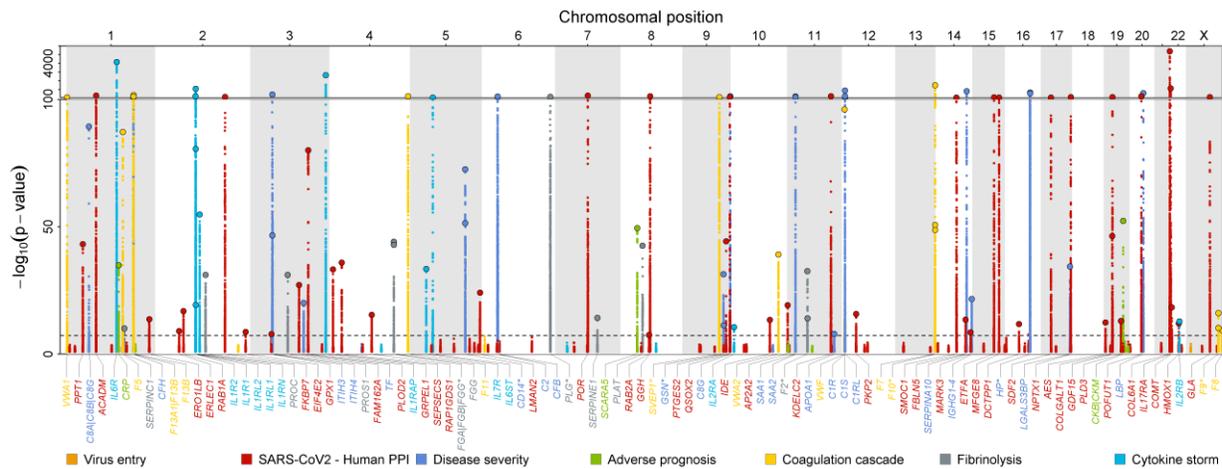
108 **Local genetic architecture of protein targets**

109 We successfully identified 220 DNA sequence variants acting in *cis* for 97 proteins recognised by
 110 106 aptamers (**Fig. 2 and Supplemental Tab. S2**). For 45 of these proteins, no pQTLs had
 111 previously been reported. Of 9 proteins recognised by more than 1 aptamer, sentinel sequence
 112 variants were concordant (identical or in high linkage disequilibrium (LD) $r^2 > 0.8$) between
 113 aptamer pairs or triplets for 7 proteins. Minor allele frequencies ranged from 0.01-49.9%, and
 114 the variance explained ranged from 0.3-70.1% for all *cis*-acting sentinel variants and 0.3-70.9%
 115 for *cis*-acting variants including 2-9 identified secondary signals at 57 targets, similar to what

116 was observed considering all *cis*- and an additional 369 *trans*-acting variants identified for 98
 117 aptamers (0.4-70.9%). Among the 97 proteins, 38 are targets of existing drugs, including 15
 118 proteins (*PLOD2*, *COMT*, *DCTPP1*, *GLA*, *ERO1LB*, *SDF2*, *MARK3*, *ERLEC1*, *FKBP7*, *PTGES2*, *EIF4E2*,
 119 *MFGE8*, *IL17RA*, *COL6A1*, and *PLAT*) (8 with no known pQTL) that were previously identified³ as
 120 interacting with structural or non-structural proteins encoded in the SARS-CoV-2 genome and
 121 16 proteins (*CD14*, *F2*, *F5*, *F8*, *F9*, *F10*, *FGB*, *IL1R1*, *IL2RA*, *IL2RB*, *IL6R*, *IL6ST*, *PLG*, *SERPINC1*,
 122 *SERPINE1*, and *VWF*) (7 with no known pQTL) that encode biomarkers related to COVID-19
 123 severity⁷, prognosis, or outcome.

124

125



126

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

131

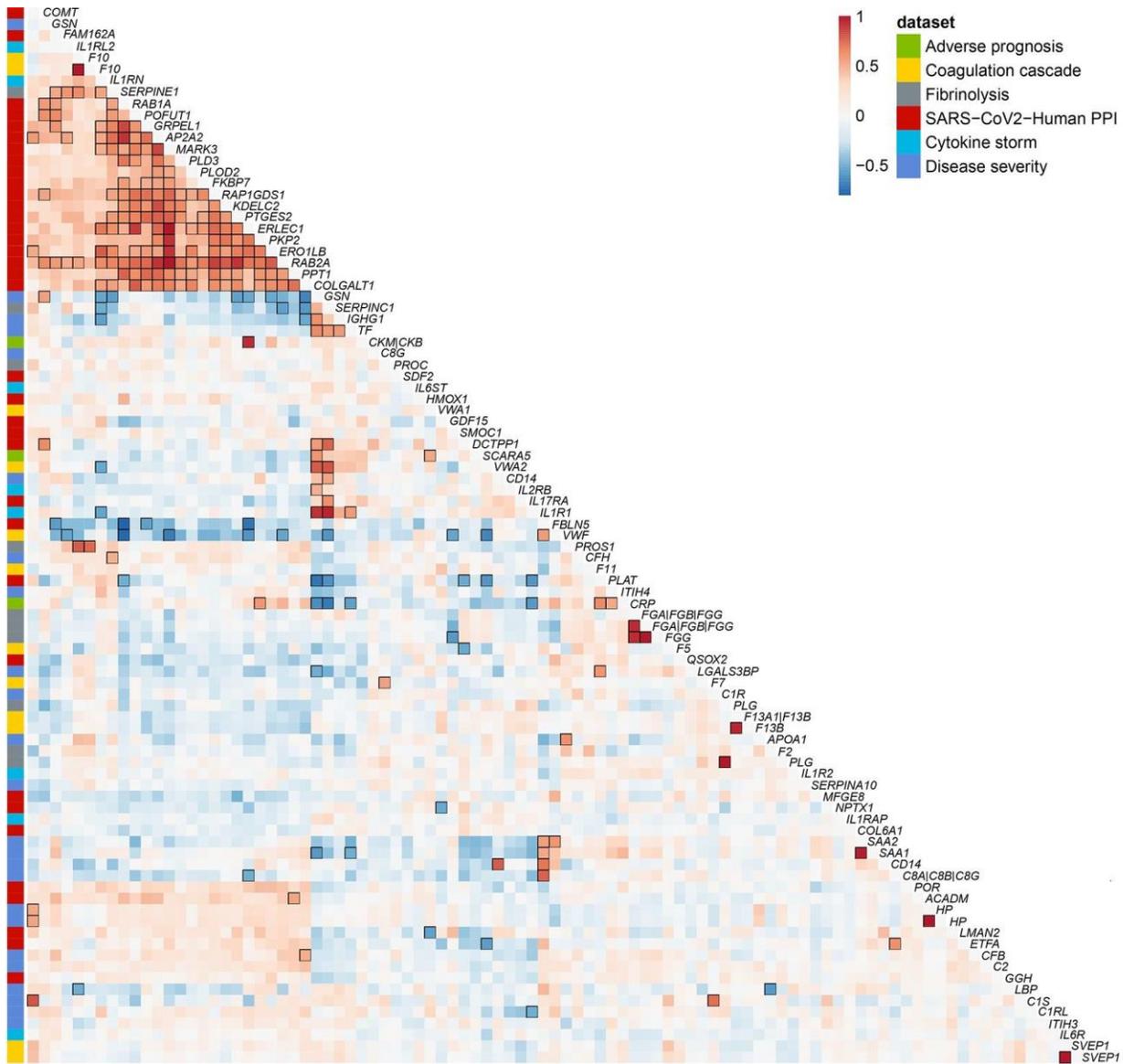
132 Proteins are known to act in a cascade-like manner. To classify such ‘vertical’ pleiotropy, i.e.
 133 associations within a pathway, as well as ‘horizontal’ pleiotropy where proteins are acting
 134 through distinct pathways, we investigated associations of identified lead *cis*-pQTLs with all
 135 measured aptamers ($N=4,776$ unique protein targets, see Methods). For 38 *cis*-pQTLs mapping
 136 to druggable targets, we found evidence for a) protein specific effects for 23 regions, b)
 137 possible vertical pleiotropy for 6, and c) horizontal pleiotropy for 9 lead *cis*-pQTLs. A similar

138 distribution across those categories was seen for the remaining *cis*-pQTLs (Fishers exact test p-
139 value=0.49).

140 To test for dependencies between host proteins predicted to interact with the virus and those
141 related to the maladaptive host response we computed genetic correlations for all proteins
142 with at least one *cis*-pQTL and reliable heritability estimates (see **Methods**). Among 86
143 considered proteins, we identified a highly connected subgroup of 24 proteins including 19
144 SARS-CoV-2-human protein interaction partners (e.g. RAB1A, RAB2A, AP2A2, PLD3, KDELB2,
145 GDP/GTP exchange protein, PPT1, GT251 or PKP2) and 5 proteins related to cytokine storm (IL-
146 1Rrp2 and IL-1Ra), fibrinolysis (PAI-1), coagulation (coagulation factor X(a)), and severity of
147 COVID-19 (GSN (gelsolin)) (**Fig. 3**). The cluster persisted in different sensitivity analyses, such as
148 omitting highly pleiotropic genomic regions (associated with >20 aptamers) or lead *cis*-pQTLs
149 (**Supplementary Fig. S1**). Manual curation highlighted protein modification and vesicle
150 trafficking involving the endoplasmic reticulum as highly represented biological processes
151 related to this cluster. Among these proteins, nine are the targets of known drugs (e.g. COMT,
152 PGES2, PLOD2, ERO1B, XTP3B, FKBP7, or MARK3). The high genetic correlation between these
153 proteins indicates shared polygenic architecture acting in *trans*, which is unlikely to be driven by
154 selected pleiotropic loci identified in the present study.

155 Apart from this cluster, we identified strong genetic correlations ($|r|>0.5$) between smaller sets
156 of proteins related to COVID-19 severity, and host proteins relevant to viral replication such as
157 between IL-6 induced proteins (SAA1, SAA2, and CD14) and fibulin 5 (FBLN5).

158



159

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

165

166 **A tiered system for *trans*-pQTLs**

167 In the absence of an accepted gold standard for the characterization of *trans*-pQTLs, we created
 168 a pragmatic, tiered system to guide selection of *trans*-pQTLs for downstream analyses. We
 169 defined as a) 'specific' *trans*-pQTLs those solely associated with a single protein or protein

170 targets creating a protein complex, b) ‘vertically’ pleiotropic *trans*-pQTLs those associated only
171 with aptamers belonging to the same common biological process (GO-term), and c) as
172 ‘horizontally’ pleiotropic *trans*-pQTLs all remaining ones, i.e. those associated with aptamers
173 across diverse biological processes. We used the entire set of aptamers available on the
174 SomaScan v4 platform, N=4,979, to establish those tiers.

175 Among 451 SNPs acting solely as *trans*-pQTLs, 114 (25.3%) were specific for a protein target, 29
176 (6.4%) showed evidence of vertical pleiotropy, and 308 (68.3%) evidence of horizontal
177 pleiotropy, indicating that *trans*-pQTLs exert their effects on the circulating proteome through
178 diverse mechanisms. As an extreme example, the most pleiotropic *trans*-pQTL (rs4648046,
179 minor allele frequency (MAF)=0.39) showed associations with over 2,000 aptamers and is in
180 high LD ($r^2=0.99$) with a known missense variant at *CFH* (rs1061170). This missense variant was
181 shown, among others, to increase DNA-binding affinity of complement factor H¹⁴, which may
182 introduce unspecific binding of complement factor H to a variety of aptamers, being small DNA-
183 fragments, and may therefore interfere with the method of measurement more generally,
184 rather than presenting a biological effect on these proteins. A similar example is the *trans*-pQTL
185 rs71674639 (MAF=0.21) associated with 789 aptamers and in high LD ($r^2=0.99$) with a missense
186 variant in *BCHE* (rs1803274).

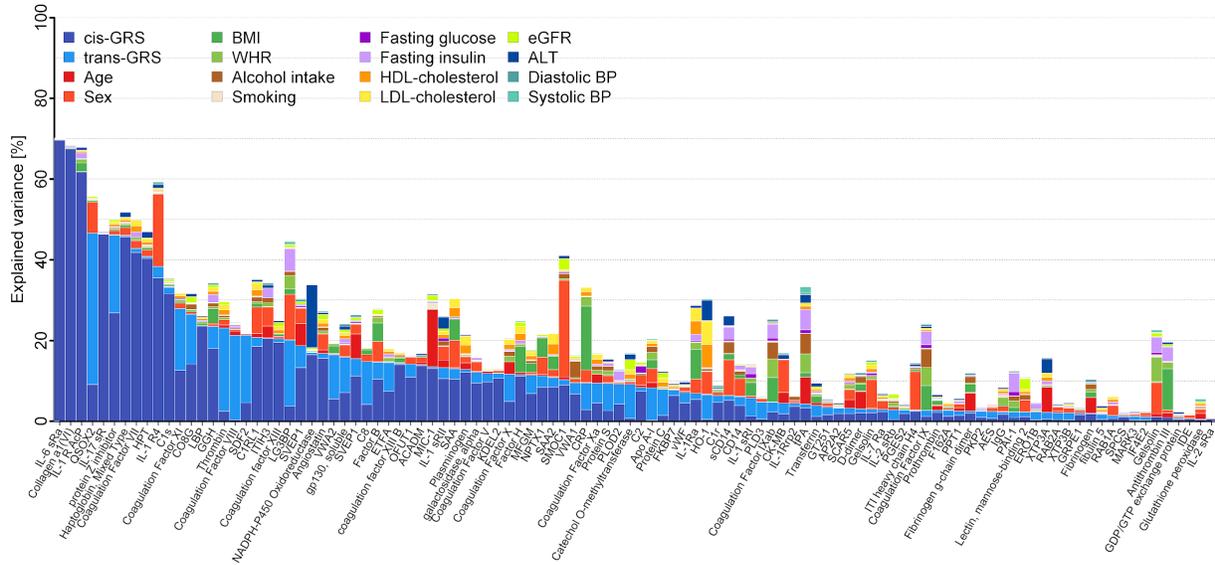
187 Sample handling is an important contributor to the identification of non-specific *trans*-pQTL
188 associations. Blood cells secrete a wide variety of biomolecules, including proteins, following
189 activation or release such as consequence of stress-induced apoptosis or lysis. Interindividual
190 genetic differences in blood cell composition can hence result in genetic differences in protein
191 profiles depending on sample handling or delays in time-to-spin. A prominent example seen in
192 our results and reported in a previous study¹⁵ is variant rs1354034 in *ARHGEF3*, associated with
193 over 1,000 aptamers (on the full SomaScan platform). *ARHGEF3* is a known locus associated
194 with platelet counts¹⁶, albeit its exact function has yet to be determined, either genetically
195 determined higher platelet counts or higher susceptibility to platelet activation may result in
196 the secretion of proteins into plasma during sample preparation. While we report such
197 examples, the extremely standardised and well controlled sample handling of the
198 contemporary and large Fenland cohort has minimised the effects of delayed sample handling

199 on proteomic assessment, as compared to historical cohorts or convenience samples such as
200 from blood donors, evidenced by the fact that previously reported and established sample
201 handling related loci, such as rs62143194 in *NLRP12*¹⁵ are not significant in our study.

202 Finally, for 27 out of 98 aptamers with at least one *cis*- and *trans*-pQTL, we identified no or only
203 very weak evidence for horizontal pleiotropy, i.e. associations in *trans* for no more than 1
204 aptamer, suggesting that those might be used as additional instruments to genetically predict
205 protein levels in independent cohorts for causal assessment.

206 ***Host factors related to candidate proteins***

207 We investigated host factors that may explain variance in the plasma abundances of aptamers
208 targeting high-priority candidate proteins using a variance decomposition approach (see
209 **Methods**). Genetic factors explained more variance compared to any other tested host factors
210 for 63 out of 106 aptamers with IL-6 sRa, collagen a1(VI), or QSOX2 being the strongest
211 genetically determined examples (**Fig. 4**). The composition of non-genetic host factors
212 contributing most to the variance explained appeared to be protein specific (**Fig. 4**). For SMOC1
213 and Interleukin-1 receptor-like 1, for example, sex explained 23.8% and 17.9% of their variance,
214 respectively, indicating different distributions in men and women. Other examples for single
215 factors with large contributions included plasma ALT (15.4% in the variance of NADPH-P450
216 oxidoreductase) or age (14.2% in the variance of GDF-15/MIC-1). We observed a strong and
217 diverse contribution from different non-genetic factors for proteins such as LG3BP, SAA, IL-1Ra,
218 or HO-1 implicating multiple, in part modifiable, factors with independent contributions to
219 plasma levels of those proteins.



220

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

229

230 Patients with multiple chronic conditions are at higher risk of getting severe COVID-19
 231 disease^{2,17,18} and to investigate the influence of disease susceptibility on protein targets of
 232 interest, we generated weighted genetic risk scores (GRS) for major metabolic (e.g. type 2
 233 diabetes and body mass index (BMI)), respiratory (e.g. asthma), and cardiovascular (e.g.
 234 coronary artery disease (CAD)) phenotypes to investigate the association with all COVID-19-
 235 related proteins (**Supplemental Fig. S2**).

236 Plasma abundances of QSOX2 were positively associated with GRS for lung function and
 237 coronary artery disease (CAD), however, as described below these disease score to protein
 238 associations were likely driven by genetic confounding. Specifically, (*cis*) variants in proximity
 239 (± 500 kb) to the protein encoding gene (*QSOX2*) were genome-wide significant for forced
 240 expiratory volume (FEV1) and forced vital capacity (FVC) and exclusion of this region from the

241 lung function genetic score abolished the score to QSOX2 association. None of the three lead
242 *cis*-pQTLs were in strong LD with the lead lung function variant ($r^2 < 0.4$) and genetic
243 colocalization of QSOX2 plasma levels and lung function¹⁹ showed strong evidence for distinct
244 genetic signals (posterior probability of near 100%). The association with the CAD-GRS was
245 attributed to the large contribution of the *ABO* locus to plasma levels of QSOX2, and exclusion
246 of this locus from the CAD score led to the loss of association with QSOX2.

247 The GRSs for BMI (N=10), estimated glomerular filtration rate (eGFR; N=7), and CAD (N=4) were
248 associated with higher as well as lower abundance of different aptamers, and the asthma-GRS
249 was specifically and positively associated with *IL1RL1*. Individuals with higher genetic
250 susceptibility to BMI had higher abundances of three putative viral interaction partners
251 (*LMAN2*, *ETFA*, and *SELENOS*), and lower levels of albumin, *GSN*, and *ITIH3*. Lower plasma
252 abundances of albumin and *GSN* have been associated with severity of COVID-19⁷. Plasma
253 abundance of *LMAN2* (or *VIP36*) was associated with the BMI-GRS (positively) and the eGFR-
254 GRS (inversely). *VIP36* is shed from the plasma membrane upon inflammatory stimuli and has
255 been shown to enhance phagocytosis by macrophages²⁰. The higher plasma levels among
256 individuals with genetically higher BMI and lower kidney function, however, do not reflect the
257 fact that both of these are considered to be risk factors for COVID-19.

258 ***Integration of gene expression data***

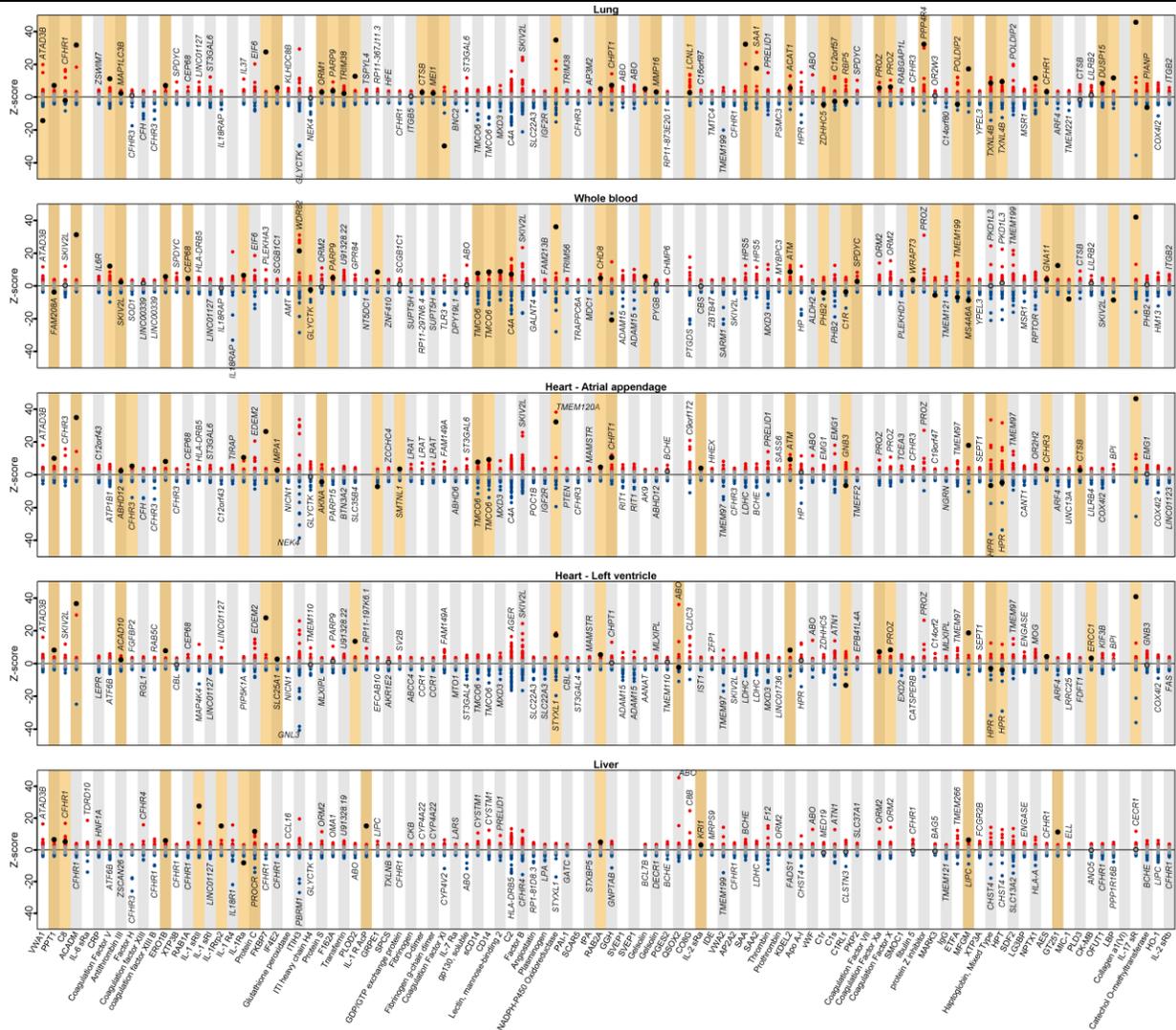
259 We integrated gene expression data across five tissues of direct or indirect relevance to SARS-
260 Cov-2 infection and COVID-19 (lung, whole blood, heart - left ventricle, heart - atrial appendage,
261 and liver) from the GTEx project^{21,22} (version 8) to identify tissues and RNA expression traits
262 contributing to protein targets. Genetically-anchored gene expression models could be
263 established using PrediXcan²³ for at least one of these tissues for 72 of the 102 high-priority
264 aptamers with at least one *cis*-pQTL located on the autosomes. Protein and gene expression
265 were significantly associated for 65 of those aptamers ($p < 0.05$) with varying tissue specificity
266 (Fig. 5), similar to previous reports^{15,24}. Predicted gene expression (druggable targets in bold) of
267 *ACADM*, ***SERPINC1***, ***EROLB1***, *POR*, *RAB2A*, *KDELC2*, *C1RL*, *AES*, ***IL17RA***, ***FKBP7***, and ***EIF4E2***, for
268 example, was consistently associated with corresponding protein levels in plasma across at

269 least three tissues, whereas gene expression in lung only was associated with plasma levels of
270 *SAA1*, *SAA2*, and *SERPINA10*.

271 Plasma levels of proteins depend on multiple biological processes rather than solely on the
272 expression of the encoding genes. Testing for enriched biological terms²⁵ across all significantly
273 associated genes ($p < 10^{-6}$) in lung highlighted 'signal peptide' (false discovery rate
274 (FDR)= 2.5×10^{-5}), 'glycoproteins' (FDR= 1.7×10^{-4}), or 'disulfide bonds' (FDR= 2.8×10^{-4}) as relevant
275 processes. These are involved in the transport and posttranslational modification of proteins
276 before secretion and highlight the complexity of plasma proteins beyond a linear dose-response
277 relationship with tissue abundance of the corresponding mRNA.

278

279



280

281 **Figure 5** Results of predicted gene expression in each of five tissues and plasma abundances of
 282 102 aptamers with at least one *cis*-pQTL on one of the autosomes using PrediXcan. Each panel
 283 displays results for a tissue. Each column contains results across successful gene expression
 284 models for the association with the aptamer listed on the x-axis. Red indicates nominally
 285 significant ($p < 0.05$) positive z-scores (y-axis) and blue nominally significant inverse z-scores for
 286 associated aptamers. Protein encoding genes are highlighted by larger black circles. Orange
 287 background indicates all examples of significant associations between the protein encoding
 288 gene and protein abundance in plasma regardless if this was the most significant one. Top
 289 genes were annotated if those differed from the protein encoding gene.

290

291 **Cross-platform comparison**

292 We tested cross-platform consistency of identified pQTLs using data on 33 protein targets also
 293 captured across 12 Olink protein panels and available in a subset of 485 Fenland participants. In

294 brief, Olink's proximity extension assays use polyclonal antibodies and protein measurements
295 are therefore expected to be less affected by the presence of protein altering variants (PAVs)
296 and so-called epitope effects, since they are likely to affect epitope binding only for a subset of
297 the antibody populations, if any.

298 We compared effect estimates for 29 *cis*- and 96 *trans*-pQTLs based on a reciprocal look-up
299 across both platforms (see **Methods, Supplemental Tab. S5**). We observed strong correlation of
300 effect estimates among 29 *cis*-pQTLs ($r=0.75$, **Fig. S3**) and slightly lower correlation for *trans*-
301 pQTLs ($r=0.54$) indicating good agreement between platforms. In detail, 36 pQTLs (30%)
302 discovered using the far larger SOMAscan-based effort were replicated ($p<0.05$ and
303 directionally consistent) in the smaller subset of participants with overlapping measurements.

304 We identified evidence for inconsistent lead *cis*-pQTLs for two of these 33 protein targets. The
305 lead *cis*-pQTL for GDF-15 from SomaScan (rs75347775) was not significantly associated with
306 GDF-15 levels measured using the Olink assay despite a clear and established signal in *cis* for
307 the Olink measure²⁶ (rs1227731, $\beta=0.59$, $p<6.5\times 10^{-16}$). However, rs1227731 was a secondary
308 signal for the SomaScan assay ($\beta=0.29$, $p<5.8\times 10^{-66}$) highlighting the value of conditional
309 analyses to recover true signals for cases where these are 'overshadowed' by potential false
310 positive lead signals caused by epitope effects. Another protein, the poliovirus receptor (PVR),
311 did not have a *cis*-pQTL in the SomaScan but in the Olink-based discovery (rs10419829,
312 $\beta=-0.84$, $p<2.9\times 10^{-33}$), which in the context of an observational correlation of $r=0.02$ suggests
313 that the two technologies target different protein targets or isoforms. A similar example is
314 ACE2, the entry receptor for SARS-CoV-2, with a correlation of $r=0.05$ between assays and for
315 which we identified only *trans*-pQTLs with evidence for horizontal pleiotropy (**Supplemental**
316 **Tab. S3**). The SCALLOP consortium investigates genetic association data focused on Olink
317 protein measures, and can be a useful and complementary resource for the subset of proteins
318 of interest that are captured (<https://www.olink.com/scallop/>).

319 **Drug target analysis**

320 We identified pQTLs for 105 proteins already the target of existing drugs or known to be
321 druggable which are implicated in the pathogenesis of COVID-19 either through interactions

322 with SARS-CoV-2 proteins, untargeted proteomic analysis of plasma in affected patients, or as
323 candidate proteins in the potentially maladaptive host inflammatory and pro-coagulant
324 responses. Of these, 18 are targets of licensed or clinical phase compounds in the ChEMBL
325 database. Thirteen of these were targets of drugs affecting coagulation or fibrinolytic pathways
326 and five were targets of drugs influencing the inflammatory response. Drugs mapping to targets
327 in the coagulation system included inhibitors of factor 2 (e.g. dabigatran and bivalirudin), factor
328 5 (drotrecogin alfa), factor 10 (e.g. apixaban, rivaroxaban), von Willebrand factor
329 (caplacizumab), plasminogen activator inhibitor 1 (aleplasinin), and tissue plasminogen
330 activator. Drugs mapping to inflammation targets included tocilizumab and satralizumab
331 (targeting the interleukin 6 receptor), brodalumab (targeting the soluble interleukin-17
332 receptor) and anakinra (targeting interleukin-1 receptor type 1). Two targets with pQTLs
333 (catechol O-methyltransferase and alpha-galactosidase-A) were identified as potential virus-
334 host interacting proteins. The former is the target for a drug for Parkinson's disease
335 (entacapone) and the latter is deficient in Fabry's disease, a lysosomal disorder for which
336 migalastat (a drug that stabilises certain mutant forms of alpha-galactosidase-A) is a treatment.

337 Out of the 105 proteins, 24 have no current licensed medicines but are deemed to be druggable
338 including multiple additional targets related to the inflammatory response, prioritised by
339 untargeted proteomics analysis of COVID-19 patient plasma samples. These included multiple
340 components of the complement cascade (e.g. Complement C2, Complement component C8,
341 Complement component C8 gamma chain, and Complement factor H). A number of inhibitors
342 of the complement cascade are licensed (e.g. the C5 inhibitor eculizumab) or in development,
343 although none target the specific complement components prioritised in the current analysis.

344 The effect of drug action on COVID-19 for the targets identified in this analysis requires careful
345 analysis. For example, one target identified through analysis of host-virus protein interactions is
346 prostaglandin E synthase 2 (PGES2) involved in prostaglandin biosynthesis. Non-steroidal anti-
347 inflammatory drugs (NSAIDs) are also known to suppress synthesis of prostaglandins and,
348 though the evidence is weak, concerns have been raised that NSAIDs may worsen outlook in
349 patients with COVID-19²⁷. The *cis*-pQTLs we identified for PGES2 might be useful to explore this
350 further.

351 **Linking cis-pQTLs to clinical outcomes**

352 We first tested whether any of the 220 *cis*-pQTLs or proxies in high LD ($r^2 > 0.8$) have been
353 reported in the GWAS catalogue and identified links between genetically verified drug targets
354 and corresponding indications for lead *cis*-pQTLs at *F2* (rs1799963 associated with venous
355 thrombosis²⁸), *IL6R* (rs2228145 with rheumatoid arthritis²⁹), and *PLG* (rs4252185 associated
356 with coronary artery disease³⁰).

357 To systematically evaluate whether higher plasma levels of candidate proteins are associated
358 with disease risk, we tested genetic risk scores (*cis*-GRS) for all 106 aptamers for their
359 associations with 633 ICD-10 coded outcomes in UK Biobank. We identified 9 significant
360 associations (false discovery rate <10%), including the druggable example of a thrombin-*cis*-GRS
361 (2 *cis*-pQTLs as instruments) and increased risk of pulmonary embolism (ICD-10 code: I26) as
362 well as phlebitis and thrombophlebitis (ICD-10 code: I80) (**Supplemental Table S6**).

363 To maximise power for disease outcomes, include clinically relevant risk factors, and allow for
364 variant-specific effects we complemented the phenome-wide strategy with a comprehensive
365 look-up for genome-wide significant associations in the MR-Base platform³¹.

366 Out of the 220 variants queried, 74 showed at least one genome-wide significant association,
367 20 of which were *cis*-pQTLs for established drug targets. We obtained high posterior
368 probabilities (PP > 75%) for a shared genetic signals between 25 *cis*-pQTLs and at least one
369 phenotypic trait using statistical (conditional) colocalisation (**Fig. 6 and Supplemental Tab. S7**).

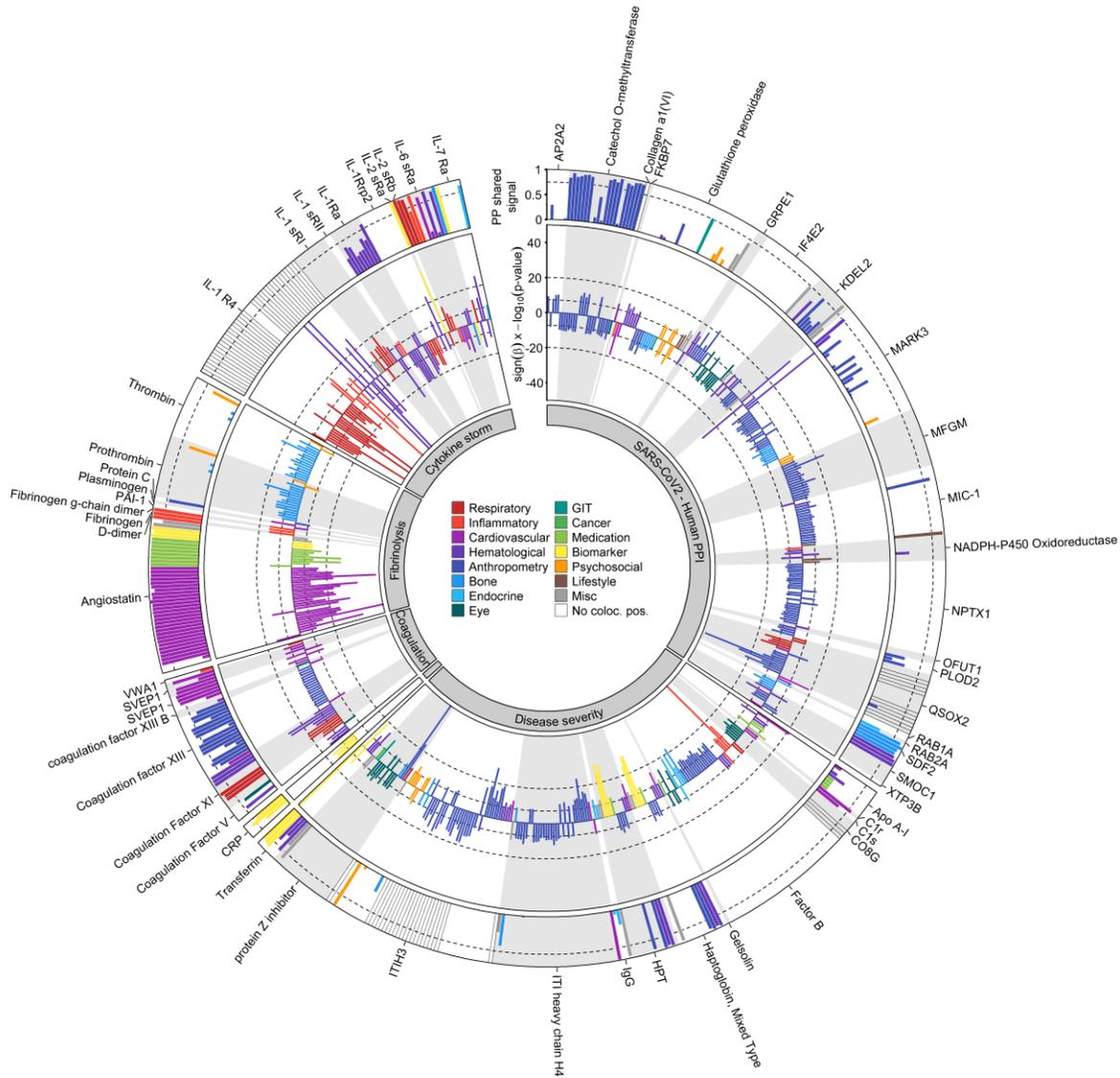
370 Among these was rs8022179, a novel *cis*-pQTL for microtubule affinity-regulating kinase 3
371 (MARK3), a regional lead signal for monocyte count and granulocyte percentage of myeloid
372 white cells¹⁶. The variant showed associations with higher plasma levels of MARK3 and
373 monocyte count and therefore suppression of MARK3 expression with protein kinase inhibitors
374 such as midostaurin may affect the protein host response to the virus. The important role of
375 monocytes and macrophages in the pathology of COVID-19 has been recognised⁴, and a range
376 of immunomodulatory agents are currently evaluated in clinical trials, with a particular focus on
377 the blockade of IL-6 and IL-1 β . Our findings indicate that proteins utilized by the virus itself,
378 such as MARK3, SMOC1, or IL-6 receptor, may increase the number of innate immune cells
379 circulating in the blood and thereby contribute to a hyperinflammatory or hypercoagulable

380 state. Stratification of large COVID-19 patient populations by *cis*-pQTL genotypes that
381 contribute to stimulation/repression of a specific immune signalling pathway is one potential
382 application of our results. However, such investigations would need to be large, i.e. include
383 thousands of patients, and results need to be interpreted with caution as targeting those
384 proteins can have effects not anticipated by the genetic analysis, which cannot mimic short
385 term and dose-dependent 'drug' exposure.

386 We observed general consistency among phenotypic traits colocalising with *cis*-pQTLs, i.e. traits
387 were closely related and effect estimates were consistent with phenotypic presentations
388 (**Supplemental Tab. S7 and Fig. 6**). For instance, rs165656, a lead *cis*-pQTL increasing catechol
389 o-methyltransferase plasma abundances, is a regional lead variant for BMI³² and specifically
390 colocalised with adiposity related traits, i.e. inversely associated with overall measures of body
391 size such as BMI, weight, and fat-free mass. In general, phenotypic characterization of potential
392 genetic instruments to simulate targeting abundances or activities of proteins can help to
393 distinguish those with narrow and well-defined or target-specific from those with undesirable
394 or broad phenotypic effects. Notable exceptions included the IL-6 receptor variant rs2228145,
395 for which the protein increasing C allele was inversely associated with the risk of coronary heart
396 disease and rheumatoid arthritis but positively with the risk for allergic disease, such as asthma.

397 ***A variant at the ABO locus links susceptibility of respiratory failure in COVID-19 to protein*** 398 ***targets***

399 A recent GWAS identified two independent genomic loci to be associated with an increased risk
400 of respiratory failure in COVID-19 patients¹⁰. We observed six proteins to be associated
401 positively with the lead signal (rs657152) at the *ABO* locus (coagulation factor VIII, sulfhydryl
402 oxidase 2 (QSOX2), von Willebrand factor, SVEP1, and heme oxygenase 1) and one inverse
403 association (interleukin-6 receptor subunit beta), but did not observe significantly associated
404 proteins with the lead variant (rs11385942) at 3p21.31. We identified a cluster of ten aptamers
405 (targeting SVEP1, coagulation factor VIII, ferritin, heme oxygenase 1, van Willebrand factor,
406 plasminogen, PLOD2, and CD14) sharing a genetic signal (regional probability: 0.88; rs941137;
407 **Supplemental Fig. S4**), which was in high LD ($r^2=0.85$) with the lead *ABO* signal associated with
408 a higher risk for respiratory failure among COVID-19 patients.



409

410

411 **Figure 6** Circos plot summarizing genome-wide significant associations between 74 *cis*-pQTLs
 412 and 239 traits³¹ in the inner ring and results from statistical colocalisation in the outer ring. The
 413 dashed line in the outer ring indicates a posterior probability of 75% of shared genetic signal
 414 between the protein and a phenotypic trait. Protein targets are classified on the basis of their
 415 reported relation to SARS-CoV-2 and COVID-19. Each slice contains any *cis*-pQTLs associated
 416 with the target protein annotated and effect estimates were aligned to the protein increasing
 417 allele, i.e. bars with a positive $-\log_{10}(p\text{-values})$ indicate positive associations with a trait from
 418 the database and *vice versa*. Clinical traits are grouped by higher-level categories and coloured
 419 accordingly. GIT = gastrointestinal tract, Misc = Miscellaneous, No coloc. pos. = colocalisation
 420 for secondary signals was not possible

421

422 **Webserver**

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

432 **DISCUSSION**

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

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

450 The successful identification of druggable targets for COVID-19 provides an insight both on
451 potential therapies but also on medications that might worsen outlook, depending on the
452 direction of the genetic effect, and whether any associated compound inhibits or activates the
453 target. We also found genetic evidence that selected protein targets, such as for MARK3 and
454 monocyte count, have potential for adverse effects on other health outcomes, but note that
455 this was not a general characteristic of all tested 'druggable' targets. Further, in-depth
456 characterization of the targets identified will be required as a first step in gauging the likely
457 success of any new or repurposed drugs identified via this analysis³³.

458 We exemplify the value of the data resource generated by being the first that links a genomic
459 risk variant for poor prognosis among COVID-19 patients, i.e. respiratory failure, at the *ABO*
460 locus¹⁰ to proteins related to the maladaptive response of the host, namely hypercoagulation,
461 as well as two putative viral interaction partners (heme oxygenase 1 and PLOD2). The risk
462 increasing A allele of rs657152 was consistently associated with higher plasma levels of
463 coagulation factor VIII and von Willebrand factor. Anticoagulation is associated with a better
464 outcome in patients with severe COVID-19³⁴, and randomised controlled trails are underway to
465 properly evaluate the benefit or harms of anticoagulant therapies.

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

474 This study is the largest genetic discovery of protein targets highly relevant to the current
475 COVID-19 pandemic and was designed to provide a rapid open access platform to help prioritise
476 drug discovery and repurposing efforts. However, important limitations apply. Firstly, protein
477 abundances have been measured in plasma, which may differ from the intracellular role of
478 proteins, and include purposefully secreted as well as leaked proteins. Secondly, while

479 aptamer-based techniques provide the broadest coverage of the plasma proteome, specificity
480 can be compromised for specific protein targets and evidence using complementary techniques
481 such as Olink or mass spectrometry efforts is useful for validation of signals. Thirdly, in-depth
482 phenotypic characterization of the high-priority *cis*-pQTLs requires appropriate formal and
483 statistical follow-up, such as colocalisation, where the genomic architecture permits existing
484 approaches not yet optimised for multiple secondary signals and outcomes, and *cis*-GRS
485 evaluation in independent and adequately powered studies for the trait of interest.
486

487 **Materials and Methods**

488 *Study participants*

489 The Fenland study is a population-based cohort of 12,435 participants born between 1950 and
490 1975 who underwent detailed phenotyping at the baseline visit from 2005-2015. Participants
491 were recruited from general practice surgeries in the Cambridgeshire region in the UK.
492 Exclusion criteria were: clinically diagnosed diabetes mellitus, inability to walk unaided,
493 terminal illness, clinically diagnosed psychotic disorder, pregnancy or lactation. The study was
494 approved by the Cambridge Local Research Ethics Committee (ref. 04/Q0108/19) and all
495 participants provided written informed consent. Population characteristics and proteomic
496 measures have previously been described in detail³⁵.

497 *Mapping of protein targets across platforms*

498 We mapped each candidate protein to its UniProt-ID (<https://www.uniprot.org/>) and used
499 those to select mapping aptamers and Olink measures based on annotation files provided by
500 the vendors.

501 *Proteomic profiling*

502 Proteomic profiling of fasted EDTA plasma samples from 12,084 Fenland Study participants
503 collected at baseline was performed by SomaLogic Inc. (Boulder, US) using an aptamer-based
504 technology (SOMAscan proteomic assay). Relative protein abundances of 4,775 human protein
505 targets were evaluated by 4,979 aptamers (SomaLogic V4), as previously described³⁵. To
506 account for variation in hybridization within runs, hybridization control probes are used to
507 generate a hybridization scale factor for each sample. To control for total signal differences
508 between samples due to variation in overall protein concentration or technical factors such as
509 reagent concentration, pipetting or assay timing, a ratio between each aptamer's measured
510 value and a reference value is computed, and the median of these ratios is computed for each
511 of the three dilution sets (40%, 1% and 0.005%) and applied to each dilution set. Samples were
512 removed if they were deemed by SomaLogic to have failed or did not meet our acceptance
513 criteria of 0.25-4 for all scaling factors. In addition to passing SomaLogic QC, only human
514 protein targets were taken forward for subsequent analysis (4,979 out of the 5284 aptamers).

515 Aptamers' target annotation and mapping to UniProt accession numbers as well as Entrez gene
516 identifiers were provided by SomaLogic.

517 Plasma samples for a subset of 500 Fenland participants were additionally measured using 12
518 Olink 92-protein panels using proximity extension assays³⁶. Of the 1104 Olink proteins, 1069
519 were unique (n=35 on >1 panel, average correlation coefficient 0.90). We imputed values below
520 the detection limit of the assay using raw fluorescence values. Protein levels were normalized
521 ('NPX') and subsequently log₂-transformed for statistical analysis. A total of 15 samples were
522 excluded based on quality thresholds recommended by Olink, leaving 485 samples for analysis.

523 *Genotyping and imputation*

524 Fenland participants were genotyped using three genotyping arrays: the Affymetrix UK Biobank
525 Axiom array (OMICS, N=8994), Illumina Infinium Core Exome 24v1 (Core-Exome, N=1060) and
526 Affymetrix SNP5.0 (GWAS, N=1402). Samples were excluded for the following reasons: 1) failed
527 channel contrast (DishQC <0.82); 2) low call rate (<95%); 3) gender mismatch between reported
528 and genetic sex; 4) heterozygosity outlier; 5) unusually high number of singleton genotypes or
529 6) impossible identity-by-descent values. Single nucleotide polymorphisms (SNPs) were
530 removed if: 1) call rate < 95%; 2) clusters failed Affymetrix SNPolisher standard tests and
531 thresholds; 3) MAF was significantly affected by plate; 4) SNP was a duplicate based on
532 chromosome, position and alleles (selecting the best probeset according to Affymetrix
533 SNPolisher); 5) Hardy-Weinberg equilibrium $p < 10^{-6}$; 6) did not match the reference or 7)
534 MAF=0.

535 Autosomes for the OMICS and GWAS subsets were imputed to the HRC (r1) panel using
536 IMPUTE4³⁷, and the Core-Exome subset and the X-chromosome (for all subsets) were imputed
537 to HRC.r1.1 using the Sanger imputation server (<https://imputation.sanger.ac.uk/>)³⁸. All three
538 arrays subsets were also imputed to the UK10K+1000Gphase3³⁹ panel using the Sanger
539 imputation server in order to obtain additional variants that do not exist in the HRC reference
540 panel. Variants with MAF < 0.001, imputation quality (info) < 0.4 or Hardy Weinberg Equilibrium
541 $p < 10^{-7}$ in any of the genotyping subsets were excluded from further analyses.

542 *GWAS and meta-analysis*

543 After excluding ancestry outliers and related individuals, 10,708 Fenland participants had both
544 phenotypes and genetic data for the GWAS (OMICS=8,350, Core-Exome=1,026, GWAS=1,332).
545 Within each genotyping subset, aptamer abundances were transformed to follow a normal
546 distribution using the rank-based inverse normal transformation. Transformed aptamer
547 abundances were then adjusted for age, sex, sample collection site and 10 principal
548 components and the residuals used as input for the genetic association analyses. Test site was
549 omitted for protein abundances measured by Olink as those were all selected from the same
550 test site. Genome-wide association was performed under an additive model using BGENIE
551 (v1.3)³⁷. Results for the three genotyping arrays were combined in a fixed-effects meta-analysis
552 in METAL⁴⁰. Following the meta-analysis, 17,652,797 genetic variants also present in the largest
553 subset of the Fenland data (Fenland-OMICS) were taken forward for further analysis.

554 *Definition of genomic regions (including cis/trans)*

555 For each aptamer, we used a genome-wide significance threshold of 5×10^{-8} and defined non-
556 overlapping regions by merging overlapping or adjoining 1Mb intervals around all genome-wide
557 significant variants (500kb either side), treating the extended MHC region (chr6:25.5–34.0Mb)
558 as one region. For each region we defined a regional sentinel variant as the most significant
559 variant in the region. We defined genomic regions shared across aptamers if regional sentinels
560 of overlapping regions were in strong LD ($r^2 > 0.8$).

561 *Conditional analysis*

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

568 We performed a forward stepwise selection procedure to identify secondary signals at each
569 locus on the X-chromosome using SNPTTEST v.2.5.2 to compute conditional GWAS based on
570 individual level data in the largest subset. Briefly, we defined conditionally independent signals

571 as those emerging after conditioning on all previously selected signals in the locus until no
572 signal was genome-wide significant.

573 *Explained variance*

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

578 *Annotation of pQTLs*

579 For each identified pQTL we first obtained all SNPs in at least moderate LD ($r^2 > 0.1$) and queried
580 comprehensive annotations using the variant effect predictor software⁴¹ (version 98.3) using
581 the *pick* option. For each *cis*-pQTL we checked whether either the variant itself or a proxy in the
582 encoding gene ($r^2 > 0.1$) is predicted to induce a change in the amino acid sequence of the
583 associated protein, so-called protein altering variants (PAVs).

584 *Mapping of cis-pQTLs to drug targets*

585 To annotate druggable targets we merged the list of proteins targeted by the SomaScan V4
586 platform with the list of druggable genes from Finan et al.¹³ based on common gene entries. We
587 further added protein – drug combinations as recommended by Gordon et al.³.

588 *Identification of relevant GWAS traits*

589 To enable linkage to reported GWAS-variants we downloaded all SNPs reported in the GWAS
590 catalog (19/12/2019, <https://www.ebi.ac.uk/gwas/>) and pruned the list of variant-outcome
591 associations manually to omit previous protein-wide GWAS. For each SNP identified in the
592 present study (N=671) we tested whether the variant or a proxy in LD ($r^2 > 0.8$) has been
593 reported to be associated with other outcomes previously.

594 *Definition of novel pQTLs*

595 To test whether any of the identified regional sentinel pQTLs has been reported previously, we
596 obtained a list of published pQTLs^{15,24,26,42,43} and defined novel pQTLs as those not in LD ($r^2 < 0.1$)
597 with any previously identified variant. We note that this approach is rather conservative, since

598 it only asks whether or not any of the reported SNPs has ever been reported to be associated
599 with any protein measured with multiplex methods.

600 *Assessment of pleiotropy*

601 To evaluate possible protein-specific pleiotropy of pQTLs we computed association statistics for
602 each of the 671 unique SNPs across 4,979 aptamers (N=4,775 unique protein targets) with the
603 same adjustment set as in the GWAS. This resulted in a protein profile for each variant defined
604 as all aptamers significantly associated ($p < 5 \times 10^{-8}$). For all aptamers we retrieved all GO-terms
605 referring to biological processes from the UniProt database using all possible UniProt-IDs as a
606 query. GO-term annotation within the UniProt database has the advantage of being manually
607 curated while aiming to omit unspecific parent terms. We tested for each pQTL if the associated
608 aptamers fall into one of the following criteria: 1) solely associated with a specific protein, 2) all
609 associated aptamers belong to a single GO-term, 3) the majority (>50%) of associated aptamers
610 but at least two belong to a single GO-term, and 4) no single GO-term covers more than 50% of
611 the associated aptamers. We refer to category 1 as protein-specific association, categories 2
612 and 3 as vertical pleiotropy, and category 4 as horizontal pleiotropy.

613 *Heritability estimates and genetic correlation*

614 We used genome-wide genotype data from 8,350 Fenland participants (Fenland-OMICs) to
615 determine SNP-based heritability and genetic correlation estimates among the 102 protein
616 targets with at least one *cis*-pQTLs and excluding proteins encoded in the X-chromosome. We
617 generated a genetic relationship matrix (GRM) using GCTA v.1.90⁴⁴ from all variants with MAF >
618 1% to calculate SNP-based heritability as implemented by biMM⁴⁵. Genetic correlations were
619 computed between all 4273 possible pairs among 93 protein targets with heritability estimates
620 larger than 1.5 times its standard error, using the generated GRM by a bivariate linear mixed
621 model as implemented by biMM. We further conducted two sensitivity analyses to evaluate
622 whether the estimated genetic correlation could be largely attributable to the top *cis*-pQTL or
623 to shared pleiotropic *trans* regions. To evaluate contribution of the top *cis* variant, each protein
624 target was regressed against its sentinel *cis* variant in addition to age, sex, sample collection
625 site, 10 principal components and the residuals were used as phenotypes to compute

626 heritability and genetic correlation estimates. To assess the contribution of 29 pleiotropic *trans*
627 regions, we excluded 2Mb genomic regions around pleiotropic *trans*-pQTLs (associated with
628 >20 aptamers) from the GRM to compute heritability and genetic correlation estimates. Genetic
629 correlations could not be computed for pairs involving IL1RL1 in the main analysis and were
630 therefore excluded. However, upon regressing out the sentinel *cis*-variant, genetic correlations
631 with this protein could be computed probably due to its large contribution to heritability.

632 *Variance decomposition*

633 We used linear mixed models as implemented in the R package *variancePartition* to decompose
634 inverse rank-normal transformed plasma abundances of 106 aptamers with at least one *cis*-
635 pQTL. To this end, we computed weighted genetic scores for each aptamer separating SNPs
636 acting in *cis* (*cis*-GRS) and *trans* (*trans*-GRS). In addition to the GRS we used participants' age,
637 sex, body mass index, waist-to-hip ratio, systolic and diastolic blood pressure, reported alcohol
638 intake, smoking consumption and fasting plasma levels of glucose, insulin, high-density
639 lipoprotein cholesterol, low-density lipoprotein cholesterol, alanine aminotransaminase as well
640 as a creatinine-based estimated glomerular filtration rate as explanatory factors. We
641 implemented this analysis in the Fenland-OMICs data set leaving 8,004 participants without any
642 missing values in the factors considered.

643 *Genetic risk scores associations*

644 We computed weighted GRS for metabolic (Insulin resistance⁴⁶, type 2 diabetes⁴⁷ and BMI⁴⁸),
645 respiratory (forced expiratory volume, forced vital capacity¹⁹ and asthma⁴⁹) and cardiovascular
646 traits (eGFR⁵⁰, systolic blood pressure⁵¹, diastolic blood pressure⁵¹ and coronary artery
647 disease³⁰) for Fenland-OMICs participants (N = 8,350) to evaluate their association with plasma
648 protein abundances. GRSs were computed from previously reported genome-wide significant
649 variants and weighted by their reported beta coefficients for continuous outcomes or log(OR)
650 for binary outcomes. Variants not available among Fenland genotypes, strand ambiguous or
651 with low imputation quality (INFO < 0.6) were excluded from the GRSs. Associations between
652 each scaled GRS and log₁₀ transformed and scaled protein levels were computed by linear
653 regressions adjusted by age, sex, 10 genetic principal components and sample collection site.

654 We implemented this analysis for the 186 proteins with at least one associated cis or trans-
655 pQTL. Associations with p-values $< 0.05/186$ were deemed significant according to Bonferroni
656 correction for multiple comparisons.

657 *Incorporation of GTEx v8 data*

658 We leveraged gene expression data in five human tissues (lung, whole blood, heart - left
659 ventricle, heart - atrial appendage, and liver), of relevance to COVID-19 and its potential
660 adverse effects and complications, from the Genotype-Tissue Expression (GTEx) project^{21,22}. For
661 the 102 Somamers with at least one *cis*-pQTL located on the autosomes and available gene
662 expression models trained in GTEx v8⁵², we performed summary-statistics based PrediXcan²³
663 analysis to identify tissue-dependent genetically determined gene expression traits that
664 significantly predict plasma protein levels. We used the standardized effect size (z-score) to
665 investigate the tissue specificity or the consistency of the association across the tissues
666 between the genetic component of the expression of the encoding gene and the corresponding
667 protein. We performed DAVID functional enrichment analyses on all the genes significantly
668 associated (Bonferroni-adjusted $p < 0.05$) with plasma levels of the proteins to identify biological
669 processes (Benjamini-Hochberg adjusted $p < 0.05$) that may explain the associations found
670 beyond the protein encoding genes.

671 *Cross-platform comparison*

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

679 *Phenome-wide scan among UK Biobank and look-up*

680 We obtained all ICD-10 codes-related genome-wide summary statistics from the most recent
681 release of the Neale lab (<http://www.nealelab.is/uk-biobank>) with at least 100 cases resulting

682 in 633 distinct ICD-10 codes. Among the 220 *cis*-pQTLs identified in the present study, 215 were
683 included in the UK Biobank summary statistics (3 aptamers had to be excluded due to
684 unavailable lead *cis*-pQTLs or proxies in LD). We next aligned effect estimates between *cis*-
685 pQTLs and UK Biobank statistics and used the *grs.summary()* function from the 'gtx' R package
686 to compute the effect of a weighted *cis*-GRS for an aptamer across all 633 ICD-codes. We
687 applied a global testing correction across all *cis*-GRS – ICD-10 code combinations using the
688 Benjamini-Hochberg procedure and declared a false discovery rate of 10% as a significance
689 threshold.

690 We queried all 220 *cis*-pQTLs for genome-wide association results using the *phewas()* function
691 of the R package 'ieugwas' linked to the IEU GWAS database. We selected all variants in strong
692 LD ($r^2 > 0.8$) with any of the *cis*-pQTLs to incorporate information on proxies. We restricted the
693 search in the ieugwas tool to the batches "ebi-a", "ieu-a", and "ukb-b" to minimize redundant
694 phenotypes.

695 *Colocalisation analysis*

696 We used statistical colocalisation⁵³ to test for a shared genetic signal between a protein target
697 and a phenotype with evidence of a significant effect of the *cis*-pQTL (see above). We obtained
698 posterior probabilities (PP) of: H0 – no signal; H1 – signal unique to the protein target; H2 –
699 signal unique to the trait; H3 – two distinct causal variants in the same locus and H4 – presence
700 of a shared causal variant between a protein target and a given trait. PPs above 75% were
701 considered highly likely. In case the *cis*-pQTL was a secondary signal we computed conditional
702 association statistics using the *cond* option from GCTA-cojo to align with the identification of
703 secondary signals. We conditioned on all other secondary signals in the locus. We note that
704 conditioning on all other secondary variants in the locus failed to produce the desired
705 conditional association statistics in a few cases probably due to moderate LD ($r^2 > 0.1$) between
706 selected secondary variants and other putative secondary variants.

707 *Multi-trait colocalization at the ABO locus*

708 We used hypothesis prioritisation in multi-trait colocalization (HyPrColoc)⁵⁴ at the *ABO* locus
709 (± 200 kb) 1) to identify protein targets sharing a common causal variant over and above what

710 could be identified in the meta-analysis to increase statistical power, and 2) to identify possible
711 multiple causal variants with distinct associated protein clusters. Briefly, HyPrColoc aims to test
712 the global hypothesis that multiple traits share a common genetic signal at a genomic location
713 and further uses a clustering algorithm to partition possible clusters of traits with distinct causal
714 variants within the same genomic region. HyPrColoc provides for each cluster three different
715 types of output: 1) a posterior probability (PP) that all traits in the cluster share a common
716 genetic signal, 2) a regional association probability, i.e. that all the metabolites share an
717 association with one or more variants in the region, and 3) the proportion of the PP explained
718 by the candidate variant. We considered a highly likely alignment of a genetic signal across
719 various traits if the regional association probability > 80%. This criterion takes to some extent
720 into account that metabolites may share multiple causal variants at the same locus and
721 provides some robustness against violation of the single causal variant assumption. We note
722 that several protein targets had multiple independent signals at the ABO locus (**Supplementary**
723 **Tab. S4**). We further filtered protein targets with no evidence of a likely genetic signal ($p > 10^{-5}$)
724 in the region before performing HyPrColoc, which improved clustering across traits due to
725 minimizing noise.

726

727

728 **ACKNOWLEDGMENTS AND FUNDING**

729 The Fenland Study (10.22025/2017.10.101.00001) is funded by the Medical Research Council
730 (MC_UU_12015/1). We are grateful to all the volunteers and to the General Practitioners and
731 practice staff for assistance with recruitment. We thank the Fenland Study Investigators,
732 Fenland Study Co-ordination team and the Epidemiology Field, Data and Laboratory teams. We
733 further acknowledge support for genomics from the Medical Research Council (MC_PC_13046).
734 Proteomic measurements were supported and governed by a collaboration agreement
735 between the University of Cambridge and Somalogic. JCZ and VPWA are supported by a 4-year
736 Wellcome Trust PhD Studentship and the Cambridge Trust, CL, EW, and NJW are funded by the
737 Medical Research Council (MC_UU_12015/1). NJW is an NIHR Senior Investigator. GK is
738 supported by grants from the National Institute on Aging (NIA): R01 AG057452, RF1 AG058942,
739 RF1 AG059093, U01 AG061359, and U19 AG063744. MR acknowledges funding from the Francis
740 Crick Institute, which receives its core funding from Cancer Research UK (FC001134), the UK
741 Medical Research Council (FC001134), and the Wellcome Trust (FC001134). ERG is supported by
742 the National Institutes of Health Genomic Innovator Award ([R35 HG010718](#)). JR is supported by
743 the German Federal Ministry of Education and Research (BMBF) within the framework of the
744 e:Med research and funding concept (grant no. 01ZX1912D).

745 **AUTHOR CONTRIBUTIONS**

746 MP, ADH, and CL designed the analysis and drafted the manuscript. MP, EW, JCSZ, VPWA, and
747 JL analysed the data. NK and EO performed quality control of proteomic measurements. JR and
748 GK designed and implemented the webserver. RO and SW advised proteome measurements
749 and assisted in quality control. EG did the gene expression analysis and interpretation of results.
750 JPC and MR provided critical review and intellectual contribution to the discussion of results.
751 NJW is PI of the Fenland cohort. All authors contributed to the interpretation of results and
752 critically reviewed the manuscript.

753 **COMPETING INTERESTS**

754 SW and RO are employees of SomaLogic.

755 **DATA AVAILABILITY**

756 All genome-wide summary statistics are made available through an interactive webserver
757 (<https://omicscience.org/apps/covidpgwas/>).

758 **CODE AVAILABILITY**

759 Each use of software programs has been clearly indicated and information on the options that were
760 used is provided in the Methods section. Source code to call programs is available upon request.

761

762 REFERENCES

- 763 1. Banerjee, A. *et al.* Articles Estimating excess 1-year mortality associated with the COVID-
764 19 pandemic according to underlying conditions and age : a population-based cohort
765 study. *Lancet* **6736**, 1–11 (2020).
- 766 2. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-
767 19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
- 768 3. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug
769 repurposing. *Nature* 1–13 (2020) doi:10.1038/s41586-020-2286-9.
- 770 4. Merad, M. & Martin, J. C. Pathological inflammation in patients with COVID-19: a key role
771 for monocytes and macrophages. *Nat. Rev. Immunol.* **2**,
- 772 5. Zhang, L. *et al.* D-dimer levels on admission to predict in-hospital mortality in patients
773 with Covid-19. *J. Thromb. Haemost.* **18**, 1324–1329 (2020).
- 774 6. Violi, F., Pastori, D., Cangemi, R., Pignatelli, P. & Loffredo, L. Hypercoagulation and
775 Antithrombotic Treatment in Coronavirus 2019: A New Challenge. (2020) doi:10.1055/s-
776 0040-1710317.
- 777 7. Messner, C. B. *et al.* Clinical classifiers of COVID-19 infection from novel ultra-high-
778 throughput proteomics. 1–35 (2020).
- 779 8. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug
780 indications. *Nat. Genet.* **47**, 856–860 (2015).
- 781 9. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as
782 likely to be approved? Revised estimates of the impact of genetic support for drug
783 mechanisms on the probability of drug approval. *PLOS Genet.* **15**, e1008489 (2019).
- 784 10. Ellinghaus, D. *et al.* The ABO blood group locus and a chromosome 3 gene cluster
785 associate with SARS-CoV-2 respiratory failure in an Italian-Spanish genome-wide
786 association analysis. doi:10.1101/2020.05.31.20114991.
- 787 11. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked
788 by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
- 789 12. Jose, R. J. & Manuel, A. COVID-19 cytokine storm: the interplay between inflammation
790 and coagulation. doi:10.1016/S2213-2600(20)30216-2.
- 791 13. Finan, C. *et al.* *The druggable genome and support for target identification and validation*
792 *in drug development*. <http://stm.sciencemag.org/>.
- 793 14. Sjöberg, A. P. *et al.* The factor H variant associated with age-related macular
794 degeneration (His-384) and the non-disease-associated form bind differentially to C-
795 reactive protein, fibromodulin, DNA, and necrotic cells. *J. Biol. Chem.* **282**, 10894–900
796 (2007).
- 797 15. Bansal, N. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79
798 (2018).

- 799 16. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to
800 Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
- 801 17. WHO. Coronavirus disease. *World Heal. Organ.* **2019**, 2633 (2020).
- 802 18. Mehra, M. R., Desai, S. S., Kuy, S., Henry, T. D. & Patel, A. N. Cardiovascular Disease, Drug
803 Therapy, and Mortality in Covid-19. *N. Engl. J. Med.* 1–8 (2020)
804 doi:10.1056/NEJMoa2007621.
- 805 19. Shrine, N. *et al.* New genetic signals for lung function highlight pathways and chronic
806 obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**,
807 481–493 (2019).
- 808 20. Shirakabe, K., Hattori, S., Seiki, M., Koyasu, S. & Okada, Y. VIP36 Protein Is a Target of
809 Ectodomain Shedding and Regulates Phagocytosis in Macrophage Raw 264.7 Cells * □ S.
810 (2011) doi:10.1074/jbc.M111.275586.
- 811 21. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in
812 humans. *Science (80-.).* **348**, 648 LP – 660 (2015).
- 813 22. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform
814 complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
- 815 23. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using
816 reference transcriptome data. **47**, (2015).
- 817 24. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood
818 plasma proteome. *Nat. Commun.* **8**, (2017).
- 819 25. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of
820 large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
- 821 26. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular
822 disease. (2017) doi:10.1371/journal.pgen.1006706.
- 823 27. Little, P. Non-steroidal anti-inflammatory drugs and covid-19. *BMJ* **368**, 1–2 (2020).
- 824 28. Klarin, D. *et al.* Genome-wide association analysis of venous thromboembolism identifies
825 new risk loci and genetic overlap with arterial vascular disease. *Nat. Genet.* **51**, 1574–
826 1579 (2019).
- 827 29. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug
828 discovery. *Nature* **506**, 376–381 (2014).
- 829 30. Nikpay, M. *et al.* A comprehensive 1000 Genomes-based genome-wide association meta-
830 analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
- 831 31. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the
832 human phenome. (2018) doi:10.7554/eLife.34408.001.
- 833 32. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power.
834 *Am. J. Hum. Genet.* **104**, 65–75 (2019).
- 835 33. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the

- 836 plasma proteome on complex diseases. *bioRxiv* 627398 (2019) doi:10.1101/627398.
- 837 34. Tang, N. *et al.* Anticoagulant treatment is associated with decreased mortality in severe
838 coronavirus disease 2019 patients with coagulopathy. *J. Thromb. Haemost.* **18**, 1094–
839 1099 (2020).
- 840 35. Williams, S. A. *et al.* Plasma protein patterns as comprehensive indicators of health. *Nat.*
841 *Med.* **25**, 1851–1857 (2019).
- 842 36. Assarsson, E. *et al.* Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity ,
843 Specificity , and Excellent Scalability. **9**, (2014).
- 844 37. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
845 *Nature* **562**, 203–209 (2018).
- 846 38. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat.*
847 *Genet.* **48**, 1279–1283 (2016).
- 848 39. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K
849 haplotype reference panel. *Nat. Commun.* **6**, 1–9 (2015).
- 850 40. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of
851 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
- 852 41. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* 1–14 (2016)
853 doi:10.1186/s13059-016-0974-4.
- 854 42. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to
855 disease. *Science (80-.).* **361**, 1–12 (2018).
- 856 43. Enroth, S. B. S., Johansson, Å., Enroth, S. B. S. & Gyllenstein, U. Strong effects of genetic
857 and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat.*
858 *Commun.* **5**, 4684 (2014).
- 859 44. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide
860 complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
- 861 45. Pirinen, M. *et al.* BiMM: Efficient estimation of genetic variances and covariances for
862 cohorts with high-dimensional phenotype measurements. *Bioinformatics* **33**, 2405–2407
863 (2017).
- 864 46. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic
865 traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–
866 1005 (2012).
- 867 47. Trompet, S. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using
868 high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513
869 (2018).
- 870 48. Lotta, L. A. *et al.* Association of Genetic Variants Related to Gluteofemoral vs Abdominal
871 Fat Distribution with Type 2 Diabetes, Coronary Disease, and Cardiovascular Risk Factors.
872 *JAMA - J. Am. Med. Assoc.* **320**, 2553–2563 (2018).

- 873 49. Olafsdottir, T. A. *et al.* Eighty-eight variants highlight the role of T cell regulation and
874 airway remodeling in asthma pathogenesis. *Nat. Commun.* **11**, (2020).
- 875 50. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses
876 of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
- 877 51. Ehret, G. B. *et al.* The genetics of blood pressure regulation and its target organs from
878 association studies in 342,415 individuals. *Nat. Genet.* **48**, 1171–1184 (2016).
- 879 52. Barbeira, A. N., Bonazzola, R., Gamazon, E. R. & Liang, Y. Exploiting the GTEx resources to
880 decipher the mechanisms at GWAS loci. (2020).
- 881 53. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic
882 association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
- 883 54. Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared
884 genetic risk factors across multiple traits. **44**, 1–47 (2019).
- 885
- 886