METHODS

Study participants. The INTERVAL study comprises about 50,000 participants nested within a randomized trial of varying blood donation intervals[9]. Between mid-2012 and mid-2014, blood donors aged 18 years and older were recruited at 25 centres of England's National Health Service Blood and Transplant (NHSBT). All participants gave informed consent before joining the study and the National Research Ethics Service approved this study (11/EE/0538). Participants completed an online questionnaire including questions about demographic characteristics (for example, age, sex, ethnicity), anthropometry (height, weight), lifestyle (for example, 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 SomaLogic assays, we randomly selected two non-overlapping subcohorts of 2,731 and 831 participants from INTERVAL. After genetic quality control, 3,301 participants (2,481 and 820 in the two subcohorts) remained for analysis (Supplementary Table 17). No statistical methods were used to determine sample size. The experiments were not randomized. Laboratory staff conducting proteomic assays were blinded to the genotypes of participants.

Plasma sample preparation. Sample collection procedures for INTERVAL have been described previously[38]. In brief, blood samples for research purposes were collected in 6-ml EDTA tubes using standard venepuncture protocols. The tubes were inverted three times and transferred at ambient temperature to UK Biocentre (Stockport, UK) for processing. Plasma was extracted into two 0.8-ml plasma aliquots by centrifugation and subsequently stored at −80 °C before use.

Protein measurements. We used a multiplexed, aptamer-based approach (SOMAscan assay) to measure the relative concentrations of 3,622 plasma proteins or 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 (for example, 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[8,39] (Extended Data Fig. 10a). The proteins cover a wide range of molecular functions (Extended Data Fig. 10b). 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 the 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[39,40] 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 levels using control aptamers and calibrator samples. At the sample level, hybridization controls on the microarray were used to correct for systematic variability in hybridization, 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 hybridization scale factors and median scale factors were used to normalize 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 the coefficient of variation (CV) 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 $\leq$ 20% in both subcohorts. Eight non-human protein targets were also excluded, leaving 3,283 SOMAmers (mapping to 2,994 unique proteins or 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)[41]. Protein subcellular locations were determined by exporting the subcellular location annotations from UniProt[42]. 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[43] 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.[44,45]. We used Bonferroni-corrected P value thresholds of $P = 1.1 \times 10^{-3}$ (0.05/45) and $P = 1.2 \times 10^{-3}$ (0.05/40), respectively. Relative protein abundances were rank-inverse normalized within each subcohort and linear regression was performed using age, sex, body mass index, 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 ($n \approx 50{,}000$) have been described previously in detail[10]. 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. Relatedness was removed by excluding one participant from each pair of close (first- or second-degree) relatives, defined as $\widehat{\pi} > 0.187$. Identity-by-descent was estimated using a subset of variants with a call rate >99% and MAF > 5% in the merged data set of both subcohorts, pruned for linkage disequilibrium (LD) using PLINK v1.9[46]. Numbers of participants excluded at each stage of the genetic QC are summarized in Extended Data Fig. 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) and resulted 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 in which 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 Fig. 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, $\leq$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 normalized 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[47].

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[48]. Genetic associations were considered to be genome-wide significant based on a conservative strategy requiring associations to have (i) a meta-analysis P value $< 1.5 \times 10^{-11}$ (genome-wide threshold of $P = 5 \times 10^{-8}$ 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 Fig. 3d).

Refinement of significant regions. To identify distinct non-overlapping regions associated with a given SOMAmer, we first defined a 1-Mb 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 1-Mb regions remained. The variant with the lowest P value for each region was assigned as the 'regional sentinel variant'. Owing to the complexity of the 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 ($r^2 \geq 0.8$) with each other were combined together into a single region.

Conditional analyses. To identify conditionally significant associations, we performed approximate genome-wide stepwise conditional analysis using GCTA v1.25.2[49] using the 'cojo-slct' option. We used the same conservative significance threshold of $P = 1.5 \times 10^{-11}$ as for the univariable analysis. As inputs for GCTA, we used the summary statistics (that is, 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 data set, and only variants also passing the univariable genome-wide threshold ($P < 1.5 \times 10^{-11}$) were considered for stepwise selection. As the conditional analyses use different data inputs to the univariable analysis (that is, summarized rather than individual-level data), there were some instances where the conditional analysis failed to include in the stepwise 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 stepwise 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[50] (https://www.ebi.ac.uk/gwas/, downloaded November 2017), which has all phenotypes mapped to the Experimental Factor Ontology (EFO)[51], by restricting to those with EFO annotations relevant to protein biomarkers (for example, '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 was *cis* or *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 whether the sentinel SNP or a proxy ($r^2 > 0.6$) was associated with the same protein (or 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, that is, a proximity extension assay method (Olink Bioscience, Uppsala, Sweden)[13] 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. Normalized 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 normalized 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 ($P < 3.1 \times 10^{-4}$; 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 1 Mb of the TSS of the gene(s) encoding the protein. pQTLs lying outside of this region were defined as *trans*. When considering the distance of the lead *cis*-associated variant from the relevant TSS, only proteins that mapped to single genes on the primary assembly in Ensembl v83 were considered.

For *trans* pQTLs, we sought to prioritize candidate genes in the region that might underpin the genotype–protein association. We applied the ProGeM framework[22], which 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[28] (http://www.gtexportal.org/home/datasets). Third, we also queried promoter capture Hi-C data in 17 human primary haematopoietic cell types[52] 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 (that is, 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 ± 5 kb 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[52] 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[53] 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[54] (http://www.omim.org/); (2) the Kyoto Encyclopedia of Genes and Genomes (KEGG)[55] (http://www.genome.jp/kegg/); and (3) STRINGdb[56] (http://string-db.org/; v10.0). We accessed OMIM data via HumanMine web tool[57] (http://www.humanmine.org/; accessed June 2016), whereby we extracted all OMIM IDs for (i) our *trans*-affected proteins and (ii) genes local (± 500 kb) 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[58], 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 using 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 prioritized the relevant local genes as being potential mediators of the *trans* association(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 equally sized sets of null permuted variants were sampled matching for MAF (bins of 5%), distance to TSS (bins of 0–0.5 kb, 0.5–2 kb, 2–5 kb, 5–10 kb, 10–20 kb, 20–100 kb and >100 kb 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 (for example, 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 owing 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)[59].

Testing for regulatory and functional enrichment. We tested whether our pQTLs were enriched for functional and regulatory characteristics using GARFIELD v1.2.0[60]. 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 1-Mb 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 $P < 5 \times 10^{-5}$ (0.05/998) to be statistically significant.

Disease annotation. To identify diseases with which our pQTLs have been associated, we queried our sentinel variants and their strong proxies ($r^2 \geq 0.8$) against publicly available disease GWAS data using PhenoScanner[61]. A list of data sets queried is available at http://www.phenoscanner.medschl.cam.ac.uk/information.html. For disease GWAS, results were filtered to $P < 5 \times 10^{-8}$ and then manually curated to retain only the entry with the strongest evidence for association (that is, 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 ($r^2 \geq 0.8$) were queried against publicly available eQTL association data using PhenoScanner. *cis* eQTL results were filtered to retain only variants with $P < 1.5 \times 10^{-11}$. 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 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* (that is, within 1 Mb) of genes that encode proteins analysed in our study and only variants in both data sets were used.

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 ($P < 1.5 \times 10^{-11}$), conventional genome-wide significance ($P < 5 \times 10^{-8}$) or a nominal *P* value threshold ($P < 1 \times 10^{-5}$) for the same protein or 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 ($P < 1.5 \times 10^{-11}$), 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.

Colocalization analysis. Colocalization testing was performed using the coloc package[62]. For testing colocalization of pQTLs and disease associations, colocalization testing was necessarily limited to disease traits for which full GWAS summary statistics had been made available. We obtained GWAS summary statistics through PhenoScanner. For testing colocalization of pQTLs with eQTLs, we used publically available summary statistics for expression traits from GTEx[28]. We used the default priors. Regions for testing were determined by dividing the genome into 0.1-cM chunks using recombination data. Evidence for colocalization was assessed using the posterior probability (PP) for hypothesis 4 (that there is an association for both traits and they are driven by the same causal variant(s)). Associations with PP4 > 0.5 were deemed likely to colocalize as this gives hypothesis 4 the highest likelihood of being correct, while PP4 > 0.8 was deemed to be 'highly likely to colocalize'.

Selection of genetic instruments for Mendelian randomization. In MR, genetic variants are used as 'instrumental variables' (IVs) for assessing the causal effect of the exposure (here a plasma protein) on the outcome (here a disease)[11,63] (Extended Data Fig. 9).

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)[31], 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 $P < 1 \times 10^{-4}$ 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 of AD[31]. Where the relevant variant was not available, the strongest proxy with $r^2 \geq 0.8$ 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 $P < 5 \times 10^{-8}$ 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 consisted mostly (77%) of individuals of European ancestry[64].

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[65,66]. For each of *G* genetic variants ($g = 1, \ldots, G$) having per-allele estimate of the association with the protein β_{Xg} and standard error σ_{Xg}, and per-allele estimate of the association with the outcome (here, AD or CHD) β_{Yg} and standard error σ_{Yg}, the IV estimate ($\widehat{\theta}_{XY}$) is obtained from generalized weighted linear regression of the genetic associations with the outcome (β_Y) on the genetic associations with the protein (β_X) weighting for the precisions of the genetic associations with the outcome and accounting for correlations between the variants according to the regression model:

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

where β_Y and β_X are vectors of the univariable (marginal) genetic associations, and the weighting matrix Ω has terms $\Omega_{g_1g_2} = \sigma_{Yg_1}\sigma_{Yg_2}\rho_{g_1g_2}$, and $\rho_{g_1g_2}$ is the correlation between the g_1th and g_2th variants.

The IV estimate from this method is:

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

and the standard error is:

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

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

To account for potential effects of functional pleiotropy[67], we performed multivariable MR using the weighted regression-based method proposed by

Burgess et al.[68]. For each of K risk factors in the model ($k = 1,\ldots,K$), the weighted regression-based method is performed by multivariable generalized weighted linear regression of the association estimates β_Y on each of the association estimates with each risk factor β_{Xk} in a single regression model:

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

where β_{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[13]) 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[3]. 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 organized into a hierarchy with five levels. We mapped each GWAS trait and indication onto the 'lowest level terms' (that is, the most specific terms available). All matching terms were recorded for each trait or indication. We matched GWAS traits to drug indications on the basis of 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.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

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 are also publically available at http://www.phpc.cam.ac.uk/ceu/proteins/, through PhenoScanner (http://www.phenoscanner.medschl.cam.ac.uk) and from the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/downloads/summary-statistics).

38. 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).
39. Gold, L. et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE* **5**, e15004 (2010).
40. Sattlecker, M. et al. Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. *Alzheimers Dement*. **10**, 724–734 (2014).
41. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
42. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. **43**, D204–D212 (2015).
43. Ashburner, M. et al.; The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet*. **25**, 25–29 (2000).
44. Menni, C. et al. Circulating proteomic signatures of chronological age. *J. Gerontol. A Biol. Sci. Med. Sci*. **70**, 809–816 (2015).
45. Ngo, D. et al. Aptamer-based proteomic profiling reveals novel candidate biomarkers and pathways in cardiovascular disease. *Circulation* **134**, 270–285 (2016).
46. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
47. 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).
48. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
49. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet*. **44**, 369–375 (2012).
50. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. **42**, D1001–D1006 (2014).
51. Malone, J. et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).
52. 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).
53. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLOS Comput. Biol*. **9**, e1003118 (2013).
54. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. **43**, D789–D798 (2015).
55. 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).
56. Szklarczyk, D. et al. STRINGv10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. **43**, D447–D452 (2015).
57. 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).
58. Franceschini, A. et al. STRINGv9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. **41**, D808–D815 (2013).
59. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
60. Iotchkova, V. et al. GARFIELD—GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. Preprint at https://www.biorxiv.org/content/early/2016/11/07/085738 (2016).
61. Staley, J. R. et al. PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
62. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. **10**, e1004383 (2014).
63. Hingorani, A. & Humphries, S. Nature's randomised trials. *Lancet* **366**, 1906–1908 (2005).
64. Nikpay, M. et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet*. **47**, 1121–1130 (2015).
65. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol*. **37**, 658–665 (2013).
66. 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**, 1880–1906 (2016).
67. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol*. **181**, 251–260 (2015).
68. 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**, 290–291 (2015).