Consequences of natural perturbations in the human plasma proteome

Benjamin B. Sun[1]*, Joseph C. Maranville[2]*, James E. Peters[1,3]*, David Stacey[1], James R. Staley[1], James Blackshaw[1], Stephen Burgess[1,4], Tao Jiang[1], Ellie Paige[1], Praveen Surendran[1], Clare Oliver-Williams[1], Mihir A. Kamat[1], Bram P. Prins[1], Sheri K. Wilcox[5], Erik S. Zimmerman[5], An Chi[2], Narinder Bansal[1,6], Sarah L. Spain[7], Angela M. Wood[1], Nicholas W. Morrell[8], John R. Bradley[9], Nebojsa Janjic[5], David J. Roberts[10,11], Willem H. Ouwehand[3,12,13,14,15], John A. Todd[16], Nicole Soranzo[3,12,14,15], Karsten Suhre[17], Dirk S. Paul[1], Caroline S. Fox[2], Robert M. Plenge[2], John Danesh[1,3,14,15], Heiko Runz[2]*, Adam S. Butterworth[1,15]*

1. MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK.
2. MRL, Merck & Co., Inc., Kenilworth, New Jersey, USA.
3. British Heart Foundation Cambridge Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK.
4. MRC Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, UK.
5. SomaLogic Inc., Boulder, Colorado 80301, USA.
6. Perinatal Institute, Birmingham B15 3BU, UK.
7. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1RQ, UK.
8. Division of Respiratory Medicine, Department of Medicine, University of Cambridge, Cambridge CB2 0QQ, UK.
9. NIHR Cambridge Biomedical Research Centre / BioResource, Cambridge University Hospitals, Cambridge CB2 0QQ, UK.
10. National Health Service (NHS) Blood and Transplant and Radcliffe Department of Medicine, NIHR Oxford Biomedical Research Centre, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK.
11. BRC Haematology Theme and Department of Haematology, Churchill Hospital, Oxford OX3 7LE, UK.
12. Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0PT, UK.
13. National Health Service (NHS) Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK.
14. Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1RQ, UK.
15. NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK.
16. JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, NIHR Oxford Biomedical Research Centre, University of Oxford, Oxford OX3 7BN, UK.
17. Department of Physiology and Biophysics, Weill Cornell Medicine - Qatar, PO 24144 Doha, Qatar.

* These authors contributed equally to this work.

Corresponding authors: asb38@medschl.cam.ac.uk (A.S.B.), jd292@medschl.cam.ac.uk (J.D.)

Abstract

Proteins are the primary functional units of biology and the direct targets of most drugs, yet there is limited knowledge of the genetic factors determining inter-individual variation in protein levels. Here we reveal the genetic architecture of the human plasma proteome, testing 10.6 million DNA variants against levels of 2,994 proteins in 3,301 individuals. We identify 1,927 genetic associations with 1,478 proteins, a 4-fold increase on existing knowledge, including *trans* associations for 1,104 proteins. To understand consequences of perturbations in plasma protein levels, we introduce an approach that links naturally occurring genetic variation with biological, disease, and drug databases. We provide insights into pathogenesis by uncovering the molecular effects of disease-associated variants. We identify causal roles for protein biomarkers in disease through Mendelian randomization analysis. Our results reveal new drug targets, opportunities for matching existing drugs with new disease indications, and potential safety concerns for drugs under development.

Introduction

Plasma proteins play key roles in biological processes such as signalling, transport, growth, repair, and defence against infection. They are frequently dysregulated in disease and are the targets of many drugs. Detailed characterisation of the genetic factors that determine inter-individual protein variability will, therefore, furnish both fundamental and applied insights[1]. Despite evidence of the heritability of plasma protein abundance[2], systematic genome-wide study of such 'protein quantitative trait loci' (pQTLs) has been constrained by inability to measure large numbers of proteins reliably in large numbers of individuals[1,3–5].

Here we combine several technical and conceptual advances to create and interrogate a rich genetic atlas of the human plasma proteome. First, we use a markedly expanded version of an aptamer-based multiplex protein assay (SOMAscan)[6] to quantify 3,620 plasma proteins in 3,301 healthy individuals, representing several-fold increases both in breadth of protein panel and in cohort size. Second, we exploit improvements in genotype imputation panels to achieve 10-fold denser genotypic coverage than in previous proteomic studies[7]. Third, we draw on studies with multi-dimensional data (e.g., transcriptomics and clinical phenotypes) and on new bioinformatics tools to help understand mechanisms and clinical consequences of perturbations in protein pathways[8]. Fourth, we use two-sample "Mendelian randomization" techniques to evaluate the causal relevance of protein biomarkers to disease[9]. Finally, we cross-reference genomic-proteomic information with disease and drug databases to identify and prioritise therapeutic targets.

Our study characterises the genetic architecture of the human plasma proteome, identifying 1,927 genotype-protein associations, including *trans*-associated loci for 1,104 proteins that provide key insights into protein regulation. More than 150 pQTLs overlap with disease

3

86   susceptibility loci, elucidating the molecular effects of disease-associated variants. We find

87   strong evidence to support causal roles in disease for several protein pathways, highlighting

88   novel therapeutic targets as well as potential safety concerns for drugs in development.

89

90   Genetic architecture of the plasma proteome

91   After stringent quality control, we performed genome-wide testing of 10.6 million autosomal

92   variants against levels of 2,994 plasma proteins in 3,301 healthy European-ancestry

93   individuals (Methods, Extended Data Figure 1). Genotypes were measured using the

94   Affymetrix Axiom UK Biobank array and imputed against a combined 1000 Genomes and

95   UK10K reference panel. Protein levels were measured using the SOMAscan assay. We

96   evaluated the robustness of protein measurements in several ways (Methods, Supplementary

97   Note). Measurements in replicate samples were highly consistent; the median coefficient of

98   variation across all proteins was 0.064 (interquartile range 0.049-0.092). We also showed

99   temporal consistency in protein levels in samples obtained two years apart from the same

100   individuals, reproduced known associations with non-genetic factors, and verified selected

101   protein measurements with multiple different assay methods.

102

103   We found 1,927 genome-wide significant ($p<1.5\text{x}10^{-11}$) associations between 1,478 proteins

104   and 764 genomic regions (Figure 1a, Supplementary Table 1, Supplementary Video 1), with

105   89% of pQTLs reported here for the first time. Of these 764 regions, 502 (66%) had *cis*

106   associations only, 228 (30%) had *trans* associations only and 34 (4%) had both *cis* and *trans*.

107   95% and 87% of our *cis* pQTL variants were located within 200Kb and 100Kb, respectively,

108   of the relevant gene's transcription start site (TSS) (Figure 1b), and 44% were within the

109   gene itself. The *p*-values for *cis* pQTL associations increased with increasing distance from

110   the TSS, mirroring findings for *cis* expression QTLs (eQTLs) in transcriptomic studies[10,11].

4

111 Of the proteins for which we identified a pQTL, 88% had either *cis* (n=374) or *trans* (n=925)

112 associations only, while 12% (n=179) had both. The majority of significantly associated

113 proteins (75%; n=1,113) had a single pQTL, while 20% had two and 5% had more than two

114 (Figure 1c). To detect multiple signals at the same locus we used stepwise conditional

115 analysis, identifying 2,658 conditionally significant associations (Supplementary Table 2). Of

116 the 1,927 locus-protein associations, 414 (21%) had multiple conditionally significant signals

117 (Figure 1d), of which 255 were *cis*.

118

119 Genetic variants that change the protein structure may result in apparent pQTLs due to

120 altered aptamer binding rather than true quantitative differences in protein levels. Using

121 bioinformatic approaches to evaluate the extent to which *cis* pQTLs might reflect technical

122 effects (Methods) we found that 91% of pQTLs were unlikely to be influenced by differential

123 aptamer binding (Supplementary Tables 1 and 3). Furthermore, in sub-studies that involved

124 immunoassays, we found strongly concordant genotype-protein associations (Supplementary

125 Note). Even where apparent pQTLs do arise from alternative protein structure that may affect

126 aptamer binding rather than differences in protein abundance, such signals may be

127 functionally significant.

128

129 The median variation in protein levels explained by our pQTLs was 5.8% (in-sample

130 estimate; interquartile range: 2.6%-12.4%, Figure 1e). For 193 proteins, however, genetic

131 variants explained more than 20% of the variation, such as for teratocarcinoma-derived

132 growth factor 1 (65%) and haptoglobin (55%). We found a strong inverse relationship

133 between effect size and minor allele frequency (MAF) (Figure 1f), consistent with previous

134 genome-wide association studies (GWAS) of quantitative traits[12–14]. We found 25

135 associations with rare (MAF <1%) variants and a further 207 associations with low-

136 frequency (MAF 1-5%) variants. Of the 31 strongest pQTLs (per-allele effect size >1.5

137 standard deviations), 25 were rare or low-frequency variants.

138

139 pQTLs were strongly enriched for missense variants ($p<0.0001$) and for location in 3'

140 untranslated ($p=0.0025$) or splice regions ($p=0.0004$) (Figure 1g, Extended Data Figure 2).

141 To assess whether pQTLs were enriched within regulatory elements from a wide range of cell

142 types and tissues[15–17], we used GARFIELD[18] (Methods). We found strong ($\geq$ 3-fold,

143 $p<5\times10^{-5}$) enrichment of pQTLs in blood cells – unsurprisingly given our use of plasma - at

144 features indicative of transcriptional activation (Extended Data Figure 3, Supplementary

145 Table 4). We also found enrichment at hepatocyte regulatory elements, consistent with the

146 liver's role in producing many secreted proteins.

147

Overlap of loci for gene expression and protein levels

149 A fundamental biological question is the extent to which genetic associations with plasma

150 protein levels are driven by effects at the transcription level, rather than other mechanisms,

151 such as altered protein clearance or cleavage of membrane receptor proteins from the cell

152 surface. We therefore cross-referenced our *cis* pQTLs with a wide range of publicly available

153 eQTL studies involving >30 tissues or cell types (Supplementary Table 5) using

154 PhenoScanner[8]. 40% (n=224) of *cis* pQTLs had an association with expression of the same

155 gene in at least one tissue (Supplementary Table 6), consistent with similar comparisons

156 within lymphoblastoid cell lines (LCLs)[19]. The greatest overlaps were found in whole blood

157 (n=117), liver (n=70) and LCLs (n=52), consistent with biological expectation, but also likely

158 driven by the larger sample sizes for expression studies of these cell types. *Cis* pQTLs were

159 significantly enriched ($p<0.0001$) for eQTLs for the corresponding gene (mean 20% versus

160 2% in a background permuted set; Methods, Supplementary Table 7). To address the

6

161 converse (i.e., to what extent do eQTLs translate into pQTLs), we used a subset of well-

162 powered eQTL studies in relevant tissues (whole blood, LCLs, liver and monocytes[20–23]). Of

163 the strongest *cis* eQTLs ($p < 1.5 \times 10^{-11}$), 12.2% of those in whole blood were also *cis* pQTLs,

164 21.3% for LCLs, 14.8% for liver and 14.7% for monocytes.

165

166 Comparisons between eQTL and pQTL studies have inherent limitations due to differences in

167 the tissues, sample sizes and technological platforms used. Moreover, plasma protein levels

168 may not reflect levels within tissues or cells. Nevertheless, our data suggest that plasma

169 protein abundance is often, but not exclusively, driven by regulation of mRNA. Our finding

170 that pQTLs are enriched in gene regulatory regions supports the notion that transcription has

171 a major role. *Cis* pQTLs without corresponding *cis* eQTLs may be underpinned by effects not

172 captured at the mRNA level, such as differences in protein stability, degradation, binding,

173 secretion, or clearance from circulation.

174

175 ## Using *trans* pQTLs to illuminate biological pathways

176 *Trans* pQTLs are particularly useful for understanding biological relationships between

177 proteins if a causal gene at the *trans*-associated genetic locus can be identified. To this end,

178 we used a combination of databases of molecular pathways and protein-protein interaction

179 networks, and functional genomic data that include variant annotation, eQTL and

180 chromosome conformation capture, to link *trans*-associated variants to potential causal genes

181 (Methods, Supplementary Table 8, Extended Data Figure 4). Of the 764 protein-associated

182 regions, 262 had *trans* associations with 1,104 proteins. We replicated previously reported

183 *trans* associations including *TMPRSS6* with transferrin receptor protein 1[24] and *SORT1* with

184 granulins[25]. Most (82%) *trans* loci were associated with fewer than four proteins. However,

185 we observed twelve regions with more than 20 associated proteins (Figure 1a, Extended Data

186   Figure 5), including well-known pleiotropic loci (e.g., *ABO, CFH, APOE, KLKB1*) and loci

187   associated with many correlated proteins (e.g., the locus containing the transcription factor

188   gene, *ZFPM2*).

189

190   We identified a number of novel associations with strong biological plausibility

191   (Supplementary Table 9, Supplementary Note). Growth differentiation factor 8 (GDF8, more

192   commonly known as myostatin) provides one such example. Because insufficient myostatin

193   has been shown to result in excessive muscle growth[26], myostatin inhibition has emerged as a

194   promising therapeutic strategy for treatment of conditions characterised by muscle weakness,

195   such as muscular dystrophy[27]. We identified a common allele (rs11079936:C) near

196   *WFIKKN2* that is associated in *trans* with lower levels of plasma GDF11/8 ($p=7.9 \times 10^{-12}$), as

197   well as in *cis* with lower levels of plasma WFIKKN2 ($p=6.9 \times 10^{-136}$, Supplementary Table 1,

198   Figure 2). The *trans* association attenuated completely upon adjustment for levels of

199   WFIKKN2 ($p=0.7$), while the *cis* association remained significant after adjustment for

200   GDF11/8 ($p=7.2 \times 10^{-113}$), suggesting that WFIKKN2 regulates GDF11/8. This observation is

201   supported by *in vitro* evidence suggesting that WFIKKN2 has high affinity for GDF8 and

202   GDF11[28]. This finding illustrates how delineation of causal genes at *trans*-associated loci

203   may help identify additional drug targets in pathways known to be relevant to disease.

204

Identifying causal pathways underlying disease susceptibility loci

206   GWAS have identified thousands of loci associated with common diseases, but the pathways

207   by which most variants influence disease susceptibility await discovery. To identify

208   intermediate links between genotype and disease, we overlapped pQTLs with disease-

209   associated genetic variants identified through GWAS ($p<5 \times 10^{-8}$). 152 of our pQTLs were

210   strongly correlated ($r^2 \geq 0.8$) with variants significantly associated with disease

211   (Supplementary Table 10), including 38 with *cis* associations, 109 with *trans* associations

212   and 5 with both. In the examples below, we illustrate how our findings provide novel insights

213   spanning a wide range of disease domains including autoimmunity, cancer and

214   cardiovascular disease.

215

216   Our *trans* pQTL data implicate previously unsuspected proteins as mediators through which

217   genetic loci exert their influence on disease risk. For example, GWAS have identified a

218   missense allele (rs3197999:A, p.Arg703Cys) in *MST1* on chromosome 3 that increases risk

219   of inflammatory bowel disease (IBD) (Figure 3)[29,30], and decreases plasma MST1 levels[31].

220   We show that this polymorphism acts in *trans* to reduce abundance of BLIMP1, encoded by

221   *PRDM1* on chromosome 6 (Figure 3, Supplementary Table 1). BLIMP1 is a transcriptional

222   master regulator that plays a critical role in the terminal differentiation of immune cells.

223   Intriguingly, there is another IBD association signal in the intergenic region adjacent to

224   *PRDM1* on chromosome 6 (Figure 3)[30]. Both *PRDM1* and its neighbour *ATG5*, which

225   encodes a protein involved in autophagy, are plausible candidate genes at this locus. Our data

226   implicate BLIMP1 as a previously unidentified mediator of the IBD association in *MST1* on

227   chromosome 3. In addition, our results provide indirect support for the hypothesis that

228   *PRDM1* is the causal gene underlying the chromosome 6 association.

229

230   We next show how pQTL data can elucidate pathogenic mechanisms. Anti-neutrophil

231   cytoplasmic antibody-associated vasculitis (AAV) is an autoimmune disease characterised by

232   autoantibodies to the neutrophil proteases proteinase-3 (PR3) or myeloperoxidase (MPO).

233   Clinico-pathological features mirror antibody specificity, with granulomatous inflammation

234   typically correlating with anti-PR3 antibodies (PR3+ AAV). GWAS have identified signals

235   in *PRTN3* (encoding PR3) and *SERPINA1* (encoding alpha1-antitrypsin, an inhibitor of PR3)

236   specific to PR3+ AAV[32]. We identified a *cis* pQTL immediately upstream of *PRTN3*

237   (Supplementary Table 1). By linking the risk allele at *PRTN3* to higher plasma levels of the

238   autoantigen PR3, our data strongly suggest a pathogenic role of anti-PR3 antibodies in this

239   disease.

240

241   The vasculitis risk allele at the *SERPINA1* locus (rs28929474:T, also known as the 'Z' allele)

242   is a missense variant (p.Glu366Lys) which results in defective secretion of alpha1-

243   antitrypsin. We found that the Z allele was not only associated with lower plasma alpha1-

244   antitrypsin, but was also a pQTL "hotspot" associated with 13 proteins using our

245   conservative significance threshold ($p<1.5 \times 10^{-11}$) (Figure 4) and 19 at $p<5 \times 10^{-8}$. This finding

246   illustrates how a single mutation can lead to widespread perturbation of downstream proteins.

247   One of these proteins was NCF2 (neutrophil cytosolic factor 2), which plays a key role in the

248   neutrophil oxidative burst. Mutations in *NCF2* can result in a rare condition known as chronic

249   granulomatous disease, which like PR3+ AAV exhibits granulomatous inflammation. Our

250   results suggest NCF2 may mediate this characteristic feature of PR3+ AAV.

251

252   Causal evaluation of candidate proteins in disease

253   Association of plasma protein levels with disease does not necessarily imply causation. To

254   help establish causality, we employed the principle of Mendelian randomization (MR)[33]. In

255   contrast with observational studies, which are liable to confounding and/or reverse causation,

256   MR analysis is akin to a "natural" randomised controlled trial, exploiting the random

257   allocation of alleles at conception (Extended Data Figure 6). Consequently, if a genetic

258   variant that specifically influences levels of a protein is also associated with disease risk, then

259   it provides strong evidence of the protein's causal role. For example, serum levels of PSP-94

260   (MSMB), a hormone synthesized by the prostate gland and secreted into seminal fluid, are

10

261 lower in patients with prostate cancer; PSP-94 is therefore a candidate biomarker for prostate

262 cancer[34]. However, it is debated whether PSP-94 plays a causal role in tumorigenesis. We

263 identified a *cis* pQTL for PSP-94 at a prostate cancer susceptibility locus[35] and showed that

264 the risk allele is associated with lower PSP-94 plasma levels, supporting a protective role for

265 the protein.

266

267 In an approach that extends classic MR analysis, we leveraged multi-variant MR analysis

268 methods to distinguish causal genes among multiple plausible candidates at disease loci

269 (Methods), exemplified by the *IL1RL1-IL18R1* locus. We identify four proteins that each had

270 *cis* pQTLs at this locus (Supplementary Table 1), which has been associated with a range of

271 immune-mediated diseases including atopic dermatitis[36]. We created a genetic score for each

272 protein using multiple protein-increasing alleles. Initial "one-protein-at-a-time" analysis

273 identified associations of the scores for IL18R1 (p=9.3x10^{-72}) and IL1RL1 (p=5.7x10^{-27})

274 with atopic dermatitis risk (Figure 5a). In addition, we found a weak association for IL1RL2

275 (p=0.013). We then mutually adjusted these associations for one another to account for the

276 effects of the variants on multiple proteins. While the association of IL18R1 remained

277 significant (p=1.5x10^{-28}), the association of IL1RL1 (p=0.01) was attenuated. In contrast, the

278 association of IL1RL2 (p=1.1x10^{-69}) became much stronger, suggesting that IL1RL2 and

279 IL18R1 are the causal proteins influencing risk of atopic dermatitis at this locus.

280

281 We also used multiple *cis*-associated variants to evaluate whether macrophage

282 metalloelastase (MMP-12) plays a causal role in coronary disease. Since MMP-12 is known

283 to play a role in lung damage, MMP-12 inhibitors are being tested in chronic obstructive

284 pulmonary disease (COPD). Observational studies report associations of higher levels of

285 plasma MMP-12 with recurrent cardiovascular events[37,38]. In contrast, our multi-allelic

286   genetic score, which explains 14% of the variation in plasma MMP-12 levels (Methods),

287   indicates that genetic predisposition to higher MMP-12 levels is associated with *decreased*

288   coronary disease risk (p=2.8x10^{-13}) (Figure 5b) and *decreased* large artery atherosclerotic

289   stroke risk[39]. As MMP-12 is released from macrophages in response to cardiovascular

290   injury[40], it is possible that higher MMP-12 levels are cardioprotective. Hence, our results

291   identify a potential cardiovascular safety concern for MMP-12 inhibitors, particularly as

292   patients with COPD are at high baseline cardiovascular risk due to smoking history.

293

Drug target prioritisation

295   Because therapeutic targets implicated by human genetic data are likely to play causal roles

296   in disease, drugs directed at such targets have greater likelihood of success[41]. According to

297   the Informa Pharmaprojects database (Citeline), 244 of the proteins linked to a disease-

298   associated variant by our pQTL data are drug targets (Supplementary Table 11). Of the

299   proteins we identified as associated with disease susceptibility loci, 49 are targets of already

300   approved drugs such as tocilizumab and ustekinumab (Supplementary Table 12).

301

302   To identify additional indications for existing drugs, we investigated disease associations of

303   pQTLs for proteins already targeted by licensed drugs. Our results suggest potential drug "re-

304   purposing" opportunities (Supplementary Table 12). For example, we identified a *cis* pQTL

305   for RANK (encoded by *TNFRSF11A*) at a locus associated with Paget's disease[42], a condition

306   characterised by excessive bone turnover leading to deformity and fracture. Standard

307   treatment consists of osteoclast inhibition with bisphosphonates, originally developed as anti-

308   osteoporotic drugs. Denosumab, another anti-osteoporosis drug, is a monoclonal antibody

309   targeting RANKL, the ligand for RANK. Our data suggest denosumab may be a useful

310   alternative for patients with Paget's disease in whom bisphosphonates are contra-indicated, a

311   hypothesis supported by clinical case reports[43,44].

312

313   To evaluate targets for drugs currently under development, we considered the example of

314   glycoprotein Ib platelet alpha subunit (GP1BA), the receptor for von Willebrand factor.

315   Drugs directed at GP1BA are in pre-clinical development as anti-thrombotic agents and in

316   phase 2 trials to treat thrombotic thrombocytopenic purpura, a life-threatening disorder. We

317   identify a *trans* pQTL for GP1BA at the pleiotropic *SH2B3/BRAP* locus, which is associated

318   with platelet count[45], myocardial infarction and stroke. The risk allele for cardiovascular

319   disease increases both plasma GP1BA and platelet count, suggesting a mechanism by which

320   this locus affects disease susceptibility. As a confirmation of the link between GP1BA and

321   the platelet count, we found a directionally concordant *cis* pQTL for GP1BA at a platelet

322   count-associated variant (Supplementary Table 12). Collectively, these results suggest that

323   targeting GP1BA may be efficacious in conditions characterised by platelet aggregation such

324   as arterial thrombosis. More generally, our data provide a foundation for generating

325   hypotheses about targets for new drug development through the approach of linking genetic

326   factors to disease via specific proteins (Supplementary Table 12).

327

Discussion

329   We conducted protein measurements of unprecedented scope and scale to reveal genetic

330   control of the human plasma proteome. Our discoveries enabled identification of important

331   consequences of natural perturbations in the plasma proteome. First, we identified the

332   downstream effects of alterations in specific protein levels, revealing novel regulators of

333   protein pathways. Second, we used plasma proteins to uncover intermediate molecular

334   pathways that connect the genome to disease endpoints. Previous investigation has focused

13

335 on genes in the vicinity of disease susceptibility loci. By contrast, we made key advances

336 through use of *trans* pQTLs to implicate previously unsuspected proteins encoded by distant

337 genes. Third, we established causal roles for protein biomarkers in vascular, neoplastic, and

338 autoimmune diseases using the principle of Mendelian randomization (MR). Proteins provide

339 an ideal paradigm for MR analysis because they are under proximal genetic control. Whereas

340 MR studies of plasma proteins have been constrained by availability of few suitable genetic

341 instruments, our data remedy this bottleneck by furnishing an extensive toolkit. Fourth, we

342 introduced an approach that should help reduce the unsustainably high attrition rates of drugs

343 in pharmaceutical pipelines. Overall, our study foreshadows major advances in post-genomic

344 science through increasing application of novel bioassay technologies to major population

345 biobanks.

Methods

Study participants

The INTERVAL study comprises approximately 50,000 participants nested within a randomised trial of blood donation intervals[46]. Between mid-2012 and mid-2014, whole-blood donors aged 18 years and older were consented and recruited at 25 centers of England's National Health Service Blood and Transplant. All participants completed an online questionnaire including questions about demographic characteristics (e.g., age, sex, ethnic group), anthropometry (height, weight), lifestyle (e.g., alcohol and tobacco consumption) and diet. Participants were generally in good health because blood donation criteria exclude people with a history of major disease (such as myocardial infarction, stroke, cancer, HIV, and hepatitis B or C) and those who have had recent illness or infection. For protein assays, we randomly selected two non-overlapping subcohorts of 2,731 and 831 participants from INTERVAL. Participant characteristics are shown in Supplementary Table 13.

Plasma sample preparation

Sample collection procedures for INTERVAL have been described previously[46]. In brief, blood samples for research purposes were collected in 6ml EDTA tubes using standard venepuncture protocols. The tubes were inverted three times and transferred at room temperature to UK Biocentre (Stockport, UK) for processing. Plasma was extracted into two 0.8ml plasma aliquots by centrifugation and subsequently stored at -80°C prior to use.

Protein measurements

369   We used a multiplexed, aptamer-based approach (SOMAscan assay) to measure the relative

370   concentrations of 3,620 plasma proteins/protein complexes assayed using 4,034 aptamers

371   ("SOMAmer reagents", hereafter referred to as "SOMAmers"). The assay extends the lower

372   limit of detectable protein abundance afforded by conventional approaches (e.g.,

373   immunoassays), measuring both extracellular and intracellular proteins (including soluble

374   domains of membrane-associated proteins), with a bias towards proteins likely to be found in

375   the human secretome (Extended Data Figure 7a)[6,47]. The proteins cover a wide range of

376   molecular functions (Extended Data Figure 7b). The selection of proteins on the platform

377   reflects both the availability of purified protein targets and a focus on proteins known to be

378   involved in pathophysiology of human disease.

379

380   Aliquots of 150µl of plasma were sent on dry ice to SomaLogic Inc. (Boulder, Colorado, US)

381   for protein measurement. Assay details have been previously described[47–49] and a technical

382   white paper with further information can be found at the manufacturer's website

383   (http://info.somalogic.com/lp_campaign_somascan-white-paper_jan-2016). In brief, modified

384   single-stranded DNA SOMAmers are used to bind to specific protein targets that are then

385   quantified using a DNA microarray. Protein concentrations are quantified as relative

386   fluorescent units.

387

388   Quality control (QC) was performed at the sample and SOMAmer level using control

389   aptamers, as well as calibrator samples. At the sample level, hybridization controls on the

390   microarray were used to correct for systematic variability in hybridization, while the median

391   signal over all features assigned to one of three dilution sets (40%, 1% and 0.005%) was used

392   to correct for within-run technical variability. The resulting hybridization scale factors and

393   median scale factors were used to normalise data across samples within a run. The acceptance

394   criteria for these values are between 0.4 and 2.5 based on historical runs. SOMAmer-level

395   QC made use of replicate calibrator samples using the same study matrix (plasma) to correct

396   for between-run variability. The acceptance criteria for each SOMAmer is that the calibration

397   scale factor be less than 0.4 from the median for each of the plates run. In addition, at the

398   plate level, the acceptance criteria are that the median of the calibration scale factors is

399   between 0.8 and 1.2, while 95% of individual SOMAmers must be less than 0.4 from the

400   median within the plate.

401

402   In addition to QC processes routinely conducted by SomaLogic, we measured protein levels

403   of 30 and 10 pooled plasma samples randomly distributed across plates for subcohort 1 and

404   subcohort 2, respectively. This approach, which involved masking laboratory technicians to

405   the presence of pooled samples, enabled estimation of the reproducibility of the protein

406   assays in directly relevant samples. We calculated coefficients of variation (CVs) for each

407   SOMAmer within each subcohort by dividing the standard deviation by the mean of the

408   pooled plasma sample protein readouts. In addition to passing SomaLogic QC processes, we

409   required SOMAmers to have a $CV \leqslant 20\%$ in both subcohorts. Eight non-human protein

410   targets were also excluded, leaving 3,283 SOMAmers (mapping to 2,994 unique

411   proteins/protein complexes) for inclusion in the GWAS.

412

413   Protein mapping to UniProt identifiers and gene names was provided by SomaLogic.

414   Mapping to Ensembl gene IDs and genomic positions was performed using Ensembl Variant

415   Effects Predictor v83 (VEP)[50]. Protein subcellular locations were determined by exporting

416   the subcellular location annotations from UniProt[51]. If the term "membrane" was included in

417   the descriptor, the protein was considered to be a membrane protein, whereas if the term

418   "secreted" (but not "membrane") was included in the descriptor, the protein was considered

419   to be a secreted protein. Proteins not annotated as either membrane or secreted proteins were

420   classified (by inference) as intracellular proteins. Proteins were mapped to molecular

421   functions using gene ontology annotations[52] from UniProt.

422

Non-genetic associations of proteins

424   To validate the protein assays, we attempted to replicate the associations with age or sex of

425   45 proteins previously reported by Ngo *et al* and 40 reported by Menni *et al*[48,53]. We used

426   Bonferroni-corrected *p*-value thresholds of $p=1.1\times10^{-3}$ (0.05/45) and $p=1.2\times10^{-3}$ (0.05/40)

427   respectively. Relative protein abundances were rank-inverse normalised within each

428   subcohort and linear regression was performed using age, sex, BMI, natural log of estimated

429   glomerular filtration rate (eGFR) and subcohort as independent variables.

430

Genotyping and imputation

432   The genotyping protocol and QC for the INTERVAL samples (n~50,000) have been

433   described previously in detail[14]. Briefly, DNA extracted from buffy coat was used to assay

434   approximately 830,000 variants on the Affymetrix Axiom UK Biobank genotyping array at

435   Affymetrix (Santa Clara, California, US). Genotyping was performed in multiple batches of

436   approximately 4,800 samples each. Sample QC was performed including exclusions for sex

437   mismatches, low call rates, duplicate samples, extreme heterozygosity and non-European

438   descent. An additional exclusion made for this study was of one participant from each pair of

439   close (first- or second-degree) relatives, defined as $\hat{\pi} > 0.187$. Identity-by-descent was

440   estimated using a subset of variants with call rate >99% and MAF >5% in the merged dataset

441   of both subcohorts, pruned for linkage disequilibrium (LD) using PLINK v1.9[54]. Numbers of

442   participants excluded at each stage of the genetic QC are summarized in Extended Data

443   Figure 1. Multi-dimensional scaling was performed using PLINK v1.9 to create components

444   to account for ancestry in genetic analyses.

445

446   Prior to imputation, additional variant filtering steps were performed to establish a high

447   quality imputation scaffold. In summary, 654,966 high quality variants (autosomal, non-

448   monomorphic, bi-allelic variants with Hardy Weinberg Equilibrium (HWE) $p{>}5x10^{-6}$, with a

449   call rate of >99% across the INTERVAL genotyping batches in which a variant passed QC,

450   and a global call rate of >75% across all INTERVAL genotyping batches) were used for

451   imputation. Variants were phased using SHAPEIT3 and imputed using a combined 1000

452   Genomes Phase 3-UK10K reference panel. Imputation was performed via the Sanger

453   Imputation Server (https://imputation.sanger.ac.uk) resulting in 87,696,888 imputed variants.

454

455   Prior to genetic association testing, variants were filtered in each subcohort separately using

456   the following exclusion criteria: (1) imputation quality (INFO) score<0.7, (2) minor allele

457   count<8, (3) HWE $p{<}5x10^{-6}$. In the small number of cases where imputed variants had the

458   same genomic position (GRCh37) and alleles, the variant with the lowest INFO score was

459   removed. 10,572,788 variants passing all filters in both subcohorts were taken forward for

460   analysis (Extended Data Figure 1).

461

462   **Genome-wide association study**

463   Within each subcohort, relative protein abundances were first natural log-transformed. Log

464   transformed protein levels were then adjusted in a linear regression for age, sex, duration

465   between blood draw and processing (binary, ≤1 day/>1day) and the first three principal

466   components of ancestry from multi-dimensional scaling. The protein residuals from this

467   linear regression were then rank-inverse normalized and used as phenotypes for association

468   testing. Univariate linear regression using an additive genetic model was used to test genetic

469   associations. Association tests were carried out on allelic dosages to account for imputation

470   uncertainty ("-method expected" option) using SNPTEST v2.5.2[55].

471

Meta-analysis and statistical significance

473   Association results from the two subcohorts were combined via fixed-effects inverse-

474   variance meta-analysis combining the betas and standard errors using METAL[56]. Genetic

475   associations were considered to be genome-wide significant based on a conservative strategy

476   requiring associations to have (i) a meta-analysis p-value$<1.5\times10^{-11}$ (genome-wide threshold

477   of $p=5\times10^{-8}$ Bonferroni corrected for 3,283 aptamers tested), (ii) at least nominal significance

478   ($p<0.05$) in both subcohorts, and also (iii) consistent direction of effect across subcohorts.

479   We did not observe significant genomic inflation (mean inflation factor was 1.0, standard

480   deviation=0.01) (Extended Data Figure 8).

481

Refinement of significant regions

483   To identify distinct non-overlapping regions associated with a given SOMAmer, we first

484   defined a 1Mb region around each significant variant for that SOMAmer. Starting with the

485   region containing the variant with the smallest p-value, any overlapping regions were then

486   merged and this process was repeated until no more overlapping 1Mb regions remained. The

487   variant with the lowest p-value for each region was assigned as the "regional sentinel

488   variant". Due to the complexity of the Major Histocompatibility Region (MHC) region, we

489   treated the extended MHC region (chr6:25.5-34.0Mb) as one region. To identify whether a

490   region was associated with multiple SOMAmers, we used an LD-based clumping approach.

491   Regional sentinel variants in high LD ($r^2\geq0.8$) with each other were combined together into a

492   single region.

493

Conditional analyses

To identify conditionally significant signals, we performed approximate genome-wide step-wise conditional analysis using GCTA v1.25.2[57] using the "cojo-slct" option. We used the same conservative significance threshold of $p=1.5\times10^{-11}$ as for the univariate analysis. As inputs for GCTA, we used the summary statistics (i.e. betas and standard errors) from the meta-analysis. Correlation between variants was estimated using the hard-called genotypes (where a genotype was called if it had a posterior probability of >0.9 following imputation or set to missing otherwise) in the merged genetic dataset, and only variants also passing the univariate genome-wide threshold ($p<1.5\times10^{-11}$) were considered for step-wise selection. As the conditional analyses use different data inputs (summarised rather than individual-level data), there were some cases where the conditional analysis failed to include sentinel variants that had borderline significant univariate associations in the step-wise selection. In these instances (n=28), we re-conducted the joint model estimation without step-wise selection in GCTA with variants identified by the conditional analysis in addition to the regional sentinel variant. We report and highlight these cases in Supplementary Table 2.

Replication of previous pQTLs

We attempted to identify all previously reported pQTLs from GWAS and to assess whether they replicated in our study. We used the NCBI Entrez programming utility in R (rentrez) to perform a literature search for pQTL studies published from 2008 onwards. We searched for the following terms: "pQTL", "pQTLs", and "protein quantitative trait locus". We supplemented this search by filtering out GWAS associations from the NHGRI-EBI GWAS Catalog v.1.0.1[58] (https://www.ebi.ac.uk/gwas/, downloaded April 2016), which has all phenotypes mapped to the Experimental Factor Ontology (EFO)[59], by restricting to those with EFO annotations relevant to protein biomarkers (e.g. "protein measurement",

21

519   EFO_0004747). Studies identified through both approaches were manually filtered to include

520   only studies that profiled plasma or serum samples and to exclude studies not assessing

521   proteins. We recorded basic summary information for each study including the assay used,

522   sample size and number of proteins with pQTLs (Supplementary Table 14). To reduce the

523   impact of ethnic differences in allele frequencies on replication rate estimates, we filtered

524   studies to include only associations reported in European-ancestry populations. We then

525   manually extracted summary data on all reported associations from the manuscript or the

526   supplementary material. This included rsID, protein UniProt ID, *p*-values, and whether the

527   association is *cis/trans* (Supplementary Table 15).

528

529   To assess replication we first identified the set of unique UniProt IDs that were also assayed

530   on the SOMAscan panel. For previous studies that used SomaLogic technology, we refined

531   this match to the specific aptamer used. We then clumped associations into distinct loci using

532   the same method applied to our pQTL (see **Refinement of significant regions**). For each

533   locus, we asked if the sentinel SNP or a proxy (r^2>0.6) was associated with the same

534   protein/aptamer in our study at a defined significance threshold. For our primary assessment,

535   we used a *p*-value threshold of 10^{-4}. We also performed sensitivity analyses to explore factors

536   that influence replication rate (Supplementary Note).

537

538   **Candidate gene annotation**

539   We defined a pQTL as *cis* when the most significantly associated variant in the region was

540   located within 1Mb of the transcription start site (TSS) of the gene(s) encoding the protein.

541   pQTLs lying outside of the region were defined as *trans*. When considering the distance of

542   the top *cis*-associated variant from the relevant TSS, only proteins that map to single genes

543   on the primary assembly in Ensembl v83 were considered.

544

545   For *trans* pQTLs, we sought to prioritise candidate genes in the region that might underpin

546   the genotype-protein association. In addition to reporting the nearest gene to the sentinel

547   variant, we employed "bottom up" and "top down" approaches, starting from the variant and

548   protein respectively For the "bottom up" approach, the sentinel variant and corresponding

549   proxies (r^2>0.8) for each *trans* pQTL were first annotated using Ensembl VEP v83 (using the

550   "pick" option) to determine whether variants were (1) protein-altering coding variants; (2)

551   synonymous coding or 5'/3' untranslated region (UTR); (3) intronic or up/downstream; or (4)

552   intergenic. Second, we queried all sentinel variants and proxies against significant *cis* eQTL

553   variants (defined by beta distribution-adjusted empirical *p*-values using a FDR threshold of

554   0.05, see http://www.gtexportal.org/home/documentationPage for details) in any cell type or

555   tissue   from   the   Genotype-Tissue   Expression   (GTEx)   project   v6[60]

556   (http://www.gtexportal.org/home/datasets). Third, we also queried promoter capture Hi-C

557   data in 17 human primary hematopoietic cell types[61] to identify contacts (with a CHICAGO

558   score >5 in at least one cell type) involving chromosomal regions containing a sentinel

559   variant. We considered gene promoters annotated on either fragment (i.e., the fragment

560   containing the sentinel variant or the other corresponding fragment) as potential candidate

561   genes. Using these three sources of information, we generated a list of candidate genes for the

562   *trans* pQTLs. A gene was considered a candidate if it fulfilled at least one of the following

563   criteria: (1) it was proximal (intragenic or ±5Kb from the gene) or nearest to the sentinel

564   variant; (2) it contained a sentinel or proxy variant (r^2>0.8) that was protein-altering; (3) it

565   had a significant *cis* eQTL in at least one GTEx tissue overlapping with a sentinel pQTL

566   variant (or proxy); or (4) it was regulated by a promoter annotated on either fragment of a

567   chromosomal contact[61] involving a sentinel variant.

568

569 For the "top down" approach, we first identified all genes with a TSS located within the

570 corresponding pQTL region using the GenomicRanges Bioconductor package[62] with

571 annotation from a GRCh37 GTF file from Ensembl

572 (ftp://ftp.ensembl.org/pub/grch37/update/gtf/homo_sapiens/; file:

573 "Homo_sapiens.GRCh37.82.gtf.gz", downloaded June 2016). We then identified any local

574 genes that had previously been linked with the corresponding *trans*-associated protein(s)

575 according to the following open source databases: (1) the Online Mendelian Inheritance in

576 Man (OMIM) catalogue[63] (http://www.omim.org/); (2) the Kyoto Encyclopedia of Genes and

577 Genomes (KEGG)[64] (http://www.genome.jp/kegg/); and (3) STRINGdb[65] (http://string-

578 db.org/; v10.0). We accessed OMIM data via HumanMine web tool[66]

579 (http://www.humanmine.org/; accessed June 2016), whereby we extracted all OMIM IDs for

580 (i) our *trans*-affected proteins and (ii) genes local (±500Kb) to the corresponding *trans*-acting

581 variant. We extracted all human KEGG pathway IDs using the KEGGREST Bioconductor

582 package[67] (https://bioconductor.org/packages/release/bioc/html/KEGGREST.html). In cases

583 where a *trans*-associated protein shared either an OMIM ID or a KEGG pathway ID with a

584 gene local to the corresponding *trans*-acting variant, we took this as evidence of a potential

585 functional involvement of that gene. We interrogated protein-protein interaction data by

586 accessing STRINGdb data using the STRINGdb Bioconductor package[68], whereby we

587 extracted all pairwise interaction scores for each *trans*-affected protein and all proteins with

588 genes local to the corresponding *trans*-acting variants. We took the default interaction score

589 of 400 as evidence of an interaction between the proteins, therefore indicating a possible

590 functional involvement for the local gene. In addition to using data from open source

591 databases in our top-down approach we also adopted a "guilt-by-association" (GbA)

592 approach utilising the same plasma proteomic data used to identify our pQTLs. We first

593 generated a matrix containing all possible pairwise Pearson's correlation coefficients between

24

594   our 3,283 SOMAmers. We then extracted the coefficients relating to our *trans*-associated

595   proteins and any proteins encoded by genes local to their corresponding *trans*-acting variants

596   (where available). Where the correlation coefficient was ≥0.5 we prioritised the relevant

597   local genes as being potential mediators of the *trans*-signal(s) at that locus.

598

599   We report the potential candidate genes for our *trans* pQTLs from both the "bottom up" and

600   "top down" approaches, highlighting cases where the same gene was highlighted by both

601   approaches.

602

603   **Functional annotation of pQTLs**

604   Functional annotation of variants was performed using Ensembl VEP v83 using the "pick"

605   option. We tested the enrichment of significant pQTL variants for certain functional classes

606   by comparing to permuted sets of variants showing no significant association with any

607   protein (*p*>0.0001 for all proteins tested). First the regional sentinel variants were LD-pruned

608   at r^2 of 0.1. Each time the sentinel variants were LD-pruned, one of the pair of correlated

609   variants was removed at random and for each set of LD-pruned sentinel variants and 100 sets

610   of equally sized null permuted variants were sampled matching for MAF (bins of 5%),

611   distance to TSS (bins of 0-0.5Kb, 0.5Kb-2Kb, 2Kb-5Kb, 10Kb-20Kb, 20Kb-100Kb and

612   >100Kb in each direction) and LD (± half the number of variants in LD with the sentinel

613   variant at r^2 of 0.8). This procedure was repeated 100 times resulting in 10,000 permuted sets

614   of variants. An empirical *p*-value was calculated as the proportion of permuted variant sets

615   where the proportion that are classified as a particular functional group exceeded that of the

616   test set of sentinel pQTL variants, and we used a significance threshold of *p*=0.005 (0.05/10

617   functional classes tested).

618

25

Evidence against aptamer-binding effects at *cis* pQTLs

All protein assays that rely on binding (e.g. of antibodies or SOMAmers) are susceptible to the possibility of binding-affinity effects, where protein-altering variants (PAVs) (or their proxies in LD) are associated with protein measurements due to differential binding rather than differences in protein abundance. To account for this potential effect, we performed conditional analysis at all *cis* pQTLs where the sentinel variant was in LD ($0.1 \leq r^2 \leq 0.9$) with a PAV in the gene(s) encoding the associated protein. First, variants were annotated with Ensembl VEP v83 using the "per-gene" option. Variant annotations were considered protein altering if they were annotated as coding sequence variant, frameshift variant, in-frame deletion, in-frame insertion, missense variant, protein altering variant, splice acceptor variant, splice donor variant, splice region variant, start lost, stop gained, or stop lost. To avoid multi-collinearity, PAVs were LD-pruned ($r^2 > 0.9$) using PLINK v1.9 before including them as covariates in the conditional analysis on the meta-analysis summary statistics using GCTA v1.25.2. Any variants with an $r^2 \geq 0.9$ with any of the PAVs were removed. Coverage of known common (MAF>5%) PAVs in our data was checked by comparison with exome sequences from ~60,000 individuals in the Exome Aggregation Consortium[69] (ExAC [http://exac.broadinstitute.org], downloaded June 2016).

Testing for regulatory and functional enrichment

We tested whether our pQTLs were enriched for functional and regulatory characteristics using GARFIELD v1.2.0[18]. GARFIELD is a non-parametric permutation-based enrichment method that compares input variants to permuted sets matched for number of proxies ($r^2 \geq 0.8$), MAF and distance to the closest TSS. It first applies "greedy pruning" ($r^2 < 0.1$) within a 1Mb region of the most significant variant. GARFIELD annotates variants with more than a thousand features, drawn predominantly from the GENCODE, ENCODE and ROADMAP

644   projects, which includes genic annotations, histone modifications, chromatin states and other

645   regulatory features across a wide range of tissues and cell types.

646

647   The enrichment analysis was run using all variants that passed our Bonferroni-adjusted

648   significance threshold $(p<1.5x10^{-11})$ for association with any protein. For each of the

649   matching criteria (MAF, distance to TSS, number of LD proxies), we used five bins. In total

650   we tested 25 combinations of features (classified as transcription factor binding sites, FAIRE-

651   seq, chromatin states, histone modifications, footprints, hotspots, or peaks) with up to 190

652   cell types from 57 tissues, leading to 998 tests. Hence, we considered enrichment with a $p<5$

653    $x10^{-5}$ (0.05/998) to be statistically significant.

654

Disease annotation

656   To identify diseases that our pQTLs have been associated with, we queried our sentinel

657   variants and their strong proxies ($r^2 \geq 0.8$) against publicly available disease GWAS data using

658   PhenoScanner[8].   A   list   of   datasets   queried   is   available   at

659   http://www.phenoscanner.medschl.cam.ac.uk/information.html. For disease GWAS, results

660   were filtered to $p<5x10^{-8}$ and then manually curated to retain only the entry with the strongest

661   evidence for association (i.e. smallest p-value) per disease. Non-disease phenotypes such as

662   anthropometric traits, intermediate biomarkers and lipids were excluded manually.

663

Cis eQTL overlap and enrichment of *cis* pQTLs for *cis* eQTLs

665   Each regional sentinel *cis* variant its strong proxies ($r^2 \geq 0.8$) were queried against publicly

666   available eQTL association data using PhenoScanner. *Cis* eQTL results were filtered to retain

667   only variants with $p<1.5x10^{-11}$. Only *cis* eQTLs for the same gene as the *cis* pQTL protein

668   were retained. To assess whether our *cis* pQTLs were more likely to also be *cis* eQTLs than

669 non-pQTL variants, we used data from the GTEx project v6, due to the availability of

670 genome-wide association results across a wide range of tissues and cell-types. GTEx results

671 were filtered to contain only variants lying in *cis* (i.e., within 1Mb) of genes that encode

672 proteins analysed in our study and only variants in both datasets were utilised.

673

674 For the enrichment analysis, the *cis* pQTL sentinel variants were first LD-pruned (r^2 <0.1)

675 and the proportion of sentinel *cis* pQTL variants that are also eQTLs (at $p<1.5 \times 10^{-11}$) for the

676 same protein/gene was compared to a permuted set of variants that were not pQTLs

677 *(p>0.0001 for all proteins)*. We generated 10,000 permuted sets of null variants matched for

678 MAF, distance to TSS and LD (as described for functional annotation enrichment in

679 **Functional annotation of pQTLs**). An empirical *p*-value was calculated as the proportion of

680 permuted variant sets where the proportion that are also *cis* eQTLs exceeded that of the test

681 set of sentinel *cis* pQTL variants. Results were similar in sensitivity analyses using the

682 standard genome-wide significance threshold of $p<5 \times 10^{-8}$ for the eQTLs (13.0% for whole

683 blood, 18.7% for LCLs, 17.1% for liver and 13.4% for monocytes) as well as also using only

684 the sentinel variants at *cis* pQTLs that were robust to adjusting for PAVs, suggesting our

685 results are robust to choice of threshold and potential differential binding effects.

686

Mediation of the GDF11/8 *trans* pQTL by WFIKKN2 levels

688 To assess whether the *trans* pQTL for GDF11/8 in the *WFIKKN2* gene region was mediated

689 by WFIKKN2 levels, we first regressed the rank inverse transformed residuals for GDF11/8

690 used for the GWAS against the WFIKKN2 residuals, adjusting for subcohort. The residuals

691 from this regression were subsequently regressed against allelic dosages for each variant in

692 the *WFIKKN2* region. As there were two SOMAmers targeting WFIKKN2, we tested both to

693 see if similar results were obtained. Regional association plots were made using Gviz[70].

694

Selection of genetic instruments for Mendelian randomization

696 In Mendelian randomization (MR), genetic variants are used as "instrumental variables" (IV)

697 for assessing the causal effect of the exposure (here a plasma protein) on the outcome (here

698 disease).

699

Proteins in the *IL1RL1-IL18R1* locus and atopic dermatitis

701 To identify the likely causal proteins that underpin the previous genetic association of the

702 *IL1RL1-IL18R1* locus (chr11:102.5-103.5Mb) with atopic dermatitis (AD)[36], we used a

703 multivariable MR approach. For each protein encoded by a gene in the *IL1RL1-IL18R1* locus,

704 we took genetic variants that had a *cis* association at $p<1\times10^{-4}$ and 'LD-pruned' them at

705 $r^2<0.1$. We then used these variants as instrumental variables for their respective proteins in

706 univariate MR. For multivariable MR, association estimates for all proteins in the locus were

707 extracted for all instruments. We used PhenoScanner to obtain association statistics for the

708 selected variants in the European-ancestry population of a recent large-scale GWAS meta-

709 analysis[36]. Where the relevant variant was not available, the strongest proxy with $r^2 \geqslant 0.8$ was

710 used.

711

MMP-12 and coronary heart disease (CHD)

713 To test whether plasma MMP-12 levels have a causal effect on risk of CHD, we selected

714 genetic variants in the *MMP12* gene region to use as instrumental variables. We constructed a

715 genetic score comprising 17 variants that had a *cis* association with MMP-12 levels at

716 $p<5\times10^{-8}$ and that were not highly correlated with one another ($r^2<0.2$). To perform

717 multivariable MR, we used association estimates for these variants with other MMP proteins

718 in the locus (MMP-1, MMP-7, MMP-8, MMP-10, MMP-13). Summary associations for

29

719   variants in the score with CHD were obtained through PhenoScanner from a recent large-

720   scale 1000 Genomes-based GWAS meta-analysis which consists mostly (77%) of individuals

721   of European ancestry[71].

722

MR analysis

723

724   Two-sample univariate MR was performed for each protein separately using summary

725   statistics in the inverse-variance weighted method adapted to account for correlated

726   variants[72,73]. For each of G genetic variants ($g = 1, ..., G$) having per-allele estimate of the

727   association with the protein β_{Xg} and standard error σ_{Xg}, and per-allele estimate of the

728   association with the outcome (here, AD or CHD) β_{Yg} and standard error σ_{Yg}, the IV estimate

729   ($\hat{\theta}_{XY}$) is obtained from generalised weighted linear regression of the genetic associations with

730   the outcome (β_Y) on the genetic associations with the protein (β_X) weighting for the

731   precisions of the genetic associations with the outcome and accounting for correlations

732   between the variants according to the regression model:

733

734    $$\beta_Y = \theta_{XY}\,\beta_X + \varepsilon, \quad \varepsilon \sim N(0, \Omega)$$

735

736   where β_Y and β_X are vectors of the univariable (marginal) genetic associations, and the

737   weighting matrix Ω has terms $\Omega_{g_1 g_2} = \sigma_{Y g_1}\sigma_{Y g_2}\rho_{g_1 g_2}$, and $\rho_{g_1 g_2}$ is the correlation between

738   the g_1th and g_2th variants.

739

740   The IV estimate from this method is:

741

742    $$\hat{\theta}_{XY} = (\beta_X{}^T \Omega^{-1} \beta_X)^{-1} \beta_X{}^T \Omega^{-1} \beta_Y$$

743

744 and the standard error is:

745

746
$$se(\hat{\theta}_{XY}) = \sqrt{(\beta_X{}^T \Omega^{-1} \beta_X)^{-1}}$$

747

748 where ^{T} is a matrix transpose. This is the estimate and standard error from the regression

749 model fixing the residual standard error to 1 (equivalent to a fixed-effects model in a meta-

750 analysis).

751

752 Genetic variants in univariate MR need to satisfy three key assumptions to be valid

753 instruments:

754   (1) the variant is associated with the risk factor of interest (i.e., the protein level),

755   (2) the variant is not associated with any confounder of the risk factor-outcome

756   association,

757   (3) the variant is conditionally independent of the outcome given the risk factor and

758   confounders.

759

760 To account for potential effects of functional pleiotropy[74], we performed multivariable MR

761 using the weighted regression-based method proposed by Burgess *et al*[75]. For each of K risk

762 factors in the model ($k = 1, ..., K$), the weighted regression-based method is performed by

763 multivariable generalized weighted linear regression of the association estimates β_Y on each

764 of the association estimates with each risk factor β_{Xk} in a single regression model:

765

766
$$\beta_Y = \theta_{XY1} \beta_{X1} + \theta_{XY2} \beta_{X2} + \cdots + \theta_{XYK} \beta_{XK} + \varepsilon, \quad \varepsilon \sim N(0, \Omega)$$

767

31

768   where β_{X1} is the vectors of the univariable genetic associations with risk factor 1, and so on.

769   This regression model is implemented by first pre-multiplying the association vectors by the

770   Cholesky decomposition of the weighting matrix, and then applying standard linear

771   regression to the transformed vectors. Estimates and standard errors are obtained fixing the

772   residual standard error to be 1 as above.

773

774   The multivariable MR analysis allows the estimation of the causal effect of a protein on

775   disease outcome accounting for the fact that genetic variants may be associated with multiple

776   proteins in the region. Causal estimates from multivariable MR represent direct causal

777   effects, representing the effect of intervening on one risk factor in the model while keeping

778   others constant.

779

780   **MMP-12 genetic score sensitivity analyses**

781   We performed two sensitivity analyses to determine the robustness of the Mendelian

782   randomization findings. First, we measured plasma MMP-12 levels using a different method

783   (proximity extension assay; Olink Bioscience, Uppsala, Sweden[76,77]) in a sub-sample of 141

784   individuals, and used this to derive genotype-MMP12 effect estimates for the 17 variants in

785   our genetic score. Second, we obtained effect estimates from a pQTL study based on

786   SOMAscan assay measurements in an independent sample of ~1,000 individuals[7]. In both

787   cases the genetic score reflecting higher plasma MMP-12 was associated with lower risk of

788   CHD.

789

790   **Overlap of pQTLs with drug targets**

791   We used the Informa Pharmaprojects database from Citeline to obtain information on drugs

792   that target proteins assayed on the SOMAscan platform. This is a manually curated database

793   that maintains profiles for >60,000 drugs. For our analysis, we focused on the following

794   information for each drug: protein target, indications, and development status. We included

795   drugs across the development pipeline, including those in pre-clinical studies or with no

796   development reported, drugs in clinical trials (all phases), and launched/registered drugs. For

797   each protein assayed, we identified all drugs in the Informa Pharmaprojects with a matching

798   protein target based on UniProt ID. When multiple drugs targeted the same protein, we

799   selected the drug with the latest stage of development.

800

801   For drug targets with significant pQTLs, we identified the subset where the sentinel variant or

802   proxy variants in LD (r^2>0.6) are also associated with disease risk through PhenoScanner.

803   We used an internal Merck auto-encoding method to map GWAS traits and drug indications

804   to a common set of terms from the Medical Dictionary for Regulatory Activities (MedDRA).

805   MedDRA terms are organized into a hierarchy with five levels. We mapped each GWAS trait

806   and indication onto the 'Lowest Level Terms' (i.e. the most specific terms available). All

807   matching terms were recorded for each trait or indication. We matched GWAS traits to drug

808   indications based on the highest level of the hierarchy, called 'System Organ Class' (SOC).

809   We designated a protein as 'matching' if at least one GWAS trait term matched with at least

810   one indication term for at least one drug.

811

812   **Data availability**

813   Summary association results will be made available on publication through

814   http://www.phenoscanner.medschl.cam.ac.uk.

33

References

1. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16,** 197–212 (2015).

2. Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11,** 786 (2015).

3. Melzer, D. *et al.* A Genome-Wide Association Study Identifies Protein Quantitative Trait Loci (pQTLs). *PLoS Genet* **4,** e1000072 (2008).

4. Enroth, S., Johansson, Å., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat Commun* **5,** (2014).

5. Deming, Y. *et al.* Genetic studies of plasma analytes identify novel potential biomarkers for several complex traits. *Sci. Rep.* **6,** 18092 (2016).

6. Rohloff, J. C. *et al.* Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. *Mol. Ther. - Nucleic Acids* **3,** e201 (2014).

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

8. Staley, J. R. *et al.* PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32,** 3207–3209 (2016).

9. Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30,** 543–52 (2015).

10. Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* **8,** e1002639 (2012).

11. Montgomery, S. B. & Dermitzakis, E. T. From expression QTLs to personalized transcriptomics. *Nat. Rev. Genet.* **12,** 277–282 (2011).

12. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24,** 14–24 (2014).

13. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).

14. 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).

15. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).

16. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

17. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–8 (2010).

18. Iotchkova, V. *et al.* GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. *bioRxiv* (2016). at http://biorxiv.org/content/early/2016/11/07/085738.abstract

19. Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499,** 79–82 (2013).

20. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45,** 1238–1243 (2013).

21. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

22. Schadt, E. E. *et al.* Mapping the Genetic Architecture of Gene Expression in Human Liver. *PLoS Biol.* **6,** e107 (2008).

863 23. Zeller, T. *et al.* Genetics and Beyond – The Transcriptome of Human Monocytes and
864   Disease Susceptibility. *PLoS One* **5,** e10693 (2010).
865 24. Nai, A. *et al.* TMPRSS6 rs855791 modulates hepcidin transcription in vitro and serum
866   hepcidin levels in normal individuals. *Blood* **118,** (2011).
867 25. Carrasquillo, M. M. *et al.* Genome-wide Screen Identifies rs646776 near Sortilin as a
868   Regulator of Progranulin Levels in Human Plasma. *Am. J. Hum. Genet.* **87,** 890–897
869   (2010).
870 26. Schuelke, M. *et al.* Myostatin Mutation Associated with Gross Muscle Hypertrophy in
871   a Child. *N. Engl. J. Med.* **350,** 2682–2688 (2004).
872 27. Malik, V., Rodino-Klapac, L. R. & Mendell, J. R. Emerging drugs for Duchenne
873   muscular dystrophy. *Expert Opin. Emerg. Drugs* **17,** 261–277 (2012).
874 28. Kondás, K., Szláma, G., Trexler, M. & Patthy, L. Both WFIKKN1 and WFIKKN2
875   have high affinity for growth and differentiation factors 8 and 11. *J. Biol. Chem.* **283,**
876   23677–84 (2008).
877 29. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of
878   inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
879 30. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory
880   bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47,**
881   979–986 (2015).
882 31. Di Narzo, A. F. *et al.* High-Throughput Characterization of Blood Serum Proteomics
883   of IBD Patients with Respect to Aging and Genetic Factors. *PLOS Genet.* **13,**
884   e1006565 (2017).
885 32. Lyons, P. A. *et al.* Genetically Distinct Subsets within ANCA-Associated Vasculitis.
886   *N. Engl. J. Med.* **367,** 214–223 (2012).
887 33. Hingorani, A. & Humphries, S. Nature's randomised trials. *Lancet* **366,** 1906–8
888   (2005).
889 34. Grönberg, H. *et al.* Prostate cancer screening in men aged 50–69 years (STHLM3): a
890   prospective population-based diagnostic study. *Lancet Oncol.* **16,** 1667–1676 (2015).
891 35. Eeles, R. A. *et al.* Multiple newly identified loci associated with prostate cancer
892   susceptibility. *Nat. Genet.* **40,** 316–321 (2008).
893 36. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases
894   and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47,**
895   1449–56 (2015).
896 37. Ganz, P. *et al.* Development and Validation of a Protein-Based Risk Score for
897   Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA*
898   **315,** 2532 (2016).
899 38. Goncalves, I. *et al.* Elevated Plasma Levels of MMP-12 Are Associated With
900   Atherosclerotic Burden and Symptomatic Cardiovascular Disease in Subjects With
901   Type 2 DiabetesSignificance. *Arterioscler. Thromb. Vasc. Biol.* **35,** 1723–1731 (2015).
902 39. Traylor, M. *et al.* A Novel MMP12 Locus Is Associated with Large Artery
903   Atherosclerotic Stroke Using a Genome-Wide Age-at-Onset Informed Approach.
904   *PLoS Genet.* **10,** e1004469 (2014).
905 40. Pinto, A. R., Godwin, J. W. & Rosenthal, N. A. Macrophages in cardiac homeostasis,
906   injury responses and progenitor cell mobilisation. *Stem Cell Res.* **13,** 705–714 (2014).
907 41. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug
908   indications. *Nat. Genet.* **47,** 856–860 (2015).
909 42. Albagha, O. M. E. *et al.* Genome-wide association study identifies variants at CSF1,
910   OPTN and TNFRSF11A as genetic risk factors for Paget's disease of bone. *Nat.*
911   *Genet.* **42,** 520–524 (2010).
912 43. Schwarz, P., Rasmussen, A. Q., Kvist, T. M., Andersen, U. B. & Jørgensen, N. R.

Paget's disease of the bone after treatment with Denosumab: a case report. *Bone* **50,** 1023–5 (2012).

44. Polyzos, S. A. *et al.* Denosumab Treatment for Juvenile Paget's Disease: Results From Two Adult Patients With Osteoprotegerin Deficiency ('Balkan' Mutation in the TNFRSF11B Gene). *J. Clin. Endocrinol. Metab.* **99,** 703–707 (2014).

45. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480,** 201–208 (2011).

46. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15,** 363 (2014).

47. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLoS One* **5,** e15004 (2010).

48. Menni, C. *et al.* Circulating Proteomic Signatures of Chronological Age. *J Gerontol A Biol Sci Med Sci* (2014). doi:10.1093/gerona/glu121

49. Sattlecker, M. *et al.* Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. *Alzheimer's Dement.* **10,** 724–734 (2014).

50. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–70 (2010).

51. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2015).

52. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

53. Ngo, D. *et al.* Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease. *Circulation* **134,** (2016).

54. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).

55. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39,** 906–913 (2007).

56. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–1 (2010).

57. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44,** 369–75, S1-3 (2012).

58. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-6 (2014).

59. Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26,** 1112–1118 (2010).

60. Aguet, F. *et al.* Local genetic effects on gene expression across 44 human tissues. *bioRxiv* (2016). at http://biorxiv.org/content/early/2016/09/09/074450.abstract

61. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).

62. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9,** e1003118 (2013).

63. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43,** D789–D798 (2015).

64. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–D462

963   (2016).

964   65.   Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated
965   over the tree of life. *Nucleic Acids Res.* **43,** D447–D452 (2015).

966   66.   Smith, R. N. *et al.* InterMine: a flexible data warehouse system for the integration and
967   analysis of heterogeneous biological data. *Bioinformatics* **28,** 3163–3165 (2012).

968   67.   Tenenbaum, D. KEGGREST: Client-side REST access to KEGG. (2016).

969   68.   Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with
970   increased coverage and integration. *Nucleic Acids Res.* **41,** D808–D815 (2013).

971   69.   Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*
972   **536,** 285–291 (2016).

973   70.   Hahne, F. & Ivanek, R. in *Methods in molecular biology (Clifton, N.J.)* **1418,** 335–351
974   (2016).

975   71.   Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association
976   meta-analysis of coronary artery disease. *Nat. Genet.* **47,** 1121–30 (2015).

977   72.   Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis
978   with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37,** 658–665
979   (2013).

980   73.   Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple
981   instrumental variables in Mendelian randomization: comparison of allele score and
982   summarized data methods. *Stat. Med.* **35,** 1880–906 (2016).

983   74.   Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of
984   pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181,** 251–60
985   (2015).

986   75.   Burgess, S., Dudbridge, F. & Thompson, S. G. Re: 'Multivariable Mendelian
987   randomization: the use of pleiotropic genetic variants to estimate causal effects'. *Am.*
988   *J. Epidemiol.* **181,** 290–1 (2015).

989   76.   Lundberg, M., Eriksson, A., Tran, B., Assarsson, E. & Fredriksson, S. Homogeneous
990   antibody-based proximity extension assays provide sensitive and specific detection of
991   low-abundant proteins in human blood. *Nucleic Acids Res.* **39,** e102–e102 (2011).

992   77.   Assarsson, E. *et al.* Homogenous 96-plex PEA immunoassay exhibiting high
993   sensitivity, specificity, and excellent scalability. *PLoS One* **9,** e95192 (2014).

994

37

Supplementary Information

Supplementary Information is available in the online version of the paper.

Acknowledgements

Author Contributions

1028

1029   Conceptualization and experimental design: J.D., A.S.B., B.B.S., H.R., R.M.P.;

1030   Methodology: B.B.S., A.B.S., J.C.M., J.E.P., H.R., S.B.; Analysis: B.B.S., J.C.M., J.E.P.,

1031   D.S., J.B., J.R.S., T.J., E.P., P.S., C.O-W., M.A.K., S.K.W., A.C., N.B., S.L.S.; Contributed

1032   reagents, materials, protocols or analysis tools: N.J., S.K.W., E.S.Z., J.B., M.A.K., J.R.S.,

1033   B.P.P.; Supervision: A.S.B., H.R., C.S.F., J.D., R.M.P., D.S.P., A.M.W.; Writing - principal:

1034   B.B.S., A.S.B., J.E.P., J.C.M., H.R., J.D.; Writing – review and editing: B.B.S., A.S.B.,

1035   J.E.P., J.C.M., J.D., H.R., K.S., A.M.W., N.J., D.J.R., J.A.T., D.S.P., N.S., C.S.F., R.M.P;

1036   Creation of the INTERVAL BioResource: J.R.B., D.J.R., W.H.O., N.W.M., J.D.; Funding

1037   acquisition: N.W.M., J.R.B., D.J.R., W.H.O., C.S.F., R.M.P., J.D.; all authors critically

1038   reviewed the manuscript.

Author Information

1039

1040   Reprints and permissions information is available at www.nature.com/reprints. The authors

1041   declare the following competing financial interests: J.C.M., A.C., C.S.F., R.M.P., H.R. are

1042   employees at MRL, Merck & Co., Inc. S.K.W., E.S.Z., N.J. are employees and stakeholders

1043   in SomaLogic, Inc. The other authors have nothing to disclose. Correspondence and requests

1044   for materials should be addressed to A.S.B. (asb38@medschl.cam.ac.uk) and J.D.

1045   (jd292@medschl.cam.ac.uk).

1046

Figures

Figure 1. The genetic architecture of plasma protein levels.

(a) Genomic location of pQTLs. Plot of sentinel variants for pQTLs (red= *cis*, blue= *trans*). Y-axis indicates the position of the gene that encodes the associated protein. The twelve most associated regions of the genome are annotated. (b) Plot of the statistical significance of the most associated (sentinel) *cis* variant for each protein against the distance from the transcription start site (TSS). (c) Histogram of the number of significantly associated loci per protein. (d) Histogram of the number of conditionally significant signals within each associated locus. (e) Histogram of protein variance explained (adjusted R^2) by conditionally significant variants. (f) Distribution of effect size against minor allele frequency (MAF) for *cis* and *trans* pQTLs. (g) Distribution of the predicted consequences of the sentinel pQTL variants compared to matched permuted null sets of variants. Asterisks highlight empirical enrichment *p*<0.005.

1060
1061

42
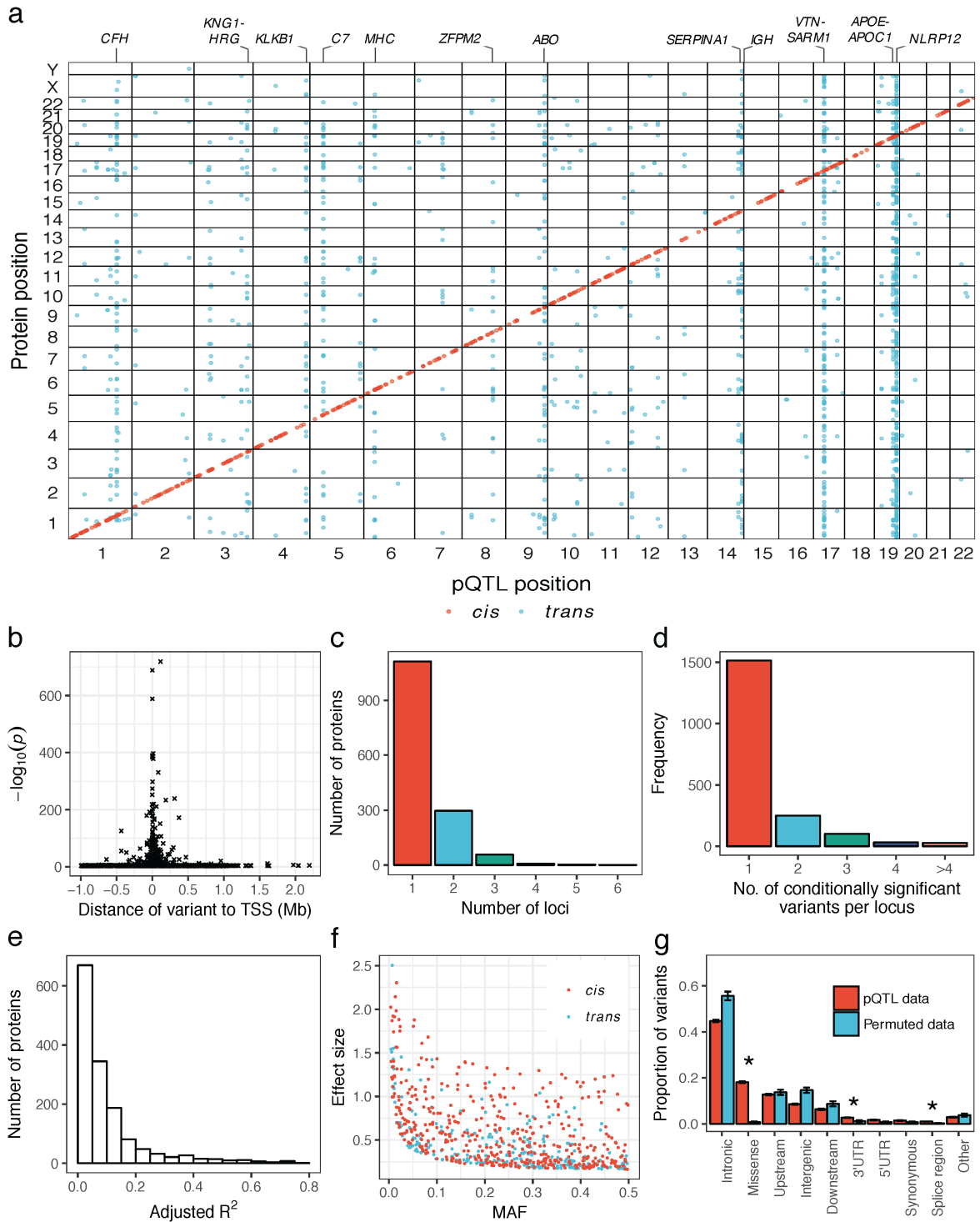
1062 **Figure 2. The GDF11/8 *trans* pQTL is mediated by genetic control of WFIKKN2 levels.**

1063 (a) Regional association plots of the *trans* pQTL (sentinel variant rs11079936) for GDF11/8

1064 before and after adjusting for levels of WFIKKN2 (upper panels), and the WFIKKN2 *cis*

1065 pQTL after adjusting for GDF11/8 levels (bottom panel). A similar pattern of association for

1066 WFIKKN2 was seen prior to GDF11/8 adjustment (not shown). (b) Proposed mechanism of

1067 how the *trans* pQTL for GDF11/8 is mediated by WFIKKN2 levels.
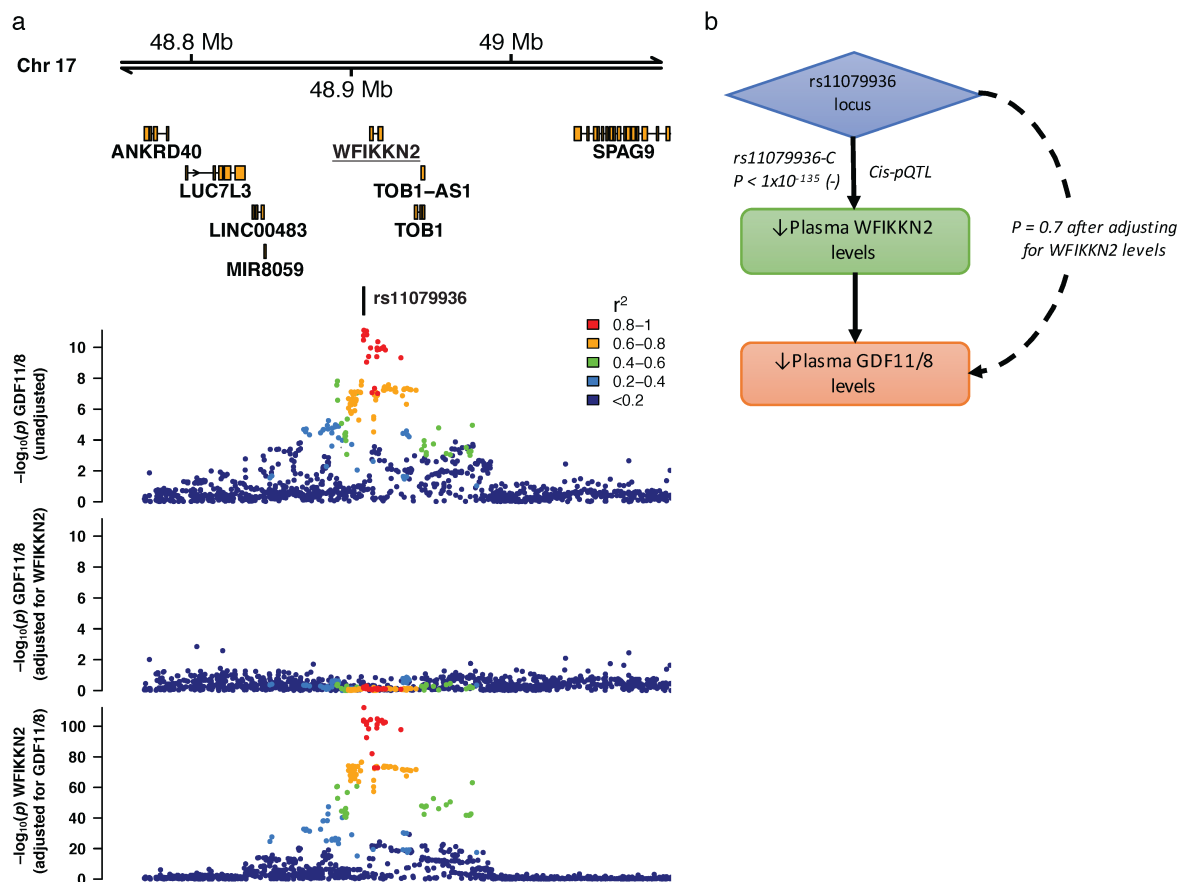
1068

1069

1070 **Figure 3. *Trans* pQTL for BLIMP1 at an inflammatory bowel disease (IBD) associated**

1071 **genetic variant in *MST1*.**

1072 (a) Missense variant rs3197999 in the *MST1* region on chromosome 3 is associated with IBD

1073 (top) and BLIMP1 levels (bottom). (b) Regional association plot of the IBD susceptibility

1074 locus on chromosome 6 adjacent to the *PRDM1* gene, which encodes BLIMP1. IBD

1075 association data are for European participants from Liu *et al.,* 2015. (c) Boxplot for relative

1076 plasma BLIMP1 levels by rs3197999 genotype, stratified by the two subcohorts used in our

1077 analysis.
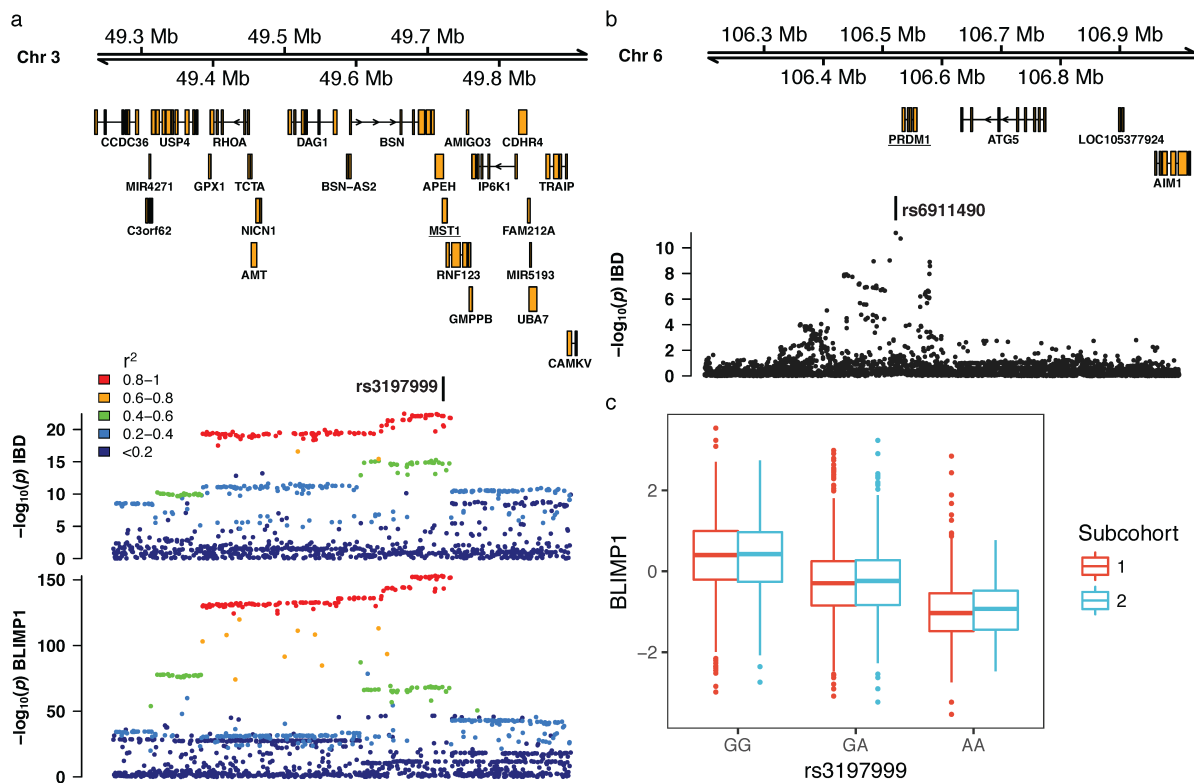
1078
1079

1080   **Figure 4. Missense variant rs28929474 in *SERPINA1* is a *trans* pQTL hotspot.**

1081   Numbers (outermost) indicate chromosomes. Interconnecting lines link the genomic location

1082   of rs28929474 and the genes encoding significantly associated ($p<1.5 \times 10^{-11}$) proteins. Line

1083   thickness is proportional to the effect size of the associations with red positive and blue

1084   negative.

1085
1086
1087

45
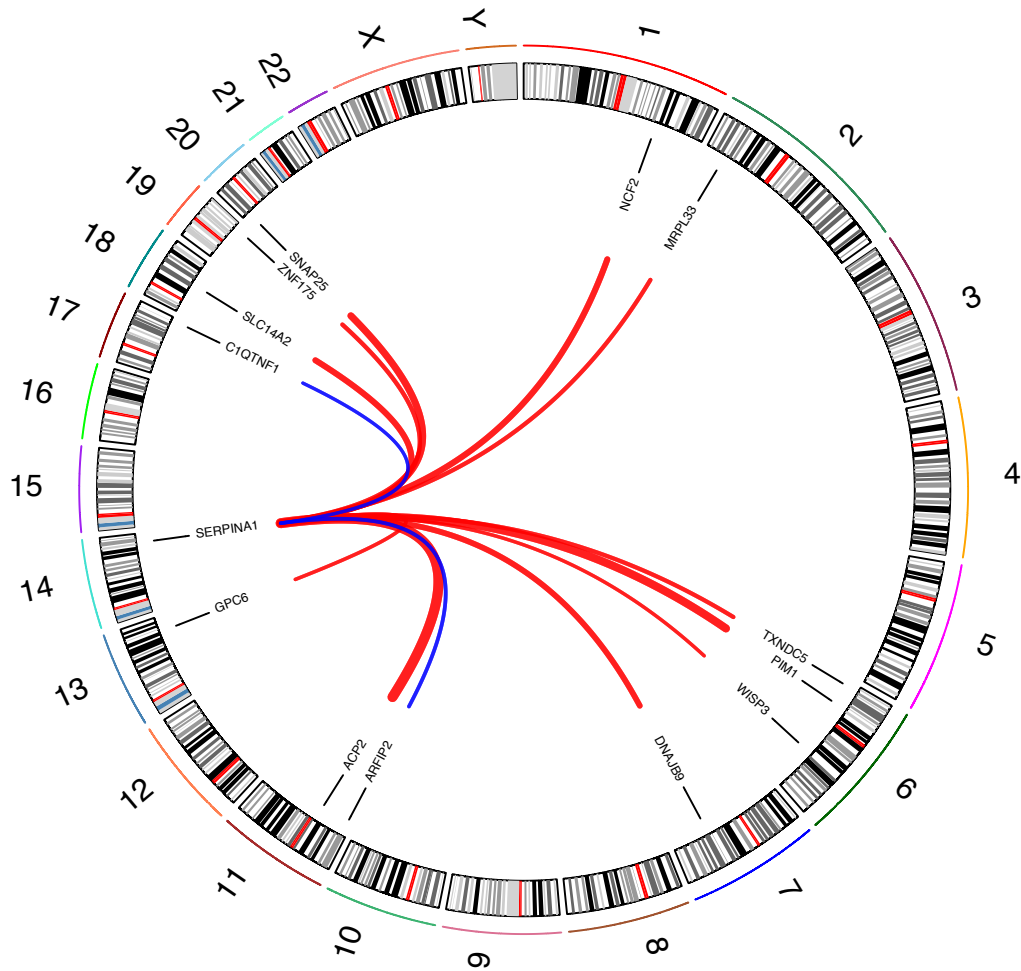
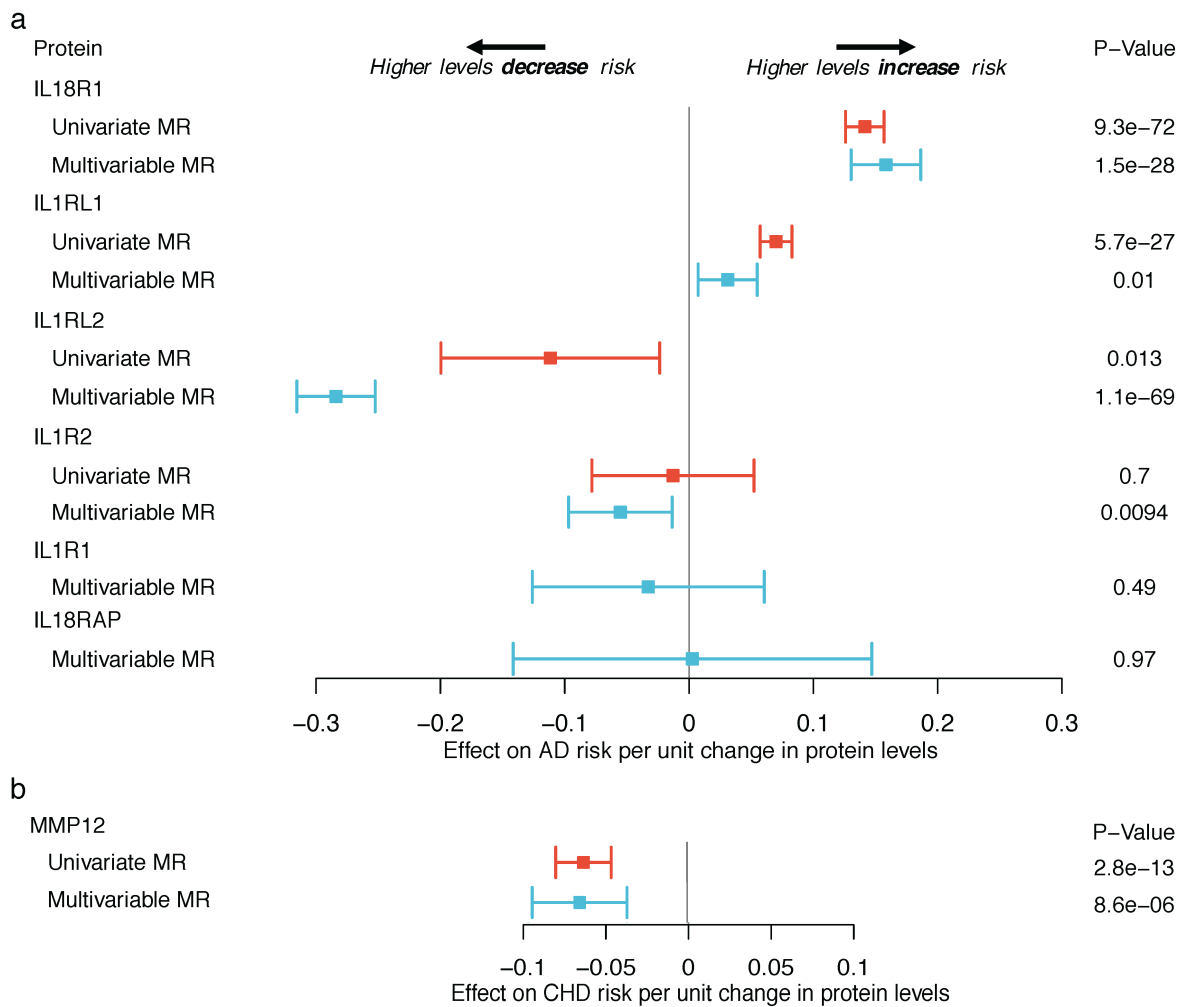1088 **Figure 5. Evaluation of causal role of proteins in disease.**

1089 Forest plot of univariate and multivariable Mendelian randomization (MR) estimates. (a)

1090 Proteins in the *IL1RL1-IL18R1* locus and risk of atopic dermatitis (AD). No univariate MR

1091 estimates available for IL1R1 and IL18RAP due to no significant pQTLs to select as a

1092 "genetic instrument". (b) MMP-12 levels and risk of coronary heart disease (CHD).

1093

1094