# 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[1,5], Praveen Surendran[1], Clare Oliver-Williams[1,6], Mihir A. Kamat[1], Bram P. Prins[1], Sheri K. Wilcox[7], Erik S. Zimmerman[7], An Chi[2], Narinder Bansal[1,8], Sarah L. Spain[9], Angela M. Wood[1], Nicholas W. Morrell[10], John R. Bradley[11], Nebojsa Janjic[7], David J. Roberts[12,13], Willem H. Ouwehand[3,14,15,16,17], John A. Todd[18], Nicole Soranzo[3,14,16,17], Karsten Suhre[19], Dirk S. Paul[1], Caroline S. Fox[2], Robert M. Plenge[2], John Danesh[1,3,16,17], Heiko Runz[2]*, Adam S. Butterworth[1,17]*

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. National Centre for Epidemiology and Population Health, The Australian National University, Canberra, ACT, Australia.
6. Homerton College, Cambridge, CB2 8PH, UK.
7. SomaLogic Inc., Boulder, Colorado 80301, USA.
8. Perinatal Institute, Birmingham B15 3BU, UK.
9. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1RQ, UK.
10. Division of Respiratory Medicine, Department of Medicine, University of Cambridge, Cambridge CB2 0QQ, UK.
11. NIHR Cambridge Biomedical Research Centre / BioResource, Cambridge University Hospitals, Cambridge CB2 0QQ, UK.
12. 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.
13. BRC Haematology Theme and Department of Haematology, Churchill Hospital, Oxford OX3 7LE, UK.
14. Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0PT, UK.
15. National Health Service (NHS) Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK.
16. Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1RQ, UK.
17. 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.
18. 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.
19. 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.)

## **Summary** (135 words)

Although proteins are the primary functional units of biology and the direct targets of most drugs, 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. 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 apply an integrated approach that links genetic variation with biological pathway, disease, and drug databases. We provide insights into pathobiology by uncovering the molecular effects of disease-associated variants and identifying causal roles for protein biomarkers in disease through Mendelian randomisation analysis. Our results reveal new drug targets, opportunities for matching existing drugs with new disease indications, and potential safety concerns for drugs under development.

(main text: 2,960 words)

Plasma proteins play key roles in a variety of biological processes including signalling, transport, growth, repair, and defence against infection. They are frequently dysregulated in disease and are important drug targets. Identifying factors that determine inter-individual protein variability should, therefore, furnish biological and medical insights[1]. Despite evidence of the heritability of plasma protein abundance[2], however, systematic assessment of how genetic variation influences plasma protein levels has been limited[1,3-5]. Studies have examined intracellular 'protein quantitative trait loci' (pQTLs)[6-8], but they have tended to be small and used cell lines rather than primary human tissues.

Here we create and interrogate a genetic atlas of the human plasma proteome, using a markedly expanded version of an aptamer-based multiplex protein assay (SOMAscan)[9] to quantify 3,622 plasma proteins in 3,301 healthy individuals. We identify 1,927 genotype-protein associations, including *trans*-associated loci for 1,104 proteins, providing new understanding of the genetic control of protein regulation. 88 pQTLs overlap with disease susceptibility loci, elucidating the molecular effects of disease-associated variants. Using the principle of Mendelian randomisation[10], we find evidence to support causal roles in disease for several protein pathways, and cross-reference our data with disease and drug databases to highlight novel potential therapeutic targets.

# RESULTS

## Genetic architecture of the plasma proteome

After stringent quality control, we performed genome-wide testing of 10.6 million imputed autosomal variants against levels of 2,994 plasma proteins in 3,301 healthy European-ancestry individuals (Methods, Extended Data Figure 1). We demonstrated robustness of protein measurements in several ways (Methods, Supplementary Note), including: highly consistent measurements in replicate samples; temporal consistency in protein levels in individuals at timepoints two years apart (Extended Data Figure 2b); replication of known associations with non-genetic factors (Supplementary Tables 1-2). To assess potential off-target cross-reactivity, we tested 920 SOMAmers for detection of proteins with ≥40% sequence homology to the target protein (Methods). Although 126 (14%) SOMAmers showed comparable binding with a homologous protein (Supplementary Table 3), nearly half of these were binding to alternative forms of the same protein.

We found 1,927 genome-wide significant ($p$<1.5x10$^{-11}$) associations between 1,478 proteins and 764 genomic regions (Figure 1a, Supplementary Table 4, Supplementary Video 1), with 89% of pQTLs previously unreported. Of the 764 associated regions, 502 (66%) had local-acting ('*cis*') associations only, 228 (30%) *trans* associations only, and 34 (4%) both *cis* and *trans* (Supplementary Note Table 1). 95% and 87% of *cis* pQTL variants were located within 200Kb and 100Kb, respectively, of the relevant gene's canonical transcription start site (TSS) (Figure 1b), and 44% were within the gene itself. The *p*-values for *cis* pQTL associations increased with distance from the TSS, mirroring findings for expression QTLs (eQTLs)[11,12]. Of the proteins for which we identified a pQTL, 88% had either *cis* (n=374) or *trans* (n=925) associations only, while 12% (n=179) had both (Supplementary Note Table 1). The majority of significantly associated proteins (75%; n=1,113) had a single pQTL, while 20% had two and 5% had >2 (Figure 1c). To detect multiple independent signals at the same locus we used stepwise conditional analysis, identifying 2,658 conditionally significant associations (Supplementary Table 5). Of the 1,927 locus-protein associations, 414 (21%) had multiple conditionally significant signals (Figure 1d), of which 255 were *cis*.

We were able to test replication of 163 pQTLs in 4,998 individuals using an alternative protein assay (Olink, Methods)[13]. Effect-size estimates for these 163 pQTLs were strongly correlated between the SOMAscan and Olink platforms (*r*=0.83; Extended Data Figure 2c). 106/163 (65% overall; 81% *cis*, 52% *trans*) pQTLs replicated after Bonferroni correction (Supplementary Tables 4,6). The lower replication rate of *trans* signals may reflect various factors, including differences between protein assays (e.g., detection of free versus complexed proteins) and the higher 'biological prior' for *cis* associations.

Of 1,927 pQTLs, 549 (28.5%) were *cis*-acting (Supplementary Table 4). Genetic variants that change protein structure may result in apparent pQTLs due to altered aptamer-binding rather than true quantitative differences in protein levels. However, we found evidence against the possibility of such artefactual associations for 371 (67.6%) *cis* pQTLs (Methods, Supplementary Tables 4, 7-8). Results were materially unchanged when we repeated downstream analyses excluding those *cis* pQTLs without evidence against binding effects.

The median variation in protein levels explained by pQTLs was 5.8% (in-sample estimate; interquartile range: 2.6-12.4%, Figure 1e). For 193 proteins, however, genetic variants explained >20% of the variation. There was a strong inverse relationship between effect-size and minor allele frequency (MAF) (Figure 1f), consistent with previous genome-wide association studies (GWAS) of quantitative traits[8,14-15]. We found 23 and 208 associations with rare (MAF <1%) variants and low-frequency (MAF 1-5%) variants, respectively (Supplementary Table 4). Of the 36 strongest pQTLs (per-allele effect-size >1.5 standard deviations), 29 were rare or low-frequency variants.

Both *cis* and *trans* pQTLs were strongly enriched for missense variants ($p<0.0001$) and for location in 3' untranslated ($p=0.0025$) or splice sites ($p=0.0004$) (Figure 1g, Extended Data Figure 3a). We found ≥3-fold enrichment ($p<5\text{x}10^{-5}$) of pQTLs at features indicative of transcriptional activation in blood cells (unsurprisingly given our use of plasma) and at hepatocyte regulatory elements, consistent with the liver's role in protein synthesis and secretion (Methods, Extended Data Figure 4, Supplementary Table 9).

## Overlap of eQTLs and pQTLs

An important question is the extent to which genetic associations with plasma protein levels are driven by effects at the transcription level, rather than other mechanisms, such as altered protein clearance or secretion. We therefore cross-referenced our *cis* pQTLs with previous eQTL studies (Supplementary Table 10), initially defining overlap between an eQTL and pQTL as high linkage disequilibrium (LD) ($r^2\geq0.8$) between the lead pQTL and eQTL variants. 40% (n=224) of *cis* pQTLs were eQTLs for the same gene in ≥1 tissue or cell-type (Supplementary Table 8). The greatest overlaps were in whole blood (n=117), liver (n=70) and lymphoblastoid cell-lines (LCLs) (n=52), consistent with biological expectation, but also likely driven by the larger eQTL study sample sizes for these cell-types. To examine whether the same causal variant was likely to underlie overlapping eQTLs and pQTLs, we performed colocalisation testing (Methods). Of 228 non-*HLA* pQTLs for which testing was possible, colocalisation in ≥1 tissue or cell-type was highly likely (posterior probability[PP]>0.8) in 179 (78.5%) and the most likely explanation (PP>0.5) in 197 (86.4%) (Supplementary Table 8). *Cis* pQTLs were significantly enriched for eQTLs for the corresponding gene ($p<0.0001$) (Methods, Supplementary Table 11). To address the converse (i.e., to what extent do eQTLs translate into pQTLs), we used a set of well-powered eQTL studies in relevant tissues (whole blood, LCLs, liver and monocytes[16-19]). Of the strongest *cis* eQTLs ($p<1.5\text{x}10^{-11}$), 12.2% of those in whole blood were also *cis* pQTLs, 21.3% for LCLs, 14.8% for liver and 14.7% for monocytes.

Comparisons between eQTL and pQTL studies have inherent limitations, including differences in the tissues, sample sizes and technological platforms used. Moreover, plasma protein levels may not reflect levels within tissues or cells. Nevertheless, our data suggest that genetic effects on plasma protein abundance are often, but not exclusively, driven by regulation of mRNA. *Cis* pQTLs without corresponding *cis* eQTLs may reflect genetic effects on processes other than transcription, including protein degradation, binding, secretion, or clearance from circulation.

## Using *trans* pQTLs to illuminate biological pathways and disease pathobiology

*Trans* pQTLs are useful for understanding biological relationships between proteins, particularly when the causal gene at the *trans*-associated locus can be identified. Of the 764 protein-associated regions, 262 had *trans* associations with 1,104 proteins (Supplementary Table 4, 12). There was no enrichment of cross-reactivity in SOMAmers with a *trans* pQTL versus those without (Supplementary Note). We replicated previously reported *trans* associations including *TMPRSS6* with transferrin receptor protein 1[20] and *SORT1* with granulins[21] and identified several novel biologically plausible *trans* associations (Supplementary Table 13), including known or presumed ligand:receptor pairs (e.g., the *CD320* gene region, which encodes the transcobalamin receptor, was associated with transcobalamin-2 levels).

Most (82%) *trans* loci were associated with <4 proteins, but 12 'hotspot' regions were associated with >20 (Figure 1a, Extended Data Figure 3b), including well-known pleiotropic loci (e.g., *ABO*, *CFH*, *APOE*, *KLKB1*) and loci associated with many correlated proteins (e.g., the *ZFPM2* locus encoding the transcription factor FOG2). Similar pleiotropy at these loci has been seen in other plasma pQTL studies[22-24], albeit with fewer proteins due to more limited assay breadth. rs28929474:T in *SERPINA1* was associated with 13 proteins at $p$<1.5x10$^{-11}$ and a further six at $p$<5x10$^{-8}$ (Figure 2). This missense variant (the 'Z-allele', p.Glu366Lys) results in defective secretion and intracellular accumulation of alpha1-antitrypsin (A1AT), an anti-protease. ZZ homozygotes have deficiency of circulating A1AT and increased risk of emphysema, liver cirrhosis and vasculitis. The 'protease-antiprotease' hypothesis posits that these clinical manifestations result from unchecked protease activity. However, our discovery of multiple *trans*-associated proteins at this locus highlights additional pathways potentially relevant to pathogenesis, a hypothesis supported by accumulating data[25].

GWAS have identified thousands of loci associated with common diseases, but the mechanisms by which most variants influence disease susceptibility await discovery. To identify intermediate links between genotype and disease, we overlapped pQTLs with disease-associated genetic variants identified through GWAS. 88 of our sentinel pQTL variants were in high LD ($r^2$≥0.8) with sentinel disease-associated variants (Supplementary Table 14), including 30 with *cis* associations, 54 with *trans* associations and 4 with both. Since some genetic loci are associated with multiple diseases, these 88 genetic loci represent 253 distinct genotype-disease associations. Overlap of a pQTL and a disease association signal does not necessarily imply that the same genetic variant underlies both traits, since there may be distinct causal variants for each trait that are in LD with one another. We therefore performed colocalisation testing (Methods). Of 108 locus-disease associations for which testing was possible (excluding the MHC region), colocalisation was highly likely (PP>0.8) for 96 (88.9%), and the most likely explanation (PP>0.5) for 106 (98.1%) (Supplementary Table 14).

*Trans* pQTLs that overlap with disease associations can highlight previously unsuspected candidate proteins through which genetic loci may influence disease risk. To help identify such candidates, we applied the ProGeM framework[26] (Methods, Supplementary Table 12, Extended Data Figure 5). We show that an inflammatory bowel disease (IBD) risk allele[27-28] (rs3197999:A, missense p.Arg703Cys) in *MST1* on chromosome 3, that decreases plasma MST1 levels[29], is a *trans* pQTL for eight additional proteins (Supplementary Table 4, Figure 3). Notably, genes that encode three of these proteins (*PRDM1*, *FASLG*, and *DOCK9*) each lie within 500kb of IBD GWAS loci where the causal gene is ambiguous[30]. For instance, the IBD-associated variant rs6911490 lies on chromosome 6 in the intergenic region between *PRDM1* (encoding BLIMP1, a master regulator of immune cell differentiation) and *ATG5* (involved in autophagy) (Figure 3c). Neither fine-mapping nor eQTL colocalisation analyses have unequivocally resolved the causal gene at this locus[30]; both *PRDM1* and *ATG5* are plausible candidates. Our data provide support for *PRDM1*.

Anti-neutrophil cytoplasmic antibody-associated vasculitis (AAV) is an autoimmune disease characterised by vascular inflammation and autoantibodies to the neutrophil proteases proteinase-3 (PR3) or myeloperoxidase. GWAS reveal distinct genetic signals according to antibody specificity[31], with variants near *PRTN3* (encoding PR3) and at the Z-allele of *SERPINA1* (encoding alpha1-antitrypsin, an inhibitor of PR3) associated specifically with PR3-antibody positive AAV. The SOMAscan assay has two SOMAmers targeting PR3; we identified a *cis* pQTL signal immediately upstream of *PRTN3* for both (Supplementary Table 4, Figures 4a-b). Conditional analysis revealed multiple independently associated variants

(Supplementary Table 5), one of which (rs7254911) was in high LD with the PR3+ vasculitis tag SNPs (Supplementary Note). We show that the vasculitis risk allele at *PRTN3* is associated with higher plasma levels of PR3 (Supplementary Note Table 4).

For one PR3 SOMAmer, we also found a *trans* pQTL at *SERPINA1*, with the Z-allele associating with lower plasma PR3 (Figure 4a). To understand the SOMAmer-specific nature of this signal, we assayed the relative affinity of these SOMAmers for the free and complexed states of PR3 and A1AT (which binds and inhibits proteases including PR3). We found that the SOMAmer showing *cis* and *trans* associations predominantly measures the PR3:A1AT complex rather than free PR3, whereas the SOMAmer with only *cis* association measures both the free and complexed forms. Importantly, neither SOMAmer bound free A1AT, demonstrating that the *SERPINA1* pQTL did not reflect non-specific cross-reactivity (Supplementary Note).

These data show that the vasculitis risk allele at *PRTN3* increases total PR3 plasma levels, consistent with its effect on *PRTN3* mRNA abundance in whole blood in GTEx data[32]. The *SERPINA1* Z-allele results in a reduced proportion of PR3 bound to A1AT. We thus demonstrate how altered availability of PR3, conferred by two independent genetic mechanisms, is a key susceptibility factor for breaking immune tolerance to PR3 and the development of PR3+ vasculitis (Figure 4c).

## Causal evaluation of candidate proteins in disease

Association of plasma protein levels with disease risk does not necessarily imply causation. To help establish causality, we employed the principle of Mendelian randomisation (MR)[10] (Extended Data Figure 6). In contrast with observational studies, which are liable to confounding and/or reverse causation, MR analysis can be akin to a ‘natural’ randomised controlled trial, exploiting the random allocation of alleles at conception. Consequently, if a genetic variant that specifically influences levels of a protein is also associated with disease risk, then it provides strong evidence of the protein’s causal role. For example, serum levels of PSP-94 (MSMB) are lower in patients with prostate cancer[33], but it is debated whether this association is correlative or causal. We identified a *cis* pQTL associated with lower PSP-94 plasma levels that overlaps with the prostate cancer susceptibility variant rs10993994[34], supporting a protective role for PSP-94 in prostate cancer (Supplementary Table 14).

Next, we leveraged multi-variant MR analysis methods to distinguish causal genes among multiple plausible candidates at disease loci, exemplified by the *IL1RL1-IL18R1* locus, which has been associated with a range of immune-mediated diseases including atopic dermatitis[35]. We identified four proteins that each had *cis* pQTLs at this locus (Supplementary Table 4), and created a genetic score for each protein (Methods). Initial ‘one-protein-at-a-time’ analysis identified associations of the scores for IL18R1 ($p$=9.3x10$^{-72}$) and IL1RL1 ($p$=5.7x10$^{-27}$) with atopic dermatitis risk (Figure 5a), and a weak association for IL1RL2 ($p$=0.013). We then mutually adjusted these associations for one another to account for the effects of the variants on multiple proteins. While the association of IL18R1 remained significant ($p$=1.5x10$^{-28}$), the association of IL1RL1 ($p$=0.01) was attenuated. In contrast, the association of IL1RL2 ($p$=1.1x10$^{-69}$) became much stronger, suggesting that IL1RL2 and IL18R1 underlie atopic dermatitis risk at this locus.

MMP-12 plays a key role in lung tissue damage, and MMP-12 inhibitors are being tested for chronic obstructive pulmonary disease[36-37]. We created a multi-allelic genetic score that explains 14% of the variation in plasma macrophage metalloelastase (MMP-12) levels

(Methods). Observational studies reveal an association of higher levels of plasma MMP-12 with recurrent cardiovascular events[38-39], stimulating interest in development of MMP-12 inhibitors for cardiovascular disease. In contrast, we found that genetic predisposition to higher MMP-12 levels is associated with *decreased* coronary disease risk ($p$=2.8x10$^{-13}$) (Figure 5b) and *decreased* large artery atherosclerotic stroke risk[40]. Understanding the discordance between the observational epidemiology and the genetic risk score will be important given the therapeutic interest in this target.

## Drug target prioritisation

Drugs directed at therapeutic targets implicated by human genetic data have a greater likelihood of success[41]. Of the proteins for which we identified a pQTL, 244 (17%) are established drug targets in the Informa Pharmaprojects database (Citeline) (Supplementary Table 15). 31 pQTLs for drug target proteins were highly likely to colocalise (posterior probability>0.8) with a disease GWAS locus, including some that are targets of approved drugs such as tocilizumab (anti-IL6R) and ustekinumab (anti-IL12/23) (Supplementary Table 16a).

To identify additional indications for existing drugs, we investigated disease associations of pQTLs for proteins already targeted by licensed drugs. Our results suggest potential drug 're-purposing' opportunities. For example, we identified a *cis* pQTL for RANK (encoded by *TNFRSF11A*) at a variant (rs884205) associated with Paget's disease, a condition characterised by excessive bone turnover, deformity and fracture (Supplementary Table 16b). Standard Paget's disease treatment consists of osteoclast inhibition with bisphosphonates, originally developed as anti-osteoporotic drugs. Denosumab, another anti-osteoporosis drug, is a monoclonal antibody targeting RANKL, the ligand for RANK. Our data suggest denosumab may be an alternative for Paget's disease patients in whom bisphosphonates are contra-indicated, a hypothesis supported by clinical case reports[43-44].

Next we evaluated targets for drugs currently under development, such as GP1BA, the receptor for von Willebrand factor. Drugs targeting GP1BA are in pre-clinical development as anti-thrombotic agents and in phase 2 trials for thrombotic thrombocytopenic purpura. We identified a *trans* pQTL for GP1BA at the pleiotropic *SH2B3/BRAP* locus, which is associated with platelet count[45], myocardial infarction (MI) and stroke (Supplementary Table 16b; $r^2$ from sentinel pQTL variant to lead platelet count, MI, and stroke variants is 0.91, 1.0, and 1.0, respectively). The risk allele for cardiovascular disease increases both plasma GP1BA and platelet count, suggesting a mechanism by which this locus affects disease susceptibility. As a confirmation of the link between GP1BA and platelet count, we found a directionally concordant *cis* pQTL for GP1BA at a platelet count-associated variant (Supplementary Table 16). Collectively, these results suggest that targeting GP1BA may be efficacious in conditions characterised by platelet aggregation such as arterial thrombosis. More generally, our data provide a substrate for generating hypotheses about potential therapeutic targets through linking genetic factors to disease via specific proteins.

# DISCUSSION

This study elucidates the genetic control of the human plasma proteome and uncovers intermediate molecular pathways that connect the genome to disease endpoints. We applied our discoveries to evaluate causal roles for proteins in important diseases using the principle of Mendelian randomisation (MR). Proteins provide an ideal paradigm for MR analysis because they are under proximal genetic control. However, application of protein-based MR has been constrained by limited availability of suitable genetic instruments, a bottleneck remedied by our data. Overall, our study foreshadows major advances in post-genomic science through increasing application of novel bioassay technologies to population biobanks.

# 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**, 4684 (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. Hause, R. J. *et al.* Identification and validation of genetic variants that influence transcription factor and cell signaling protein levels. *Am J Hum Genet* **95**, 194–208 (2014).

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

8. Battle, A. *et al.* Impact of regulatory variation from RNA to protein. *Science*. **347**, 644–7 (2015).

9. 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).

10. 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).

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

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

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

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

15. 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).

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

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

18. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*. **6**, e107 (2008).

19. Zeller, T. *et al.* Genetics and beyond – the transcriptome of human monocytes and disease susceptibility. *PLoS One* **5**, e10693 (2010).

20. Nai, A. *et al.* TMPRSS6 rs855791 modulates hepcidin transcription in vitro and serum hepcidin levels in normal individuals. *Blood* **118**, 4459-62 (2011).

21. Carrasquillo, M. M. *et al.* Genome-wide screen identifies rs646776 near sortilin as a regulator of progranulin levels in human plasma. *Am. J. Hum. Genet*. **87**, 890–897 (2010).

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

23. Yao, C. *et al.* Genome-wide association study of plasma proteins identifies putatively causal genes, proteins, and pathways for cardiovascular disease. *bioRxiv* (2017). doi:10.1101/136523

24. de Vries, P. S. *et al.* Whole-genome sequencing study of serum peptide levels: the Atherosclerosis Risk in Communities study. *Hum. Mol. Genet*. **26**, 3442–3450 (2017).

25. Gooptu, B., Dickens, J. A. & Lomas, D. A. The molecular and cellular pathology of $\alpha_1$-antitrypsin deficiency. *Trends Mol. Med*. **20**, 116–27 (2014).

26. Stacey, D. *et al.* ProGeM: A framework for the prioritisation of candidate causal genes at molecular quantitative trait loci. *bioRxiv* 230094 (2017). doi:10.1101/230094

27. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).

28. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet*. **47**, 979–986 (2015).

29. Di Narzo, A. F. *et al.* High-throughput characterization of blood serum proteomics of IBD patients with respect to aging and genetic factors. *PLoS Genet*. **13**, e1006565 (2017).

30. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

31. Lyons, P. A. *et al.* Genetically distinct subsets within ANCA-associated vasculitis. *N. Engl. J. Med*. **367**, 214–223 (2012).

32. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

33. Grönberg, H. *et al.* Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol.* **16**, 1667–1676 (2015).

34. Eeles, R. A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40**, 316–321 (2008).

35. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47**, 1449–56 (2015).

36. Dahl, R. *et al.* Effects of an oral MMP-9 and -12 inhibitor, AZD1236, on biomarkers in moderate/severe COPD: A randomised controlled trial. *Pulm. Pharmacol. Ther.* **25**, 169–177 (2012).

37. Churg, A. *et al.* Effect of an MMP-9/MMP-12 inhibitor on smoke-induced emphysema and airway remodelling in guinea pigs. *Thorax* **62**, 706–713 (2007).

38. Ganz, P. *et al.* Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA* **315**, 2532-41 (2016).

39. Goncalves, I. *et al.* Elevated plasma levels of MMP-12 are associated with atherosclerotic burden and symptomatic cardiovascular disease in subjects with type 2 diabetes. *Arterioscler. Thromb. Vasc. Biol.* **35**, 1723–1731 (2015).

40. Traylor, M. *et al.* A novel MMP12 locus is associated with large artery atherosclerotic stroke using a genome-wide age-at-onset informed approach. *PLoS Genet.* **10**, e1004469 (2014).

41. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

42. Albagha, O. M. E. *et al.* Genome-wide association study identifies variants at CSF1, OPTN and TNFRSF11A as genetic risk factors for Paget's disease of bone. *Nat. Genet.* **42**, 520–524 (2010).

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).

# ONLINE METHODS

## Study participants

The INTERVAL study comprised about 50,000 participants nested within a randomised trial of varying 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 (NHSBT). 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 diseases (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. After genetic QC, 3,301 participants (2,481 and 820 in the two subcohorts) remained for analysis (Supplementary Table 17).

## Plasma sample preparation

Sample collection procedures for INTERVAL have been described previously[47]. 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

We used a multiplexed, aptamer-based approach (SOMAscan assay) to measure the relative concentrations of 3,622 plasma proteins/protein complexes assayed using 4,034 modified aptamers ("SOMAmer reagents", hereafter referred to as 'SOMAmers'; Supplementary Table 18). The assay extends the lower limit of detectable protein abundance afforded by conventional approaches (e.g., immunoassays), measuring both extracellular and intracellular proteins (including soluble domains of membrane-associated proteins), with a bias towards proteins likely to be found in the human secretome (Extended Data Figure 7a)[9,48]. The proteins cover a wide range of molecular functions (Extended Data Figure 7b). The selection of proteins on the platform reflects both the availability of purified protein targets and a focus on proteins suspected to be involved in pathophysiology of human disease.

Aliquots of 150 µl of plasma were sent on dry ice to SomaLogic Inc. (Boulder, Colorado, US) for protein measurement. Assay details have been previously described[48-50] and a technical white paper with further information can be found at the manufacturer's website (http://somalogic.com/wp-content/uploads/2017/06/SSM-002-Technical-White-Paper_010916_LSM1.pdf). In brief, modified single-stranded DNA SOMAmers are used to bind to specific protein targets that are then quantified using a DNA microarray. Protein concentrations are quantified as relative fluorescent units.

Quality control (QC) was performed at the sample and SOMAmer level using control aptamers, as well as calibrator samples. At the sample level, hybridisation controls on the microarray were used to correct for systematic variability in hybridisation, while the median signal over all features assigned to one of three dilution sets (40%, 1% and 0.005%) was used to correct for within-run technical variability. The resulting hybridisation scale factors and median scale factors were used to normalise data across samples within a run. The acceptance criteria for these values are between 0.4 and 2.5 based on historical runs. SOMAmer-level QC made use of replicate calibrator samples using the same study matrix (plasma) to correct for between-run variability. The acceptance criterion for each SOMAmer was that the calibration scale factor be less than 0.4 from the median for each of the plates run. In addition, at the plate level, the acceptance criteria were that the median of the calibration scale factors be between 0.8 and 1.2, and that 95% of individual SOMAmers be less than 0.4 from the median within the plate.

In addition to QC processes routinely conducted by SomaLogic, we measured protein levels of 30 and 10 pooled plasma samples randomly distributed across plates for subcohort 1 and subcohort 2, respectively. Laboratory technicians were blinded to the presence of pooled samples. This approach enabled estimation of the reproducibility of the protein assays. We calculated CVs for each SOMAmer within each subcohort by dividing the standard deviation by the mean of the pooled plasma sample protein read-outs. In addition to passing SomaLogic QC processes, we required SOMAmers to have a CV≤20% in both subcohorts. Eight non-human protein targets were also excluded, leaving 3,283 SOMAmers (mapping to 2,994 unique proteins/protein complexes) for inclusion in the GWAS.

Protein mapping to UniProt identifiers and gene names was provided by SomaLogic. Mapping to Ensembl gene IDs and genomic positions was performed using Ensembl Variant Effect Predictor v83 (VEP)[51]. Protein subcellular locations were determined by exporting the subcellular location annotations from UniProt[52]. If the term ‘membrane’ was included in the descriptor, the protein was considered to be a membrane protein, whereas if the term ‘secreted’ (but not ‘membrane’) was included in the descriptor, the protein was considered to be a secreted protein. Proteins not annotated as either membrane or secreted proteins were classified (by inference) as intracellular proteins. Proteins were mapped to molecular functions using gene ontology annotations[53] from UniProt.

## Non-genetic associations of proteins

To provide confidence in the reproducibility of the protein assays, we attempted to replicate the associations with age or sex of 45 proteins previously reported by Ngo *et al* and 40 reported by Menni *et al*[49,54]. We used Bonferroni-corrected *p*-value thresholds of $p$=1.1x10$^{-3}$ (0.05/45) and $p$=1.2x10$^{-3}$ (0.05/40) respectively. Relative protein abundances were rank-inverse normalised within each subcohort and linear regression was performed using age, sex, BMI, natural log of estimated glomerular filtration rate (eGFR) and subcohort as independent variables.

## Genotyping and imputation

The genotyping protocol and QC for the INTERVAL samples (n~50,000) have been described previously in detail[15]. Briefly, DNA extracted from buffy coat was used to assay approximately 830,000 variants on the Affymetrix Axiom UK Biobank genotyping array at Affymetrix (Santa Clara, California, US). Genotyping was performed in multiple batches of approximately 4,800 samples each. Sample QC was performed including exclusions for sex mismatches, low call rates, duplicate samples, extreme heterozygosity and non-European descent. An additional exclusion made for this study was of one participant from each pair of close (first- or second-degree) relatives, defined as $\hat{\pi}$>0.187. Identity-by-descent was estimated using a subset of variants with a call rate >99% and MAF >5% in the merged dataset of both subcohorts, pruned for linkage disequilibrium (LD) using PLINK v1.9[55]. Numbers of participants excluded at each stage of the genetic QC are summarised in Extended Data Figure 1. Multi-dimensional scaling was performed using PLINK v1.9 to create components to account for ancestry in genetic analyses.

Prior to imputation, additional variant filtering steps were performed to establish a high-quality imputation scaffold. In summary, 654,966 high quality variants (autosomal, non-monomorphic, bi-allelic variants with Hardy Weinberg Equilibrium (HWE) $p > 5 \times 10^{-6}$, with a call rate of >99% across the INTERVAL genotyping batches in which a variant passed QC, and a global call rate of >75% across all INTERVAL genotyping batches) were used for imputation. Variants were phased using SHAPEIT3 and imputed using a combined 1000 Genomes Phase 3-UK10K reference panel. Imputation was performed via the Sanger Imputation Server (https://imputation.sanger.ac.uk) resulting in 87,696,888 imputed variants.

Prior to genetic association testing, variants were filtered in each subcohort separately using the following exclusion criteria: (1) imputation quality (INFO) score<0.7, (2) minor allele count<8, (3) HWE $p < 5 \times 10^{-6}$. In the small number of cases where imputed variants had the same genomic position (GRCh37) and alleles, the variant with the lowest INFO score was removed. 10,572,788 variants passing all filters in both subcohorts were taken forward for analysis (Extended Data Figure 1).

## Genome-wide association study

Within each subcohort, relative protein abundances were first natural log-transformed. Log-transformed protein levels were then adjusted in a linear regression for age, sex, duration between blood draw and processing (binary, ≤1 day/>1day) and the first three principal components of ancestry from multi-dimensional scaling. The protein residuals from this linear regression were then rank-inverse normalised and used as phenotypes for association testing. Simple linear regression using an additive genetic model was used to test genetic associations.

Association tests were carried out on allelic dosages to account for imputation uncertainty ("-method expected" option) using SNPTEST v2.5.2[56].

## Meta-analysis and statistical significance

Association results from the two subcohorts were combined via fixed-effects inverse-variance meta-analysis combining the betas and standard errors using METAL[57]. Genetic associations were considered to be genome-wide significant based on a conservative strategy requiring associations to have (i) a meta-analysis $p$-value<1.5x10$^{-11}$ (genome-wide threshold of $p$=5x10$^{-8}$ Bonferroni-corrected for 3,283 aptamers tested), (ii) at least nominal significance ($p$<0.05) in both subcohorts, and (iii) consistent direction of effect across subcohorts. We did not observe significant genomic inflation (mean inflation factor was 1.0, standard deviation=0.01) (Extended Data Figure 2d).

## Refinement of significant regions

To identify distinct non-overlapping regions associated with a given SOMAmer, we first defined a 1Mb region around each significant variant for that SOMAmer. Starting with the region containing the variant with the smallest $p$-value, any overlapping regions were then merged and this process was repeated until no more overlapping 1Mb regions remained. The variant with the lowest $p$-value for each region was assigned as the "regional sentinel variant". Due to the complexity of the Major Histocompatibility Region (MHC) region, we treated the extended MHC region (chr6:25.5-34.0Mb) as one region. To identify whether a region was associated with multiple SOMAmers, we used an LD-based clumping approach. Regional sentinel variants in high LD ($r^2$≥0.8) with each other were combined together into a single region.

## Conditional analyses

To identify conditionally significant signals, we performed approximate genome-wide step-wise conditional analysis using GCTA v1.25.2[58] using the "cojo-slct" option. We used the same conservative significance threshold of $p$=1.5x10$^{-11}$ as for the univariable 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 univariable genome-wide threshold ($p$<1.5x10$^{-11}$) were considered for step-wise selection. As the conditional analyses use different data inputs to the univariable analysis (i.e. summarised rather than individual-level data), there were some instances where the conditional analysis failed to include in the step-wise selection sentinel variants that were only just statistically significant in the univariable analysis. In these instances (n=28), we re-conducted the joint model estimation without step-wise selection in GCTA, using the variants identified by the conditional analysis in addition to the regional sentinel variant. We report and highlight these cases in Supplementary Table 5.

## 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[59] (https://www.ebi.ac.uk/gwas/, downloaded November 2017), which has all phenotypes mapped to the Experimental Factor Ontology (EFO)[60], by restricting to those with EFO annotations relevant to protein biomarkers (e.g., ‘protein measurement’, EFO_0004747). Studies identified through both approaches were manually filtered to include only studies that profiled plasma or serum samples and to exclude studies not assessing proteins. We recorded basic summary information for each study including the assay used, sample size and number of proteins with pQTLs (Supplementary Table 19). To reduce the impact of ethnic differences in allele frequencies on replication rate estimates, we filtered studies to include only associations reported in European-ancestry populations. We then manually extracted summary data on all reported associations from the manuscript or the supplementary material. This included rsID, protein UniProt ID, *p*-values, and whether the association is *cis/trans* (Supplementary Table 20).

To assess replication we first identified the set of unique UniProt IDs that were also assayed on the SOMAscan panel. For previous studies that used SomaLogic technology, we refined this match to the specific aptamer used. We then clumped associations into distinct loci using the same method that we applied to our pQTLs (see **Refinement of significant regions**). For each locus, we asked if the sentinel SNP or a proxy ($r^2$>0.6) was associated with the same protein/aptamer in our study at a defined significance threshold. For our primary assessment, we used a *p*-value threshold of $10^{-4}$ (Supplementary Table 21). We also performed sensitivity analyses to explore factors that influence replication rate (Supplementary Note).

## Replication study using Olink assay

To test replication of 163 pQTLs for 116 proteins, we performed protein measurements using an alternative assay, i.e., a proximity extension assay method (Olink Bioscience, Uppsala, Sweden)[4] in an additional subcohort of 4,998 INTERVAL participants. Proteins were measured using three 92-protein ‘panels’ – ‘inflammatory’, ‘cvd2’ and ‘cvd3’ (10 proteins were assayed on more than 1 panel). 4,902, 4,947 and 4,987 samples passed quality control for the 'inflammatory', 'cvd2' and 'cvd3' panels, respectively, of which, 712, 715 and 721 samples were from individuals included in our primary pQTL analysis using the SOMAscan assay. Normalised protein levels ('NPX') were regressed on age, sex, plate, time from blood draw to processing (in days), and season (categorical – 'Spring', 'Summer', 'Autumn', 'Winter'). The residuals were then rank-inverse normalized. Genotype data was processed as described earlier. Linear regression of the rank-inversed normalised residuals on genotype was carried out in SNPTEST with the first three components of multi-dimensional scaling as covariates to adjust for ancestry. pQTLs were considered to have replicated if they met a *p*-value threshold Bonferroni-corrected for the number of tests ($p<3.1\text{x}10^{-4}$; 0.05/163) and had a directionally concordant beta estimate with the SOMAscan estimate.

## Candidate gene annotation

We defined a pQTL as *cis* when the most significantly associated variant in the region was located within 1Mb of the transcription start site (TSS) of the gene(s) encoding the protein. pQTLs lying outside of the region were defined as *trans*. When considering the distance of the lead *cis*-associated variant from the relevant TSS, only proteins that map to single genes on the primary assembly in Ensembl v83 were considered.

For *trans* pQTLs, we sought to prioritise candidate genes in the region that might underpin the genotype-protein association. We applied the ProGeM framework[26] that leverages a combination of databases of molecular pathways, protein-protein interaction networks, and variant annotation, as well as functional genomic data including eQTL and chromosome conformation capture. In addition to reporting the nearest gene to the sentinel variant, ProGeM employs complementary 'bottom up' and 'top down' approaches, starting from the variant and protein respectively. For the ‘bottom up’ approach, the sentinel variant and corresponding proxies ($r^2$>0.8) for each *trans* pQTL were first annotated using Ensembl VEP v83 (using the ‘pick’ option) to determine whether variants were (1) protein-altering coding variants; (2) synonymous coding or 5’/3’ untranslated region (UTR); (3) intronic or up/downstream; or (4) intergenic. Second, we queried all sentinel variants and proxies against significant *cis* eQTL variants (defined by beta distribution-adjusted empirical *p*-values using an FDR threshold of 0.05, see http://www.gtexportal.org/home/documentationPage for details) in any cell type or tissue from the Genotype-Tissue Expression (GTEx) project v6[32] (http://www.gtexportal.org/home/datasets). Third, we also queried promoter capture Hi-C data in 17 human primary hematopoietic cell types[61] to identify contacts (with a CHICAGO score >5 in at least one cell type) involving chromosomal regions containing a sentinel variant. We considered gene promoters annotated on either fragment (i.e., the fragment containing the sentinel variant or the other corresponding fragment) as potential candidate genes. Using these three sources of information, we generated a list of candidate genes for the *trans* pQTLs. A gene was considered a candidate if it fulfilled at least one of the following criteria: (1) it was proximal (intragenic or ±5Kb from the gene) or nearest to the sentinel variant; (2) it contained a sentinel or proxy variant ($r^2$>0.8) that was protein-altering; (3) it had a significant *cis* eQTL in at least one GTEx tissue overlapping with a sentinel pQTL variant (or proxy); or (4) it was regulated by a promoter annotated on either fragment of a chromosomal contact[61] involving a sentinel variant.

For the ‘top down’ approach, we first identified all genes with a TSS located within the corresponding pQTL region using the GenomicRanges Bioconductor package[62] with annotation from a GRCh37 GTF file from Ensembl (ftp://ftp.ensembl.org/pub/grch37/update/gtf/homo_sapiens/; file:

‘Homo_sapiens.GRCh37.82.gtf.gz’, downloaded June 2016). We then identified any local genes that had previously been linked with the corresponding *trans*-associated protein(s) according to the following open source databases: (1) the Online Mendelian Inheritance in Man (OMIM) catalogue[63] (http://www.omim.org/); (2) the Kyoto Encyclopedia of Genes and Genomes (KEGG)[64] (http://www.genome.jp/kegg/); and (3) STRINGdb[65] (http://string-db.org/; v10.0). We accessed OMIM data via HumanMine web tool[66] (http://www.humanmine.org/; accessed June 2016), whereby we extracted all OMIM IDs for (i) our *trans*-affected proteins and (ii) genes local (±500Kb) to the corresponding *trans*-acting variant. We extracted all human KEGG pathway IDs using the KEGGREST Bioconductor package (https://bioconductor.org/packages/release/bioc/html/KEGGREST.html). In cases where a *trans*-associated protein shared either an OMIM ID or a KEGG pathway ID with a gene local to the corresponding *trans*-acting variant, we took this as evidence of a potential functional involvement of that gene. We interrogated protein-protein interaction data by accessing STRINGdb data using the STRINGdb Bioconductor package[67], whereby we extracted all pairwise interaction scores for each *trans*-affected protein and all proteins with genes local to the corresponding *trans*-acting variants. We took the default interaction score of 400 as evidence of an interaction between the proteins, therefore indicating a possible functional involvement for the local gene. In addition to using data from open source databases in our top down approach we also adopted a “guilt-by-association” (GbA) approach utilising the same plasma proteomic data used to identify our pQTLs. We first generated a matrix containing all possible pairwise Pearson’s correlation coefficients between our 3,283 SOMAmers. We then extracted the coefficients relating to our *trans*-associated proteins and any proteins encoded by genes local to their corresponding *trans*-acting variants (where available). Where the correlation coefficient was ≥0.5 we prioritised the relevant local genes as being potential mediators of the *trans* signal(s) at that locus.

We report the potential candidate genes for our *trans* pQTLs from both the 'bottom up' and 'top down' approaches, highlighting cases where the same gene was highlighted by both approaches.

## Functional annotation of pQTLs

Functional annotation of variants was performed using Ensembl VEP v83 using the 'pick' option. We tested the enrichment of significant pQTL variants for certain functional classes by comparing to permuted sets of variants showing no significant association with any protein ($p$>0.0001 for all proteins tested). First, the regional sentinel variants were LD-pruned at $r^2$ of 0.1. Each time the sentinel variants were LD-pruned, one of the pairs of correlated variants was removed at random and for each set of LD-pruned sentinel variants, 100 sets of equally sized null permuted variants were sampled matching for MAF (bins of 5%), distance to TSS (bins of 0-0.5Kb, 0.5-2Kb, 2-5Kb, 5-10Kb, 10-20Kb, 20-100Kb and >100Kb in each direction) and LD (± half the number of variants in LD with the sentinel variant at $r^2$ of 0.8). This procedure was repeated 100 times resulting in 10,000 permuted sets of variants. An empirical $p$-value was calculated as the proportion of permuted variant sets where the proportion that is classified as a particular functional group exceeded that of the test set of sentinel pQTL variants, and we used a significance threshold of $p$=0.005 (0.05/10 functional classes tested).

## 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 ($r^2 \geq 0.1$ and $r^2 \leq 0.9$) 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. 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 (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[69]. 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 projects, which includes genic annotations, histone modifications, chromatin states and other regulatory features across a wide range of tissues and cell types.

The enrichment analysis was run using all variants that passed our Bonferroni-adjusted significance threshold ($p < 1.5 \times 10^{-11}$) for association with any protein. For each of the matching criteria (MAF, distance to TSS, number of LD proxies), we used five bins. In total we tested

25 combinations of features (classified as transcription factor binding sites, FAIRE-seq, chromatin states, histone modifications, footprints, hotspots, or peaks) with up to 190 cell types from 57 tissues, leading to 998 tests. Hence, we considered enrichment with a $p$<5 x10$^{-5}$ (0.05/998) to be statistically significant.

## Disease annotation

To identify diseases that our pQTLs have been associated with, we queried our sentinel variants and their strong proxies ($r^2$≥0.8) against publicly available disease GWAS data using PhenoScanner[70]. A list of datasets queried is available at http://www.phenoscanner.medschl.cam.ac.uk/information.html. For disease GWAS, results were filtered to $p$<5x10$^{-8}$ and then manually curated to retain only the entry with the strongest evidence for association (i.e. smallest $p$-value) per disease. Non-disease phenotypes such as anthropometric traits, intermediate biomarkers and lipids were excluded manually.

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

For each regional sentinel *cis* pQTL variant, its strong proxies ($r^2$≥0.8) were queried against publicly available eQTL association data using PhenoScanner. *Cis* eQTL results were filtered to retain only variants with $p$<1.5x10$^{-11}$. Only *cis* eQTLs for the same gene as the *cis* pQTL protein were retained. We tested whether *cis* pQTLs were significantly enriched for eQTLs for the corresponding gene compared to null sets of variants appropriately matched for MAF and distance to nearest TSS. For this analysis, we restricted eQTL data to the GTEx project v6, since this project provided complete summary statistics across a wide range of tissues and cell-types, in contrast to many other studies which only report $p$-values below some significance level. GTEx results were filtered to contain only variants lying in *cis* (i.e., within 1Mb) of genes that encode proteins analysed in our study and only variants in both datasets were utilised.

For the enrichment analysis, the *cis* pQTL sentinel variants were first LD-pruned ($r^2$ <0.1) and the proportion of sentinel *cis* pQTL variants that are also eQTLs (at our pQTL significance threshold [$p$<1.5x10$^{-11}$], conventional genomewide significance [$p$<5x10$^{-8}$] or a nominal $p$-value threshold [$p$<1x10$^{-5}$]) for the same protein/gene was compared to a permuted set of variants that were not pQTLs *(p*>0.0001 for all proteins). We generated 10,000 permuted sets of null variants for each significance threshold matched for MAF, distance to TSS and LD (as described for functional annotation enrichment in **Functional annotation of pQTLs**). An empirical $p$-value was calculated as the proportion of permuted variant sets where the proportion that are also *cis* eQTLs exceeded that of the test set of sentinel *cis* pQTL variants.

At a stringent eQTL significance threshold ($p$<1.5x10$^{-11}$), we found significant enrichment of *cis* pQTLs for eQTLs ($p$<0.0001) (Supplementary Table 11) with 19.5% overlap observed compared to a mean overlap of 1.8% in the null sets. Results were similar in sensitivity analyses using the standard genome-wide or nominal significance thresholds as well as when using only the sentinel variants at *cis* pQTLs that were robust to adjusting for PAVs (Supplementary Table 7), suggesting our results are robust to the choice of threshold and potential differential binding effects.

## Colocalisation analysis

Colocalisation testing was performed using the coloc package[71]. For testing colocalisation of pQTLs and disease association signals, colocalisation testing was necessarily limited to disease traits where full GWAS summary statistics had been made available. We obtained GWAS summary statistics obtained through PhenoScanner. For testing colocalisation of pQTLs with eQTLs, we used publically available summary statistics for expression traits from GTEx[32]. We used the default priors. Regions for testing were determined by dividing the genome into 0.1cM chunks using recombination data. Evidence for colocalisation was assessed using the posterior probability (PP) for hypothesis 4 (that there is an association signal for both traits and they are driven by the same causal variant[s]). Signals with PP4>0.5 were deemed likely to colocalise as this gives hypothesis 4 the highest likelihood of being correct, while PP4>0.8 was deemed to be 'highly likely to colocalise'.

## Selection of genetic instruments for Mendelian randomisation

In Mendelian randomisation (MR), genetic variants are used as 'instrumental variables' (IV) for assessing the causal effect of the exposure (here a plasma protein) on the outcome (here disease)[10,72] (Extended Data Figure 6).

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

To identify the likely causal proteins that underpin the previous genetic association of the *IL1RL1-IL18R1* locus (chr11:102.5-103.5Mb) with atopic dermatitis (AD)[35], we used the following approach. For each protein encoded by a gene in the *IL1RL1-IL18R1* locus, we took genetic variants that had a *cis* association at $p<1\text{x}10^{-4}$ and 'LD-pruned' them at $r^2<0.1$ to leave largely independent variants. We then used these genetic variants to construct a genetic score for each protein. Formally, we used these variants as instrumental variables for their respective proteins in univariable MR. For multivariable MR, association estimates for all proteins in the locus were extracted for all instruments. We used PhenoScanner to obtain association statistics for the selected variants in the European-ancestry population of a recent large-scale GWAS meta-analysis[35]. Where the relevant variant was not available, the strongest proxy with $r^2 \geqslant 0.8$ was used.

**MMP-12 and coronary heart disease (CHD)**

To test whether plasma MMP-12 levels have a causal effect on risk of CHD, we selected genetic variants in the *MMP12* gene region to use as instrumental variables. We constructed a genetic score comprising 17 variants that had a *cis* association with MMP-12 levels at $p$<5x10$^{-8}$ and that were not highly correlated with one another ($r^2$<0.2). To perform multivariable MR, we used association estimates for these variants with other MMP proteins in the locus (MMP-1, MMP-7, MMP-8, MMP-10, MMP-13). Summary associations for variants in the score with CHD were obtained through PhenoScanner from a recent large-scale GWAS meta-analysis which consists mostly (77%) individuals of European ancestry[73].

## MR analysis

Two-sample univariable MR was performed for each protein separately using summary statistics in the inverse-variance weighted method adapted to account for correlated variants[74-75]. For each of $G$ genetic variants ($g = 1, \dots, G$) having per-allele estimate of the association with the protein $\beta_{Xg}$ and standard error $\sigma_{Xg}$, and per-allele estimate of the association with the outcome (here, AD or CHD) $\beta_{Yg}$ and standard error $\sigma_{Yg}$, the IV estimate ($\hat{\theta}_{XY}$) is obtained from generalised weighted linear regression of the genetic associations with the outcome ($\beta_Y$) on the genetic associations with the protein ($\beta_X$) weighting for the precisions of the genetic associations with the outcome and accounting for correlations between the variants according to the regression model:

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

where $\beta_Y$ and $\beta_X$ are vectors of the univariable (marginal) genetic associations, and the weighting matrix $\Omega$ has terms $\Omega_{g_1 g_2} = \sigma_{Yg_1}\sigma_{Yg_2}\rho_{g_1 g_2}$, and $\rho_{g_1 g_2}$ is the correlation between the $g_1$th and $g_2$th variants.

The IV estimate from this method is:

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

and the standard error is:

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

where $^T$ is a matrix transpose. This is the estimate and standard error from the regression model fixing the residual standard error to 1 (equivalent to a fixed-effects model in a meta-analysis).

Genetic variants in univariable MR need to satisfy three key assumptions to be valid instruments:

(1) the variant is associated with the risk factor of interest (i.e., the protein level),
(2) the variant is not associated with any confounder of the risk factor-outcome association,
(3) the variant is conditionally independent of the outcome given the risk factor and confounders.

To account for potential effects of functional pleiotropy[76], we performed multivariable MR using the weighted regression-based method proposed by Burgess *et al*[77]. For each of $K$ risk factors in the model ($k = 1, \dots, K$), the weighted regression-based method is performed by multivariable generalized weighted linear regression of the association estimates $\beta_Y$ on each of the association estimates with each risk factor $\beta_{Xk}$ in a single regression model:

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

where $\beta_{X1}$ is the vectors of the univariable genetic associations with risk factor 1, and so on. This regression model is implemented by first pre-multiplying the association vectors by the Cholesky decomposition of the weighting matrix, and then applying standard linear regression to the transformed vectors. Estimates and standard errors are obtained fixing the residual standard error to be 1 as above.

The multivariable MR analysis allows the estimation of the causal effect of a protein on disease outcome accounting for the fact that genetic variants may be associated with multiple proteins in the region. Causal estimates from multivariable MR represent direct causal effects, representing the effect of intervening on one risk factor in the model while keeping others constant.

**MMP-12 genetic score sensitivity analyses**

We performed two sensitivity analyses to determine the robustness of the MR findings. First, we measured plasma MMP-12 levels using a different method (proximity extension assay; Olink Bioscience, Uppsala, Sweden[4]) in 4,998 individuals, and used this to derive genotype-MMP12 effect estimates for the 17 variants in our genetic score. Second, we obtained effect estimates from a pQTL study based on SOMAscan assay measurements in an independent sample of ~1,000 individuals[22]. In both cases the genetic score reflecting higher plasma MMP-12 was associated with lower risk of CHD.

## Overlap of pQTLs with drug targets

We used the Informa Pharmaprojects database from Citeline to obtain information on drugs that target proteins assayed on the SOMAscan platform. This is a manually curated database that maintains profiles for >60,000 drugs. For our analysis, we focused on the following information for each drug: protein target, indications, and development status. We included drugs across the development pipeline, including those in pre-clinical studies or with no development reported, drugs in clinical trials (all phases), and launched/registered drugs. For each protein assayed, we identified all drugs in the Informa Pharmaprojects with a matching protein target based on UniProt ID. When multiple drugs targeted the same protein, we selected the drug with the latest stage of development.

For drug targets with significant pQTLs, we identified the subset where the sentinel variant or proxy variants in LD ($r^2$>0.8) are also associated with disease risk through PhenoScanner. We used an internal Merck auto-encoding method to map GWAS traits and drug indications to a common set of terms from the Medical Dictionary for Regulatory Activities (MedDRA). MedDRA terms are organised into a hierarchy with five levels. We mapped each GWAS trait and indication onto the 'Lowest Level Terms' (i.e. the most specific terms available). All matching terms were recorded for each trait or indication. We matched GWAS traits to drug indications based on the highest level of the hierarchy, called 'System Organ Class' (SOC). We designated a protein as 'matching' if at least one GWAS trait term matched with at least one indication term for at least one drug.

## Data availability

Participant-level genotype and protein data, and full summary association results from the genetic analysis, are available through the European Genotype Archive (accession number

EGAS00001002555). Summary association results will also be made available via FTP and through PhenoScanner (http://www.phenoscanner.medschl.cam.ac.uk).

# Online References

46. Di Angelantonio, E. *et al.* Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* **390**, 2360-2371 (2017).

47. 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).

48. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004 (2010).

49. Menni, C. *et al.* Circulating proteomic signatures of chronological age. *J Gerontol A Biol Sci Med Sci* **70**, 809-16 (2014).

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

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

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

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

54. Ngo, D. *et al.* Aptamer-based proteomic profiling reveals novel candidate biomarkers and pathways in cardiovascular disease. *Circulation* **134**, 270-285 (2016).

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

56. 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).

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

58. 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).

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

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

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 (2016).

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

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

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

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

69. Iotchkova, V. *et al.* GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. *bioRxiv* (2016). doi:10.1101/085738

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

71. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. **10**, e1004383 (2014).

72. Hingorani, A. & Humphries, S. Nature's randomised trials. *Lancet* **366**, 1906–8 (2005).

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

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

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

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

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

78. Merkel, P. A. *et al.* Identification of functional and expression polymorphisms associated with risk for anti-neutrophil cytoplasmic autoantibody-associated vasculitis. *Arthritis Rheumatol*. **69**, 1054-1066 (2016).

## Supplementary Information

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

## Acknowledgements

Aaron Day-Williams, Joshua McElwee, Dorothee Diogo, William Astle, Emanuele Di Angelantonio, Ewan Birney, Arianne Richard, Justin Mason and Michael Inouye commented helpfully on the manuscript; Mark Sharp helped with mapping drug indications to GWAS traits.

We thank the following: participants in INTERVAL and its study co-ordination team (at the Universities of Cambridge and Oxford and at NHSBT), including the staff at 25 NHSBT blood donation centres; the INTERVAL Operations Team led by Dr Richard Houghton and Dr Carmel Moore; the INTERVAL Data Management Team led by Dr Matthew Walker; and staff at SomaLogic for processing and running the proteomic assays.

The Cardiovascular Epidemiology Unit at the University of Cambridge is supported by the UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), UK National Institute for Health Research Cambridge Biomedical Research Centre, European Research Council (268834), and European Commission Framework Programme 7 (HEALTH-F2-2012-279233). B.B.S. was funded by the Cambridge School of Clinical Medicine MRC/Sackler Prize PhD Studentship (MR/K50127X/1) and supported by the Cambridge School of Clinical Medicine MB-PhD programme. J.E.P. was funded by a British Heart Foundation Clinical Research Fellowship through the BHF Cambridge Centre of Excellence [RE/13/6/30180] and a UK Research Innovation Clinician Scientist Fellowship (MR/S004068/1). P.S. was supported by a Rutherford Fund Fellowship (MR/S003746/1). D.S.P. and D.S. were funded by the Wellcome Trust (105602/Z/14/Z). N.S. was supported by the Wellcome Trust (WT098051 and WT091310), the EU FP7 (EPIGENESYS 257082 and BLUEPRINT HEALTH-F5-2011-282510). J.A.T was supported by the supported by the Wellcome Trust (091157) and JDRF (9-

2011-253). K.S. was funded by the Biomedical Research Program funds at Weill Cornell Medicine in Qatar, a program funded by the Qatar Foundation. J.D. was supported by a British Heart Foundation Professor, European Research Council Senior Investigator, and National Institute for Health Research (NIHR) Senior Investigator. The INTERVAL study was funded by NHSBT (11-01-GEN) and has been supported by the NIHR-BTRU in Donor Health and Genomics (NIHR BTRU-2014-10024) at the University of Cambridge in partnership with NHSBT. This study was partially funded by Merck and Co., Kenilworth, NJ, USA. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health of England, or NHSBT.

## Author Contributions

Conceptualization and experimental design: J.D., A.S.B., B.B.S., H.R., R.M.P.; 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., 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 reagents, materials, protocols or analysis tools: N.J., S.K.W., E.S.Z., J.B., M.A.K., J.R.S., B.P.P.; Supervision: A.S.B., H.R., J.D., R.M.P., C.S.F., D.S.P., A.M.W.; Writing - principal: 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., 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; Creation of the INTERVAL BioResource: J.R.B., D.J.R., W.H.O., N.W.M., J.D.; Funding acquisition: N.W.M., J.R.B., D.J.R., W.H.O.,H.R., R.M.P., J.D.; all authors critically reviewed the manuscript.

## Author Information

Reprints and permissions information is available at www.nature.com/reprints. The authors declare the following competing financial interests: J.C.M., A.C., C.S.F., R.M.P., H.R. are

employees at MRL, Merck & Co., Inc. S.K.W., E.S.Z., N.J. are employees and stakeholders in SomaLogic, Inc. The other authors have nothing to disclose. Correspondence and requests for materials should be addressed to A.S.B. (asb38@medschl.cam.ac.uk) and J.D. (jd292@medschl.cam.ac.uk).

# 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 12 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.

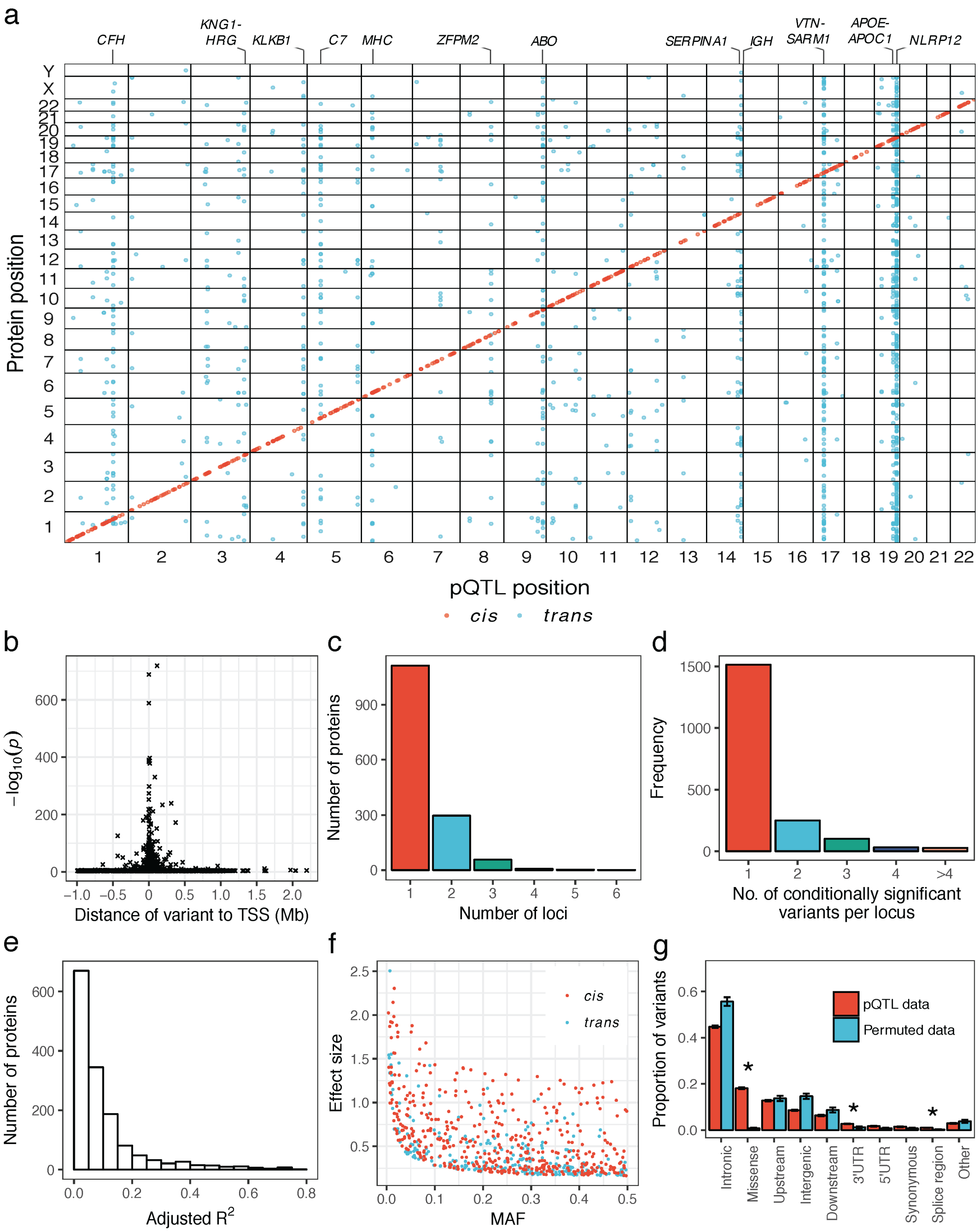

a
CFH
KNG1-
HRG
KLKB1
C7
MHC
ZFPM2
ABO
SERPINA1
IGH
VTN-
SARM1
APOE-
APOC1
NLRP12
Protein position
Y
X
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
pQTL position
cis
trans
b
−log10(p)
0
200
400
600
−1.0 −0.5 0 0.5 1.0 1.5 2.0
Distance of variant to TSS (Mb)
c
Number of proteins
0
300
600
900
1 2 3 4 5 6
Number of loci
d
Frequency
0
500
1000
1500
1 2 3 4 >4
No. of conditionally significant
variants per locus
e
Number of proteins
0
200
400
600
0.0 0.2 0.4 0.6 0.8
Adjusted R2
f
Effect size
0.5
1.0
1.5
2.0
2.5
0.0 0.1 0.2 0.3 0.4 0.5
MAF
cis
trans
g
Proportion of variants
0.0
0.2
0.4
0.6
pQTL data
Permuted data
*
*
*
Intronic
Missense
Upstream
Intergenic
Downstream
3'UTR
5'UTR
Synonymous
Splice region
Other

**Figure 2. Missense variant rs28929474 in *SERPINA1* is a *trans* pQTL hotspot.**

Numbers (outermost) indicate chromosomes. Interconnecting lines link the genomic location of rs28929474 and the genes encoding significantly associated ($p$<1.5x10$^{-11}$) proteins. Line thickness is proportional to the effect-size of the associations with red positive and blue negative. Genes with an asterisk indicate *trans* pQTLs that reached conventional genome-wide significance ($p$<5x10$^{-8}$).

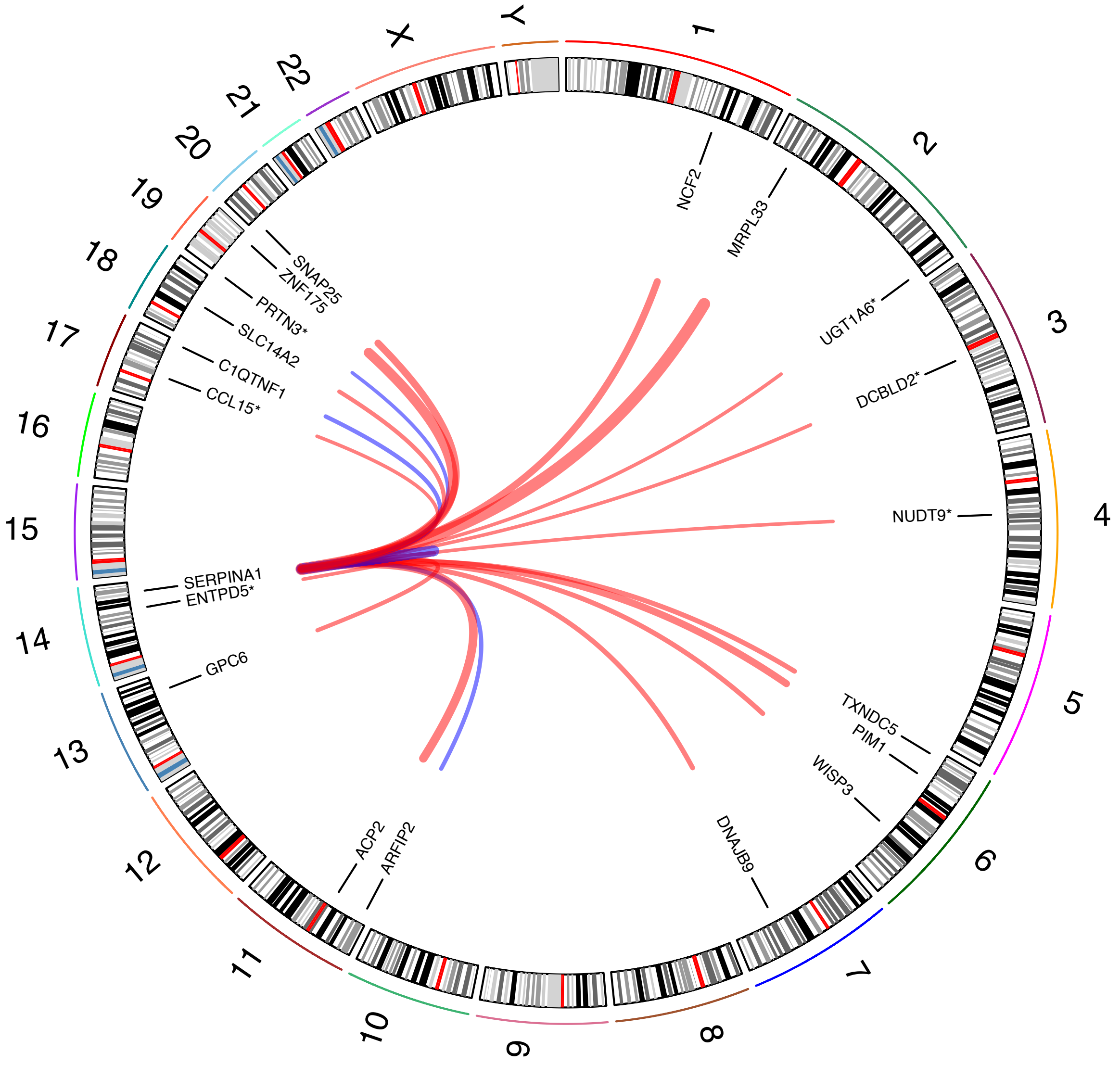


**Figure 3. *Trans* pQTL for BLIMP1 at an inflammatory bowel disease (IBD) associated genetic variant in *MST1*.**

(a) IBD-associated missense variant (rs3197999:A) in the *MST1* region on chromosome 3 is associated with abundance of multiple proteins in plasma. Interconnecting lines link the genomic location of rs3197999 and the genes encoding significantly associated ($p$<1.5x10$^{-11}$) proteins. Line thickness is proportional to the effect size. Red and blue lines indicate positive and negative effects of the IBD risk allele, respectively. * highlights genes in IBD GWAS loci. (b) Regional association plots of the IBD susceptibility locus at *MST1*, showing IBD association signal (top) and *trans* pQTLs for BLIMP1, DOCK9 and FASLG (bottom 3 panels). Colour key indicates $r^2$ with rs3197999. (c) Regional association plot of the IBD susceptibility locus on chromosome 6 adjacent to the *PRDM1* gene, which encodes BLIMP1. All IBD association data are for European participants from Liu *et al.,* 2015.

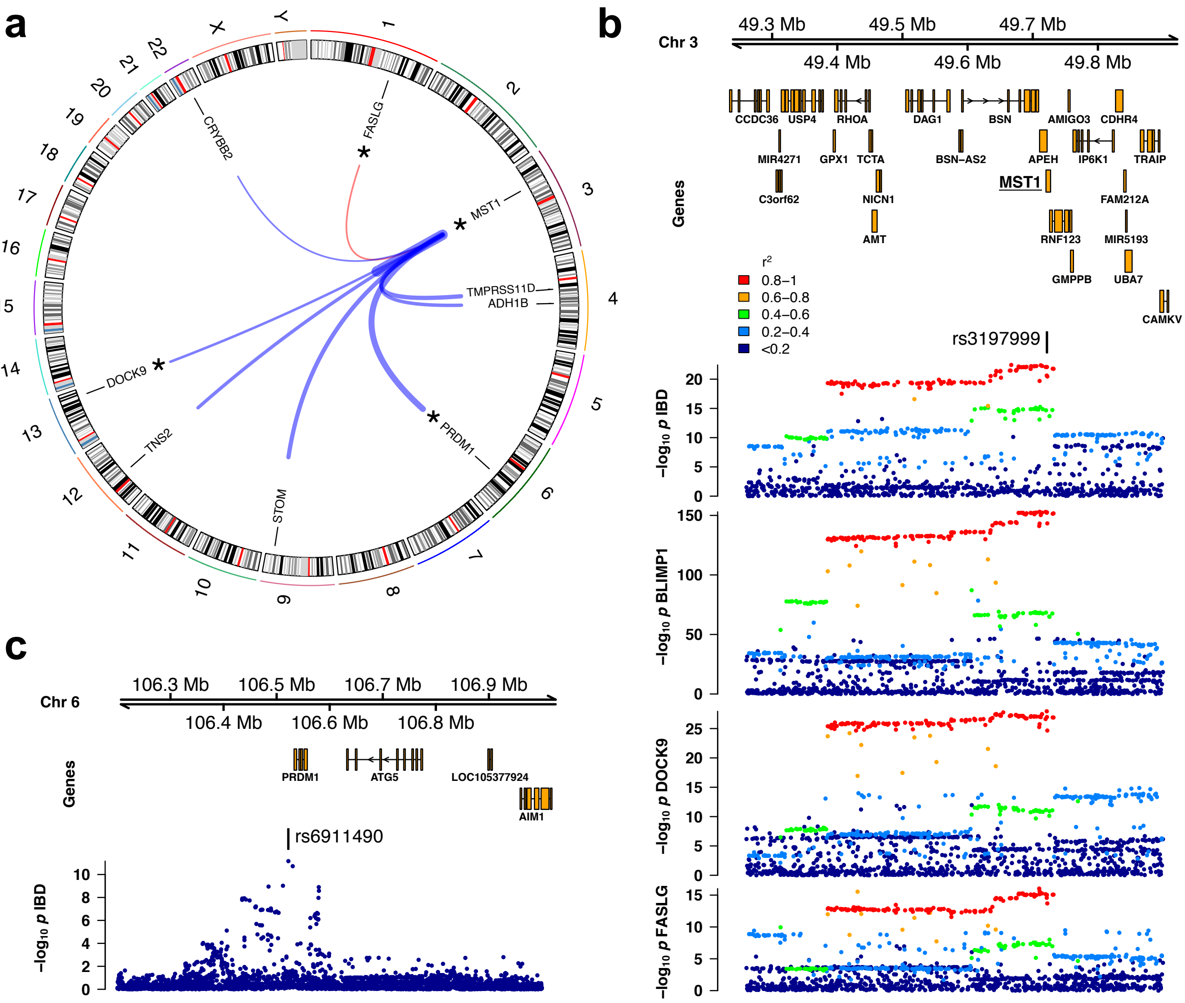


**Figure 4. *SERPINA1*, *PRTN3* and vasculitis.**

a) Manhattan plots for GWAS of plasma PR3 measured with the two SOMAmers and the Olink assay, showing the *cis* pQTL at *PRTN3* (which encodes PR3) for all three PR3 assays and the *SERPINA1 trans* pQTL for SOMAmer PRTN3.3514.49.2.
b) Regional association plots at the *PRTN3* region for the two PR3 SOMAmers and the Olink PR3 assay. LD to the sentinel variant rs10425544 is indicated by the colour key. 'Vasculitis GWAS' track shows the variants reported in GWASs of ANCA-associated vasculitis. VCRCi= rs138303849, most significant imputed variant from the Vasculitis Clinical Research Consortium[78]; VCRCt = rs62132293, directly genotyped SNP reported by the VCRC; EVGC= rs62132295, variant reported by the European Vasculitis Genetics Consortium[39] (see Supplementary Note). 'Independent pQTL' track indicates the position of conditionally independent PR3 pQTL variants (black lettering = lead variant for both SOMAmers; purple and green = conditionally independent variants for SOMAmer PRTN3.3514.49.2 and PRTN3.13720.95.3, respectively).
c) Proposed mechanisms by which *PRTN3* and *SERPINA1* impact PR3 levels and therefore influence vasculitis risk. Left panel: individuals without either the *PRTN3* or the *SERPINA1* vasculitis risk alleles. Middle panel: in individuals with the *SERPINA1* Z-allele, A1AT polymerises and is accumulated intracellularly resulting in reduced secretion into the circulation. As a consequence of reduced circulating A1AT, plasma free PR3 is increased. Right panel: individuals with rs7254911:G, a *cis* pQTL upstream of *PRTN3,* have higher circulating levels of total PR3. Increases in either free or total PR3 predispose to loss of immune tolerance, with increased formation of anti-PR3 antibodies and risk of vasculitis.

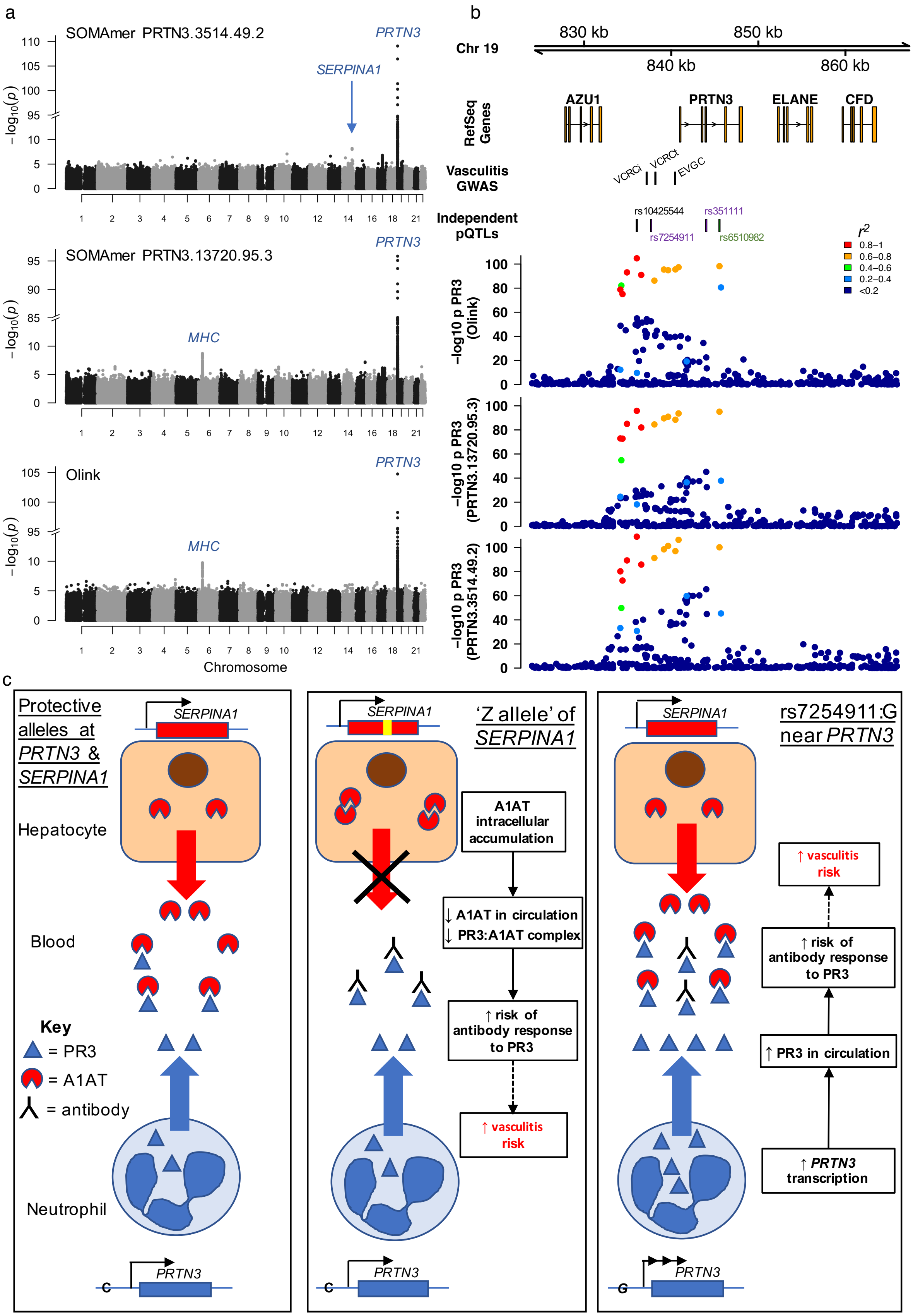
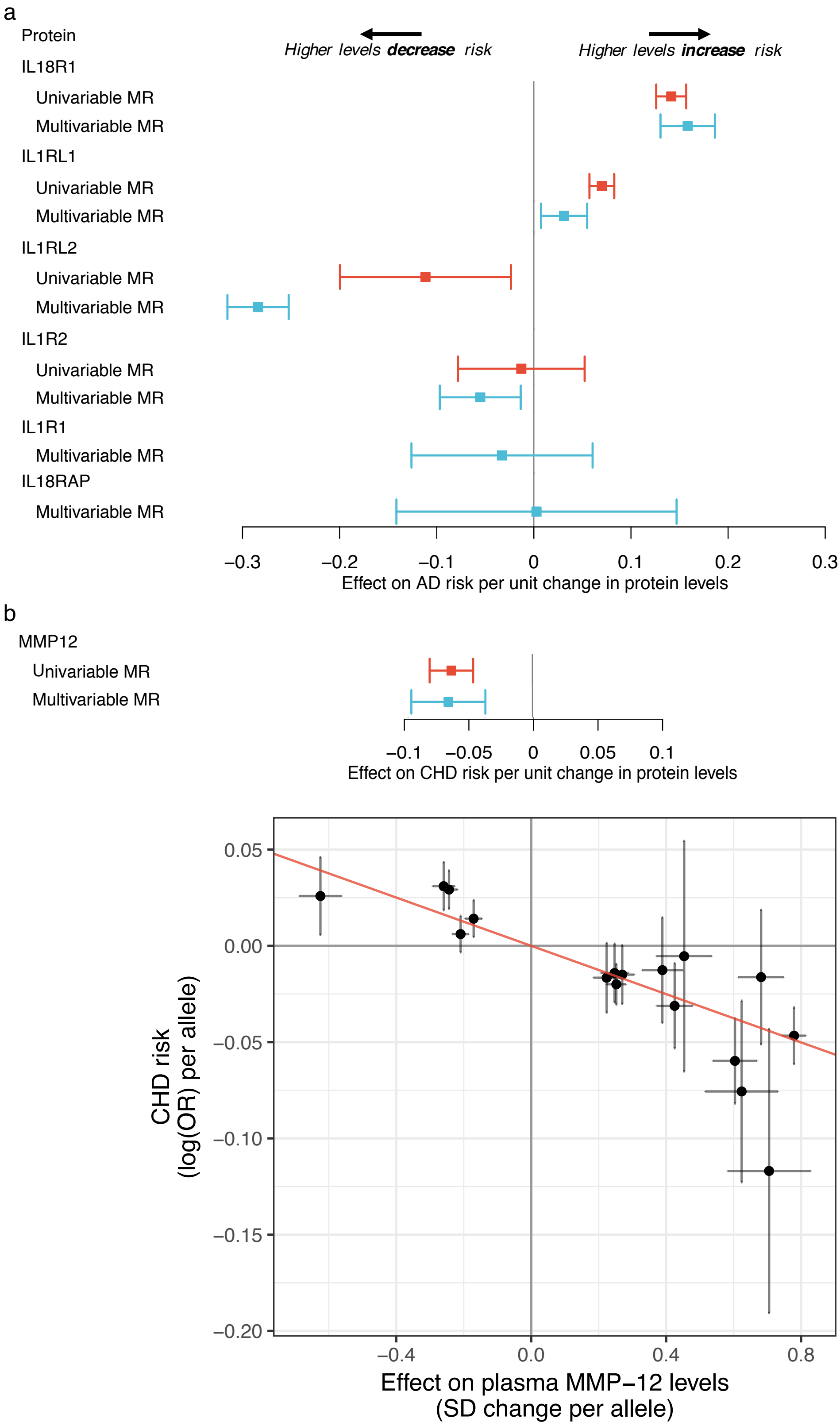

a
SOMAmer PRTN3.3514.49.2
PRTN3
SERPINA1
−log10(p)
110
105
100
95
10
5
0
1
2
3
4
5
6
7
8
9
10
12
14
16
18
21
SOMAmer PRTN3.13720.95.3
PRTN3
MHC
95
90
85
10
5
0
Olink
PRTN3
MHC
105
100
95
10
5
0
Chromosome
b
Chr 19
830 kb
850 kb
840 kb
860 kb
RefSeq Genes
AZU1
PRTN3
ELANE
CFD
Vasculitis GWAS
VCRCi
VCRCi
EVGC
Independent pQTLs
rs10425544
rs7254911
rs351111
rs6510982
$r^2$
0.8–1
0.6–0.8
0.4–0.6
0.2–0.4
<0.2
−log10 p PR3 (Olink)
−log10 p PR3 (PRTN3.13720.95.3)
−log10 p PR3 (PRTN3.3514.49.2)
100
80
60
40
20
0
c
Protective alleles at PRTN3 & SERPINA1
SERPINA1
Hepatocyte
Blood
Key
= PR3
= A1AT
= antibody
Neutrophil
PRTN3
C
'Z allele' of SERPINA1
A1AT intracellular accumulation
↓ A1AT in circulation
↓ PR3:A1AT complex
↑ risk of antibody response to PR3
↑ vasculitis risk
rs7254911:G near PRTN3
↑ vasculitis risk
↑ risk of antibody response to PR3
↑ PR3 in circulation
↑ PRTN3 transcription
G

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

Forest plot of univariable and multivariable Mendelian randomization (MR) estimates. (a) Proteins in the *IL1RL1-IL18R1* locus and risk of atopic dermatitis (AD). No univariable MR estimates available for IL1R1 and IL18RAP due to no significant pQTLs to select as a “genetic instrument”. (b) MMP-12 levels and risk of coronary heart disease (CHD). Above: MR estimates. Below: estimated effects (with 95% confidence intervals) on plasma MMP-12 and CHD risk for each variant used in the genetic score.


a
Protein
Higher levels decrease risk
Higher levels increase risk
P−Value
IL18R1
Univariable MR
9.3e−72
Multivariable MR
1.5e−28
IL1RL1
Univariable MR
5.7e−27
Multivariable MR
0.01
IL1RL2
Univariable MR
0.013
Multivariable MR
1.1e−69
IL1R2
Univariable MR
0.7
Multivariable MR
0.0094
IL1R1
Multivariable MR
0.49
IL18RAP
Multivariable MR
0.97
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
Effect on AD risk per unit change in protein levels
b
MMP12
P−Value
Univariable MR
2.8e−13
Multivariable MR
8.6e−06
−0.1
−0.05
0
0.05
0.1
Effect on CHD risk per unit change in protein levels
CHD risk
(log(OR) per allele)
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.4
0.0
0.4
0.8
Effect on plasma MMP−12 levels
(SD change per allele)