RESEARCH ARTICLE SUMMARY

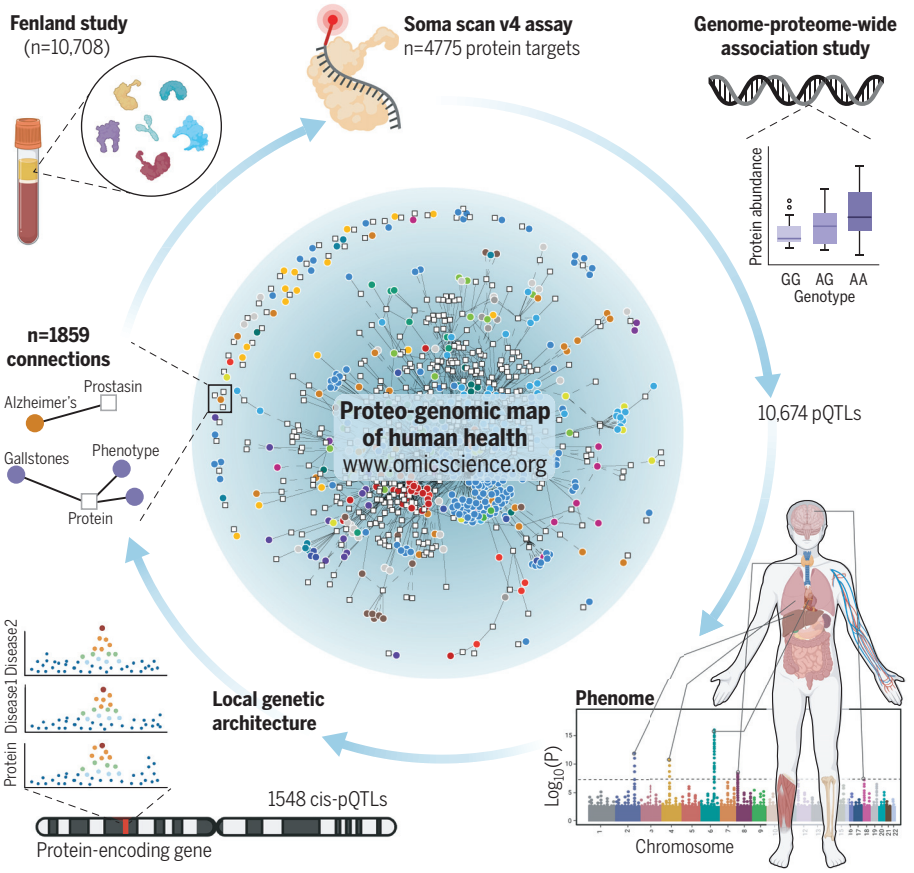
DISEASE GENOMICS

Mapping the proteo-genomic convergence of human diseases

Maik Pietzner†, Eleanor Wheeler†, Julia Carrasco-Zanini, Adrian Cortes, Mine Koprulu, Maria A. Wörheide, Erin Oerton, James Cook, Isobel D. Stewart, Nicola D. Kerrison, Jian'an Luan, Johannes Raffler, Matthias Arnold, Wiebke Arlt, Stephen O'Rahilly, Gabi Kastenmüller, Eric R. Gamazon, Aroon D. Hingorani, Robert A. Scott, Nicholas J. Wareham, Claudia Langenberg*

INTRODUCTION: Proteins are essential functional units of the human body and represent the largest class of drug targets.

RATIONALE: Broad-capture proteomics has the potential to identify causal disease genes, mechanisms, and candidate drug targets through systematically integrating knowledge about genetic signals that are shared among the protein-encoding gene, the resulting protein abundance or function, and common complex diseases. Although technological advances now enable such enquiry at scale, the genetic architecture of most proteins and its relevance for human health remains unknown. We performed a genome-proteome–wide association study including 4775 protein targets measured in plasma from 10,708 European-descent individuals (mean age 48.6 years, 53.3% women). We used the identified protein–quantitative trait loci (pQTLs) to create a proteo-genomic map of human health based on shared, colocalized genetic architecture tested across thousands of phenotypes at protein-encoding loci (cis-pQTLs).



Summary of the study design (outer circle) to construct a proteo-genomic map (inner circle) of human health. Connections between protein-encoding genes, proteins, diseases, and phenotypes were drawn for all examples with strong evidence of a shared genetic signal based on statistical colocalization (posterior probability > 80%). Parts of the figure were generated using BioRender.com.

RESULTS: We identified 10,674 genetic variant–protein target associations ($P < 1.004 \times 10^{-11}$) distributed across 2548 genomic regions (1097 unreported) and covering 3892 distinct protein targets. Of 1538 protein targets with at least one cis-pQTL, we found that half share a genetic signal with gene expression in at least one of 49 tissues; alternative splicing events account for about one-fifth of those, demonstrating the utility of plasma proteomics as a means to infer tissue effects. We demonstrated that cis-pQTLs helped to prioritize candidate causal genes at 558 established risk loci for 537 collated phenotypes. For one-fourth of these (24.6%), this included genes not reported or different from those prioritized by gene expression QTLs, including *PRSS8* (encoding prostasin) for Alzheimer's disease or *RSPO1* (encoding R-spondin–1) for endometrial cancer. We created a cis-anchored proteo-genomic map of human health including 1859 gene-protein-phenotype connections comprising 412 proteins and 506 curated traits. The map highlighted strong cross-disease biological convergence. For example, the genetic signal at *EFEMP1* (EGF-containing fibulin-like extracellular matrix protein 1) was shared across diverse connective tissue disorders consistent with abnormal elastic fiber morphology of the *Efemp1* knockout mouse. Integration of diverse "omic" layers identified a supersaturated bile to promote cholesterol crystallization and gallstone formation as the mode of action at *SULT2A1*. We developed an approach to classify pQTLs by integrating ontology mapping with a data-derived protein network. This showed that 39% (n = 2302) of trans-pQTLs (i.e., those distant from the protein-encoding gene) were protein- or pathway-specific and identified established risk loci, such as rs738409 (*PNPLA3*), an established liver fibrosis locus, to act on several proteins that are all part of a specific protein community. We developed an interactive web resource (www.omicscience.org/apps/pgwas) to facilitate rapid access and interrogation to our results.

CONCLUSION: Genetically anchored plasma proteomics identifies shared etiologies across diseases, enables prioritization of drug targets, and provides a systems biology context for gene-to-phenotype and protein-to-phenotype connections. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: claudia.langenberg@mrc-epid.cam.ac.uk
†These authors contributed equally to this work.
Cite this article as M. Pietzner *et al.*, *Science* **374**, eabj1541 (2021). DOI: 10.1126/science.abj1541

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abj1541

RESEARCH ARTICLE

DISEASE GENOMICS

Mapping the proteo-genomic convergence of human diseases

Maik Pietzner[1,2]†, Eleanor Wheeler[1]†, Julia Carrasco-Zanini[1], Adrian Cortes[3], Mine Koprulu[1], Maria A. Wörheide[4], Erin Oerton[1], James Cook[1], Isobel D. Stewart[1], Nicola D. Kerrison[1], Jian'an Luan[1], Johannes Raffler[4,5], Matthias Arnold[4,6], Wiebke Arlt[7], Stephen O'Rahilly[8], Gabi Kastenmüller[4,9], Eric R. Gamazon[10,11], Aroon D. Hingorani[12,13,14], Robert A. Scott[3], Nicholas J. Wareham[1,13], Claudia Langenberg[1,2,13]*

Characterization of the genetic regulation of proteins is essential for understanding disease etiology and developing therapies. We identified 10,674 genetic associations for 3892 plasma proteins to create a cis-anchored gene-protein-disease map of 1859 connections that highlights strong cross-disease biological convergence. This proteo-genomic map provides a framework to connect etiologically related diseases, to provide biological context for new or emerging disorders, and to integrate different biological domains to establish mechanisms for known gene-disease links. Our results identify proteo-genomic connections within and between diseases and establish the value of cis-protein variants for annotation of likely causal disease genes at loci identified in genome-wide association studies, thereby addressing a major barrier to experimental validation and clinical translation of genetic discoveries.

Proteins are the central layer of information transfer from the genome to the phenome, and recent studies have started to elucidate how natural sequence variation in the human genome affects protein concentrations measured from readily available biofluids such as blood (*1*–*6*). Investigation of the clinical consequences of these so-called protein–quantitative trait loci (pQTLs) can help to better explain disease mechanisms and provide insights into the shared genetic architecture across diseases within a translational framework that puts humans as the model organism at the center (*2*, *4*). This approach is now pursued at scale by pharmaceutical companies for the discovery of drug targets or repurposing opportunities (*7*, *8*). Earlier studies have used bespoke panels (*3*, *6*, *9*) or larger proteomic platforms (*1*, *2*, *4*, *5*) to characterize the genetic architecture of proteins, and their results have shown how these strategies can provide insight into the pathogenesis of specific diseases. Less attention has been given to (i) providing a framework to assess the protein specificity of genetic variation residing outside (trans) the protein-encoding gene; (ii) understanding the clinical relevance of pQTLs for proteins detected in plasma but not known to be actively secreted (*7*); (iii) classifying thousands of proteins according to their genetic architecture as explained by cis variants, specific trans variants, or unspecific trans variants; (iv) demonstrating the specific utility of pQTLs for the prioritization of candidate genes at established risk loci; and (v) systematically mapping shared gene-protein-disease signals to uncover connections among thousands of considered diseases and other phenotypes.

Profiling thousands of proteins circulating in blood at population scale is currently possible only with the use of large libraries of affinity reagents—antibodies or short oligonucleotides called aptamers—because gold-standard methods such as mass spectrometry lack throughput. We previously provided a detailed comparison of 871 overlapping proteins measured in 485 individuals (*10*) using the two most comprehensive platforms, the aptamer-based SomaScan v4 assay and the antibody-based Olink proximity extension assay. We demonstrated that the majority of pQTLs are consistent across platforms (64%), in line with smaller-scale efforts (*4*), but also highlighted the need to triangulate pQTLs with gene expression and phenotypic information to derive tangible biological hypotheses. Here, we present a genome-proteome–wide association study targeting 4775 distinct proteins measured from plasma samples of 10,708 generally healthy European-descent individuals who were participants in the Fenland study (table S1) (*11*). We identified 10,674 variant-protein associations and developed a framework to systematically identify protein- and pathway-specific pQTLs augmenting current ontology-based classifications in a data-driven manner. We found that half of all pQTLs close to the protein-encoding gene—cis-pQTLs—colocalize with gene expression or splicing QTLs in various tissues, allowing us to derive functional insights within tissues by integrating genetics with plasma proteomics. Moreover, cis-pQTLs have the specific ability to prioritize candidate causal genes at established genetic risk loci. By means of phenome-wide colocalization screens, we generated a proteo-genomic map of human health covering 1859 gene-protein-phenotype triplets, which provides insights into the shared etiology across diseases and the identification of pathophysiological pathways through cross-domain integration.

Genetic associations for protein targets

We performed a genome-proteome–wide association analysis by testing 10.2 million genotyped or imputed autosomal and X-chromosomal genetic variants with minor allele frequency (MAF) of >1% among 10,708 participants in the Fenland study measuring 4775 distinct proteins (*12*). We identified 2584 genomic regions (1543 within ±500 kb of the protein-encoding genes, i.e., cis) associated with at least one of 3892 protein targets at $P < 1.004 \times 10^{-11}$. Of these regions, 1097 covered variants that have not previously been reported to be associated with plasma proteins (*1*–*6*, *9*) ($r^2 < 0.1$), of which 64% (867 of 1356 pQTLs) available in (*4*) replicated ($P < 0.05$, directionally consistent). Further, 61% of pQTLs (488 of 797; table S2) replicated using the complementary Olink technique (see supplementary materials), with a higher proportion of replication for variants in cis (81.2%) relative to trans (44.2%). Most regions (79.3%, $n = 2050$) were associated with a single protein target, but we observed pleiotropy (≥2 protein targets) at the remaining regions, including association with up to five (16.1%, $n = 418$), 6 to 20 (3.4%, $n = 88$), or 21 to 50 (0.7%, $n = 19$) associated protein targets, and substantial pleiotropy at eight regions (*CFH*, *ARF4-ARHGEF3*, *C4A-CFB*, *BCHE*, *VTN*, *CFD*, *ABO*, *GCKR*) associated with 59 to 1539 protein targets (Fig. 1). The 194 pleiotropic regions harboring a cis-pQTL identified master regulators of the plasma proteome, including glycosyltransferases such as the histo-blood

[1]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge CB2 0QQ, UK. [2]Computational Medicine, Berlin Institute of Health at Charité–Universitätsmedizin Berlin, 10117 Berlin, Germany. [3]GlaxoSmithKline, Stevenage SG1 2NY, UK. [4]Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany. [5]Institut für Digitale Medizin, Universitätsklinikum Augsburg, 86156 Augsburg, Germany. [6]Department of Psychiatry and Behavioral Sciences, Duke University, Durham, NC 27710, USA. [7]Institute of Metabolism and Systems Research, University of Birmingham, Birmingham B15 2TT, UK. [8]MRC Metabolic Diseases Unit, Wellcome Trust–Medical Research Council Institute of Metabolic Science, University of Cambridge, Cambridge CB2 0QQ, UK. [9]German Centre for Diabetes Research (DZD), 85764 Neuherberg, Germany. [10]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN 37203, USA. [11]Clare Hall, University of Cambridge, Cambridge CB3 9AL, UK. [12]UCL British Heart Foundation Research Accelerator, Institute of Cardiovascular Science, University College London, London WC1E 6BT, UK. [13]Health Data Research UK, Gibbs Building, London NW1 2BE, UK. [14]Institute of Health Informatics, University College London, London NW1 2DA, UK.
*Corresponding author. Email: claudia.langenberg@mrc-epid.cam.ac.uk
†These authors contributed equally to this work.

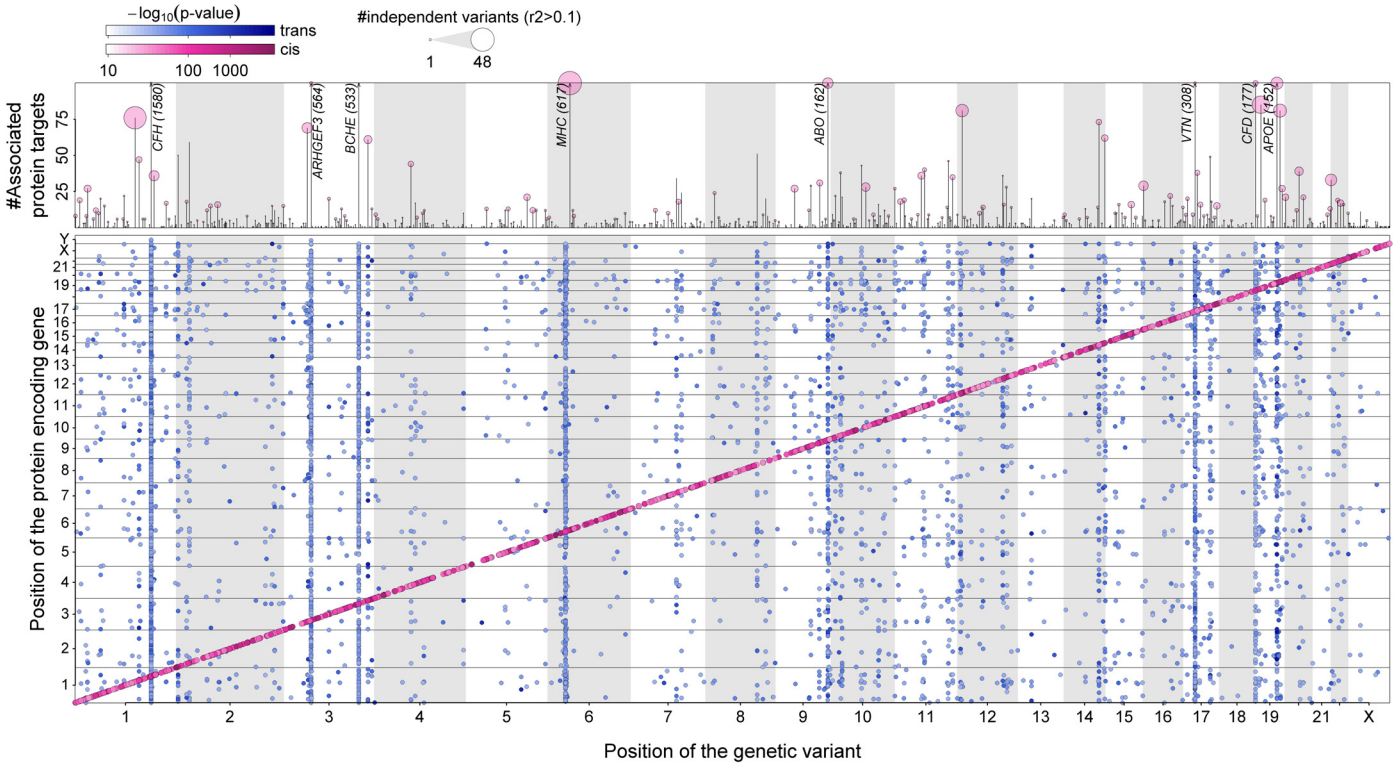


Fig. 1. Regional sentinel genetic variants associated ($P < 1.004 \times 10^{-11}$) with at least one protein target in up to 10,708 participants from the Fenland Study. The lower panel maps the genomic locations of the genetic variants against the genomic locations of the protein-encoding genes. Genetic variants close to the protein-encoding gene (±500 kb) are highlighted in pink (cis-pQTLs); all others are shown in blue (trans-pQTLs). Darker shades indicate more significant *P* values. The upper panel shows the number of associated protein targets for each genomic region (vertical line), with circles above representing the number of approximately independent genetic variants ($r^2 < 0.1$), such that larger circles indicate more genetic variants in the region.

group ABO system transferase (*ABO*), key metabolic enzymes such as glucokinase regulatory protein (*GCKR*), or lipid mediators such as apolipoprotein E, establishing a network-like structure of the circulating proteome (*1*).

Of the 3892 protein targets, 26.8% (n = 1046) had pQTLs in both cis and trans, 13.4% (n = 523) in cis only, and 59.6% (n = 2323) in trans only, among a total of 8328 sentinel variant-protein target associations (Fig. 1 and tables S2 and S3). We identified another 2346 secondary pQTLs at those loci by means of an adapted stepwise conditional analysis (median, 1; range, 1 to 13), indicating widespread allelic heterogeneity in cis (68.8%) and trans (31.2%). The majority of the 5442 distinct variants were located in introns (~44%) or were in high linkage disequilibrium (LD) ($r^2 > 0.6$) with a missense variant (~21%), with similar distributions across cis- and trans-pQTLs (fig. S1). We observed 663 cis-pQTLs with direct consequences for the structure of the protein target (protein-altering variants, PAVs), including important substructures such as disulfide bonds (4.2%), α helices (3.1%), and β strands (2.6%) (fig. S1). Such variants are predicted to affect correct folding of protein targets, including diminished secretion or reduced half-life in the bloodstream, rather than expression of the protein-encoding gene (*13*). For example, we observed an enrichment of PAVs among actively secreted proteins (*14*) (39.6% versus 33.7%, P = 0.04, χ^2 test), possibly indicating modulation of common posttranslational modifications such as glycosylation.

An integrated classification system for pathway-specific pQTLs

We integrated a data-driven protein network with ontology mapping [Gene Ontology (GO) terms; Fig. 2, A and B, and fig. S2] to distinguish pathway-specific pQTLs from those exerting effects on multiple unrelated targets (see supplementary materials) (*15*). We successfully assigned 40.8% (n = 1790 in cis, n = 423 in trans) of the 5442 genetic variants as protein-specific and 5.9% (n = 236 in cis, n = 86 in trans) as pathway-specific on the basis of converging evidence from the network and ontology mapping, and another 16.5% (n = 498 in cis, n = 402 in trans) to be likely pathway-specific based on either source. In total, 1802 protein targets had at least one (likely) specific pQTL in cis (n = 1385) or trans (n = 417). We classified 648 variants that would have been missed by ontology mapping as protein community–specific through our data-driven network approach. One example is rs738408 (*PNPLA3*), a non-alcoholic fatty liver disease variant (*16*) associated with 22 of 70 aptamers from the same protein community (Fig. 2C). *PNPLA3* encodes patatin-like phospholipase domain–containing protein 3 (PNPLA3), and rs738408 tags the missense variant rs738409 (I148M), which renders PNPLA3 resistant to ubiquitylation-mediated degradation and results in subsequent accumulation on hepatic lipid droplets, which in turn causes fatty liver disease (*17*). The associated protein targets included multiple metabolic and detoxification enzymes highly expressed in the liver, such as alcohol dehydrogenases, arginosuccinate lyase, bile salt sulfotransferase, or aminoacylase-1. Our results support the hypothesis that these enzymes might appear in plasma of otherwise healthy individuals only as a result of lipid overload–induced lysis of hepatocytes. The putative liver damage–specific effect, anchored on the *PNPLA3* trans-pQTL, makes those protein targets potential biomarker candidates, as opposed to the tissue-unspecific proteins currently used to identify fatty liver disease or liver injury in the clinic (*18*).

Contribution of cis and trans genetic architecture

We observed three major categories of protein targets based on the contribution of genetic

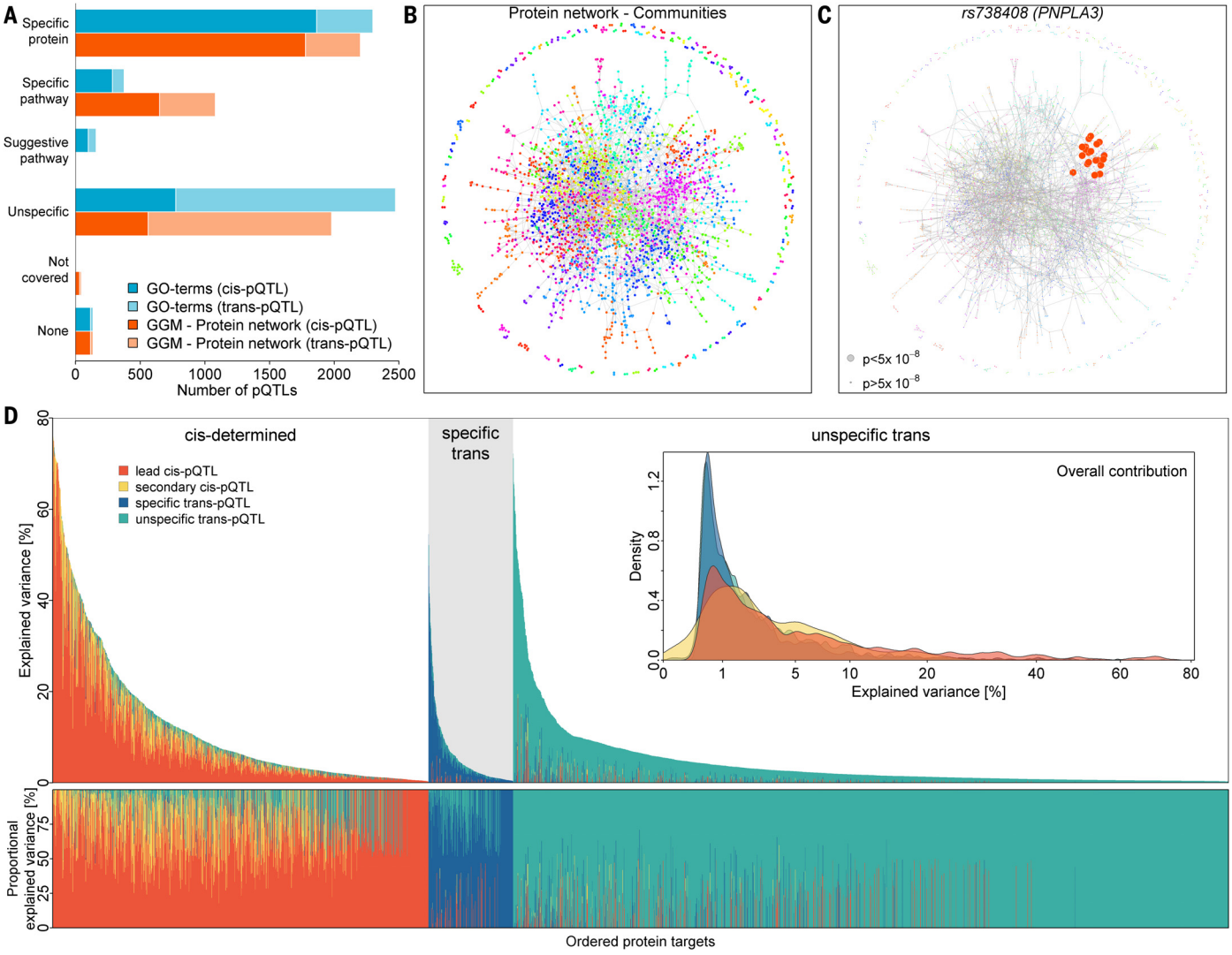


Fig. 2. Classification of protein quantitative trait loci (pQTLs, cis and trans) and subsequent partition of the explained variance in plasma abundances of protein targets. (**A**) Bar chart of pQTL classification based on GO term mapping (blue) or community mapping in a protein network derived by Gaussian graphical modeling (GGM; orange) of associated protein targets. Darker shades indicate cis-pQTLs and lighter colors trans-pQTLs. (**B**) Data-driven protein network colored according to 191 identified protein communities. (**C**) A community-specific pQTL (*PNPLA3*) that was not captured by GO term mapping. Gene annotation was as reported in the supplementary materials. (**D**) Absolute (top) and relative (bottom) explained variance in plasma abundances of protein targets by identified pQTLs. Coloring indicates contribution of the lead cis-pQTL (orange), secondary cis-pQTLs (yellow), protein- or pathway-specific trans-pQTLs (blue), and unspecific trans-pQTLs (green). Protein targets have been grouped by underlying genetic architecture as mostly explained by cis-pQTLs ("cis-determined"), mostly explained by specific trans-pQTLs ("specific trans"), and mostly explained by unspecific trans-pQTLs ("unspecific trans"). The inset displays the overall distribution of explained variance by each of the four categories. The variance explained was computed using linear regression models. See fig. S3 for a graphical display of effect size distributions.

variation to plasma concentrations (Fig. 2D, fig. S3, and table S3). For about one-third of the protein targets (n = 1249), >50% of genetic variance was explained by one or more cis-pQTLs, whereas for 7.2% of the targets (n = 282), protein- or pathway-specific trans-pQTLs accounted for most of the genetic variation; this left two-thirds of the targets (n = 2361) mainly explained by unspecific trans-pQTLs (*12*). Overall, we observed a median genetic contribution of 2.7% [interquartile range (IQR), 1.0% to 7.6%], reaching values above 70% for proteins such as vitronectin (rs704, MAF = 47.3%) or sialic acid–binding Ig-like lectin 9 (rs2075803, MAF = 44.1%) which were often driven by only a single common cis-pQTL. PAVs, which affect the binding epitope of the protein target, are the likely explanation for such strong and isolated genetic effects. Although more than two-thirds of the protein targets with at least one cis-pQTL were unrelated to PAVs, we found that 158 of the protein targets (32.9%) linked to a PAV ($r^2 > 0.6$) shared a genetic signal with at least one disease or risk factor (see below). This suggests that the conformation and possibly the function of the protein target, rather than the plasma abundance of the protein target, might be more relevant as mediators of downstream phenotypic consequences, and that aptamers are able to detect such probably dysfunctional proteins.

Our approach to identify protein-/pathway-specific trans-pQTLs allowed us to uncover biologically relevant information, which was otherwise hidden by strong and unspecific trans-pQTLs that possibly interfere with the measurement technique rather than the biology of the protein target. For example, rs704, a missense variant within *VTN* that associated with a higher fraction of single-chain vitronectin with altered binding properties (*19*, *20*) explained 72% of the variance in MICOS complex

subunit MIC10 (MOS1), far outperforming the contribution of the specific trans-pQTL rs398041972 (0.7%). rs398041972 resides about 1 Mb upstream of *TMEM11*, encoding transmembrane protein 11, a physical interaction partner of MOS1 as part of the MICOS complex (*21*). In general, we observed that the median contribution of specific trans-pQTLs to the variance in plasma concentrations was 1.1% (IQR, 0.6% to 2.6%) across 687 protein targets, reaching values as high as 38.3% for catenin β-1 via two trans-pQTLs (rs1392446 and rs35024584) within the same region for which we prioritized *CDH6* as a candidate causal gene. *CDH6* encodes cadherin 6, which physically interacts with catenin β-1 (*22*). We systematically tested for an enrichment of putative protein interaction partners among the 20 closest genes at each specific trans locus and observed a factor of 1.53 enrichment ($P = 1.8 \times 10^{-10}$, χ^2 test) of first- and second-degree neighbors from the STRING network (*23*), highlighting the ability of our classification system to identify biologically meaningful trans-pQTLs.

Shared genetic architecture with gene expression and splicing

We integrated plasma pQTL results with both gene expression and splicing QTL data (eQTL and sQTL, respectively) from the GTEx version 8 release (*24*) using statistical colocalization [posterior probability (PP) > 80%] for all 1584 protein targets with at least one cis-pQTL (*12*). There was strong evidence that half (50.1%) of these had a shared signal with gene expression in at least one tissue, with a median of 4.5 tissues (IQR, 2 to 12; Fig. 3A), vastly expanding our previous knowledge of gene expression contribution across tissues (*4*, *9*). The majority of cis-pQTLs ($n = 584$, 73.4%) showed plasma protein and gene expression effects in the same direction in all tissues (Fig. 3A), but 26.6% ($n = 212$) showed evidence of at least one pair with opposite effects, including 108 where the protein effect was opposite to the direction observed for gene expression across all tissues with evidence for colocalization. For example, the A-allele of the lead cis-pQTL rs2295621 for immunoglobulin superfamily member 8 (*IGSF8*) was inversely associated with plasma abundance of the protein target ($\beta = -0.19$, $P < 1.65 \times 10^{-32}$) but positively associated with expression of the corresponding mRNA across 33 tissues (table S4). Uncoupling of gene and protein expression, even within the same cell, is a frequently described phenomenon, and possible mechanisms include differential translation, protein degradation, contextual confounders such as time and developmental state, or protein-level buffering (*25*). For 145 protein targets, we identified strong evidence of a tissue-specific contribution to plasma abundances based on a single tissue strongly outweighing all others (Fig. 3A and table S4). These included known tissue-specific examples such as vitamin K–dependent protein C in liver tissue, but also less obvious ones, such as hepatitis A virus cellular receptor 1 (or TIM-1), an entry receptor for multiple human viruses, for which the cis-pQTL and cis-eQTL specifically colocalized in tissue from the transverse colon. To maximize power for the most closely aligned tissue compartment,

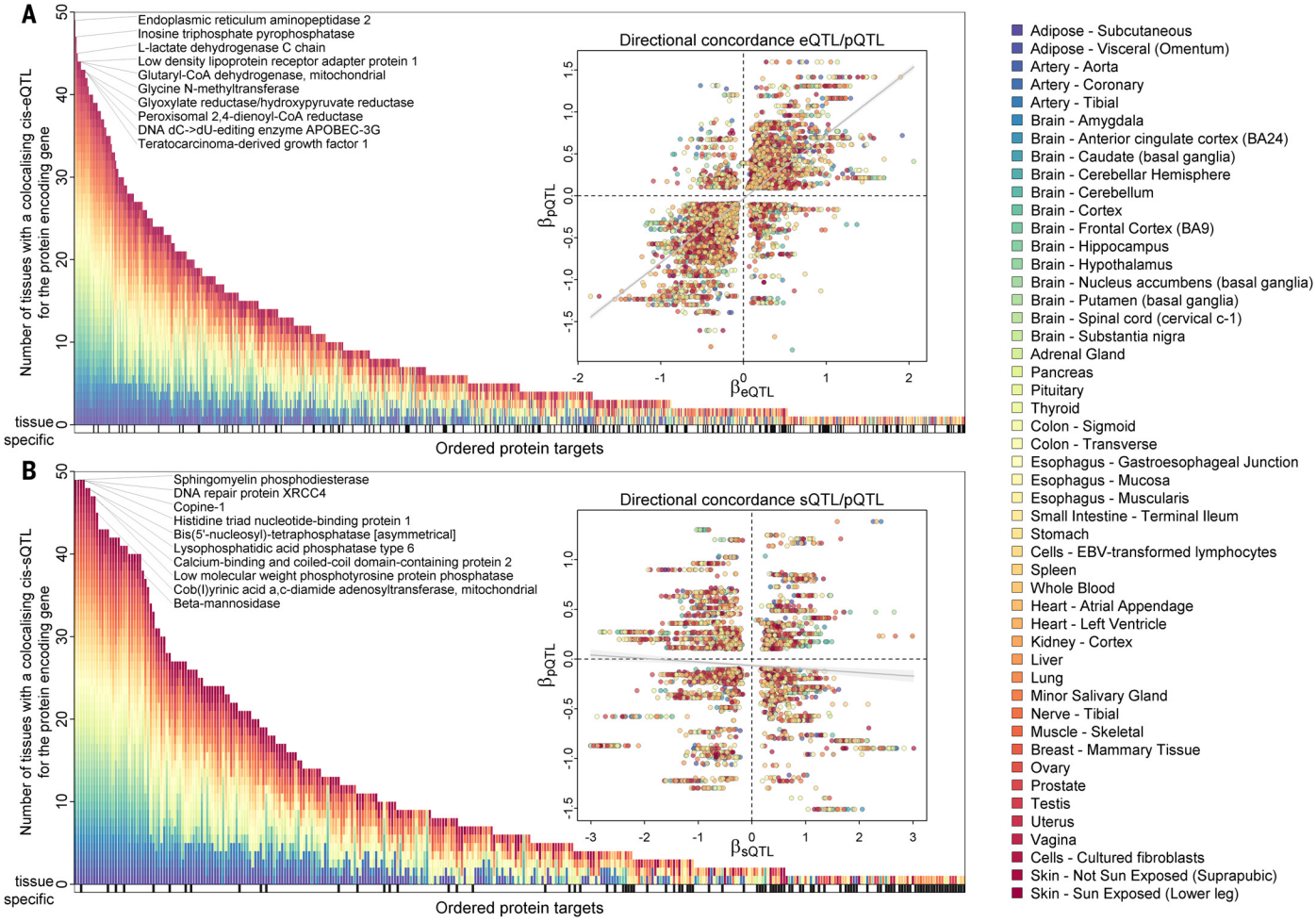


Fig. 3. Integration of gene and splicing quantitative trait loci (eQTLs and sQTLs). (**A**) Protein targets ordered by the number of tissues for which at least one of the cis-pQTLs was also a cis-eQTL as determined by statistical colocalization (posterior probability >80% for a shared signal). Protein targets for which the eQTL showed evidence for a tissue-specific effect are indicated by black vertical lines underneath. (**B**) Same as (A) but considering cis-sQTLs.

whole blood, we integrated gene expression data from the eQTLGen consortium (*26*), which confirmed 140 cis-eQTL/pQTL pairs and revealed another 38 cis-eQTL/pQTL pairs not seen in the GTEx resource, including immune cell–specific mediators of the inflammatory response such as leukocyte immunoglobulin-like receptor subfamily A member 3 (table S4).

To obtain insights beyond the average readout across all transcript species, we examined alternative splicing as a source of protein target variation (*12*). One-fifth (20.1%) of cis signals were shared with a cis-sQTL in at least one tissue (median, 6 tissues; IQR, 2 to 15) (Fig. 3B); 84 of these were not seen with eQTL data, which suggests that the pQTL-relevant transcript isoform was masked from the bulk of assayed transcripts. In contrast to the eQTL colocalization, we did not observe an overall pattern of aligning effect directions (Fig. 3B). This might be best explained by the intron-usage quantification of splicing events within GTEx version 8, which does not allow straightforward mapping of the eventually transcribed isoforms, and the expression of an alternative protein isoform with less affinity to the SOMAmer reagent. The latter may have accounted for the 90 protein target examples where the colocalizing cis-sQTL explained more than 10% of the variance in plasma concentrations (table S4) and emphasizes the ability of splicing QTLs to determine the underlying sources of variation in plasma abundances of protein targets. In summary, our results demonstrate that proteins measured in plasma can be used as proxies for tissue processes when anchored on a shared genetic variation with tissue-specific gene expression or alternative splicing data.

cis-pQTLs enable identification of candidate causal genes at GWAS loci

We used the inherent biological specificity of cis-pQTLs to systematically identify candidate causal genes for genome-wide significant variants reported in the genome-wide association studies (GWAS) catalog as of 25 January 2021 ($P < 5 \times 10^{-8}$) by assessing 558 cis-regions for which the pQTL was in strong LD ($r^2 > 0.8$) with at least one variant for 537 collated traits and diseases (Fig. 4 and table S5) (see supplementary materials) (*12*). For one-fourth of these (24.6%), we annotated a gene different from the reported or mapped gene, and for another 79 cis-regions (14.2%), our predicted causal gene was reported as part of a longer list of potential causal genes.

Among the genes we identified are candidates with strong biological plausibility, such as *AGRP*, encoding Agouti-related protein, a neuropeptide involved in appetite regulation (*27*), suggesting a possible mechanism for measures of body fat distribution associated at this locus. Another example was *NSF*, encoding *N*-ethylmaleimide–sensitive factor (NSF), which may be involved in the fusion of vesicles with membranes, enabling the release of

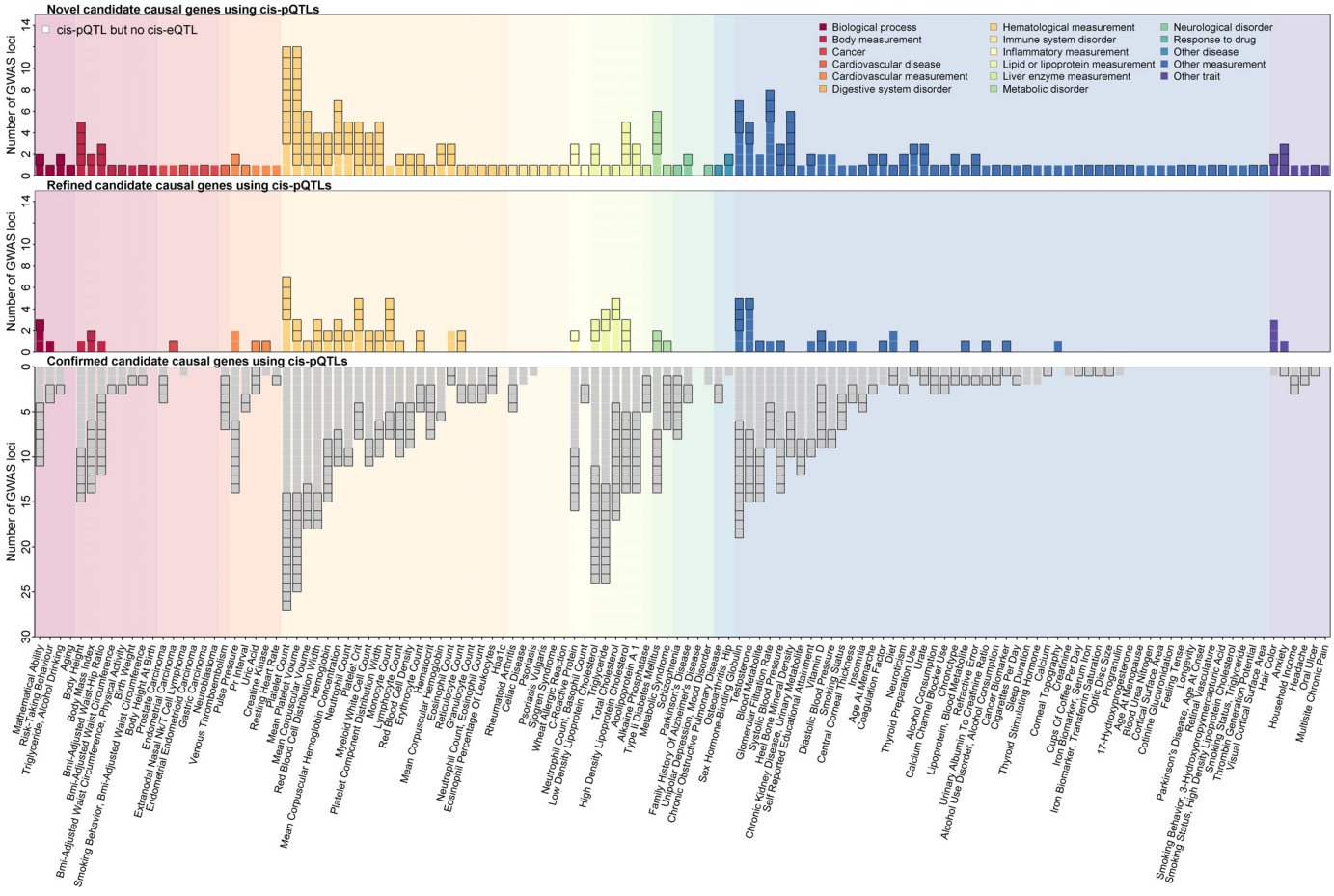


Fig. 4. Causal gene assignment for associations reported in the GWAS catalog using identified cis-pQTLs. Each panel displays the number of loci that have been reported in the GWAS catalog for a curated phenotype and were identified as protein quantitative trait in close proximity (±500 kb) to the protein-encoding gene (cis-pQTL) in the current study. Mapping of GWAS loci and cis-pQTLs was done using the LD between the reported variants ($r^2 > 0.8$). The upper panel displays the number of GWAS loci for which cis-pQTLs provided candidate causal genes. The middle panel displays the number of GWAS loci for which cis-pQTLs refined the list of candidate causal genes at the locus. The lower panel displays the number of GWAS loci with confirmative evidence from cis-pQTLs for already assigned candidate causal genes. Examples where gene prioritization was facilitated through pQTL but not gene expression QTL (eQTL) evidence are highlighted by a border around the box. Colors represent broad trait categories.

neurotransmitters into the extracellular space (28); this locus was previously identified for Parkinson's disease (table S5).

We further assigned *PRSS8* as a candidate causal gene at the *KAT8* locus for Alzheimer's disease (AD), supported by strong LD (r^2 = 0.96) and a high posterior probability of a shared genetic signal (98%) between the lead cis-pQTL (rs368991827, MAF = 27.8%) and the common *KAT8* intronic variant (rs59735493) that has been reported for AD (fig. S4). *PRSS8* codes for prostasin, and we estimated a 13% reduction in AD risk [odds ratio, 0.87; 95% confidence interval (CI), 0.82 to 0.91; $P = 3.8 \times 10^{-8}$] for each 1 SD higher normalized plasma abundance of prostasin. The locus has been identified by multiple GWAS efforts (29), yet prioritization strategies have failed to provide conclusive evidence for a causal gene (30). Prostasin is a serine protease highly expressed in epithelial tissue, which regulates sodium channels (31) and represses TLR4-mediated inflammation in human and mouse models of inflammatory bowel disease (32), a mechanism that might also be relevant to TLR4-mediated neuroinflammation in AD (33).

We observed multiple examples in which our cis-pQTL mapping identified biologically plausible candidate genes that were not implicated by cis-eQTL mapping (Fig. 4). For example, we assigned *RSPO1* as a candidate causal gene at the eQTL-supported *CDCA8* locus for endometrial cancer (34). The intergenic variant rs113998067 is the lead signal for endometrial cancer and was a secondary cis-pQTL for R-spondin–1, encoded by *RSPO1*. Statistical colocalization confirmed a highly likely shared signal (PP = 98.2%) (fig. S5). Accordingly, we estimated a 91% increased risk for endometrial cancer per 1 SD higher plasma abundance of R-spondin–1 (odds ratio, 1.91; 95% CI, 1.52 to 2.41; $P = 3.6 \times 10^{-8}$). R-spondin–1 is a secreted activator protein that acts as an agonist for the canonical Wnt signaling pathway (35), playing a regulatory role as an adult stem cell growth factor. Work in mouse models (36), however, suggests that R-spondin–1 up-regulates the expression of estrogen receptor–α independent of Wnt/β-catenin signaling and might therefore amplify estrogen-mediated endometrial cancer risk (36). We note that the effect estimate for rs113998067 did not differ by sex (P = 0.12), and knockout models in male and female mice have shown abnormal development of testes and ovaries, respectively (37, 38), possibly indicating a wider impact on diseases of reproductive tissues.

A map of proteo-genomic connections across the phenome

We systematically assessed the sharedness of gene-protein-disease triplets through phenome-wide colocalization of cis-pQTL regions (see supplementary materials) (12) to identify and create a genetically anchored map of proteins involved in the etiology of common complex diseases, which could represent potential druggable targets. We identified 1859 gene-protein-trait triplets (network edges, Fig. 5 and fig. S6) comprising 412 protein targets and 506 curated phenotypes (fig. S7 and table S6). The mapping of these shared gene-protein-phenome connections highlights a large number of insights, as discussed below, while confirming previously established connections for known pleiotropic loci [e.g., GCKR (n = 197 traits), α-1-antitrypsin (n = 79 traits), or apolipoprotein A-V (n = 64 traits)] and established disease genes [e.g., roto-oncogene tyrosine protein kinase receptor RET (*RET*) and Hirschsprung's disease (39), or C-C motif chemokine 21 (*CCL21*) and rheumatoid arthritis (40)].

The map highlights 10 diseases for which we identified five or more colocalizing cis-pQTLs, including coronary artery disease (n = 12), hyperlipidemia indicated by lipid-lowering medication (n = 8), ulcerative colitis (n = 7), Alzheimer's disease (n = 6), and type 2 diabetes (n = 5). Statistical power was greatest for the detection of shared genetic architecture for traits for which measures were available in the largest number of people, in line with a median of 2 colocalizing cis-pQTLs (IQR, 2 to 4; maximum 32 for mean platelet volume) for blood cell parameters and biomarkers available in large-scale biobanks. For 104 of 191 curated phenotypes with at least three colocalizing protein targets, we observed significant enrichment of pathways [false discovery rate (q value) < 5%; table S7]. These reflected the known biology of the corresponding clinical entities, such as "wound healing" for platelet count, "skeletal system development" for height, "cholesterol metabolism" for coronary artery disease, or "response to virus" for Crohn's disease, as well as yet less understood ones such as "Toll-like receptor signaling" for hypothyroidism, for which two of the genes (*IRF3* and *TLR3*) have already been shown to confer virus-induced disease onset in mouse models (41).

The proteo-genomic map provides a new framework to (i) connect etiologically related diseases, (ii) provide biological context for new or emerging disorders such as COVID-19, and (iii) integrate information from different biological domains to establish mechanisms for known gene-disease links. For each of these scenarios, we provide selected examples to highlight the scientific opportunities arising from this map, both below and on the related open resource platform (www.omicscience.org).

Potential candidate genes for COVID-19 outcomes

We integrated GWAS summary statistics in our map for four different outcome definitions related to COVID-19, ranging from susceptibility to COVID-19 to severe cases requiring hospitalization (42). These GWAS differed substantially in the number of included cases (5101 to 38,984), and we observed that results were sensitive to the choice of outcome. We replicated the previously reported candidate genes *ABO* and *OAS1* (43) (fig. S8), both of which showed consistent evidence across these different outcome definitions. For *ABO*, the lead cis-pQTL (rs576125, MAF = 33.5%) also colocalized with pulmonary embolism (Fig. 5), a common complication of severe COVID-19 (44) potentially attributable to altered abundances of proteins involved in the coagulation cascade (15). We further observed suggestive evidence for *NSF* (for the risk of COVID-19 hospitalization) and *BCAT2* (for severe COVID-19), each of which shared a genetic signal with only one of these four outcomes and therefore will require external validation of their possible role in COVID-19 or associated pathologies.

Integrating multiple OMICs layers elucidates a disease mechanism for gallstones

We identified a signal at *SULT2A1*, a known gallstone locus (45), to be shared between bile salt sulfotransferase (SULT2A1) and the risk of cholelithiasis (odds ratio per 1 SD higher normalized protein abundance, 2.12; 95% CI, 1.66 to 2.70; $P = 2.1 \times 10^{-37}$) as well as cholecystectomy (odds ratio, 2.09; 95% CI, 1.86 to 2.34; $P = 7.8 \times 10^{-38}$). Multitrait colocalization (46) further identified that the signal was also shared with mRNA expression of *SULT2A1* in the liver, plasma concentrations of multiple sulfated steroids (47) including sulfate conjugates of androgen and pregnenolone metabolites, and bile acids. The high posterior probability (PP = 99%) was largely explained (63%) by rs212100, a variant in high LD (r^2 = 0.90) with the lead cis-pQTL at this locus (Fig. 6A and fig. S9). The consistent positive effect directions across all physiological entities, and in particular sulfated steroids and primary bile acid metabolites, suggest higher SULT2A1 activity as the mode of action. The concurrent inverse association with lower plasma concentrations of the secondary bile acid glycolithocholate indicates diminished formation of lithocholic acid, an essential detergent to solubilize fats, including cholesterol (48). Our vertical integration of diverse biology entities points to a supersaturated bile that promotes cholesterol crystallization and gallstone formation as a causal mechanism at a locus for which the mode of action has only been vaguely hypothesized (45).

Convergence of soft tissue disorders through FBLN3

A protein target connected to a very large number (n = 37) of diseases and other phenotypes was FBLN3 (extracellular matrix glycoprotein encoded by *EFEMP1*), which showed

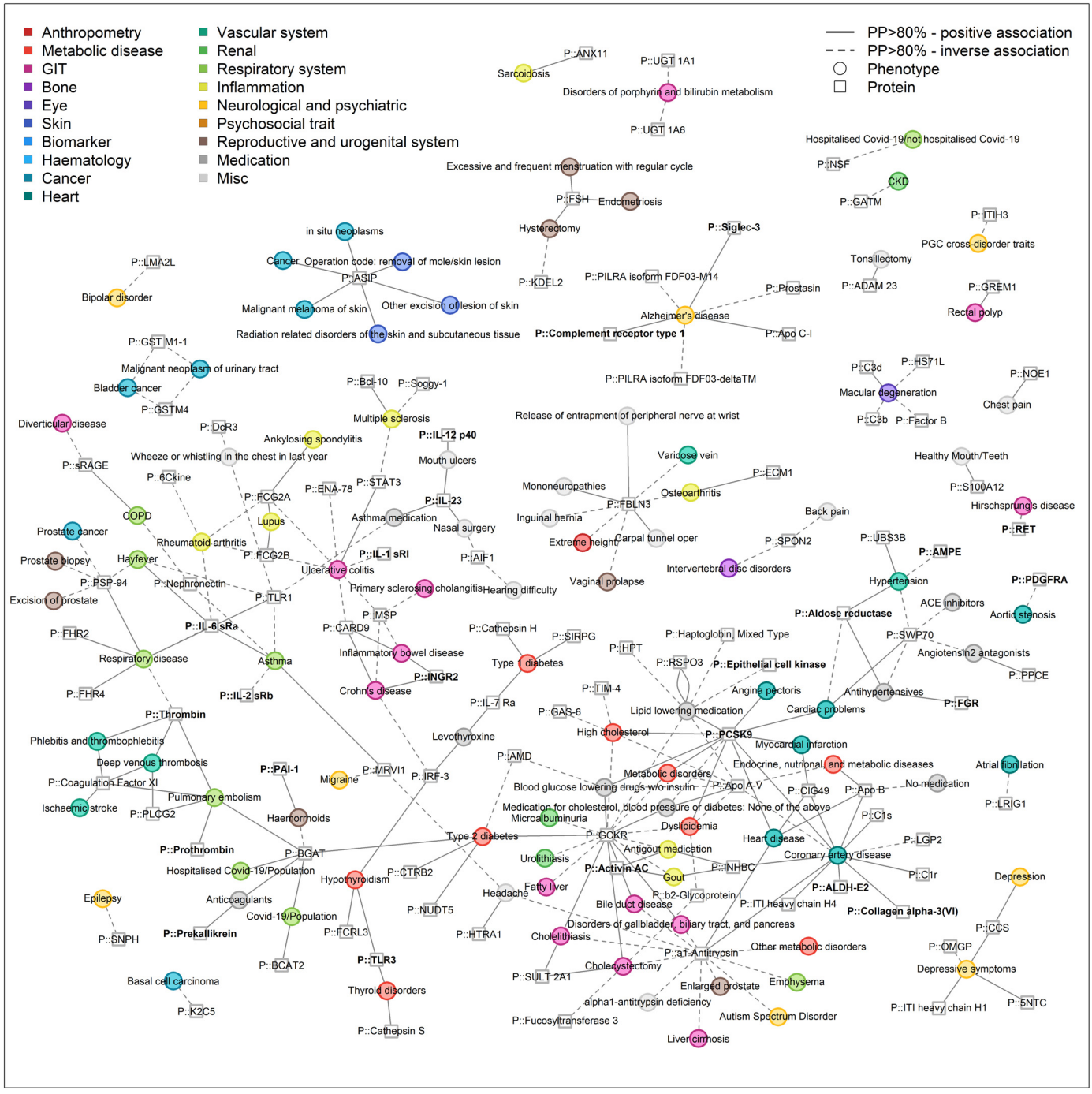


Fig. 5. Network representation of phenome-wide colocalization analysis for protein-encoding loci. This figure is restricted to connections between proteins and binary endpoints, mainly diseases, to increase visibility and show shared etiology. Only protein targets and phenotypes with at least one connection are included. Effect directions are indicated by the line type (solid = higher protein abundance, increased risk; dashed = higher protein abundance, reduced risk). Colors indicate categories of phenotypes. The entire network is composed of 412 protein targets (squares) and 506 phenotypes (circles) as nodes, which are connected (n = 1859 edges) if there is evidence of a shared genetic signal (posterior probability >80%) and is shown in fig. S6. See www.omicscience.org/apps/pgwas for an interactive version of the figure.

gene-protein convergence of diverse connective tissue disorders as well as gene expression of *EFEMP1* in subcutaneous adipose tissue, with high confidence in the lead cis-pQTL (rs3791679, MAF = 23.4%) being the causal variant in multitrait colocalization (Fig. 6B and fig. S10). The common A-allele of rs3791679 was associated with lower plasma abundance of FBLN3 and increased risk for a range of connective or soft tissue abnormalities, including hernias, varicose veins, vaginal prolapse, and hypermobility, several of which have previously been reported in individual GWAS but have not been connected (*49*–*54*). This spectrum of human clinical features suggests that lower plasma levels of A-allele carriers result in altered elastic fiber morphology and/or lower content, in line with evidence from *Efemp1*

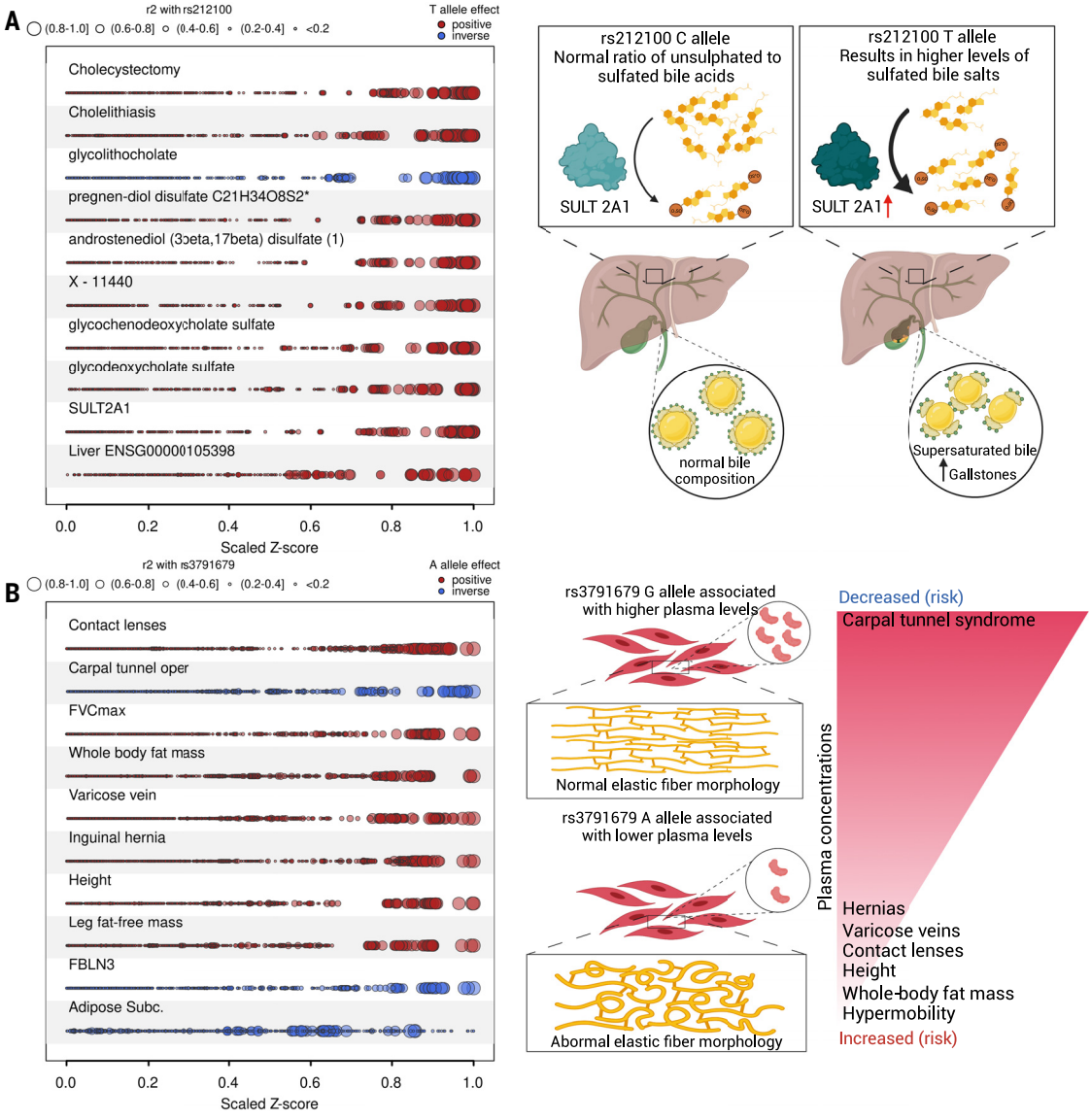


Fig. 6. Selected phenotypic examples from the proteo-genomic map. (**A**) Plot visualizing convergence of genetic variants at the *SULT2A1* locus in relation to the LD with the candidate gene variant identified by multitrait colocalization. Z-scores from GWAS for each annotated trait have been scaled by the absolute maximum, and dot size is proportional to the LD (r^2). Colors indicate the direction of effect aligned to the risk-increasing allele (red, positive; blue, inverse). The scheme on the right depicts the suggested mode of action by which higher SULT2A1 activity translates to higher risk of gallstones. (**B**) Same as (A), but for diseases and other phenotypes colocalizing at the *EFEMP1* locus. The scheme on the right depicts a proposed mechanism by which altered secretion of FBLN3 leads to the observed phenotypes. See figs. S9 and S10 for stacked regional association plots for (A) and (B).

knockout mice that display abnormal elastic fiber morphology, develop different types of hernias, and exhibit pelvic organ prolapse (*55*). FBLN3 is part of the extracellular matrix and widely expressed, but its function is incompletely understood (*56*). We provide insights about its role in the etiology of a large number of connective tissue disorders, including a potential explanation for the established link between carpal tunnel syndrome and shorter stature (*51*). Mutations in *EFEMP1* cause a rare eye disease called Doyne honeycomb retinal dystrophy (DHRD) (*57*), characterized by visual disturbances and drusenoid deposits due to accumulating intracellular FBLN3. We observed sharedness of the signal at this protein locus with vision-related phenotypes, including use of contact lenses (myopia) and decreased optic disc area, a risk factor for open-angle glaucoma (*50*), with lower protein concentrations associated with greater risk, as also observed in patients with DHRD.

Differential effect sizes of cis-pQTLs by sex and age

We systematically tested differences in the genetic associations of all protein targets included in the proteo-genomic map (N = 412) by age or sex. We identified a total of 14 protein targets that showed evidence for significant ($P < 5.9 \times 10^{-5}$) effect modification of the cis-pQTL by sex (N = 10) or age (N = 8), including four common to both (table S8). This included biologically plausible candidates, such as annexin II, where the cis-pQTL showed a stronger effect in women, albeit with a strong significant effect in either sex (women, $\beta = -0.86$, $P < 1.7 \times 10^{-467}$; men, $\beta = -0.64$, $P < 2.5 \times 10^{-231}$). This finding is in line with evidence of isoform expression of the protein-encoding gene *ANXA2* in male and female reproductive tissues, including prostate (PP = 81.9%)

and vagina (PP = 87.4%), and a possible role of the locus in puberty timing (*58*, *59*).

We noted that most of the identified cis-pQTLs showed age- and sex-differential and not dimorphic effects (*60*) and were linked to missense variation (inhibin C, vitronectin, Siglec 9, GCKR, SOD3, CPA4, and PILRA) or alternative splicing events (annexin II, BGAT, and CO8G) with very strong overall effects, enabling the detection of even small effect differences between strata more easily (*61*). In general, our results are concordant with the few sex-specific effects of molecular QTLs reported so far (*62*, *63*) and show that systematic efforts for both molecular QTLs and disease GWAS are needed to better understand the mechanisms underlying such differences. Crucially, investigating the relevance of these genetic differences for phenotypic expression depends on the availability of sex-specific GWAS results across the human phenome.

Druggable targets and repurposing opportunities

We systematically identified druggable proteins in the proteo-genomic map by linking the protein-encoding gene to the druggable genome (*64*) and identified 60 protein targets linked to at least one phenotype, including 22 protein targets linked to a disease (table S9). We replicated established examples, such as the interleukin-6 receptor for rheumatoid arthritis or thrombin for deep venous thrombosis (Fig. 5). We also identified 31 candidates with potential repurposing opportunities for one to eight diseases (for a total of 32 different indications), following a search and prioritization strategy in Open Targets (*65*).

Webserver

To enable customized and in-depth exploration of high-priority protein targets—that is, those with at least one cis-pQTL—we created an interactive online resource (www.omicscience.org/apps/pgwas). The webserver provides intuitive representations of genetic findings and enables the look-up of summary statistics for individual single-nucleotide polymorphisms (SNPs), genes, and whole genomic regions across all protein targets. To interactively assess specificity and to identify pleotropic cis-pQTLs that present strong trans-like association profiles, we generated an interactive heatmap of genetic associations of all cis-pQTLs across all high-priority candidate proteins. We further provide detailed annotations of the protein targets, including links to external databases such as UniProt or Reactome, information on currently available drugs, characterization of associated SNPs, as well as results from our colocalization analysis with eQTLs, sQTLs, diseases, and other phenotypes. An interactive version of the proteo-genomic map allows a deep dive into proteins or phenotypes of particular interest to explore cross-disease connections within subnetworks.

Discussion

The promise of proteomic technologies and their integration with genomic data lies in their application to rare and common human diseases. Although previous studies had started to exploit the phenotypic consequences of pQTLs, they mainly focused on identifying and describing the genetic architecture of proteins measured by specific platforms (*1*–*6*, *9*). We performed a systematic integration of the phenome and created a proteo-genomic map of human health that identifies many potential causal disease genes and highlights genetically driven connections across diverse human conditions. The traditional classification of diseases relies on the aggregation of symptoms commonly presenting together and, with the exception of Mendelian disorders, is rarely based on shared etiology (*66*). Our network anchors the convergence of diseases in their shared genetic etiology, as shown for FBLN3, providing mechanistic understanding and a starting point for the identification of treatment strategies targeting underlying genetic causes.

Uncertainty in assigning causal genes and variants remains a major limitation for experimental validation and clinical translation of results from the plethora of hypothesis-free genetic association studies. We show how cis-pQTLs identify causal candidate genes at established disease risk loci, including COVID-19, providing immediate hypotheses for experimental follow-up for a large number of disease genes.

The uncertain specificity of genetic variation affecting protein content outside of the protein-encoding region, trans-pQTLs, restricts the discovery of de novo biological insights in protein regulation and instrumentation of such variants for genetic prediction, such as with polygenic scores. We show how data-driven network clustering augments ontology-based classification approaches and identifies biologically plausible examples, such as for *PNLPA3* and a community of liver-derived protein targets.

Genetic variation found for proteins circulating in blood raises the question of transferability to disease-relevant tissue processes. We show that for about half of the protein targets with a cis-pQTL, this can be linked to gene expression in various tissues and provide examples, such as for SULT2A1, that illustrate how multidomain integration can identify tissue-specific mechanism. In its most simple form, such cis-pQTLs determine the basal rate of protein production within cells and are more or less constantly released into plasma as a result of natural cell turnover (*67*). Integration of genetic information allowed us to separate out such enclosed effects from other mechanisms, leading to higher cell turnover or leakage, such as for SULT2A1 and the liver-specific effect of the *PNPLA3* variant. Although this provides a strategy to point to relevant tissues, overlapping data for tissue-specific gene and protein expression is required to quantify the contribution of various tissues to the plasma proteome.

To accelerate the use and translational potential of our findings, we generated an open-access interactive web resource that enables the scientific community to easily and rapidly capitalize on these results for future research across clinical specialties. We demonstrate for multiple examples how this resource can be used to put gene-phenotype findings into a systems biological context.

Although our study is distinguished by its comprehensive discovery and characterization of pQTLs in cis and trans along with a systematic integration of the phenome, it does have limitations. First, the nature of the technology used to measure protein concentrations is designed to maximize discovery by generating a large library of affinity reagents, which rely on a preserved shape of the target protein and hence might miss genetic effects specific to a particular isoform of the protein (*10*). The semiquantitative nature of the assay makes risk estimates based on Mendelian randomization studies challenging. A thorough discussion of assay differences can be found in our previous work (*10*), and we observed consistent cis-pQTLs for the highlighted examples, including RSPO1, SULT2A1, and FBLN3, as measured with Olink. Second, our study cohort consisted of predominantly healthy middle-aged participants of European descent, and replication of our results in ethnically diverse populations is warranted, in particular for the discovery of drug targets. Further work would also be required to investigate possible modifying effects of phenotypic characteristics on gene-protein associations, such as by sex, age, or behavioral factors. Third, our study concentrated on the common spectrum of variation in the genome. Investigation of rare variation is likely to identify pQTLs with larger effect sizes and possibly more severe phenotypic consequences. Finally, our proteo-genomic map is limited to publicly available GWAS summary statistics; additional insights will result from the inclusion of further data for additional phenotypes (in particular, cancers) and understudied diseases.

Methods summary

Detailed materials and methods are provided in the supplementary materials (*12*). We performed a genome-proteome–wide association study among 10,708 participants of European descent in the Fenland study (table S1) on 10.2 million genetic variants and plasma abundances of 4775 distinct protein targets measured in plasma using established workflows (*15*). Protein targets were measured using the SomaScan v4 assay employing 4979 single-stranded oligonucleotides (aptamers) with specific binding affinities to 4775 unique protein targets (*68*, *69*).

We used the term "protein target" to refer to proteins targeted by at least one aptamer. We define significant genetic variant–protein target associations (pQTLs) at a stringent Bonferroni threshold ($P < 1.004 \times 10^{-11}$) and performed approximate conditional analysis to detect secondary signals for each genomic region identified by distance-based clumping of association statistics. We defined cross-aptamer regions using a combined approach of multi-trait colocalization (*46*) and LD-clumping. We classified pQTLs as protein- or pathway-specific by assessing pQTL specificity across the entire proteome ($P < 5 \times 10^{-8}$) while testing whether associated protein targets were captured by a common GO term or a protein community in a data-driven protein network. We computed the variance explained in plasma abundances of protein targets by cis-pQTLs (within ±500 kb of the protein-encoding gene) or trans-pQTLs according to different specificity categories using linear regression models. We used statistical colocalization (*70*) to test for a shared genetic signal between expression or alternative splicing of the protein-encoding gene and the cis-pQTL in one of at least 49 tissues of the GTEx v8 project (*24*). We systematically cross-referenced established genetic risk loci for common complex diseases and phenotypes with pQTLs by identifying cis-pQTLs or strong proxies ($r^2 > 0.8$) in the GWAS catalog (www.ebi.ac.uk/gwas/). We finally performed phenome-wide colocalization screens at 1548 protein-encoding loci using publicly available (*71*) as well as in-house–curated genome-wide association statistics for thousands of phenotypes. We applied stringent priors and conservative filters to derive high-confidence protein-phenotype links. We used basic functions of R (v.3.6.0), the R package *igraph*, and the BioRender web application (https://biorender.com/) to create figures. The Fenland study was approved by the National Health Service (NHS) Health Research Authority Research Ethics Committee (NRES Committee–East of England Cambridge Central, ref. 04/Q0108/19), and all participants provided written informed consent.

REFERENCES AND NOTES

1. V. Emilsson *et al*., Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018). doi: 10.1126/science.aaq1327; pmid: 30072576
2. K. Suhre *et al*., Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017). doi: 10.1038/ncomms14357; pmid: 28240269
3. L. Folkersen *et al*., Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLOS Genet.* **13**, e1006706 (2017). doi: 10.1371/journal.pgen.1006706; pmid: 28369058
4. B. B. Sun *et al*., Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018). doi: 10.1038/s41586-018-0175-2; pmid: 29875488
5. C. Yao *et al*., Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018). doi: 10.1038/s41467-018-05512-x; pmid: 30111768
6. A. Gilly *et al*., Whole-genome sequencing analysis of the cardiometabolic proteome. *Nat. Commun.* **11**, 6336 (2020). doi: 10.1038/s41467-020-20079-2; pmid: 33303764
7. K. Suhre, M. I. McCarthy, J. M. Schwenk, Genetics meets proteomics: Perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021). doi: 10.1038/s41576-020-0268-2; pmid: 32860016
8. J. Zheng *et al*., Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020). doi: 10.1038/s41588-020-0682-6; pmid: 32895551
9. L. Folkersen *et al*., Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020). doi: 10.1038/s42255-020-00287-2; pmid: 33067605
10. M. Pietzner *et al*., Cross-platform proteomics to advance genetic prioritisation strategies. bioRxiv [preprint]. 19 March 2021. doi: 10.1101/2021.03.18.435919; pmid: 435919
11. T. Lindsay *et al*., Descriptive epidemiology of physical activity energy expenditure in UK adults (The Fenland study). *Int. J. Behav. Nutr. Phys. Act.* **16**, 126 (2019). doi: 10.1186/s12966-019-0882-6; pmid: 31818302
12. Associated code is available on GitHub. doi: 10.5281/zenodo.5385532
13. M. Narayan, Disulfide bonds: Protein folding and subcellular protein trafficking. *FEBS J.* **279**, 2272–2282 (2012). doi: 10.1111/j.1742-4658.2012.08636.x; pmid: 22594874
14. M. Uhlén *et al*., The human secretome. *Sci. Signal.* **12**, eaaz0274 (2019). doi: doi:
15. M. Pietzner *et al*., Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020). doi: 10.1038/s41467-020-19996-z; pmid: 33328453
16. M. Eslam, L. Valenti, S. Romeo, Genetics and epigenetics of NAFLD and NASH: Clinical impact. *J. Hepatol.* **68**, 268–279 (2018). doi: 10.1016/j.jhep.2017.09.003; pmid: 29122391
17. S. BasuRay, Y. Wang, E. Smagris, J. C. Cohen, H. H. Hobbs, Accumulation of PNPLA3 on lipid droplets is the basis of associated hepatic steatosis. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 9521–9526 (2019). doi: 10.1073/pnas.1901974116; pmid: 31019090
18. P. N. Newsome *et al*., Guidelines on the management of abnormal liver blood tests. *Gut* **67**, 6–19 (2018). doi: 10.1136/gutjnl-2017-314924; pmid: 29122851
19. D. M. Tollefsen, C. J. Weigel, M. H. Kabeer, The presence of methionine or threonine at position 381 in vitronectin is correlated with proteolytic cleavage at arginine 379. *J. Biol. Chem.* **265**, 9778–9781 (1990). doi: 10.1016/S0021-9258(19)38738-1; pmid: 1693616
20. D. I. Leavesley *et al*., Vitronectin—Master controller or micromanager? *IUBMB Life* **65**, 807–818 (2013). pmid: 24030926
21. V. Guarani *et al*., QIL1 is a novel mitochondrial protein required for MICOS complex stability and cristae morphology. *eLife* **4**, e06265 (2015). doi: 10.7554/eLife.06265; pmid: 25997101
22. P. B. Maguire *et al*., Proteomic Analysis Reveals a Strong Association of β-Catenin With Cadherin Adherens Junctions in Resting Human Platelets. *Proteomics* **18**, e1700419 (2018). doi: 10.1002/pmic.201700419; pmid: 29510447
23. D. Szklarczyk *et al*., STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019). doi: 10.1093/nar/gky1131; pmid: 30476243
24. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020). doi: 10.1126/science.aaz1776; pmid: 32913098
25. C. Buccitelli, M. Selbach, mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644 (2020). doi: 10.1038/s41576-020-0258-4; pmid: 32709985
26. U. Võsa *et al*., Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv 447367 [preprint]. 19 October 2018. doi: 10.1101/447367; pmid: 447367
27. S. M. Sternson, D. Atasoy, Agouti-related protein neuron circuits that regulate appetite. *Neuroendocrinology* **100**, 95–102 (2014). doi: 10.1159/000369072; pmid: 25402352
28. R. W. Baker, F. M. Hughson, Chaperoning SNARE assembly and disassembly. *Nat. Rev. Mol. Cell Biol.* **17**, 465–479 (2016). doi: 10.1038/nrm.2016.65; pmid: 27301672
29. I. E. Jansen *et al*., Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019). doi: 10.1038/s41588-018-0311-9; pmid: 30617256
30. J. Schwartzentruber *et al*., Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat. Genet.* **53**, 392–402 (2021). doi: 10.1038/s41588-020-00776-w; pmid: 33589840
31. S. Aggarwal, P. K. Dabla, S. Arora, Prostasin: An Epithelial Sodium Channel Regulator. *J. Biomark.* **2013**, 179864 (2013). doi: 10.1155/2013/179864; pmid: 26317012
32. Y. Sugitani *et al*., Sodium absorption stimulator prostasin (PRSS8) has an anti-inflammatory effect via downregulation of TLR4 signaling in inflammatory bowel disease. *J. Gastroenterol.* **55**, 408–417 (2020). doi: 10.1007/s00535-019-01660-z; pmid: 31916038
33. M. Calvo-Rodriguez, C. García-Rodríguez, C. Villalobos, L. Núñez, Role of Toll Like Receptor 4 in Alzheimer's Disease. *Front. Immunol.* **11**, 1588 (2020). doi: 10.3389/fimmu.2020.01588; pmid: 32983082
34. T. A. O'Mara *et al*., Identification of nine new susceptibility loci for endometrial cancer. *Nat. Commun.* **9**, 3166 (2018). doi: 10.1038/s41467-018-05427-7; pmid: 30093612
35. M. E. Binnerts *et al*., R-Spondin1 regulates Wnt signaling by inhibiting internalization of LRP6. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14700–14705 (2007). doi: 10.1073/pnas.0702305104; pmid: 17804805
36. A. Geng *et al*., A novel function of R-spondin1 in regulating estrogen receptor expression independent of Wnt/β-catenin signaling. *eLife* **9**, e56434 (2020). doi: 10.7554/eLife.56434; pmid: 32749219
37. A.-A. Chassot *et al*., WNT4 and RSPO1 together are required for cell proliferation in the early mouse gonad. *Development* **139**, 4461–4472 (2012). doi: 10.1242/dev.078972; pmid: 23095882
38. A.-A. Chassot *et al*., Activation of beta-catenin signaling by Rspo1 controls differentiation of the mammalian ovary. *Hum. Mol. Genet.* **17**, 1264–1277 (2008). doi: 10.1093/hmg/ddn016; pmid: 18250098
39. P. Edery *et al*., Mutations of the RET proto-oncogene in Hirschsprung's disease. *Nature* **367**, 378–380 (1994). doi: 10.1038/367378a0; pmid: 8114939
40. E. A. Stahl *et al*., Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010). doi: 10.1038/ng.582; pmid: 20453842
41. N. Harii *et al*., Thyrocytes express a functional toll-like receptor 3: Overexpression can be induced by viral infection and reversed by phenylmethimazole and is associated with Hashimoto's autoimmune thyroiditis. *Mol. Endocrinol.* **19**, 1231–1250 (2005). doi: 10.1210/me.2004-0100; pmid: 15661832
42. COVID-19 Host Genetics Initiative, Mapping the human genetic architecture of COVID-19. *Nature* 10.1038/s41586-021-03767-x (2021). doi: 10.1038/s41586-021-03767-x
43. S. Zhou *et al*., A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659–667 (2021). doi: 10.1038/s41591-021-01281-1; pmid: 33633408
44. M. B. Whyte, P. A. Kelly, E. Gonzalez, R. Arya, L. N. Roberts, Pulmonary embolism in hospitalised patients with COVID-19. *Thromb. Res.* **195**, 95–99 (2020). doi: 10.1016/j.thromres.2020.07.025; pmid: 32682004
45. A. D. Joshi *et al*., Four Susceptibility Loci for Gallstone Disease Identified in a Meta-analysis of Genome-Wide Association Studies. *Gastroenterology* **151**, 351–363.e28 (2016). doi: 10.1053/j.gastro.2016.04.007; pmid: 27094239
46. C. N. Foley *et al*., A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* **12**, 764 (2021). doi: 10.1038/s41467-020-20885-8; pmid: 33536417
47. S.-Y. Y. Shin *et al*., An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014). doi: 10.1038/ng.2982; pmid: 24816252
48. F. Lammert *et al*., Gallstones. *Nat. Rev. Dis. Primers* **2**, 16024 (2016). doi: 10.1038/nrdp.2016.24; pmid: 27121416
49. A. R. Wood *et al*., Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014). doi: 10.1038/ng.3097; pmid: 25282103
50. H. Springelkamp *et al*., New insights into the genetics of primary open-angle glaucoma based on meta-analyses of intraocular pressure and optic disc characteristics. *Hum. Mol. Genet.* **26**, 438–453 (2017). pmid: 28073927
51. A. Wiberg *et al*., A genome-wide association analysis identifies 16 novel susceptibility loci for carpal tunnel syndrome. *Nat. Commun.* **10**, 1030 (2019). doi: 10.1038/s41467-019-08993-6; pmid: 30833571
52. E. Jorgenson *et al*., A genome-wide association study identifies four novel susceptibility loci underlying inguinal hernia.

Nat. Commun. **6**, 10130 (2015). doi: 10.1038/ncomms10130; pmid: 26686553
53. N. Shrine *et al.*, New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493 (2019). doi: 10.1038/s41588-018-0321-7; pmid: 30804560
54. J. K. Pickrell *et al.*, Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016). doi: 10.1038/ng.3570; pmid: 27182965
55. P. J. McLaughlin *et al.*, Lack of fibulin-3 causes early aging and herniation, but not macular degeneration in mice. *Hum. Mol. Genet.* **16**, 3059–3070 (2007). doi: 10.1093/hmg/ddm264; pmid: 17872905
56. I. Livingstone, V. N. Uversky, D. Furniss, A. Wiberg, The Pathophysiological Significance of Fibulin-3. *Biomolecules* **10**, 1294 (2020). doi: 10.3390/biom10091294; pmid: 32911658
57. L. Y. Marmorstein *et al.*, Aberrant accumulation of EFEMP1 underlies drusen formation in Malattia Leventinese and age-related macular degeneration. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13067–13072 (2002). doi: 10.1073/pnas.202491599; pmid: 12242346
58. B. Hollis *et al.*, Genomic analysis of male puberty timing highlights shared genetic basis with hair colour and lifespan. *Nat. Commun.* **11**, 1536 (2020). doi: 10.1038/s41467-020-14451-5; pmid: 32210231
59. F. R. Day *et al.*, Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet.* **49**, 834–841 (2017). doi: 10.1038/ng.3841
60. E. A. Khramtsova, L. K. Davis, B. E. Stranger, The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019). doi: 10.1038/s41576-018-0083-1; pmid: 30581192
61. H. Aschard, A perspective on interaction effects in genetic association studies. *Genet. Epidemiol.* **40**, 678–688 (2016). doi: 10.1002/gepi.21989; pmid: 27390122
62. M. Oliva *et al.*, The impact of sex on gene expression across human tissues. *Science* **369,** eaba3066 (2020). doi: 10.1126/science.aba3066
63. K. Mittelstrass *et al.*, Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLOS Genet.* **7**, e1002215 (2011). doi: 10.1371/journal.pgen.1002215; pmid: 21852955
64. C. Finan *et al.*, The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, eaag1166 (2017). doi: 10.1126/scitranslmed.aag1166; pmid: 28356508
65. D. Ochoa *et al.*, Open Targets Platform: Supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021). doi: 10.1093/nar/gkaa1027; pmid: 33196847
66. I. Kola, J. Bell, A call to reform the taxonomy of human disease. *Nat. Rev. Drug Discov.* **10**, 641–642 (2011). doi: 10.1038/nrd3534; pmid: 21878965
67. R. Sender, R. Milo, The distribution of cellular turnover in the human body. *Nat. Med.* **27**, 45–48 (2021). doi: 10.1038/s41591-020-01182-9; pmid: 33432173
68. S. A. Williams *et al.*, Plasma protein patterns as comprehensive indicators of health. *Nat. Med.* **25**, 1851–1857 (2019). doi: 10.1038/s41591-019-0665-2; pmid: 31792462
69. L. Gold *et al.*, Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLOS ONE* **5**, e15004 (2010). doi: 10.1371/journal.pone.0015004; pmid: 21165148
70. C. Giambartolomei *et al.*, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genet.* **10**, e1004383 (2014). doi: 10.1371/journal.pgen.1004383; pmid: 24830394
71. B. Elsworth *et al.*, The MRC IEU OpenGWAS data infrastructure. bioRxiv 244293 [preprint]. 10 August 2020.doi: 10.1101/2020.08.10.244293; pmid: 244293

ACKNOWLEDGMENTS

We are grateful to all Fenland volunteers and to the General Practitioners and practice staff for assistance with recruitment. We thank the Fenland Study Investigators, Fenland Study Co-ordination team and the Epidemiology Field, Data and Laboratory teams. Proteomic measurements were supported and governed by a collaboration agreement between the University of Cambridge and SomaLogic. This research has been conducted using the UK Biobank Resource (application no. 20361 and 44448). **Funding:** The Fenland Study (10.22025/2017.10.101.00001) is funded by the Medical Research Council (MC_UU_12015/1). We further acknowledge support for genomics from the Medical Research Council (MC_PC_13046). This work was supported in part by the UKRI/NIHR Strategic Priorities Award in Multimorbidity Research for the Multimorbidity Mechanism and Therapeutics Research Collaborative (MR/V033867/1); NIH awards R35HG010718, R01HG011138, R01GM140287, and NIH/NIA AG068026 (E.R.G.); National Institute on Aging grants U01 AG061359, RF1 AG057452, and RF1 AG059093 (M.A.W., M.A., and G.K.); a 4-year Wellcome Trust PhD Studentship and the Cambridge Trust (J.C.-Z.); a Gates Fellowship (M.K.); and the Medical Research Council (MC_UU_00006/1 - Etiology and Mechanisms) (C.L., E.W., M.P., J.L., E.O., I.S., N.K., and N.J.W.). **Author contributions:** Conceptualization: C.L., M.P., E.W.; data curation/software: E.O., N.D.K., J.L., M.A.W., J.R.; formal analysis: M.P., E.W., J.C.-Z., A.C., M.K., I.D.S., J.C.; methodology: R.A.S., E.R.G.; visualization: M.P., J.C.-Z., M.K., C.L.; funding acquisition: C.L., N.J.W.; project administration: C.L., N.J.W.; supervision: C.L., R.A.S., G.K.; writing–original draft: M.P., C.L., E.W., J.C.-Z., M.K.; writing–review and editing: E.O., J.C., I.D.S., A.D.H., N.D.K., M.A., W.A., S.O., N.J.W. **Competing interests:** R.A.S. and A.C. are current employees and/or stockholders of GlaxoSmithKline. E.R.G. receives an honorarium from the journal *Circulation Research* of the American Heart Association as a member of the Editorial Board. S.O. has received remuneration for consultancy services provided to Pfizer Inc., Astra Zeneca, ERX Pharmaceuticals, GSK, Third Rock Ventures, and LG Life Sciences. All other authors declare that they have no competing interests. **Data and materials availability:** Data from the Fenland cohort can be requested by bona fide researchers for specified scientific purposes via the study website (www.mrc-epid.cam.ac.uk/research/studies/fenland/information-for-researchers/). Summary statistics can be obtained from www.omicscience.org/apps/pgwas. Publicly available summary statistics for look-up and colocalization of pQTLs were obtained from https://gwas.mrcieu.ac.uk/ and www.ebi.ac.uk/gwas/. Associated code and scripts for the analysis are available on GitHub (https://github.com/MRC-Epid/pGWAS_discovery) and have been permanently archived using Zenodo (*12*).

SUPPLEMENTARY MATERIALS

23 April 2021; accepted 29 September 2021
Published online 14 October 2021
10.1126/science.abj1541

Science

Mapping the proteo-genomic convergence of human diseases

Maik PietznerEleanor WheelerJulia Carrasco-ZaniniAdrian CortesMine KopruluMaria A. WörheideErin OertonJames CookIsobel D. StewartNicola D. KerrisonJian'an LuanJohannes RafflerMatthias ArnoldWiebke ArltStephen O'RahillyGabi KastenmüllerEric R. GamazonAroon D. HingoraniRobert A. ScottNicholas J. WarehamClaudia Langenberg

Science, 374 (6569), eabj1541. • DOI: 10.1126/science.abj1541

Detangling gene-disease connections

Many diseases are at least partially due to genetic causes that are not always understood or targetable with specific treatments. To provide insight into the biology of various human diseases as well as potential leads for therapeutic development, Pietzner *et al*. undertook detailed, genome-wide proteogenomic mapping. The authors analyzed thousands of connections between potential disease-associated mutations, specific proteins, and medical conditions, thereby providing a detailed map for use by future researchers. They also supplied some examples in which they applied their approach to medical contexts as varied as connective tissue disorders, gallstones, and COVID-19 infections, sometimes even identifying single genes that play roles in multiple clinical scenarios. —YN