Mapping the proteo-genomic convergence of human diseases

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

DOI:
10.1126/science.abj1541

Title: Mapping the proteo-genomic convergence of human diseases

Authors: 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#]

Affiliations:

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 (BIH) 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 Behavioural Sciences, Duke University, Durham, NC 27710, USA
7. Institute of Metabolism and Systems Research, University of Birmingham, Birmingham B15 2TT, United Kingdom
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, United Kingdom
12. UCL British Heart Foundation Research Accelerator, Institute of Cardiovascular Science, University College London, WC1E 6BT, UK.
13. Health Data Research UK, Gibbs Building, 215 Euston Road, London NW1 2BE, UK
14. Institute of Health Informatics, University College London, 222 Euston Road, London NW1 2DA, UK

*these authors contributed equally

#Correspondence to Dr. Claudia Langenberg (claudia.langenberg@mrc-epid.cam.ac.uk)

ABSTRACT:

Characterization of the genetic regulation of proteins is essential for understanding disease etiology and developing therapies. We identified 10,674 genetic associations for 3,892 plasma proteins to create a cis-anchored gene-protein-disease map of 1,859 connections that highlights strong cross-disease biological convergence. This proteo-genomic map provides a framework to 1) connect etiologically related diseases, 2) provide biological context for emerging disorders, and 3) integrate different biological domains to establish mechanisms for known gene-disease link. Our results establish the value of cis-protein variants for annotation of likely causal disease genes at GWAS loci, addressing a major barrier for experimental validation and clinical translation of genetic discoveries, and identify proteo-genomic connections within and between diseases.

One Sentence summary: A genetically anchored map of protein – disease links identifies shared etiology across diverse diseases and possible therapeutic directions.

MAIN TEXT

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 impacts on 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 understand disease mechanisms and provide insights into the shared genetic architecture across diseases within a translational framework that puts humans as the model organisms at the center (*2*, *4*). This is now pursued at scale by pharmaceutical companies for the discovery of drug targets or repurposing opportunities (*7*, *8*). Earlier studies have started to characterize the genetic architecture of proteins using bespoke panels (*3*, *6*, *9*) or larger proteomic platforms (*1*, *2*, *4*, *5*), and have demonstrated how this can provide insight into the pathogenesis of specific diseases. There has been less attention on: a) providing a framework to assess the protein specificity of genetic variation residing outside (trans) the protein encoding gene, b) understanding the clinical relevance of pQTLs for proteins detected in plasma but known to not be actively secreted (*7*), c) classifying thousands of proteins based on their genetic architecture as explained by merely cis variants, specific trans variants, or unspecific trans variants, d) demonstrating the specific utility of pQTLs for the prioritization of candidate genes at established risk loci, and e) systematically mapping shared gene-protein-disease signals to uncover connections among thousands of considered diseases and phenotypes.

Profiling thousands of proteins circulating in blood at population-scale is currently only possible using large libraries of affinity reagents, namely antibodies or alternatively short oligonucleotides, called aptamers, since gold standard methods such as mass spectrometry lack throughput. We have previously provided a detailed comparison of 871 overlapping proteins measured in 485 individuals (*10*) of 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 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 4,775 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 show that half of all pQTLs close to the protein-encoding gene, cis-pQTLs, colocalize with gene expression or splicing QTLs in various tissues allowing to derive functional insights within tissues by integrating genetics with

plasma proteomics. We demonstrate the specific ability of cis-pQTLs to prioritize candidate causal genes at established genetic risk loci. By means of phenome-wide colocalization screens we generate a proteo-genomic map of human health covering 1,859 gene-protein-trait triplets providing insights into the shared etiology across diseases and the identification of pathophysiological pathways through cross-domain integration.

RESULTS

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 a minor allele frequency (MAF) >1% among 10,708 participants in the Fenland study targeting 4,775 distinct proteins (*12*). We identified 2,584 genomic regions (1,543 within ±500 kb of the protein-encoding genes, cis) associated with at least one of 3,892 protein targets at $p<1.004x10^{-11}$. 1,097 of these regions covered variants that have not been reported to be associated with plasma proteins so far (*1–6*, *9*) ($r^2<0.1$), of which 64% (867 out of 1,356 pQTLs) available in *(4)* replicated ($p<0.05$, directionally consistent). Further, 61% (488 out of 797, **Table S2**) of pQTLs replicated using the complementary Olink technique (*12*), with a higher proportion for variants in cis (81.2%) compared to trans (44.2%). Most regions (79.3%, n=2,050) were associated with a single protein target, but we observed substantial pleiotropy (≥2 protein targets) at the remaining regions, including up to five (16.1%, n=418), 6-20 (3.4%, n=88), or 21-50 (0.7%, n=19) associated protein targets, and with eight regions (*CFH, ARF4-ARHGEF3, C4A-CFB, BCHE, VTN, CFD, ABO, GCKR*) associated with 59 to 1,539 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 group ABO system transferase (*ABO*), key metabolic enzymes like glucokinase regulatory protein (*GCKR*), or lipid mediators such as apolipoprotein E, establishing a network-like structure of the circulating proteome (*1*).

Out of the 3,892 protein targets, 26.8% (n=1,046) had pQTLs in cis and trans, 13.4% (n=523) in cis only, and 59.6% (n=2,323) in trans only, among a total of 8,328 sentinel variant-protein target associations (**Fig. 1** and **Tables S2 and S3**). We identified another 2,346 secondary pQTLs at those loci using an adapted stepwise conditional analysis (median: 1, range: 1 - 13) indicating widespread allelic heterogeneity in cis (68.8%) and trans (31.2%). The majority of the 5,442 distinct variants were located in introns (~44%) or were in high 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% vs 33.7%, p=0.04, X^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 (GO terms, **Fig. 2A-B and S2**) to distinguish pathway-specific pQTLs from those exerting effects on multiple unrelated targets (*12*, *15*). We successfully assigned 40.8% (n=1,790 in cis, n=423 in trans) of the 5,442 genetic variants as protein-specific and 5.9% (n=236 in *cis*, n=86 in trans) as pathway-specific based on 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, 1,802 protein targets had at least one (likely) specific pQTL in cis (n=1,385) 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*) which was associated with 22 out 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) rendering PNPLA3 resistant to ubiquitylation-mediated degradation and resulting in subsequent accumulation on hepatic lipid droplets causing 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 those might only appear in plasma of otherwise healthy individuals 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 compared to 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 variation to plasma concentrations (**Fig. 2D and S3, and Table S3**). For about a third (n=1,249) of the protein targets, genetic variance was mainly explained by one or more cis-pQTLs, whereas for 7.2% (n=282) protein- or pathway-specific trans-pQTLs accounted for most of the genetic variation, leaving two-thirds (n=2,361) mainly explained by unspecific trans-pQTLs (*12*). Overall, we observed a median genetic contribution of 2.7% (IQR: 1.0% - 7.6%) reaching values above 70% for proteins like 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, affecting the binding epitope of the protein target, are the likely explanation for such strong and isolated genetic effects. While more than two-thirds of the protein targets with at least one cis-pQTL were unrelated to PAVs, we provide evidence that 158 (32.9%) of the protein targets 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 conformation and hence function of the protein target, rather than 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 biological 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* 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%-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 1.53-fold enrichment (Chi-square test p-value=1.8×10^{-10}) 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) from the GTEx version 8 release (*24*) using statistical colocalization (posterior probability (PP) > 80%) for all 1,584 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 and a median of 4.5 tissues (IQR: 2-12; **Fig. 3A**), vastly expanding 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 of 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, $p<1.65x10^{-32}$) 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 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, 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: 2-15) (**Fig. 3B**), and 84 of these were not seen with eQTL data, suggesting 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 GWAS catalog ($p<5x10^{-8}$; download:

25/01/2021) 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**) (*12*). For a quarter 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 neurotransmitters into the extracellular space (*28*); a locus that was 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%-CI: 0.82-0.91, p=3.8x10^{-8}) for each 1 s.d. 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 (*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 which might also be relevant to TLR4-mediated neuroinflammation in AD (*33*).

We observed multiple examples in which our cis-pQTL mapping identified biologically plausible candidates 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 s.d. higher plasma abundance of R-spondin-1 (odds ratio: 1.91, 95%-CI: (1.52-2.41), p-value=3.6x10^{-8}). R-spondin-1 is a secreted activator protein which 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 upregulates estrogen receptor alpha 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 testis and ovary, 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 sharedness of gene-protein-disease triplets through phenome-wide colocalization of cis-pQTL regions (*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 1,859 gene-protein-trait triplets (network edges, **Fig. 5 and S6**) comprising 412 protein targets and 506 curated traits (**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 (for example GCKR (n=197 traits), alpha-1-antitrypsin (n=79 traits), or apolipoprotein A-V (n=64 traits)) and established disease genes (for example 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 ten 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 - 4, maximum 32 for mean platelet volume) for blood cell parameters and biomarkers available in large-scale biobanks. For 104 out of 191 curated phenotypes with at least 3 colocalizing protein targets, we observed significant enrichment of pathways (false discovery rate (q-value) < 5%; **Table S7**). These reflected 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 1) connect etiologically related diseases, 2) provide biological context for new or emerging disorders, such as COVID-19, and 3) integrate information from different biological domains to establish mechanisms for known gene-disease links. For each of these scenarios, we provide selected examples below to highlight the scientific opportunities arising from this map and 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 (*42*), that differed substantially in the number of included cases (5,101 – 38,984), and observed that results were sensitive to the choice of outcome. We replicated *ABO* and *OAS1* as two candidate causal genes (*43*) (**Fig. S8**) with both showing consistent evidence across outcomes, ranging from susceptibility to COVID-19 to severe cases requiring hospitalization. The lead cis-pQTL for BGAT (rs576125, MAF=33.5%, within *ABO*) also colocalized with pulmonary embolism (**Fig. 5**), a common complication of severe COVID-19 (*44*), which might be 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) which each shared a genetic signal with only one of these four outcomes, and therefore requiring 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 shared signal at *SULT2A1*, a known gallstone locus (*45*), between bile salt sulfotransferase (SULT2A1) and risk of cholelithiasis (odds ratio per 1 s.d. higher normalized protein abundances: 2.12, 95%-CI: 1.66 – 2.70, p-value=2.1×10^{-37}) and cholecystectomy (odds ratio: 2.09, 95%-CI: 1.86 – 2.34, p-value=7.8×10^{-38}). We next used multi-trait colocalization (*46*) and further identified that 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 shared the same signal with high posterior probability (PP=99%) 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, clearly favor higher SULT2A1 activity as a mode of action. The concurrent inverse association with lower plasma concentrations of the secondary bile acid glycholithocholate indicates diminished formation of lithocholic acid, an essential detergent to solubilize fats, including cholesterol (*48*). Our vertical integration of diverse biology entities indicates 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 phenotypes was FBLN3 (extracellular matrix glycoprotein encoded by *EFEMP1*), showing 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=33.6%) being the causal variant in

multi-trait colocalization (**Fig. 6B** and **Fig. S10**). The locus has previously been reported but not connected across separate GWASs conducted for height (*49*), optic disc area (*50*), carpal tunnel syndrome (*51*), inguinal hernia (*52*), and lung function (*53*). *Efemp1* knock-out mice display abnormal elastic fiber morphology, develop different types of hernias, and have smaller body size and lower body fat (*54*), in line with the human spectrum of clinical features. FBLN3 is part of the extracellular matrix and widely expressed but its function is incompletely understood (*55*). 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) (*56*), characterized by visual disturbances and drusenoid deposits due to accumulating intracellular FBLN3. We observed sharedness of the common 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.

Differences of cis-pQTLs by sex and age

We systematically tested differences in the genetic associations of all protein targets included in the proteogenomic map (N=412) by age or sex. We identified a total of 14 protein targets that showed evidence for significant ($p<5.9x10^{-5}$) effect modification of the cis-pQTL by sex (N=10) or age (N=8), including four common to both (**Table S8**). This included biological 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-value$<1.7x10^{-467}$; men: beta=-0.64, p-value$<2.5x10^{-231}$). 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 (*57*, *58*).

We noted that most of the identified cis-pQTLs showed age- and sex-differential and not dimorphic effects (*59*) 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 (*60*). In general, our results are concordant with the few sex-specific effects of molecular QTLs reported so far (*61*, *62*) 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 (*63*) and identified 60 protein targets linked to at least one phenotype, including 22 protein targets linked to a disease (**Table S9**). We replicate established examples, such as the IL-6 receptor for rheumatoid arthritis or thrombin for deep venous thrombosis (**Fig. 5**). We also identified 31 candidates with potential repurposing opportunities for 1 to 8 diseases (total of 32 different indications), following a search and prioritization strategy in Open targets (*64*).

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.omiscience.org/apps/pgwas). The webserver provides intuitive representations of genetic findings and enables the look-up of summary statistics for individual SNPs, genes, and whole genomic regions across all protein targets. To interactively assess specificity and 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 provided 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, and disease 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. While previous studies started to exploit the phenotypic consequences of pQTLs, they have 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 (*65*). 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 demonstrate 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 multi-domain 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 due to natural cell turnover (*66*). Integration of genetic information allowed us to separate out such enclosed effects from other mechanisms leading to higher cell turn over or leakage, such as for SULT2A1 and the liver-specific effect of the *PNPLA3* variant. While this provides a tangible 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 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.

While 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. Firstly, 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 semi-quantitative 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, and we observed consistent cis-pQTLs for examples highlighted, including RSPO1, SULT2A1, and FBLN3, as

measured with Olink. Secondly, our study cohort consisted of predominantly healthy middle-aged participants of European-decent 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. Thirdly, 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 and inclusion of further data for additional phenotypes, in particular cancers, and understudied diseases, will provide additional insights.

MATERIALS and METHODS

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-decent in the Fenland study (**Table S1**) on 10.2 million genetic variants and plasma abundances of 4,775 distinct protein targets measured in plasma using established workflows (*15*). Protein targets were measured using the SomaScan v4 assay employing 4,979 single-stranded oligonucleotides (aptamers) with specific binding affinities to 4,775 unique protein targets (*67*, *68*). 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 ($p<1.004x10^{-11}$) 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 ($p<5x10^{-8}$) 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- (within ±500kb of the protein-encoding gene) or trans-pQTLs according to different specificity categories using linear regression models. We used statistical colocalization (*69*) to test for a shared genetic signal between expression or alternative splicing of the protein-encoding gene and the cis-pQTL in one out 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 (https://www.ebi.ac.uk/gwas/). We finally performed phenome-wide colocalization screens at 1,548 protein-encoding loci using publicly available (*70*) 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.

SUPPLEMENTARY MATERIALS

Materials and methods

Tables S1-S9

Fig. S1-S10

References (66 - 80)

REFERENCES AND NOTES

1. V. Emilsson, M. Ilkov, J. R. Lamb, N. Finkel, E. F. Gudmundsson, R. Pitts, H. Hoover, V. Gudmundsdottir, S. R. Horman, T. Aspelund, L. Shu, V. Trifonov, S. Sigurdsson, A. Manolescu, J. Zhu, Ö. Olafsson, J. Jakobsdottir, S. A. Lesley, J. To, J. Zhang, T. B. Harris, L. J. Launer, B. Zhang, G. Eiriksdottir, X. Yang, A. P. Orth, L. L. Jennings, V. Gudnason, Co-regulatory networks of human serum proteins link genetics to disease. *Science*. **361**, 769–773 (2018).

2. K. Suhre, M. Arnold, A. M. Bhagwat, R. J. Cotton, R. Engelke, J. Raffler, H. Sarwath, G. Thareja, A. Wahl, R. K. Delisle, L. Gold, M. Pezer, G. Lauc, M. A. E. D. Selim, D. O. Mook-Kanamori, E. K. Al-Dous, Y. A. Mohamoud, J. Malek, K. Strauch, H. Grallert, A. Peters, G. Kastenmüller, C. Gieger, J. Graumann, Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8** (2017), doi:10.1038/ncomms14357.

3. L. Folkersen, E. Fauman, M. Sabater-Lleal, R. J. Strawbridge, M. Frånberg, B. Sennblad, D. Baldassarre, F. Veglia, S. E. Humphries, R. Rauramaa, U. de Faire, A. J. Smit, P. Giral, S. Kurl, E. Mannarino, S. Enroth, Å. Johansson, S. B. Enroth, S. Gustafsson, L. Lind, C. Lindgren, A. P. Morris, V. Giedraitis, A. Silveira, A. Franco-Cereceda, E. Tremoli, IMPROVE study group, U. Gyllensten, E. Ingelsson, S. Brunak, P. Eriksson, D. Ziemek, A. Hamsten, A. Mälarstig, Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* **13**, e1006706 (2017).

4. B. B. Sun, J. C. Maranville, J. E. Peters, D. Stacey, J. R. Staley, J. Blackshaw, S. Burgess, T. Jiang, E. Paige, P. Surendran, C. Oliver-Williams, M. A. Kamat, B. P. Prins, S. K. Wilcox, E. S. Zimmerman, A. Chi, N. Bansal, S. L. Spain, A. M. Wood, N. W. Morrell, J. R. Bradley, N. Janjic, D. J. Roberts, W. H. Ouwehand, J. A. Todd, N. Soranzo, K. Suhre, D. S. Paul, C. S. Fox, R. M. Plenge, J. Danesh, H. Runz, A. S. Butterworth, Genomic atlas of the human plasma proteome. *Nature*. **558**, 73–79 (2018).

5. C. Yao, G. Chen, C. Song, J. Keefe, M. Mendelson, T. Huan, B. B. Sun, A. Laser, J. C. Maranville, H. Wu, J. E. Ho, P. Courchesne, A. Lyass, M. G. Larson, C. Gieger, J. Graumann, A. D. Johnson, J. Danesh, H. Runz, S.-J. J. Hwang, C. Liu, A. S. Butterworth, K. Suhre, D. Levy, Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018).

6. A. Gilly, Y.-C. Park, G. Png, A. Barysenka, I. Fischer, T. Bjørnland, L. Southam, D. Suveges, S. Neumeyer, N. W. Rayner, E. Tsafantakis, M. Karaleftheri, G. Dedoussis, E. Zeggini, Whole-genome sequencing analysis of the cardiometabolic proteome. *Nat. Commun.* **11**, 6336 (2020).

7. K. Suhre, M. I. McCarthy, J. M. Schwenk, Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.*, 1–19 (2020).

8. J. Zheng, V. Haberland, D. Baird, V. Walker, P. C. Haycock, M. R. Hurle, A. Gutteridge, P. Erola, Y. Liu, S. Luo, J. Robinson, T. G. Richardson, J. R. Staley, B. Elsworth, S. Burgess, B. B. Sun, J. Danesh, H. Runz, J. C. Maranville, H. M. Martin, J. Yarmolinsky, C. Laurin, M. V. Holmes, J. Z. Liu, K. Estrada, R. Santos, L. McCarthy, D. Waterworth, M. R. Nelson, G. D. Smith, A. S. Butterworth, G. Hemani, R. A. Scott, T. R. Gaunt, Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

9. L. Folkersen, S. Gustafsson, Q. Wang, D. H. Hansen, Å. K. Hedman, A. Schork, K. Page, D. V. Zhernakova, Y. Wu, J. Peters, N. Eriksson, S. E. Bergen, T. S. Boutin, A. D. Bretherick, S. Enroth, A. Kalnapenkis, J. R. Gådin, B. E. Suur, Y. Chen, L. Matic, J. D. Gale, J. Lee, W. Zhang, A. Quazi, M. Ala-Korpela, S. H. Choi, A. Claringbould, J. Danesh, G. Davey Smith, F. de Masi, S. Elmståhl, G. Engström, E. Fauman, C. Fernandez, L. Franke, P. W. Franks, V. Giedraitis, C. Haley, A.

Hamsten, A. Ingason, Å. Johansson, P. K. Joshi, L. Lind, C. M. Lindgren, S. Lubitz, T. Palmer, E. Macdonald-Dunlop, M. Magnusson, O. Melander, K. Michaelsson, A. P. Morris, R. Mägi, M. W. Nagle, P. M. Nilsson, J. Nilsson, M. Orho-Melander, O. Polasek, B. Prins, E. Pålsson, T. Qi, M. Sjögren, J. Sundström, P. Surendran, U. Võsa, T. Werge, R. Wernersson, H. J. Westra, J. Yang, A. Zhernakova, J. Ärnlöv, J. Fu, J. G. Smith, T. Esko, C. Hayward, U. Gyllensten, M. Landen, A. Siegbahn, J. F. Wilson, L. Wallentin, A. S. Butterworth, M. V. Holmes, E. Ingelsson, A. Mälarstig, Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).

10. M. Pietzner, E. Wheeler, J. Carrasco-Zanini, N. D. Kerrison, E. Oerton, M. Koprulu, J. Luan, A. D. Hingorani, S. A. Williams, N. J. Wareham, C. Langenberg, *bioRxiv*, in press, doi:10.1101/2021.03.18.435919.

11. T. Lindsay, K. Westgate, K. Wijndaele, S. Hollidge, N. Kerrison, N. Forouhi, S. Griffin, N. Wareham, S. Brage, Descriptive epidemiology of physical activity energy expenditure in UK adults (The Fenland study). *Int. J. Behav. Nutr. Phys. Act.* **16**, 126 (2019).

12. Associated code is available on GitHub. Extended Material and Methods are avialable as Supplemental Material (2021), , doi:10.5281/zenodo.5385532.

13. M. Narayan, Disulfide bonds: protein folding and subcellular protein trafficking. *FEBS J.* **279**, 2272–82 (2012).

14. M. Uhlén, M. J. Karlsson, A. Hober, A. Svensson, J. Scheffel, D. Kotol, W. Zhong, A. Tebani, L. Strandberg, F. Edfors, E. Sjöstedt, J. Mulder, A. Mardinoglu, A. Berling, S. Ekblad, M. Dannemeyer, S. Kanje, J. Rockberg, M. Lundqvist, M. Malm, A. Volk, P. Nilsson, A. Månberg, T. Dodig-crnkovic, E. Pin, M. Zwahlen, P. Oksvold, K. Von Feilitzen, R. S. Häussler, M. Hong, C. Lindskog, F. Ponten, B. Katona, J. Vuu, E. Lindström, J. Nielsen, J. Robinson, B. Ayoglu, D. Mahdessian, D. Sullivan, P. Thul, F. Danielsson, C. Stadler, E. Lundberg, G. Bergström, A. Gummesson, B. G. Voldborg, H. Tegel, S. Hober, B. Forsström, J. M. Schwenk, L. Fagerberg, Å. Sivertsson, The human secretome. **0274**, 1–9 (2019).

15. M. Pietzner, E. Wheeler, J. Carrasco-Zanini, J. Raffler, N. D. Kerrison, E. Oerton, V. P. W. Auyeung, J. Luan, C. Finan, J. P. Casas, R. Ostroff, S. A. Williams, G. Kastenmüller, M. Ralser, E. R. Gamazon, N. J. Wareham, A. D. Hingorani, C. Langenberg, Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020).

16. M. Eslam, L. Valenti, S. Romeo, Genetics and epigenetics of NAFLD and NASH: Clinical impact. *J. Hepatol.* **68**, 268–279 (2018).

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).

18. P. N. Newsome, R. Cramb, S. M. Davison, J. F. Dillon, M. Foulerton, E. M. Godfrey, R. Hall, U. Harrower, M. Hudson, A. Langford, A. Mackie, R. Mitchell-Thain, K. Sennett, N. C. Sheron, J. Verne, M. Walmsley, A. Yeoman, Guidelines on the management of abnormal liver blood tests. *Gut*. **67**, 6–19 (2018).

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).

20. D. I. Leavesley, A. S. Kashyap, T. Croll, M. Sivaramakrishnan, A. Shokoohmand, B. G. Hollier, Z. Upton, Vitronectin--master controller or micromanager? *IUBMB Life*. **65**, 807–18 (2013).

21. V. Guarani, E. M. McNeill, J. A. Paulo, E. L. Huttlin, F. Fröhlich, S. P. Gygi, D. Van Vactor, J. W. Harper, QIL1 is a novel mitochondrial protein required for MICOS complex stability and

cristae morphology. *Elife*. **4** (2015), doi:10.7554/eLife.06265.

22. P. B. Maguire, T. Donlon, M. Parsons, K. Wynne, E. Dillon, F. Ní Áinle, P. B. Szklanna, Proteomic Analysis Reveals a Strong Association of β-Catenin With Cadherin Adherens Junctions in Resting Human Platelets. *Proteomics*. **18** (2018), doi:10.1002/pmic.201700419.

23. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, C. von Mering, STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

24. GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. **369**, 1318–1330 (2020).

25. C. Buccitelli, M. Selbach, mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644 (2020).

26. U. Võsa, A. Claringbould, H.-J. Westra, M. Jan Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Kasela, N. Pervjakova, I. Alvaes, M.-J. Fave, M. Agbessi, M. Christiansen, R. Jansen, I. Seppälä, L. Tong, A. Teumer, K. Schramm, G. Hemani, J. Verlouw, H. Yaghootkar, R. Sönmez, A. Brown, V. Kukushkina, A. Kalnapenkis, S. Rüeger, E. Porcu, J. Kronberg-Guzman, J. Kettunen, J. Powell, B. Lee, F. Zhang, W. Arindrarto, F. Beutner, B. Consortium, H. Brugge, Q. Consortium, J. Dmitreva, M. Elansary, B. P. Fairfax, M. Georges, B. T. Heijmans, M. Kähönen, Y. Kim, J. C. Knight, P. Kovacs, K. Krohn, S. Li, M. Loeffler, U. M. Marigorta, H. Mei, Y. Momozawa, M. Müller-Nurasyid, M. Nauck, M. Nivard, B. Penninx, J. Pritchard, O. Raitakari, O. Rotzchke, E. P. Slagboom, C. DA Stehouwer, M. Stumvoll, P. Sullivan, J. Thiery, A. Tönjes, J. van Dongen, M. van Iterson, J. Veldink, U. Völker, C. Wijmenga, M. Swertz, A. Andiappan, G. W. Montgomery, S. Ripatti, M. Perola, Z. Kutalik, P. Awadalla, L. Milani, W. Ouwehand, K. Downes, O. Stegle, A. Battle, J. Yang, P. M. Visscher, M. Scholz, G. Gibson, T. Esko, L. Franke, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Kasela, N. Pervjakova, I. Alvaes, M.-J. Fave, M. Agbessi, M. Christiansen, R. Jansen, I. Seppälä, L. Tong, A. Teumer, K. Schramm, G. Hemani, J. Verlouw, H. Yaghootkar, R. Sönmez, A. Brown, V. Kukushkina, A. Kalnapenkis, S. Rüeger, E. Porcu, J. Kronberg-Guzman, J. Kettunen, J. Powell, B. Lee, F. Zhang, W. Arindrarto, F. Beutner, H. Brugge, J. Dmitreva, M. Elansary, B. P. Fairfax, M. Georges, B. T. Heijmans, M. Kähönen, Y. Kim, J. C. Knight, P. Kovacs, K. Krohn, S. Li, M. Loeffler, U. M. Marigorta, H. Mei, Y. Momozawa, M. Müller-Nurasyid, M. Nauck, M. Nivard, B. Penninx, J. Pritchard, O. Raitakari, O. Rotzchke, E. P. Slagboom, C. DA Stehouwer, M. Stumvoll, P. Sullivan, P. Hoen, J. Thiery, A. Tönjes, J. van Dongen, M. van Iterson, J. Veldink, U. Völker, C. Wijmenga, M. Swertz, A. Andiappan, G. W. Montgomery, S. Ripatti, M. Perola, Z. Kutalik, E. Dermitzakis, S. Bergmann, T. Frayling, J. van Meurs, H. Prokisch, H. Ahsan, B. Pierce, T. Lehtimäki, D. Boomsma, B. Psaty, S. Gharib, P. Awadalla, L. Milani, W. Ouwehand, K. Downes, O. Stegle, A. Battle, J. Yang, P. M. Visscher, M. Scholz, G. Gibson, T. Esko, L. Franke, Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*. **18**, 447367 (2018).

27. S. M. Sternson, D. Atasoy, Agouti-related protein neuron circuits that regulate appetite. *Neuroendocrinology*. **100**, 95–102 (2014).

28. R. W. Baker, F. M. Hughson, Chaperoning SNARE assembly and disassembly. *Nat. Rev. Mol. Cell Biol.* **17**, 465–79 (2016).

29. I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, L. Athanasiu, N. Voyle, P. Proitsi, A. Witoelar, S. Stringer, D. Aarsland, I. S. Almdahl, F. Andersen, S. Bergh, F. Bettella, S. Bjornsson, A. Brækhus, G. Bråthen, C. de Leeuw, R. S. Desikan, S. Djurovic, L. Dumitrescu, T. Fladby, T. J. Hohman, P. V. Jonsson, S. J. Kiddle, A. Rongve, I. Saltvedt, S. B. Sando, G. Selbæk, M. Shoai, N. G. Skene, J. Snaedal, E. Stordal, I. D.

Ulstein, Y. Wang, L. R. White, J. Hardy, J. Hjerling-Leffler, P. F. Sullivan, W. M. van der Flier, R. Dobson, L. K. Davis, H. Stefansson, K. Stefansson, N. L. Pedersen, S. Ripke, O. A. Andreassen, D. Posthuma, Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

30. J. Schwartzentruber, S. Cooper, J. Z. Liu, I. Barrio-Hernandez, E. Bello, N. Kumasaka, A. M. H. Young, R. J. M. Franklin, T. Johnson, K. Estrada, D. J. Gaffney, P. Beltrao, A. Bassett, Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat. Genet.* (2021), doi:10.1038/s41588-020-00776-w.

31. S. Aggarwal, P. K. Dabla, S. Arora, Prostasin: An Epithelial Sodium Channel Regulator. *J. biomarkers*. **2013**, 179864 (2013).

32. Y. Sugitani, A. Nishida, O. Inatomi, M. Ohno, T. Imai, M. Kawahara, K. Kitamura, A. Andoh, 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).

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).

34. T. A. O'Mara, D. M. Glubb, F. Amant, D. Annibali, K. Ashton, J. Attia, P. L. Auer, M. W. Beckmann, A. Black, M. K. Bolla, H. Brauch, H. Brenner, L. Brinton, D. D. Buchanan, B. Burwinkel, J. Chang-Claude, S. J. Chanock, C. Chen, M. M. Chen, T. H. T. Cheng, C. L. Clarke, M. Clendenning, L. S. Cook, F. J. Couch, A. Cox, M. Crous-Bous, K. Czene, F. Day, J. Dennis, J. Depreeuw, J. A. Doherty, T. Dörk, S. C. Dowdy, M. Dürst, A. B. Ekici, P. A. Fasching, B. L. Fridley, C. M. Friedenreich, L. Fritschi, J. Fung, M. García-Closas, M. M. Gaudet, G. G. Giles, E. L. Goode, M. Gorman, C. A. Haiman, P. Hall, S. E. Hankison, C. S. Healey, A. Hein, P. Hillemanns, S. Hodgson, E. A. Hoivik, E. G. Holliday, J. L. Hopper, D. J. Hunter, A. Jones, C. Krakstad, V. N. Kristensen, D. Lambrechts, L. Le Marchand, X. Liang, A. Lindblom, J. Lissowska, J. Long, L. Lu, A. M. Magliocco, L. Martin, M. McEvoy, A. Meindl, K. Michailidou, R. L. Milne, M. Mints, G. W. Montgomery, R. Nassir, H. Olsson, I. Orlow, G. Otton, C. Palles, J. R. B. Perry, J. Peto, L. Pooler, J. Prescott, T. Proietto, T. R. Rebbeck, H. A. Risch, P. A. W. Rogers, M. Rübner, I. Runnebaum, C. Sacerdote, G. E. Sarto, F. Schumacher, R. J. Scott, V. W. Setiawan, M. Shah, X. Sheng, X.-O. Shu, M. C. Southey, A. J. Swerdlow, E. Tham, J. Trovik, C. Turman, J. P. Tyrer, C. Vachon, D. VanDen Berg, A. Vanderstichele, Z. Wang, P. M. Webb, N. Wentzensen, H. M. J. Werner, S. J. Winham, A. Wolk, L. Xia, Y.-B. Xiang, H. P. Yang, H. Yu, W. Zheng, P. D. P. Pharoah, A. M. Dunning, P. Kraft, I. De Vivo, I. Tomlinson, D. F. Easton, A. B. Spurdle, D. J. Thompson, Identification of nine new susceptibility loci for endometrial cancer. *Nat. Commun.* **9**, 3166 (2018).

35. M. E. Binnerts, K.-A. Kim, J. M. Bright, S. M. Patel, K. Tran, M. Zhou, J. M. Leung, Y. Liu, W. E. Lomas, M. Dixon, S. A. Hazell, M. Wagle, W.-S. Nie, N. Tomasevic, J. Williams, X. Zhan, M. D. Levy, W. D. Funk, A. Abo, R-Spondin1 regulates Wnt signaling by inhibiting internalization of LRP6. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14700–5 (2007).

36. A. Geng, T. Wu, C. Cai, W. Song, J. Wang, Q. C. Yu, Y. A. Zeng, A novel function of R-spondin1 in regulating estrogen receptor expression independent of Wnt/β-catenin signaling. *Elife*. **9** (2020), doi:10.7554/eLife.56434.

37. A.-A. Chassot, S. T. Bradford, A. Auguste, E. P. Gregoire, E. Pailhoux, D. G. de Rooij, A. Schedl, M.-C. Chaboissier, WNT4 and RSPO1 together are required for cell proliferation in the early mouse gonad. *Development*. **139**, 4461–72 (2012).

38. A.-A. Chassot, F. Ranc, E. P. Gregoire, H. L. Roepers-Gajadien, M. M. Taketo, G. Camerino, D. G. de Rooij, A. Schedl, M.-C. Chaboissier, Activation of beta-catenin signaling by Rspo1

controls differentiation of the mammalian ovary. *Hum. Mol. Genet.* **17**, 1264–77 (2008).

39. P. Edery, S. Lyonnet, L. M. Mulligan, A. Pelet, E. Dow, L. Abel, S. Holder, C. Nihoul-Fékété, B. A. Ponder, A. Munnich, Mutations of the RET proto-oncogene in Hirschsprung's disease. *Nature*. **367**, 378–80 (1994).

40. E. A. Stahl, S. Raychaudhuri, E. F. Remmers, G. Xie, S. Eyre, B. P. Thomson, Y. Li, F. A. S. Kurreeman, A. Zhernakova, A. Hinks, C. Guiducci, R. Chen, L. Alfredsson, C. I. Amos, K. G. Ardlie, A. Barton, J. Bowes, E. Brouwer, N. P. Burtt, J. J. Catanese, J. Coblyn, M. J. H. Coenen, K. H. Costenbader, L. A. Criswell, J. B. A. Crusius, J. Cui, P. I. W. De Bakker, P. L. De Jager, B. Ding, P. Emery, E. Flynn, P. Harrison, L. J. Hocking, T. W. J. Huizinga, D. L. Kastner, X. Ke, A. T. Lee, X. Liu, P. Martin, A. W. Morgan, L. Padyukov, M. D. Posthumus, T. R. D. J. Radstake, D. M. Reid, M. Seielstad, M. F. Seldin, N. A. Shadick, S. Steer, P. P. Tak, W. Thomson, A. H. M. Van Der Helm-Van Mil, I. E. Van Der Horst-Bruinsma, C. E. Van Der Schoot, P. L. C. M. Van Riel, M. E. Weinblatt, A. G. Wilson, G. J. Wolbink, B. P. Wordsworth, C. Wijmenga, E. W. Karlson, R. E. M. Toes, N. De Vries, A. B. Begovich, J. Worthington, K. A. Siminovitch, P. K. Gregersen, L. Klareskog, R. M. Plenge, Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).

41. N. Harii, C. J. Lewis, V. Vasko, K. McCall, U. Benavides-Peralta, X. Sun, M. D. Ringel, M. Saji, C. Giuliani, G. Napolitano, D. J. Goetz, L. D. Kohn, 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**, 1231–50 (2005).

42. C.-19 H. G. Initiative, Mapping the human genetic architecture of COVID-19. *Nature* (2021), doi:10.1038/s41586-021-03767-x.

43. S. Zhou, G. Butler-Laporte, T. Nakanishi, D. R. Morrison, J. Afilalo, M. Afilalo, L. Laurent, M. Pietzner, N. Kerrison, K. Zhao, E. Brunet-Ratnasingham, D. Henry, N. Kimchi, Z. Afrasiabi, N. Rezk, M. Bouab, L. Petitjean, C. Guzman, X. Xue, C. Tselios, B. Vulesevic, O. Adeleye, T. Abdullah, N. Almamlouk, Y. Chen, M. Chassé, M. Durand, C. Paterson, J. Normark, R. Frithiof, M. Lipcsey, M. Hultström, C. M. T. Greenwood, H. Zeberg, C. Langenberg, E. Thysell, M. Pollak, V. Mooser, V. Forgetta, D. E. Kaufmann, J. B. Richards, A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659–667 (2021).

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).

45. A. D. Joshi, C. Andersson, S. Buch, S. Stender, R. Noordam, L.-C. Weng, P. E. Weeke, P. L. Auer, B. Boehm, C. Chen, H. Choi, G. Curhan, J. C. Denny, I. De Vivo, J. D. Eicher, D. Ellinghaus, A. R. Folsom, C. Fuchs, M. Gala, J. Haessler, A. Hofman, F. Hu, D. J. Hunter, H. L. A. Janssen, J. H. Kang, C. Kooperberg, P. Kraft, W. Kratzer, W. Lieb, P. L. Lutsey, S. Darwish Murad, B. G. Nordestgaard, L. R. Pasquale, A. P. Reiner, P. M. Ridker, E. Rimm, L. M. Rose, C. M. Shaffer, C. Schafmayer, R. M. Tamimi, A. G. Uitterlinden, U. Völker, H. Völzke, Y. Wakabayashi, J. L. Wiggs, J. Zhu, D. M. Roden, B. H. Stricker, W. Tang, A. Teumer, J. Hampe, A. Tybjærg-Hansen, D. I. Chasman, A. T. Chan, A. D. Johnson, Four Susceptibility Loci for Gallstone Disease Identified in a Meta-analysis of Genome-Wide Association Studies. *Gastroenterology*. **151**, 351-363.e28 (2016).

46. C. N. Foley, J. R. Staley, P. G. Breen, B. B. Sun, P. D. W. Kirk, S. Burgess, J. M. M. Howson, A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* **12**, 764 (2021).

47. S.-Y. Y. Shin, E. B. Fauman, A.-K. K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I.

Erte, V. Forgetta, T.-P. P. Yang, K. Walter, C. Menni, L. Chen, L. Vasquez, A. M. Valdes, C. L. Hyde, V. Wang, D. Ziemek, P. Roberts, L. Xi, E. Grundberg, M. Waldenberger, J. B. Richards, R. P. Mohney, M. V. Milburn, S. L. John, J. Trimmer, F. J. Theis, J. P. Overington, K. Suhre, M. J. Brosnan, C. Gieger, G. Kastenmüller, T. D. Spector, N. Soranzo, N. Soranzo, An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).

48. F. Lammert, K. Gurusamy, C. W. Ko, J.-F. Miquel, N. Méndez-Sánchez, P. Portincasa, K. J. van Erpecum, C. J. van Laarhoven, D. Q.-H. Wang, Gallstones. *Nat. Rev. Dis. Prim.* **2**, 16024 (2016).

49. A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, N. Amin, M. L. Buchkovich, D. C. Croteau-Chonka, F. R. Day, Y. Duan, T. Fall, R. Fehrmann, T. Ferreira, A. U. Jackson, J. Karjalainen, K. S. Lo, A. E. Locke, R. Mägi, E. Mihailov, E. Porcu, J. C. Randall, A. Scherag, A. A. E. Vinkhuyzen, H.-J. Westra, T. W. Winkler, T. Workalemahu, J. H. Zhao, D. Absher, E. Albrecht, D. Anderson, J. Baron, M. Beekman, A. Demirkan, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, R. M. Fraser, A. Goel, J. Gong, A. E. Justice, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, J. C. Lui, M. Mangino, I. Mateo Leach, C. Medina-Gomez, M. A. Nalls, D. R. Nyholt, C. D. Palmer, D. Pasko, S. Pechlivanis, I. Prokopenko, J. S. Ried, S. Ripke, D. Shungin, A. Stancáková, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, U. Afzal, J. Arnlöv, G. M. Arscott, S. Bandinelli, A. Barrett, C. Bellis, A. J. Bennett, C. Berne, M. Blüher, J. L. Bolton, Y. Böttcher, H. A. Boyd, M. Bruinenberg, B. M. Buckley, S. Buyske, I. H. Caspersen, P. S. Chines, R. Clarke, S. Claudi-Boehm, M. Cooper, E. W. Daw, P. A. De Jong, J. Deelen, G. Delgado, J. C. Denny, R. Dhonukshe-Rutten, M. Dimitriou, A. S. F. Doney, M. Dörr, N. Eklund, E. Eury, L. Folkersen, M. E. Garcia, F. Geller, V. Giedraitis, A. S. Go, H. Grallert, T. B. Grammer, J. Gräßler, H. Grönberg, L. C. P. G. M. de Groot, C. J. Groves, J. Haessler, P. Hall, T. Haller, G. Hallmans, A. Hannemann, C. A. Hartman, M. Hassinen, C. Hayward, N. L. Heard-Costa, Q. Helmer, G. Hemani, A. K. Henders, H. L. Hillege, M. A. Hlatky, W. Hoffmann, P. Hoffmann, O. Holmen, J. J. Houwing-Duistermaat, T. Illig, A. Isaacs, A. L. James, J. Jeff, B. Johansen, Å. Johansson, J. Jolley, T. Juliusdottir, J. Junttila, A. N. Kho, L. Kinnunen, N. Klopp, T. Kocher, W. Kratzer, P. Lichtner, L. Lind, J. Lindström, S. Lobbens, M. Lorentzon, Y. Lu, V. Lyssenko, P. K. E. Magnusson, A. Mahajan, M. Maillard, W. L. McArdle, C. A. McKenzie, S. McLachlan, P. J. McLaren, C. Menni, S. Merger, L. Milani, A. Moayyeri, K. L. Monda, M. A. Morken, G. Müller, M. Müller-Nurasyid, A. W. Musk, N. Narisu, M. Nauck, I. M. Nolte, M. M. Nöthen, L. Oozageer, S. Pilz, N. W. Rayner, F. Renstrom, N. R. Robertson, L. M. Rose, R. Roussel, S. Sanna, H. Scharnagl, S. Scholtens, F. R. Schumacher, H. Schunkert, R. A. Scott, J. Sehmi, T. Seufferlein, J. Shi, K. Silventoinen, J. H. Smit, A. V. Smith, J. Smolonska, A. V Stanton, K. Stirrups, D. J. Stott, H. M. Stringham, J. Sundström, M. A. Swertz, A.-C. Syvänen, B. O. Tayo, G. Thorleifsson, J. P. Tyrer, S. van Dijk, N. M. van Schoor, N. van der Velde, D. van Heemst, F. V. A. van Oort, S. H. Vermeulen, N. Verweij, J. M. Vonk, L. L. Waite, M. Waldenberger, R. Wennauer, L. R. Wilkens, C. Willenborg, T. Wilsgaard, M. K. Wojczynski, A. Wong, A. F. Wright, Q. Zhang, D. Arveiler, S. J. L. Bakker, J. Beilby, R. N. Bergman, S. Bergmann, R. Biffar, J. Blangero, D. I. Boomsma, S. R. Bornstein, P. Bovet, P. Brambilla, M. J. Brown, H. Campbell, M. J. Caulfield, A. Chakravarti, R. Collins, F. S. Collins, D. C. Crawford, L. A. Cupples, J. Danesh, U. de Faire, H. M. den Ruijter, R. Erbel, J. Erdmann, J. G. Eriksson, M. Farrall, E. Ferrannini, J. Ferrières, I. Ford, N. G. Forouhi, T. Forrester, R. T. Gansevoort, P. V Gejman, C. Gieger, A. Golay, O. Gottesman, V. Gudnason, U. Gyllensten, D. W. Haas, A. S. Hall, T. B. Harris, A. T. Hattersley, A. C. Heath, C. Hengstenberg, A. A. Hicks, L. A. Hindorff, A. D. Hingorani, A. Hofman, G. K. Hovingh, S. E. Humphries, S. C. Hunt, E. Hypponen, K. B. Jacobs, M.-R. Jarvelin, P. Jousilahti, A. M. Jula, J. Kaprio, J. J. P. Kastelein, M. Kayser, F. Kee, S. M. Keinanen-Kiukaanniemi, L. A. Kiemeney, J. S. Kooner, C. Kooperberg, S. Koskinen, P. Kovacs, A. T. Kraja, M. Kumari, J. Kuusisto, T. A. Lakka, C. Langenberg, L. Le Marchand, T. Lehtimäki, S. Lupoli, P. A. F. Madden, S. Männistö, P. Manunta, A. Marette, T. C. Matise, B. McKnight, T. Meitinger, F. L. Moll, G. W. Montgomery,

A. D. Morris, A. P. Morris, J. C. Murray, M. Nelis, C. Ohlsson, A. J. Oldehinkel, K. K. Ong, W. H. Ouwehand, G. Pasterkamp, A. Peters, P. P. Pramstaller, J. F. Price, L. Qi, O. T. Raitakari, T. Rankinen, D. C. Rao, T. K. Rice, M. Ritchie, I. Rudan, V. Salomaa, N. J. Samani, J. Saramies, M. A. Sarzynski, P. E. H. Schwarz, S. Sebert, P. Sever, A. R. Shuldiner, J. Sinisalo, V. Steinthorsdottir, R. P. Stolk, J.-C. Tardif, A. Tönjes, A. Tremblay, E. Tremoli, J. Virtamo, M.-C. Vohl, Electronic Medical Records and Genomics (eMEMERGEGE) Consortium, MIGen Consortium, PAGEGE Consortium, LifeLines Cohort Study, P. Amouyel, F. W. Asselbergs, T. L. Assimes, M. Bochud, B. O. Boehm, E. Boerwinkle, E. P. Bottinger, C. Bouchard, S. Cauchi, J. C. Chambers, S. J. Chanock, R. S. Cooper, P. I. W. de Bakker, G. Dedoussis, L. Ferrucci, P. W. Franks, P. Froguel, L. C. Groop, C. A. Haiman, A. Hamsten, M. G. Hayes, J. Hui, D. J. Hunter, K. Hveem, J. W. Jukema, R. C. Kaplan, M. Kivimaki, D. Kuh, M. Laakso, Y. Liu, N. G. Martin, W. März, M. Melbye, S. Moebus, P. B. Munroe, I. Njølstad, B. A. Oostra, C. N. A. Palmer, N. L. Pedersen, M. Perola, L. Pérusse, U. Peters, J. E. Powell, C. Power, T. Quertermous, R. Rauramaa, E. Reinmaa, P. M. Ridker, F. Rivadeneira, J. I. Rotter, T. E. Saaristo, D. Saleheen, D. Schlessinger, P. E. Slagboom, H. Snieder, T. D. Spector, K. Strauch, M. Stumvoll, J. Tuomilehto, M. Uusitupa, P. van der Harst, H. Völzke, M. Walker, N. J. Wareham, H. Watkins, H.-E. Wichmann, J. F. Wilson, P. Zanen, P. Deloukas, I. M. Heid, C. M. Lindgren, K. L. Mohlke, E. K. Speliotes, U. Thorsteinsdottir, I. Barroso, C. S. Fox, K. E. North, D. P. Strachan, J. S. Beckmann, S. I. Berndt, M. Boehnke, I. B. Borecki, M. I. McCarthy, A. Metspalu, K. Stefansson, A. G. Uitterlinden, C. M. van Duijn, L. Franke, C. J. Willer, A. L. Price, G. Lettre, R. J. F. Loos, M. N. Weedon, E. Ingelsson, J. R. O'Connell, G. R. Abecasis, D. I. Chasman, M. E. Goddard, P. M. Visscher, J. N. Hirschhorn, T. M. Frayling, Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–86 (2014).

50. H. Springelkamp, A. I. Iglesias, A. Mishra, R. Höhn, R. Wojciechowski, A. P. Khawaja, A. Nag, Y. X. Wang, J. J. Wang, G. Cuellar-Partida, J. Gibson, J. N. C. Bailey, E. N. Vithana, P. Gharahkhani, T. Boutin, W. D. Ramdas, T. Zeller, R. N. Luben, E. Yonova-Doing, A. C. Viswanathan, S. Yazar, A. J. Cree, J. L. Haines, J. Y. Koh, E. Souzeau, J. F. Wilson, N. Amin, C. Müller, C. Venturini, L. S. Kearns, J. H. Kang, NEIGHBORHOOD Consortium, Y. C. Tham, T. Zhou, E. M. van Leeuwen, S. Nickels, P. Sanfilippo, J. Liao, H. van der Linde, W. Zhao, L. M. E. van Koolwijk, L. Zheng, F. Rivadeneira, M. Baskaran, S. J. van der Lee, S. Perera, P. T. V. M. de Jong, B. A. Oostra, A. G. Uitterlinden, Q. Fan, A. Hofman, E.-S. Tai, J. R. Vingerling, X. Sim, R. C. W. Wolfs, Y. Y. Teo, H. G. Lemij, C. C. Khor, R. Willemsen, K. J. Lackner, T. Aung, N. M. Jansonius, G. Montgomery, P. S. Wild, T. L. Young, K. P. Burdon, P. G. Hysi, L. R. Pasquale, T. Y. Wong, C. C. W. Klaver, A. W. Hewitt, J. B. Jonas, P. Mitchell, A. J. Lotery, P. J. Foster, V. Vitart, N. Pfeiffer, J. E. Craig, D. A. Mackey, C. J. Hammond, J. L. Wiggs, C.-Y. Cheng, C. M. van Duijn, S. MacGregor, 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).

51. A. Wiberg, M. Ng, A. B. Schmid, R. W. Smillie, G. Baskozos, M. V Holmes, K. Künnapuu, R. Mägi, D. L. Bennett, D. Furniss, A genome-wide association analysis identifies 16 novel susceptibility loci for carpal tunnel syndrome. *Nat. Commun.* **10**, 1030 (2019).

52. E. Jorgenson, N. Makki, L. Shen, D. C. Chen, C. Tian, W. L. Eckalbar, D. Hinds, N. Ahituv, A. Avins, A genome-wide association study identifies four novel susceptibility loci underlying inguinal hernia. *Nat. Commun.* **6**, 10130 (2015).

53. N. Shrine, A. L. Guyatt, A. M. Erzurumluoglu, V. E. Jackson, B. D. Hobbs, C. A. Melbourne, C. Batini, K. A. Fawcett, K. Song, P. Sakornsakolpat, X. Li, R. Boxall, N. F. Reeve, M. Obeidat, J. H. Zhao, M. Wielscher, S. Weiss, K. A. Kentistou, J. P. Cook, B. B. Sun, J. Zhou, J. Hui, S. Karrasch, M. Imboden, S. E. Harris, J. Marten, S. Enroth, S. M. Kerr, I. Surakka, V. Vitart, T. Lehtimäki, R. J. Allen, P. S. Bakke, T. H. Beaty, E. R. Bleecker, Y. Bossé, C. A. Brandsma, Z. Chen, J. D. Crapo, J. Danesh, D. L. DeMeo, F. Dudbridge, R. Ewert, C. Gieger, A. Gulsvik, A. L. Hansell, K. Hao, J. D. Hoffman, J. E. Hokanson, G. Homuth, P. K. Joshi, P. Joubert, C. Langenberg, X. Li, L. Li, K. Lin, L.

Lind, N. Locantore, J. Luan, A. Mahajan, J. C. Maranville, A. Murray, D. C. Nickle, R. Packer, M. M. Parker, M. L. Paynton, D. J. Porteous, D. Prokopenko, D. Qiao, R. Rawal, H. Runz, I. Sayers, D. D. Sin, B. H. Smith, M. Soler Artigas, D. Sparrow, R. Tal-Singer, P. R. H. J. Timmers, M. Van den Berge, J. C. Whittaker, P. G. Woodruff, L. M. Yerges-Armstrong, O. G. Troyanskaya, O. T. Raitakari, M. Kähönen, O. Polašek, U. Gyllensten, I. Rudan, I. J. Deary, N. M. Probst-Hensch, H. Schulz, A. L. James, J. F. Wilson, B. Stubbe, E. Zeggini, M. R. Jarvelin, N. Wareham, E. K. Silverman, C. Hayward, A. P. Morris, A. S. Butterworth, R. A. Scott, R. G. Walters, D. A. Meyers, M. H. Cho, D. P. Strachan, I. P. Hall, M. D. Tobin, L. V. Wain, New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493 (2019).

54. P. J. McLaughlin, B. Bakall, J. Choi, Z. Liu, T. Sasaki, E. C. Davis, A. D. Marmorstein, L. Y. Marmorstein, Lack of fibulin-3 causes early aging and herniation, but not macular degeneration in mice. *Hum. Mol. Genet.* **16**, 3059–70 (2007).

55. I. Livingstone, V. N. Uversky, D. Furniss, A. Wiberg, The Pathophysiological Significance of Fibulin-3. *Biomolecules*. **10** (2020), doi:10.3390/biom10091294.

56. L. Y. Marmorstein, F. L. Munier, Y. Arsenijevic, D. F. Schorderet, P. J. McLaughlin, D. Chung, E. Traboulsi, A. D. Marmorstein, Aberrant accumulation of EFEMP1 underlies drusen formation in Malattia Leventinese and age-related macular degeneration. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 13067–72 (2002).

57. B. Hollis, F. R. Day, A. S. Busch, D. J. Thompson, A. L. G. Soares, P. R. H. J. Timmers, A. Kwong, D. F. Easton, P. K. Joshi, N. J. Timpson, PRACTICAL Consortium, 23andMe Research Team, K. K. Ong, J. R. B. Perry, Genomic analysis of male puberty timing highlights shared genetic basis with hair colour and lifespan. *Nat. Commun.* **11**, 1536 (2020).

58. F. R. Day, D. J. Thompson, H. Helgason, D. I. Chasman, H. Finucane, P. Sulem, K. S. Ruth, S. Whalen, A. K. Sarkar, E. Albrecht, E. Altmaier, M. Amini, C. M. Barbieri, T. Boutin, A. Campbell, E. Demerath, A. Giri, C. He, J. J. Hottenga, R. Karlsson, I. Kolcic, P.-R. Loh, K. L. Lunetta, M. Mangino, B. Marco, G. McMahon, S. E. Medland, I. M. Nolte, R. Noordam, T. Nutile, L. Paternoster, N. Perjakova, E. Porcu, L. M. Rose, K. E. Schraut, A. V Segrè, A. V Smith, L. Stolk, A. Teumer, I. L. Andrulis, S. Bandinelli, M. W. Beckmann, J. Benitez, S. Bergmann, M. Bochud, E. Boerwinkle, S. E. Bojesen, M. K. Bolla, J. S. Brand, H. Brauch, H. Brenner, L. Broer, T. Brüning, J. E. Buring, H. Campbell, E. Catamo, S. Chanock, G. Chenevix-Trench, T. Corre, F. J. Couch, D. L. Cousminer, A. Cox, L. Crisponi, K. Czene, G. Davey Smith, E. J. C N de Geus, R. de Mutsert, I. De Vivo, J. Dennis, P. Devilee, I. dos-Santos-Silva, A. M. Dunning, J. G. Eriksson, P. A. Fasching, L. Fernández-Rhodes, L. Ferrucci, D. Flesch-Janys, L. Franke, M. Gabrielson, I. Gandin, G. G. Giles, H. Grallert, D. F. Gudbjartsson, P. Guénel, P. Hall, E. Hallberg, U. Hamann, T. B. Harris, C. A. Hartman, G. Heiss, M. J. Hooning, J. L. Hopper, F. Hu, D. J. Hunter, M. Arfan Ikram, H. Kyung Im, M.-R. Järvelin, P. K. Joshi, D. Karasik, M. Kellis, Z. Kutalik, G. LaChance, D. Lambrechts, C. Langenberg, L. J. Launer, J. S. E Laven, S. Lenarduzzi, J. Li, P. A. Lind, S. Lindstrom, Y. Liu, an Luan, R. Mägi, A. Mannermaa, H. Mbarek, M. I. McCarthy, C. Meisinger, T. Meitinger, C. Menni, A. Metspalu, K. Michailidou, L. Milani, R. L. Milne, G. W. Montgomery, A. M. Mulligan, M. A. Nalls, P. Navarro, H. Nevanlinna, D. R. Nyholt, A. J. Oldehinkel, T. A. O, S. Padmanabhan, A. Palotie, N. Pedersen, A. Peters, J. Peto, P. D. P Pharoah, A. Pouta, P. Radice, I. Rahman, S. M. Ring, A. Robino, F. R. Rosendaal, I. Rudan, R. Rueedi, D. Ruggiero, C. F. Sala, M. K. Schmidt, R. A. Scott, M. Shah, R. Sorice, M. C. Southey, U. Sovio, M. Stampfer, M. Steri, K. Strauch, T. Tanaka, E. Tikkanen, N. J. Timpson, M. Traglia, T. Truong, J. P. Tyrer, A. G. Uitterlinden, D. R. Velez Edwards, V. Vitart, U. Völker, P. Vollenweider, Q. Wang, E. Widen, K. Willems van Dijk, G. Willemsen, R. Winqvist, B. H. R Wolffenbuttel, J. Hua Zhao, M. Zoledziewska, M. Zygmunt, B. Z. Alizadeh, D. I. Boomsma, M. Ciullo, F. Cucca, T. Esko, N. Franceschini, C. Gieger, V. Gudnason, C. Hayward, P. Kraft, D. A. Lawlor, P. K. E Magnusson, N.

G. Martin, D. O. Mook-Kanamori, E. A. Nohr, O. Polasek, D. Porteous, A. L. Price, P. M. Ridker, H. Snieder, T. D. Spector, D. Stöckl, D. Toniolo, S. Ulivi, J. A. Visser, H. Völzke, N. J. Wareham, J. F. Wilson, A. B. Spurdle, U. Thorsteindottir, K. S. Pollard, D. F. Easton, J. Y. Tung, J. Chang-Claude, D. Hinds, A. Murray, J. M. Murabito, K. Stefansson, K. K. Ong, J. R. B Perry, Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Publ. Gr.* **49** (2017), doi:10.1038/ng.3841.

59. 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).

60. H. Aschard, A perspective on interaction effects in genetic association studies. *Genet. Epidemiol.* **40**, 678–688 (2016).

61. M. Oliva, M. Muñoz-Aguirre, S. Kim-Hellmuth, V. Wucher, A. D. H. Gewirtz, D. J. Cotter, P. Parsana, S. Kasela, B. Balliu, A. Viñuela, S. E. Castel, P. Mohammadi, F. Aguet, Y. Zou, E. A. Khramtsova, A. D. Skol, D. Garrido-Martín, F. Reverter, A. Brown, P. Evans, E. R. Gamazon, A. Payne, R. Bonazzola, A. N. Barbeira, A. R. Hamel, A. Martinez-Perez, J. M. Soria, GTEx Consortium, B. L. Pierce, M. Stephens, E. Eskin, E. T. Dermitzakis, A. V Segrè, H. K. Im, B. E. Engelhardt, K. G. Ardlie, S. B. Montgomery, A. J. Battle, T. Lappalainen, R. Guigó, B. E. Stranger, The impact of sex on gene expression across human tissues. *Science*. **369** (2020), doi:10.1126/science.aba3066.

62. K. Mittelstrass, J. S. Ried, Z. Yu, J. Krumsiek, C. Gieger, Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genet*. **7**, 1002215 (2011).

63. C. Finan, A. Gaulton, F. A. Kruger, R. T. Lumbers, T. Shah, J. Engmann, L. Galver, R. Kelley, A. Karlsson, R. Santos, J. P. Overington, A. D. Hingorani, J. P. Casas, The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9** (2017), doi:10.1126/scitranslmed.aag1166.

64. D. Ochoa, A. Hercules, M. Carmona, D. Suveges, A. Gonzalez-Uriarte, C. Malangone, A. Miranda, L. Fumis, D. Carvalho-Silva, M. Spitzer, J. Baker, J. Ferrer, A. Raies, O. Razuvayevskaya, A. Faulconbridge, E. Petsalaki, P. Mutowo, S. Machlitt-Northen, G. Peat, E. McAuley, C. K. Ong, E. Mountjoy, M. Ghoussaini, A. Pierleoni, E. Papa, M. Pignatelli, G. Koscielny, M. Karim, J. Schwartzentruber, D. G. Hulcoop, I. Dunham, E. M. McDonagh, Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).

65. I. Kola, J. Bell, A call to reform the taxonomy of human disease. *Nat. Rev. Drug Discov.* **10**, 641–2 (2011).

66. R. Sender, R. Milo, The distribution of cellular turnover in the human body. *Nat. Med.* **27**, 45–48 (2021).

67. S. A. Williams, M. Kivimaki, C. Langenberg, A. D. Hingorani, J. P. Casas, C. Bouchard, C. Jonasson, M. A. Sarzynski, M. J. Shipley, L. Alexander, J. Ash, T. Bauer, J. Chadwick, G. Datta, R. K. DeLisle, Y. Hagar, M. Hinterberg, R. Ostroff, S. Weiss, P. Ganz, N. J. Wareham, Plasma protein patterns as comprehensive indicators of health. *Nat. Med.* **25**, 1851–1857 (2019).

68. L. Gold, D. Ayers, J. Bertino, C. Bock, A. Bock, E. N. Brody, J. Carter, A. B. Dalby, B. E. Eaton, T. Fitzwater, D. Flather, A. Forbes, T. Foreman, C. Fowler, B. Gawande, M. Goss, M. Gunn, S. Gupta, D. Halladay, J. Heil, J. Heilig, B. Hicke, G. Husar, N. Janjic, T. Jarvis, S. Jennings, E. Katilius, T. R. Keeney, N. Kim, T. H. Koch, S. Kraemer, L. Kroiss, N. Le, D. Levine, W. Lindsey, B. Lollo, W. Mayfield, M. Mehan, R. Mehler, S. K. Nelson, M. Nelson, D. Nieuwlandt, M. Nikrad, U. Ochsner, R. M. Ostroff, M. Otis, T. Parker, S. Pietrasiewicz, D. I. Resnicow, J. Rohloff, G. Sanders, S. Sattin, D. Schneider, B. Singer, M. Stanton, A. Sterkel, A. Stewart, S. Stratford, J. D. Vaught, M. Vrkljan, J. J. Walker, M. Watrobka, S. Waugh, A. Weiss, S. K. Wilcox, A. Wolfson, S.

K. Wolk, C. Zhang, D. Zichi, Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One*. **5** (2010), doi:10.1371/journal.pone.0015004.

69. C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, V. Plagnol, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

70. B. Elsworth, M. Lyon, T. Alexander, Y. Liu, P. Matthews, J. Hallett, P. Bates, T. Palmer, V. Haberland, G. D. Smith, J. Zheng, P. Haycock, T. R. Gaunt, G. Hemani, *bioRxiv*, in press, doi:10.1101/2020.08.10.244293.

71. S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, Y. Luo, C. Sidore, A. Kwong, N. Timpson, S. Koskinen, S. Vrieze, L. J. Scott, H. Zhang, A. Mahajan, J. Veldink, U. Peters, C. Pato, C. M. Van Duijn, C. E. Gillies, I. Gandin, M. Mezzavilla, A. Gilly, M. Cocca, M. Traglia, A. Angius, J. C. Barrett, D. Boomsma, K. Branham, G. Breen, C. M. Brummett, F. Busonero, H. Campbell, A. Chan, S. Chen, E. Chew, F. S. Collins, L. J. Corbin, G. D. Smith, G. Dedoussis, M. Dorr, A. E. Farmaki, L. Ferrucci, L. Forer, R. M. Fraser, S. Gabriel, S. Levy, L. Groop, T. Harrison, A. Hattersley, O. L. Holmen, K. Hveem, M. Kretzler, J. C. Lee, M. McGue, T. Meitinger, D. Melzer, J. L. Min, K. L. Mohlke, J. B. Vincent, M. Nauck, D. Nickerson, A. Palotie, M. Pato, N. Pirastu, M. McInnis, J. B. Richards, C. Sala, V. Salomaa, D. Schlessinger, S. Schoenherr, P. E. Slagboom, K. Small, T. Spector, D. Stambolian, M. Tuke, J. Tuomilehto, L. H. Van Den Berg, W. Van Rheenen, U. Volker, C. Wijmenga, D. Toniolo, E. Zeggini, P. Gasparini, M. G. Sampson, J. F. Wilson, T. Frayling, P. I. W. De Bakker, M. A. Swertz, S. McCarroll, C. Kooperberg, A. Dekker, D. Altshuler, C. Willer, W. Iacono, S. Ripatti, N. Soranzo, K. Walter, A. Swaroop, F. Cucca, C. A. Anderson, R. M. Myers, M. Boehnke, M. I. McCarthy, R. Durbin, G. Abecasis, J. Marchini, A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

72. J. Huang, B. Howie, S. McCarthy, Y. Memari, K. Walter, J. L. Min, P. Danecek, G. Malerba, E. Trabetti, H. F. Zheng, G. Gambaro, J. B. Richards, R. Durbin, N. J. Timpson, J. Marchini, N. Soranzo, S. Al Turki, A. Amuzu, C. A. Anderson, R. Anney, D. Antony, M. S. Artigas, M. Ayub, S. Bala, J. C. Barrett, I. Barroso, P. Beales, M. Benn, J. Bentham, S. Bhattacharya, E. Birney, D. Blackwood, M. Bobrow, E. Bochukova, P. F. Bolton, R. Bounds, C. Boustred, G. Breen, M. Calissano, K. Carss, J. P. Casas, J. C. Chambers, R. Charlton, K. Chatterjee, L. Chen, A. Ciampi, S. Cirak, P. Clapham, G. Clement, G. Coates, M. Cocca, D. A. Collier, C. Cosgrove, T. Cox, N. Craddock, L. Crooks, S. Curran, D. Curtis, A. Daly, I. N. M. Day, A. Day-Williams, G. Dedoussis, T. Down, Y. Du, C. M. Van Duijn, I. Dunham, S. Edkins, R. Ekong, P. Ellis, D. M. Evans, I. S. Farooqi, D. R. Fitzpatrick, P. Flicek, J. Floyd, A. R. Foley, C. S. Franklin, M. Futema, L. Gallagher, P. Gasparini, T. R. Gaunt, M. Geihs, D. Geschwind, C. Greenwood, H. Griffin, D. Grozeva, X. Guo, X. Guo, H. Gurling, D. Hart, A. E. Hendricks, P. Holmans, L. Huang, T. Hubbard, S. E. Humphries, M. E. Hurles, P. Hysi, V. Iotchkova, A. Isaacs, D. K. Jackson, Y. Jamshidi, J. Johnson, C. Joyce, K. J. Karczewski, J. Kaye, T. Keane, J. P. Kemp, K. Kennedy, A. Kent, J. Keogh, F. Khawaja, M. E. Kleber, M. Van Kogelenberg, A. Kolb-Kokocinski, J. S. Kooner, G. Lachance, C. Langenberg, C. Langford, D. Lawson, I. Lee, E. M. Van Leeuwen, M. Lek, R. Li, Y. Li, J. Liang, H. Lin, R. Liu, J. Lönnqvist, L. R. Lopes, M. Lopes, J. Luan, D. G. MacArthur, M. Mangino, G. Marenne, W. März, J. Maslen, A. Matchan, I. Mathieson, P. McGuffin, A. M. McIntosh, A. G. McKechanie, A. McQuillin, S. Metrustry, N. Migone, H. M. Mitchison, A. Moayyeri, J. Morris, R. Morris, D. Muddyman, F. Muntoni, B. G. Nordestgaard, K. Northstone, M. C. O'Donovan, S. O'Rahilly, A. Onoufriadis, K. Oualkacha, M. J. Owen, A. Palotie, K. Panoutsopoulou, V. Parker, J. R. Parr, L. Paternoster, T. Paunio, F. Payne, S. J. Payne, J. R. B. Perry, O. Pietilainen, V. Plagnol, R. C. Pollitt, S. Povey, M. A. Quail, L. Quaye, L. Raymond, K. Rehnström, C. K. Ridout, S. Ring, G. R. S. Ritchie, N. Roberts, R. L. Robinson, D. B. Savage, P. Scambler, S. Schiffels, M. Schmidts, N. Schoenmakers, R. H. Scott, R. A. Scott, R. K. Semple, E. Serra, S. I. Sharp, A. Shaw, H. A. Shihab, S. Y. Shin, D. Skuse, K. S. Small, C. Smee, G. D. Smith, L. Southam, O. Spasic-

Boskovic, T. D. Spector, D. St. Clair, B. St. Pourcain, J. Stalker, E. Stevens, J. Sun, G. Surdulescu, J. Suvisaari, P. Syrris, I. Tachmazidou, R. Taylor, J. Tian, M. D. Tobin, D. Toniolo, M. Traglia, A. Tybjaerg-Hansen, A. M. Valdes, A. M. Vandersteen, A. Varbo, P. Vijayarangakannan, P. M. Visscher, L. V. Wain, J. T. R. Walters, G. Wang, J. Wang, Y. Wang, K. Ward, E. Wheeler, P. Whincup, T. Whyte, H. J. Williams, K. A. Williamson, C. Wilson, S. G. Wilson, K. Wong, C. J. Xu, J. Yang, G. Zaza, E. Zeggini, F. Zhang, P. Zhang, W. Zhang, Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 1–9 (2015).

73. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 (2018).

74. C. J. Willer, Y. Li, G. R. Abecasis, METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. **26**, 2190–2191 (2010).

75. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

76. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).

77. T. Berisa, J. K. Pickrell, Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*. **32**, 283–5 (2016).

78. G. Hemani, J. Zheng, B. Elsworth, K. H. Wade, V. Haberland, D. Baird, C. Laurin, S. Burgess, J. Bowden, R. Langdon, V. Y. Tan, J. Yarmolinsky, H. A. Shihab, N. J. Timpson, D. M. Evans, C. Relton, R. M. Martin, G. Davey Smith, T. R. Gaunt, P. C. Haycock, The MR-Base platform supports systematic causal inference across the human phenome. *Elife*. **7** (2018), doi:10.7554/eLife.34408.

79. L. A. Lotta, M. Pietzner, I. D. Stewart, L. B. L. Wittemans, C. Li, R. Bonelli, J. Raffler, E. K. Biggs, C. Oliver-Williams, V. P. W. Auyeung, J. Luan, E. Wheeler, E. Paige, P. Surendran, G. A. Michelotti, R. A. Scott, S. Burgess, V. Zuber, E. Sanderson, A. Koulman, F. Imamura, N. G. Forouhi, K.-T. Khaw, MacTel Consortium, J. L. Griffin, A. M. Wood, G. Kastenmüller, J. Danesh, A. S. Butterworth, F. M. Gribble, F. Reimann, M. Bahlo, E. Fauman, N. J. Wareham, C. Langenberg, A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat. Genet.* **53**, 54–64 (2021).

80. G. Yu, L. G. Wang, Y. Han, Q. Y. He, ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).

ACKNOWLEDGEMENTS

We are grateful to all the 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). ERG is supported by the National Institutes of Health (NIH) Awards R35HG010718, R01HG011138, R01GM140287, and NIH/NIA AG068026. MAW, MA, and GK are supported by grants from the National Institute on Aging (NIA): U01 AG061359, RF1 AG057452, and RF1 AG059093). JCZ is supported by a 4-year Wellcome Trust PhD Studentship and the Cambridge Trust; MK is supported by a Gates Fellowship; CL, EW, MP, JL, EO, IS, NK, and NJW are funded by the Medical Research Council (MC_UU_00006/1 - Aetiology and Mechanisms). 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).

AUTHOR CONTRIBUTIONS

Conceptualization: CL, MP, EW
Data curation/Software: EO, NDK, JL, MAW, JR
Formal Analysis: MP, EW, JCZ, AC, MK, IDS, JC
Methodology: RAS, ERG
Visualization: MP, JCZ, MK
Funding acquisition: CL, NJW
Project administration: CL, NJW
Supervision: CL, RAS, GK
Writing – original draft: MP, CL, EW, JCZ, MK
Writing – review & editing: EO, JC, IDS, ADH, NDK, MA, WA, SOR, NJW

COMPETING INTERESTS

RAS and AC are current employees and/or stockholders of GlaxoSmithKline. ERG receives an honorarium from the journal Circulation Research of the American Heart Association as a member of the Editorial Board. 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 (https://www.mrc-epid.cam.ac.uk/research/studies/fenland/information-for-researchers/). Data will either be shared through an institutional data sharing agreement or arrangements will be made for analyses to be conducted remotely without the necessity for data transfer. Summary statistics can be obtained from www.omicscience.org/apps/pgwas. Publicly available summary statistics for look-up and colocalisation of pQTLs were obtained from https://gwas.mrcieu.ac.uk/ and https://www.ebi.ac.uk/gwas/. Associated code and scripts for the analysis is available on GitHub (https://github.com/MRC-Epid/pGWAS_discovery) and has been permanently archived using Zenodo (*12*).

FIGURE LEGENDS

Fig. 1 Regional sentinel genetic variants associated ($p<1.004x10^{-11}$) with at least one protein target in up to 10,708 participants in 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 in close proximity to the protein-encoding gene (±500 kb) are highlighted in pink (cis-pQTLs) and all others are shown in blue. Darker colors indicate lower p-values. The upper panel shows the number of associated protein targets for each genomic region, with dot sizes on top giving the number of approximately independent genetic variants ($r^2<0.1$), such that larger dots refer to more genetic variants in the region.

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 colors 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 as reported in the Materials and Methods. D) Absolute (upper panel) and relative (lower panel) explained variance in plasma abundances of protein targets by identified pQTLs. Coloring indicates contribution of the lead cis-pQTL (dark purple), secondary cis-pQTLs (purple), protein- or pathway-specific trans-pQTLs (pink), and unspecific trans-pQTLs (yellow). The inset displays the overall distribution of explained variance by each of the four categories. The variance explained was computed using linear regression models. A graphical display of effect size distributions can be found in Fig. S3.

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 with black lines underneath. B) Same as A) but considering cis-sQTLs.

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 trait 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 by linkage disequilibrium of 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-pQLTs 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 evidence are highlighted by a border around the box. Colors represent broad trait categories.

Fig. 5 Network representation of phenome-wide colocalization analysis for protein-encoding loci. The entire network is composed of 412 protein targets (squares) and 506 phenotypes (circles) as nodes, which are connected (n=1,859 edges) if there is evidence of a shared genetic signal (posterior probability >80%) and is shown in Fig. S6. This figure is restricted to connections between proteins and binary endpoints, mainly diseases, to increase visibility and show shared etiology among the clinical most relevant outcomes. Only protein targets and phenotypes with at least one connection are included. Effect directions are indicated by the line type aligned with the allele associated with higher amounts of the protein target (solid – positive, dashed – inverse association with the phenotype). Colors indicate categories of phenotypes. An interactive version of the figure can be found at www.omicscience.org/apps/pgwas.

Fig. 6 Selected phenotypic examples from the proteogenomic map. A Plot visualizing convergence of genetic variants at the *SULT2A1* locus in relation to the LD with the candidate gene variant identified by multi-trait colocalization. Z-scores from GWAS for each annotated trait have been scaled by the absolute maximum, and dot size is proportional to the LD. Colors indicate the direction of effect aligned to the protein-increasing allele (red – positive, blue - inverse) The scheme on the right depicts the suggest mode of action by which higher SULT2A1 activity translates to higher risk of gallstones. **B** Same as A but for phenotypes colocalizing at the *EFEMP1* locus. The scheme on the right depicts a proposed mechanisms by which altered secretion of FBLN3 leads to the observed phenotypes. Stacked regional association plots for A and B can be found in Figs. S9 and S10.

Figure 1

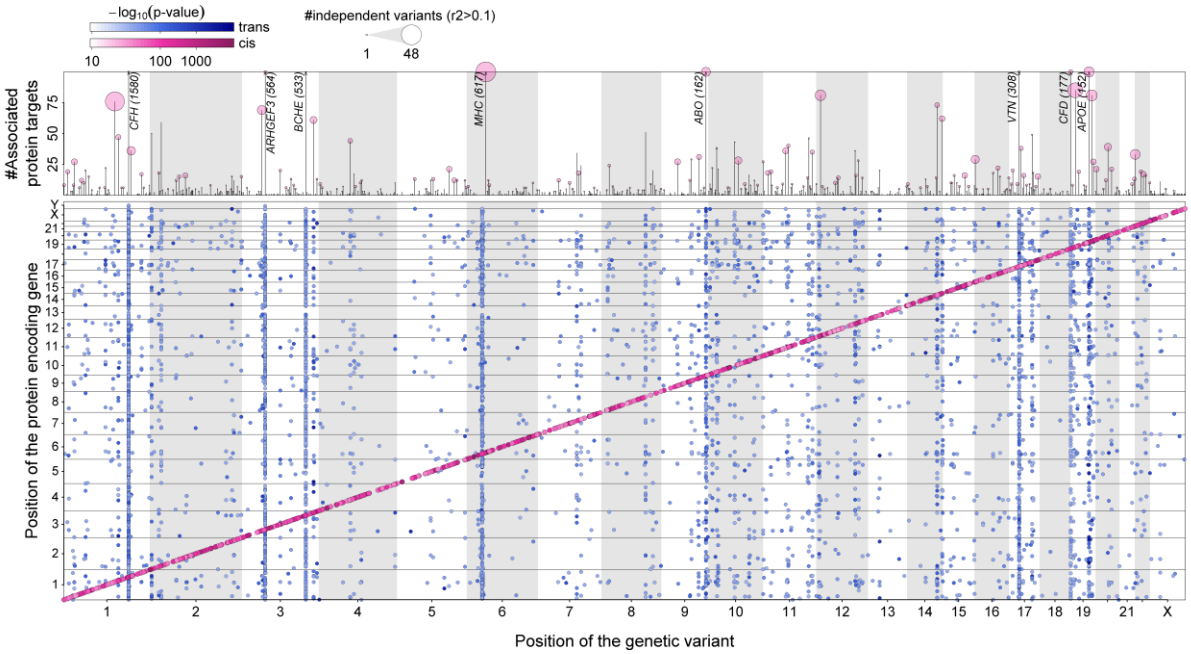

−log10(p-value)
trans
cis
10
100
1000
#independent variants (r2>0.1)
1
48
#Associated protein targets
25
50
75
CFH (1580)
ARHGEF3 (664)
BCHE (533)
MHC (617)
ABO (162)
VTN (308)
CFD (177)
APOE (152)
Position of the protein encoding gene
X
21
19
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
21
X
Position of the genetic variant

Figure 2

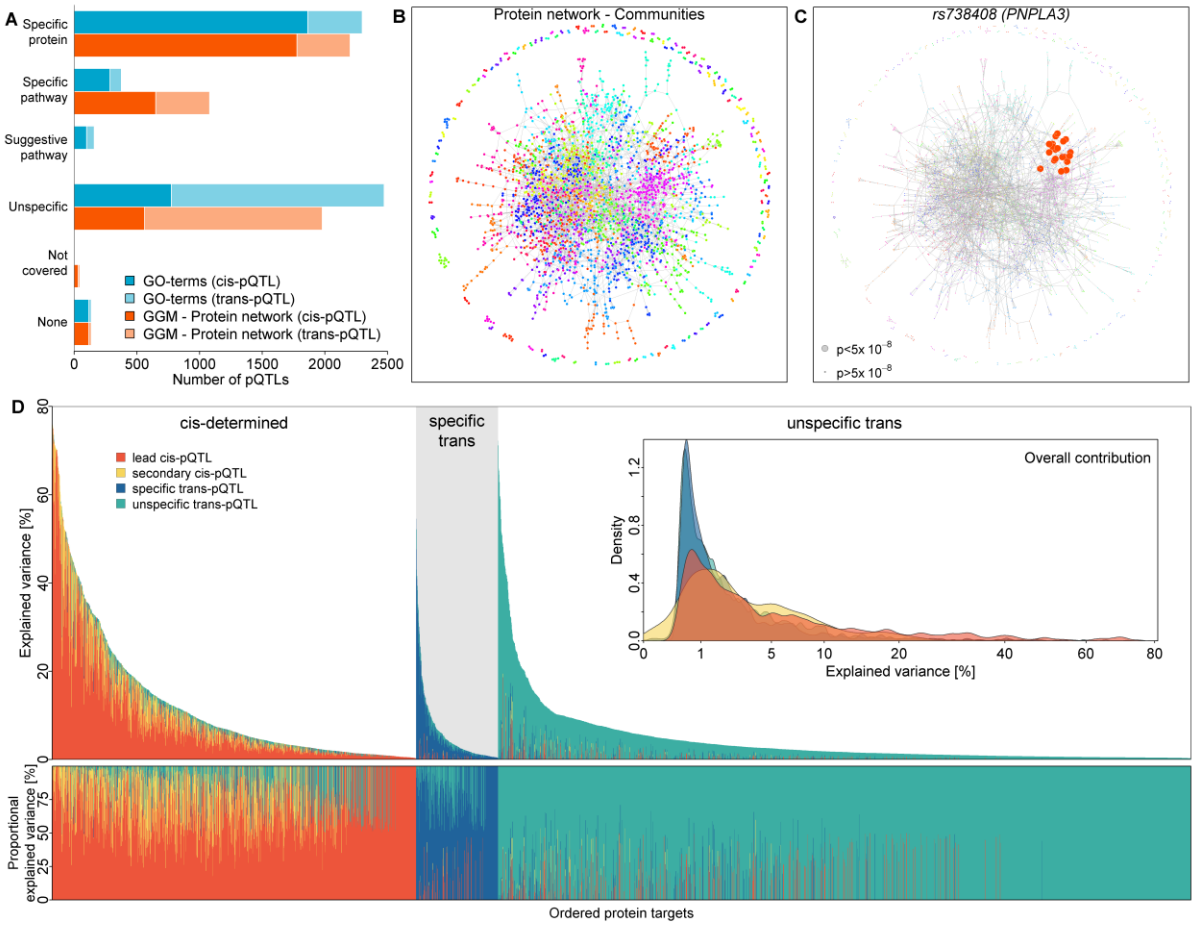

A
Specific protein
Specific pathway
Suggestive pathway
Unspecific
Not covered
None
GO-terms (cis-pQTL)
GO-terms (trans-pQTL)
GGM - Protein network (cis-pQTL)
GGM - Protein network (trans-pQTL)
0
500
1000
1500
2000
2500
Number of pQTLs
B
Protein network - Communities
C
rs738408 (PNPLA3)
p<5x 10-8
p>5x 10-8
D
cis-determined
specific trans
unspecific trans
lead cis-pQTL
secondary cis-pQTL
specific trans-pQTL
unspecific trans-pQTL
Explained variance [%]
Overall contribution
Density
Explained variance [%]
Proportional explained variance [%]
Ordered protein targets

Figure 3

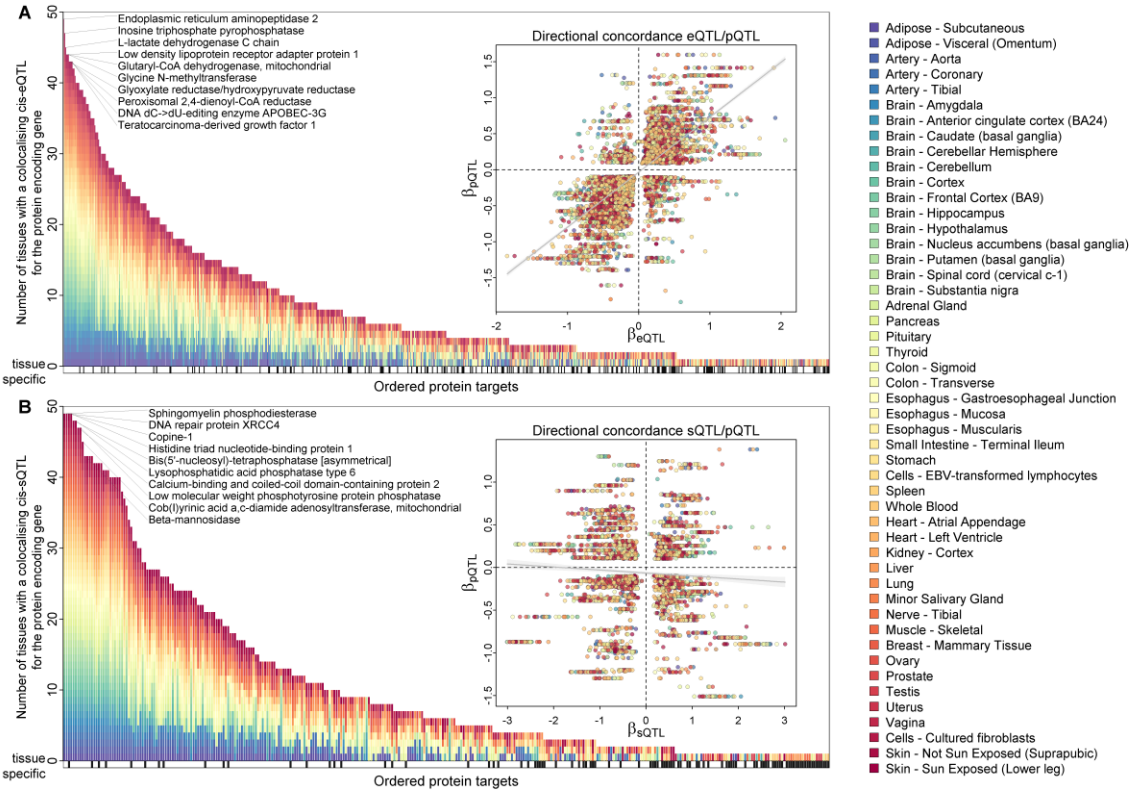

A
Endoplasmic reticulum aminopeptidase 2
Inosine triphosphate pyrophosphatase
L-lactate dehydrogenase C chain
Low density lipoprotein receptor adapter protein 1
Glutaryl-CoA dehydrogenase, mitochondrial
Glycine N-methyltransferase
Glyoxylate reductase/hydroxypyruvate reductase
Peroxisomal 2,4-dienoyl-CoA reductase
DNA dC->dU-editing enzyme APOBEC-3G
Teratocarcinoma-derived growth factor 1
Directional concordance eQTL/pQTL
Number of tissues with a colocalising cis-eQTL for the protein encoding gene
tissue specific
Ordered protein targets
B
Sphingomyelin phosphodiesterase
DNA repair protein XRCC4
Copine-1
Histidine triad nucleotide-binding protein 1
Bis(5'-nucleosyl)-tetraphosphatase [asymmetrical]
Lysophosphatidic acid phosphatase type 6
Calcium-binding and coiled-coil domain-containing protein 2
Low molecular weight phosphotyrosine protein phosphatase
Cob(I)yrinic acid a,c-diamide adenosyltransferase, mitochondrial
Beta-mannosidase
Directional concordance sQTL/pQTL
Number of tissues with a colocalising cis-sQTL for the protein encoding gene
tissue specific
Ordered protein targets
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Artery - Aorta
Artery - Coronary
Artery - Tibial
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Adrenal Gland
Pancreas
Pituitary
Thyroid
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Small Intestine - Terminal Ileum
Stomach
Cells - EBV-transformed lymphocytes
Spleen
Whole Blood
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Nerve - Tibial
Muscle - Skeletal
Breast - Mammary Tissue
Ovary
Prostate
Testis
Uterus
Vagina
Cells - Cultured fibroblasts
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)

Figure 4

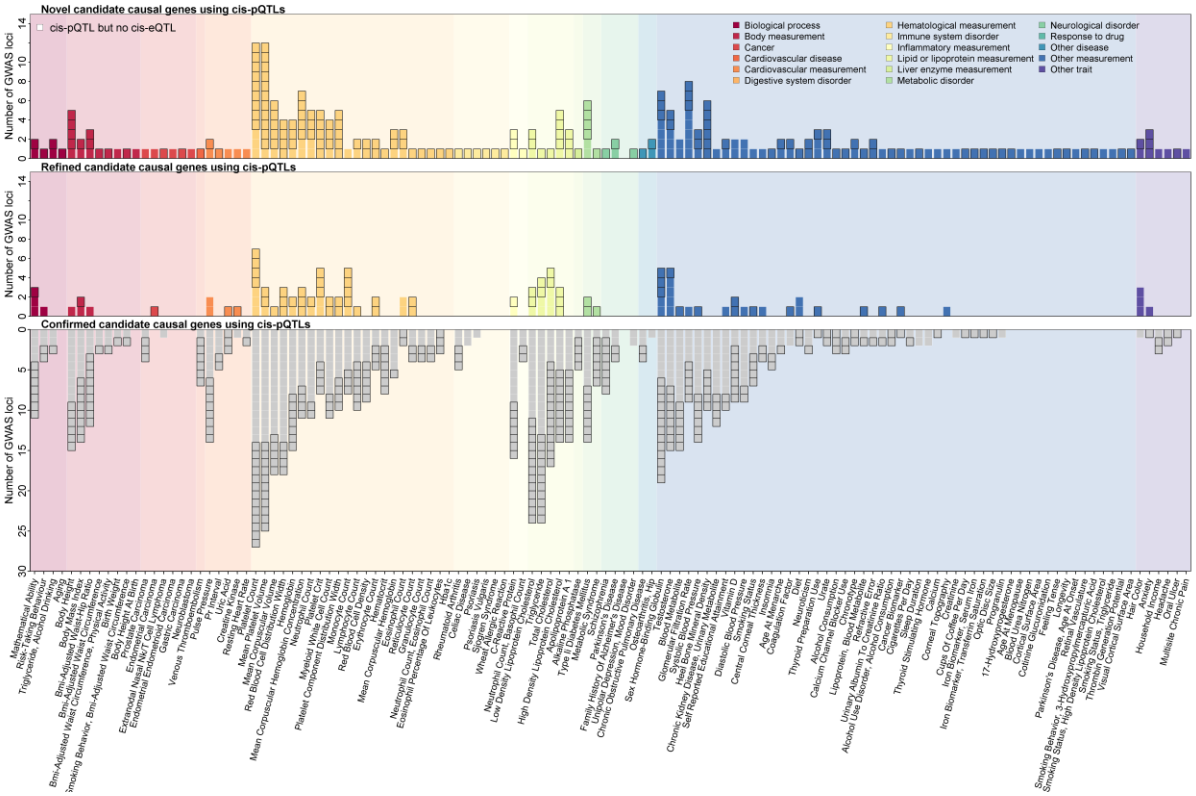

Novel candidate causal genes using cis-pQTLs
cis-pQTL but no cis-eQTL
Biological process
Body measurement
Cancer
Cardiovascular disease
Cardiovascular measurement
Digestive system disorder
Hematological measurement
Immune system disorder
Inflammatory measurement
Lipid or lipoprotein measurement
Liver enzyme measurement
Metabolic disorder
Neurological disorder
Response to drug
Other disease
Other measurement
Other trait
Refined candidate causal genes using cis-pQTLs
Confirmed candidate causal genes using cis-pQTLs
Number of GWAS loci

Figure 5

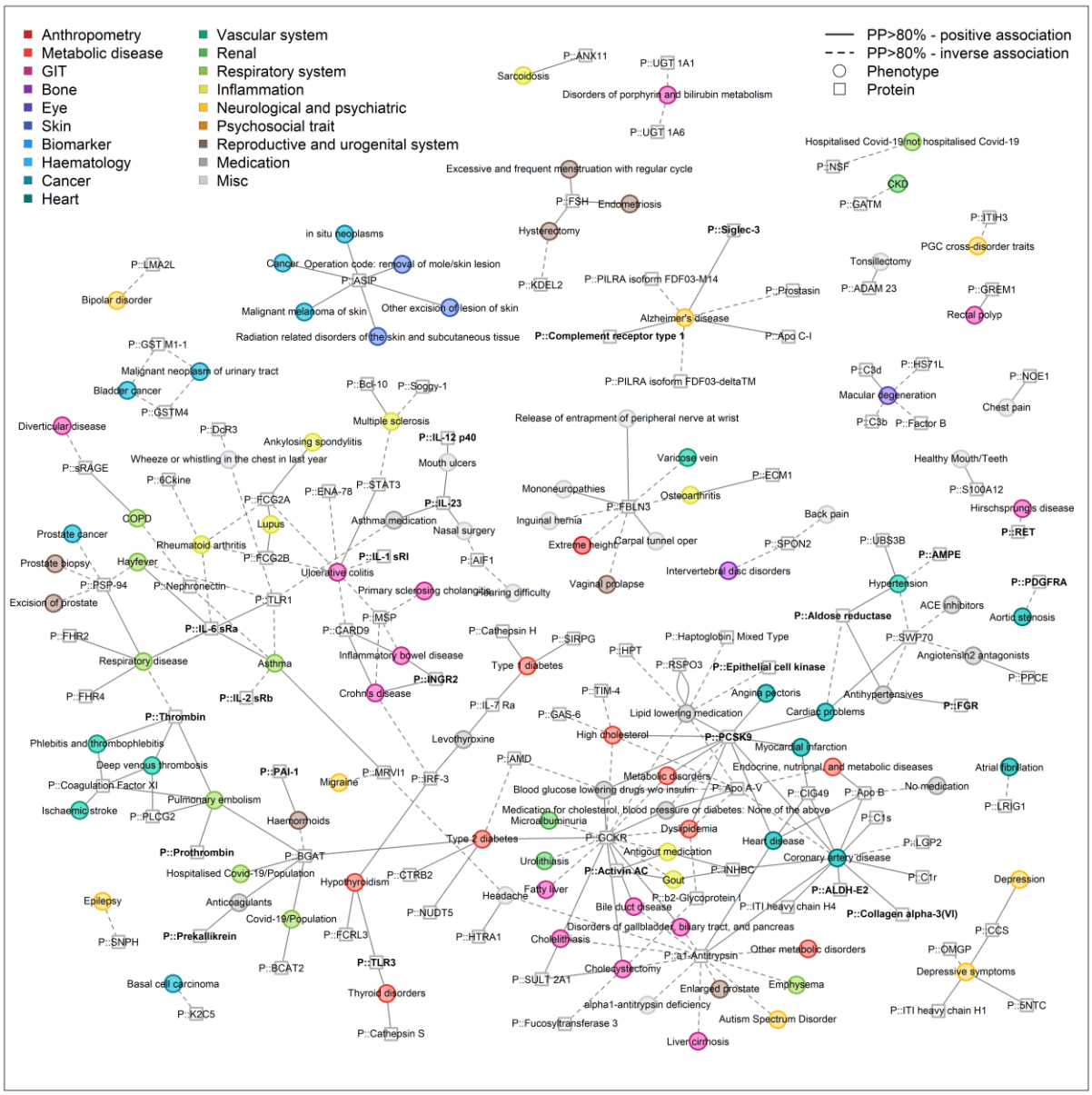
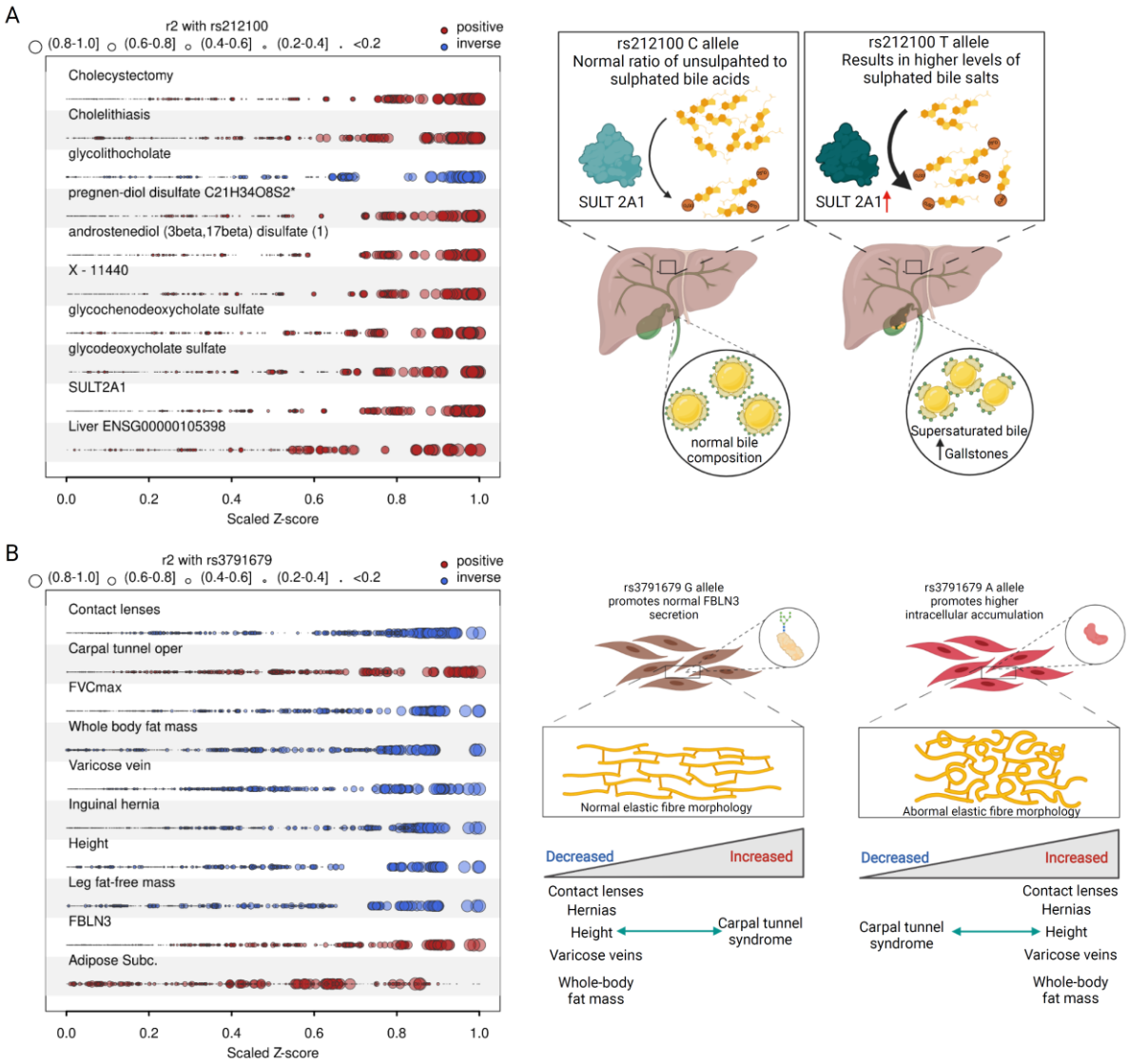

Anthropometry
Metabolic disease
GIT
Bone
Eye
Skin
Biomarker
Haematology
Cancer
Heart
Vascular system
Renal
Respiratory system
Inflammation
Neurological and psychiatric
Psychosocial trait
Reproductive and urogenital system
Medication
Misc
PP>80% - positive association
PP>80% - inverse association
Phenotype
Protein
Sarcoidosis
P::ANX11
P::UGT 1A1
Disorders of porphyrin and bilirubin metabolism
P::UGT 1A6
Excessive and frequent menstruation with regular cycle
P::FSH
Endometriosis
Hysterectomy
P::KDEL2
Hospitalised Covid-19/not hospitalised Covid-19
P::NSF
CKD
P::GATM
P::ITIH3
PGC cross-disorder traits
Tonsillectomy
P::ADAM 23
P::GREM1
Rectal polyp
in situ neoplasms
Cancer Operation code: removal of mole/skin lesion
P::ASIP
Malignant melanoma of skin
Other excision of lesion of skin
Radiation related disorders of the skin and subcutaneous tissue
P::LMA2L
Bipolar disorder
P::PILRA isoform FDF03-M14
P::Siglec-3
P::Prostasin
Alzheimer's disease
P::Complement receptor type 1
P::Apo C-I
P::PILRA isoform FDF03-deltaTM
P::GST M1-1
Malignant neoplasm of urinary tract
Bladder cancer
P::GSTM4
P::C3d
P::HS71L
Macular degeneration
P::C3b
P::Factor B
P::NOE1
Chest pain
P::Bcl-10
P::Soggy-1
Multiple sclerosis
Diverticular disease
P::DcR3
Ankylosing spondylitis
P::IL-12 p40
Release of entrapment of peripheral nerve at wrist
P::sRAGE
Wheeze or whistling in the chest in last year
P::6Ckine
Mouth ulcers
P::STAT3
P::ENA-78
P::FCG2A
P::IL-23
Asthma medication
Nasal surgery
Lupus
COPD
Prostate cancer
Rheumatoid arthritis
Hayfever
Prostate biopsy
P::FCG2B
P::IL-1 sRI
P::AIF1
Ulcerative colitis
Primary sclerosing cholangitis
Hearing difficulty
P::PSP-94
P::Nephronectin
Excision of prostate
P::TLR1
Varicose vein
Mononeuropathies
Osteoarthritis
P::ECM1
P::FBLN3
Inguinal hernia
Extreme height
Carpal tunnel oper
Vaginal prolapse
Back pain
P::SPON2
Intervertebral disc disorders
Healthy Mouth/Teeth
P::S100A12
Hirschsprungs disease
P::RET
P::UBS3B
P::AMPE
Hypertension
P::PDGFRA
ACE inhibitors
Aortic stenosis
P::Aldose reductase
P::FHR2
P::IL-6 sRa
P::IMSP
P::CARD9
P::Cathepsin H
P::SIRPG
P::HPT
P::Haptoglobin, Mixed Type
P::RSPO3
P::Epithelial cell kinase
P::SWP70
Angiotensin2 antagonists
P::PPCE
Respiratory disease
Asthma
Inflammatory bowel disease
P::INGR2
Crohn's disease
Type 1 diabetes
P::FHR4
P::IL-2 sRb
P::IL-7 Ra
P::TIM-4
P::GAS-6
Angina pectoris
Antihypertensives
P::FGR
Cardiac problems
Lipid lowering medication
P::Thrombin
Phlebitis and thrombophlebitis
Deep venous thrombosis
P::Coagulation Factor XI
Pulmonary embolism
Ischaemic stroke
P::PLCG2
Levothyroxine
High cholesterol
P::PCSK9
Myocardial infarction
P::AMD
P::PAI-1
Migraine
P::MRVI1
P::IRF-3
Metabolic disorders
Endocrine, nutritional and metabolic diseases
Blood glucose lowering drugs w/o insulin
P::Apo A-V
P::CIG49
P::App B
No medication
Atrial fibrillation
P::LRIG1
Medication for cholesterol, blood pressure or diabetes: None of the above
Microalbuminuria
Dyslipidemia
P::C1s
Haemorrhoids
P::Prothrombin
P::BGAT
Type 2 diabetes
P::GCKR
Heart disease
P::LGP2
Coronary artery disease
Urolithiasis
Antigout medication
P::Activin AC
P::NHBC
Gout
P::C1r
Hospitalised Covid-19/Population
Hypothyroidism
P::CTRB2
Fatty liver
Headache
P::b2-Glycoprotein I
P::ITI heavy chain H4
P::ALDH-E2
P::Collagen alpha-3(VI)
Depression
Epilepsy
Anticoagulants
Covid-19/Population
P::NUDT5
Bile duct disease
Disorders of gallbladder, biliary tract, and pancreas
P::SNPH
P::Prekallikrein
P::FCRL3
P::HTRA1
Cholelithiasis
P::a1-Antitrypsin
Other metabolic disorders
P::CCS
P::OMGP
Depressive symptoms
P::BCAT2
P::TLR3
Cholecystectomy
P::SULT 2A1
Enlarged prostate
Emphysema
Basal cell carcinoma
P::K2C5
Thyroid disorders
alpha1-antitrypsin deficiency
Autism Spectrum Disorder
P::ITI heavy chain H1
P::5NTC
P::Fucosyltransferase 3
P::Cathepsin S
Liver cirrhosis

Figure 6


A
r2 with rs212100
(0.8-1.0] (0.6-0.8] (0.4-0.6] (0.2-0.4] <0.2
positive
inverse
Cholecystectomy
Cholelithiasis
glycolithocholate
pregnen-diol disulfate C21H34O8S2*
androstenediol (3beta,17beta) disulfate (1)
X - 11440
glycochenodeoxycholate sulfate
glycodeoxycholate sulfate
SULT2A1
Liver ENSG00000105398
0.0 0.2 0.4 0.6 0.8 1.0
Scaled Z-score
rs212100 C allele
Normal ratio of unsulpahted to sulphated bile acids
SULT 2A1
rs212100 T allele
Results in higher levels of sulphated bile salts
SULT 2A1
normal bile composition
Supersaturated bile
Gallstones
B
r2 with rs3791679
(0.8-1.0] (0.6-0.8] (0.4-0.6] (0.2-0.4] <0.2
positive
inverse
Contact lenses
Carpal tunnel oper
FVCmax
Whole body fat mass
Varicose vein
Inguinal hernia
Height
Leg fat-free mass
FBLN3
Adipose Subc.
0.0 0.2 0.4 0.6 0.8 1.0
Scaled Z-score
rs3791679 G allele
promotes normal FBLN3 secretion
Normal elastic fibre morphology
Decreased
Increased
Contact lenses
Hernias
Height
Varicose veins
Whole-body fat mass
Carpal tunnel syndrome
rs3791679 A allele
promotes higher intracellular accumulation
Abormal elastic fibre morphology
Decreased
Increased
Contact lenses
Hernias
Carpal tunnel syndrome
Height
Varicose veins
Whole-body fat mass

Science
AAAS

Supplementary Materials for

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#]

Correspondence to: claudia.langenberg@mrc-epid.cam.ac.uk

This PDF file includes:

Materials and Methods

Study participants

The Fenland study is a population-based cohort of 12,435 participants of predominantly white-European ancestry born between 1950 and 1975 who underwent detailed phenotyping at the baseline visit from 2005-2015. Participants were recruited from general practice surgeries in the Cambridgeshire region in the UK. Exclusion criteria were: physician diagnosis of diabetes mellitus, inability to walk unaided, terminal illness, clinically diagnosed psychotic disorder, pregnancy, or lactation. The study was approved by the Cambridge Local Research Ethics Committee (NRES Committee – East of England Cambridge Central, ref. 04/Q0108/19), and all participants provided written informed consent. Use of human biological samples was in accord with the terms of the informed consents under an IRB/EC approved protocol.

Genotyping and imputation

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

Autosomes for the OMICS and GWAS subsets were imputed to the HRC (r1) panel using IMPUTE448, and the Core-Exome subset and the X-chromosome (for all subsets) were imputed to HRC.r1.1 using the Sanger imputation server (*71*). All three array subsets were also imputed to the UK10K+1000Gphase3 (*72*) panel using the Sanger imputation server to obtain additional variants that do not exist in the HRC reference panel. Variants with MAF < 0.001, imputation quality (info) < 0.4, or Hardy Weinberg Equilibrium $p < 10^{-7}$ in any of the genotyping subsets were excluded from further analyses.

Proteomic measurements

Proteomic profiling of fasted EDTA plasma samples from 12,084 Fenland Study participants collected at baseline was performed by SomaLogic Inc. using an aptamer-based technology (SOMAscan proteomic assay). Relative protein abundances of 4,775 human protein targets

were evaluated by 4,979 aptamers (SomaLogic V4), and a detailed description can be found elsewhere (*67*). Briefly, the SOMAscan assay uses a library of short single-stranded DNA molecules, which are chemically modified to specifically bind to protein targets, and the relative amount of aptamers binding to protein targets is determined using DNA microarrays. To account for variation in hybridization within runs, hybridization control probes are used to generate a hybridization scale factor for each sample. To control for total signal differences between samples due to variation in overall protein concentration or technical factors such as reagent concentration, pipetting, or assay timing, a ratio between each aptamer's measured value and a reference value was computed, and the median of these ratios was computed for each of the three dilution sets (20%, 0.5%, and 0.005%) and applied to each dilution set. Samples were removed if they were deemed by SomaLogic to have failed or did not meet our acceptance criteria of 0.25-4 for all scaling factors. In addition to passing SomaLogic QC, only human protein targets were taken forward for subsequent analysis (4,979 out of the 5,284 aptamers). Aptamers' target annotation and mapping to UniProt accession numbers as well as Entrez gene identifiers were provided by SomaLogic, and we used those to obtain genomic positions of protein-encoding genes.

GWAS and meta-analysis

After excluding ancestry outliers and related individuals, 10,708 Fenland participants had both phenotypes and genetic data for the GWAS (OMICS=8,350, Core-Exome=1,026, GWAS=1,332). Within each genotyping subset, aptamer abundances were transformed to follow a normal distribution using the rank-based inverse normal transformation. Transformed aptamer abundances were then adjusted for age, sex, sample collection site, and 10 genetic principal components and the residuals used as input for the genetic association analyses. Genome-wide association was performed under an additive model using BGENIE (v1.3) (*73*). Results for the three genotyping arrays were combined in a fixed-effects meta-analysis in METAL (*74*). Following the meta-analysis up to 10.2 million genetic variants also present in the largest subset of the Fenland data (Fenland-OMICS) with overall MAF≥1% were taken forward for further analysis.

Conditional analysis

To identify conditionally independent signals in a genomic region associated with an aptamer, we performed conditional analysis as implemented in GCTA (*75*) using the *--slct* option, with a collinear cut-off of 0.1 and a p-value threshold of 1.004×10^{-11}. As a quality control step, we fitted a final model including all identified variants for a given genomic region using individual-level data in the largest available data set ('Fenland-OMICs') and discarded all variants no longer meeting genome-wide significance.

We performed a forward stepwise selection procedure to identify secondary signals at each locus on the X-chromosome using SNPTEST v.2.5.2 to compute conditional association statistics based on individual-level data in the largest subset. Briefly, we defined conditionally independent signals as those emerging after conditioning on all previously selected signals in the locus until no signal was significant genome-wide.

Variant annotation

For each identified pQTL we first obtained all SNPs in at least moderate LD (r^2>0.1) using PLINK (version 2.0) and queried comprehensive annotations using the Variant Effect Predictor software (*76*) (version 98.3) with the --pick option. For each cis-pQTL we checked whether either the variant itself or a proxy in the encoding gene (r^2>0.6) was predicted to induce a change in the amino acid sequence of the associated protein, so-called protein-altering variants (PAVs). We further obtained domain information for each protein target via UniProt and tested whether the predicted amino acid exchange falls into any protein domain.

Locus definition

For each aptamer, we used a genome-wide significance threshold of 1.004×10^{-11} and defined non-overlapping regions by merging overlapping or adjoining 1 Mb intervals around all genome-wide significant variants (500 kb on either side), treating the extended MHC region (chr6:25.5–34.0 Mb) as one region. For each region we defined a regional sentinel variant as the most significant variant in the region. We defined genomic regions shared across aptamers based on a combination of multi-trait colocalization and LD-clumping (r^2>0.8). To this end, we first collapsed overlapping regions across all aptamers and divided those into approximately LD-independent regions (*77*). For each of the 776 regions we performed multi-trait colocalization as implemented in the R package *hyprcoloc* (*46*) to obtain clusters of aptamers sharing a common causal variant defined by a regional probability >80% and a posterior probability of at least 50%. However, some clusters included protein targets for which the lead cis-pQTL was not in LD with the variant prioritized by HyPrColoc, most likely due to violation of the one causal variant assumption. We then subdivided those clusters manually and clumped signals based on a strong LD (r^2>0.8). This procedure resulted in 2,548 genomic regions with at least one cis-pQTL. Genetic variants from conditional analysis were assigned to the same locus as the sentinel signal. We classified pQTLs as cis-acting instruments if the variant was less than 500 kb away from the gene body of the protein-encoding gene.

Replication using the Olink PEA technology

To test for replication of pQTLs identified using the SomaScan platform with a complementary technique, we used in-house GWAS results for 1,069 proteins measured using Proximity Extension Assays provided by Olink in a subset of 485 individuals of the Fenland study. A

detailed description of the GWAS can be found elsewhere (*10*) but largely followed the same protocol as described for SomaScan in the present paper. We first established which pQTLs were expected to replicate in the smaller sample size measured with both technologies by rerunning SNP – protein associations in the smaller subset. We identified 797 pQTLs that a) mapped to proteins covered on the Olink platform (N=543) and b) were significant at p-value<0.01. For each of the 797 SNP -protein pairs we repeated the same analysis but now instead of using SomaScan protein targets using the corresponding Olink protein as outcome. We considered associations to be replicated if they were directionally concordant and at least nominally significant at p-value<0.05 in this analysis.

Data-driven protein target network

We constructed a data-driven protein network using Gaussian graphical modeling similar to previous work (*2*). First, for each pair of highly correlated aptamers targeting the same protein (Pearson correlation>0.5), we dropped one at random to avoid artificial null results in the network, leaving 4,929 aptamers. We next computed residual plasma abundances by accounting for the effects of age, sex, test site, and the first three proteomic principal components using linear regression models. We finally used the R implementation ggm.estimate.pcor from the package *GeneNet* to estimate full-order partial correlations among residual aptamer abundances and only kept edges meeting a stringent Bonferroni cut-off ($p<2.05x10^{-9}$). The final network consisted of 2,936 aptamers and 4,669 edges. We performed community detection using the Girvan-Newman algorithm as implemented in the R package igraph and obtained 191 distinct protein communities in the network.

Classification of pQTLs

We classified the specificity of pQTLs based on two complementary approaches. We first derived the set of associated aptamers at a nominal GWAS-threshold of significance ($p<5x10^{-8}$) for each of the 5,442 uniquely identified variants. Next, we tested whether all aptamers 1) belonged to the same protein target (including complexes), 2) could be assigned to a common GO term (as previously described (*15*)), or 3) belonged to the same protein community in the data-driven protein network. We classified variants fulfilling 1) as protein-specific, variants fulfilling 2) and 3) as pathway-specific, variants fulfilling either 2) or 3) as suggestive pathway-specific, and any other variant as non-specific.

Variance explained

We estimated the variance explained by pQTLs for plasma levels of each aptamer with at least one associated pQTL using different sets of genetic instruments. To this end, we successively included 1) the lead cis-pQTL, 2) secondary cis-pQTLs, 3) specific trans-pQTLs, and 4) all trans-pQTLs in a linear regression model using residual aptamer abundances as outlined in

the GWAS section. We used the R^2 of the entire model as an estimate for variance explained. We did this analysis in the largest set of Fenland participants genotyped on a single array.

Candidate causal gene assignment

We integrated functional assignments with protein-protein interaction network data to assign putative causal genes for each variant - aptamer pair. Briefly, for each variant close to the protein-encoding gene (±500 kb) we assigned this gene as candidate causal gene. For pQTLs in *trans* we used a scoring system by integrating 1) search for functional variants (VEP score 1-12 and R^2>0.6), 2) LD with an eQTL (R^2>0.8), 3) distance of the gene products from the 20 closest genes to the protein target associated with the pQTL in the STRING protein-protein interaction, and 4) the closest gene. We assigned a score of two for 1) - 3) and a score of one for the closest gene and retained the gene(s) with the highest score(s) as (a) possible candidate(s). In a second step, we aligned gene assignments across all aptamers based on the definition of genomic loci, that is, for each locus shared across multiple aptamers we repeated the scoring system taking into account all possible candidate genes using a score of three for cis assignments. This procedure allowed us to refine assignment at otherwise poorly defined trans loci and to obtain higher confidence scores at each locus.

Incorporation of gene expression data

We incorporated gene expression and splicing QTL data by cross-referencing all cis-pQTLs with cis-eQTL/sQTLs identified in the GTEx version 8 release across 49 distinct tissues using an LD threshold of r^2>0.8 to identify likely similar signals (*24*). If at least on cis-pQTL mapped to a corresponding cis-eQTL or cis-sQTL for the protein-encoding gene, we used statistical colocalization (*69*) to test for a shared genetic signal between protein abundance measured in plasma and expression of the respective gene across all available tissues. We considered a PP>80% as evidence for a highly likely shared signal. We used a 500 kb window around the cis-pQTL for colocalization analysis. To identify cis-pQTL/cis-eQTL pairings which are likely to be tissue-specific, we obtained Z-scores for the candidate variant across all tissues and divided by the square root of the respective tissue sample size to normalize across tissues. We defined such pairings as tissue-specific if the normalized Z-score was more than ±5 times the median absolute deviation (MAD) away from the median normalized Z-score, which represents a robust measure of outlier detection. For transcripts expressed in fewer than five tissues, we considered the number of colocalizing tissues as a threshold for tissue specificity. To account for possible measurement artefacts, we repeated this process for 277 protein targets with evidence that a secondary signal was in LD with a cis-eQTL/sQTL using summary statistics conditioned on the lead pQTL in the region for colocalization. All GTEx variant-gene cis-eQTL and cis-sQTL associations from each tissue were downloaded in January 2020 from

https://console.cloud.google.com/storage/browser/gtex-resources. We further tested for colocalization with gene expression determined from whole blood using data from the eQTLGen consortium (*26*), which included data on more than 30,000 participants.

Annotation of GWAS catalog loci

We downloaded genome-wide significant summary statistics from the GWAS catalog (date 25/01/2021) and tested whether any of the identified pQTLs or proxies (r^2>0.8) have been reported to be associated with any non-proteomic trait, that is omitting any results that related to multiplex proteomic assays. Out of 227,631 entries, 113,618 entries passed this and additional filtering steps (missing effect estimates, missing risk allele, and not passing genome-wide significance). We next assessed whether for 3,139 lead and secondary pQTLs in cis, linkage by LD (r^2>0.8) to findings reported in the GWAS catalog may help to prioritize potential causal genes for the phenotypic trait. We compared the reported or mapped gene (closest gene assigned by the GWAS catalog) to the protein-encoding gene at the locus. This left us with 4,133 entries, including 590 cis-regions and 556 mapped GWAS traits. We collated loci into single entries to account for the variety of entries at pleiotropic loci and further dropped 32 cis-regions for which more than one protein target mapped to the signals in the region, resulting in 3,868 entries.

Phenome-wide scans at protein-encoding loci

We performed phenome-wide scans using statistical colocalization for 1,584 protein targets where we had evidence for at least one cis-pQTL. Briefly, we queried the Open GWAS database (*78*) as well as an in-house database of curated GWAS summary statistics hosted by GSK using a defined region (±500 kb) around the protein-encoding gene body and tested whether any of the traits in the databases showed a high posterior probability (PP) of shared genetic signal with plasma concentrations of the encoded protein target using statistical colocalization (*69*). We chose a cut-off of PP>80% to declare that a protein target and a phenotypic trait are highly likely to share a genetic signal at a locus. We used a conservative prior setting, prior probability of $1x10^{-6}$ that both traits have a common genetic signal, along with a check that the regional lead signals for the protein and the trait are in strong LD (r^2>0.8) to declare colocalization. We repeated this analysis using conditional statistics for the protein target accounting for a possible binding artefact introduced by the lead signal at the locus. We manually curated common trait names to reduce redundancy of phenotypes across both databases and kept the association with the largest PP for each mapped trait for a shared signal when all other definitions supported colocalization. Finally, we collapsed all pairs of protein targets and phenotypic traits with high evidence for colocalization into a protein-disease network by drawing an edge between a protein target and a phenotypic trait if there

was a high PP (>80%) for a shared signal. We used the lead signal at the locus aligned to the protein-increasing allele to indicate effect directions and visualized the network using the *igraph* R package. We report Mendelian Randomization estimates for binary outcomes derived from the UK Biobank as odds ratios, by transforming β-effect estimates according to the following formula: log(odds ratio) = β / (μ * (1 - μ)), where μ = case fraction, since analysis were performed using a linear regression frame work for computational efficacy.

COVID-19 summary statistics

We downloaded genome-wide summary statistics for four different outcome definitions of COVID-19 from the Human Genetics Initiative (https://www.covid19hg.org/). These included A2 (very severe respiratory confirmed COVID-19 vs. population), B1 (hospitalized COVID-19 vs. not hospitalized COVID-19), B2 (hospitalized COVID-19 vs. population), C2 (COVID-19 vs. population). To map the LD panel for colocalization analysis, we restricted those statistics to participants of European ancestry and excluded results contributed by 23&me.

Metabolite GWAS results for SULT2A1

We extracted summary statistics for 69 metabolites associated with the lead *cis*-pQTL ($p<1x10^{-6}$) for SULT2A1 from an in-house metabolome-wide GWAS based on the EPIC-Norfolk cohort, methods of which have been described previously(*79*).

Pathway enrichment analysis

We performed GO term enrichment as implemented in the R package *clusterProfiler* (*80*) separately for each mapped trait in the protein-phenotype network that had at least three colocalizing protein targets. We used all three GO term categories for this purpose and retained only pathways that met statistical significance after correction for multiple testing using the Benjamini-Hochberg procedure controlling the false-discovery rate at 5%.

Testing for effect modification by age and sex

We included an interaction term between the cis-pQTL and age (continuous) or sex in a linear regression model with the same adjustments as in the main analysis to test for potential differences of the 417 cis-pQTLs included in the proteogenomic map by age or sex. For interactions significant below the Bonferroni-corrected significance level of 0.05/(2*417), we estimated group-specific estimates, dichotomizing age at the median age of 49 years, by running linear regression models within each stratum using the largest set of individuals with the same genotype platform (N=8350).

Mapping of druggable targets

To annotate druggable targets we merged the list of proteins targeted by the SomaScan V4 platform with an updated list of druggable genes from Finan *et al.* (*63*) based on common gene entries. We deprioritized drugs with multiple reported side effects in ongoing or completed clinical trials, missing efficacy, discordant effect directions between the estimated effect of life-long higher protein concentrations mediated by genetic variants and action of the drug, a lack of clinical data, or having been withdrawn from major markets.

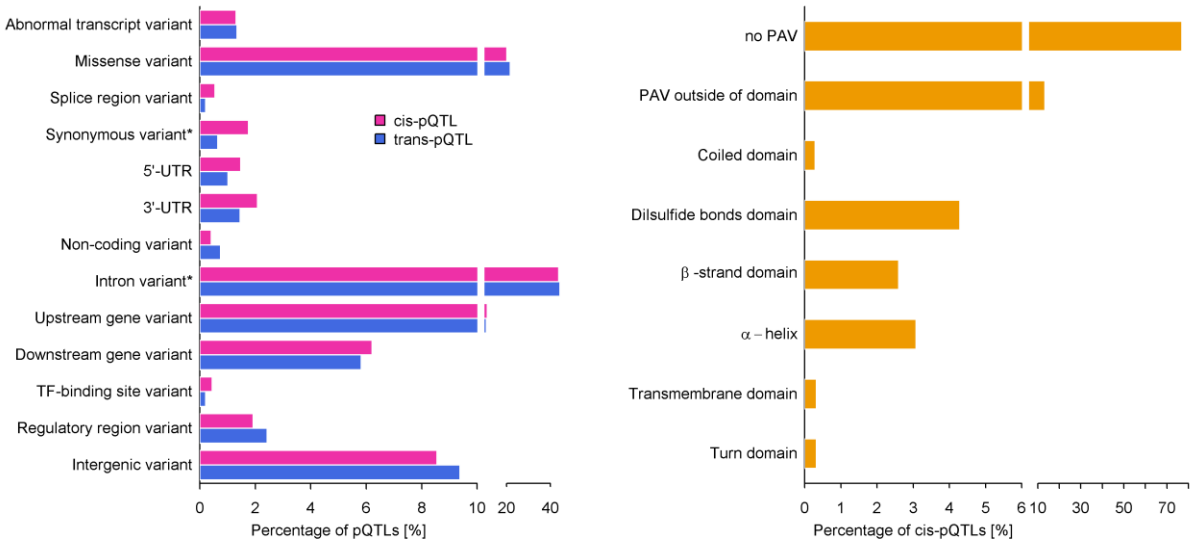


Fig. S1. Functional annotation of genetic variants associated with at least one protein target. Left. Bar chart of the distribution of annotations of 4,976 genetic variants associated with at least one protein target based on the Variant Effect Predictor tool. Colors indicate protein quantitative trait loci (pQTLs) either in close proximity to the protein-encoding gene (cis-pQTL, pink) or elsewhere in the genome (trans-pQTL, blue). Categories significantly enriched for cis-pQTLs are indicated with an asterisk (Fisher's-exact test, p-value < 0.003). The list of all 5,442 sentinel and secondary signals representing distinct variants was collapsed based on the identification of functional variants in strong LD (R^2>0.6). **Right.** Fraction of cis-pQTLs (N=3,139) harboring a protein-altering variant (PAV) and further split based on the location of the PAV in any common protein domain.

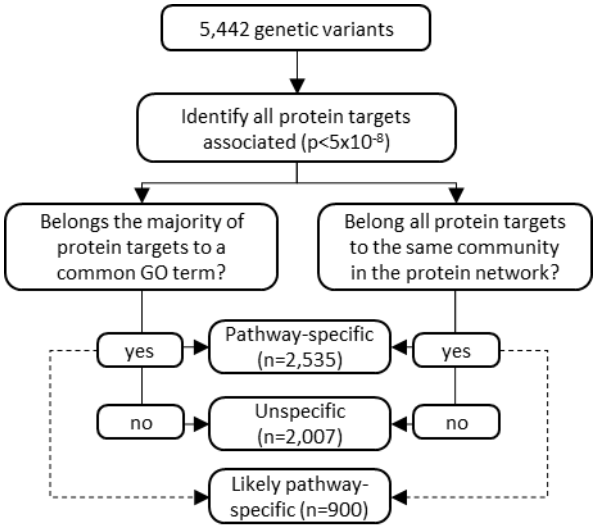


Fig. S2. Flowchart for pQTL classification. For each of the 5,442 distinct genetic variants we identified the set of significantly associated aptamers ($p<5x10^{-8}$) across the entire data. In the next step we tested whether all of the targeted proteins fall within a GO term and/or belong to the same community of protein targets in a data-driven protein network.

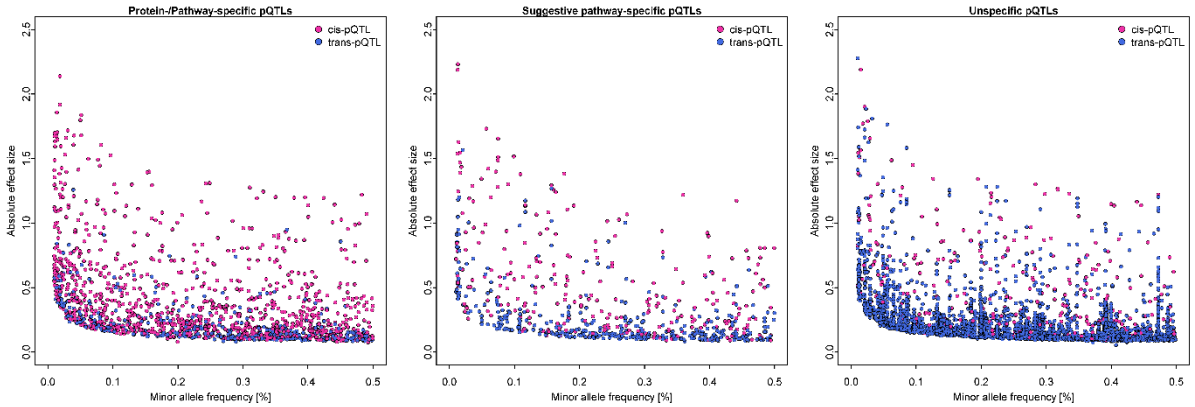


Fig. S3. Distribution of absolute effect sizes for identified protein quantitative trait loci (pQTLs). pQTLs are separated by specificity across the proteome and absolute effect sizes are plotted against the minor allele frequency.

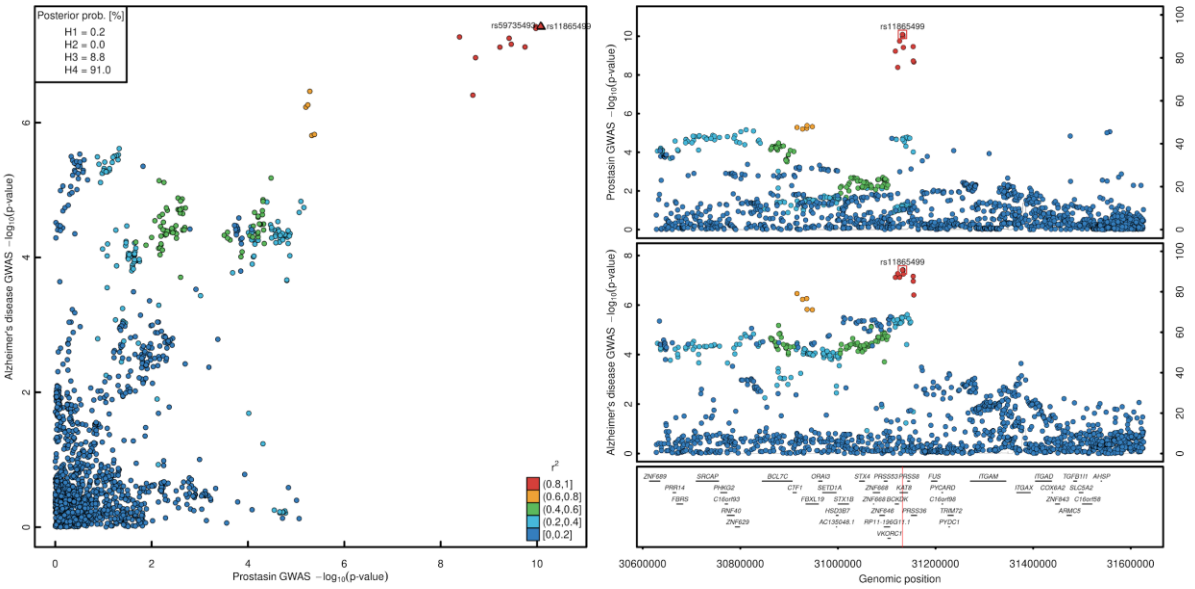


Fig. S4. Locusplot comparing association statistics for prostasin and Alzheimer's disease. The left panel displays a comparison of -$\log_{10}$-transformed p-values from GWAS summary statistics for genetic variants in a 500 kb region around the lead signal on chromosome 16. Coloring was done based on linkage disequilibrium with the lead variant for the protein. The right panel is a stacked locuszoom plot with annotation of protein-encoding genes underneath. Location of the lead variant is indicated by a red line.

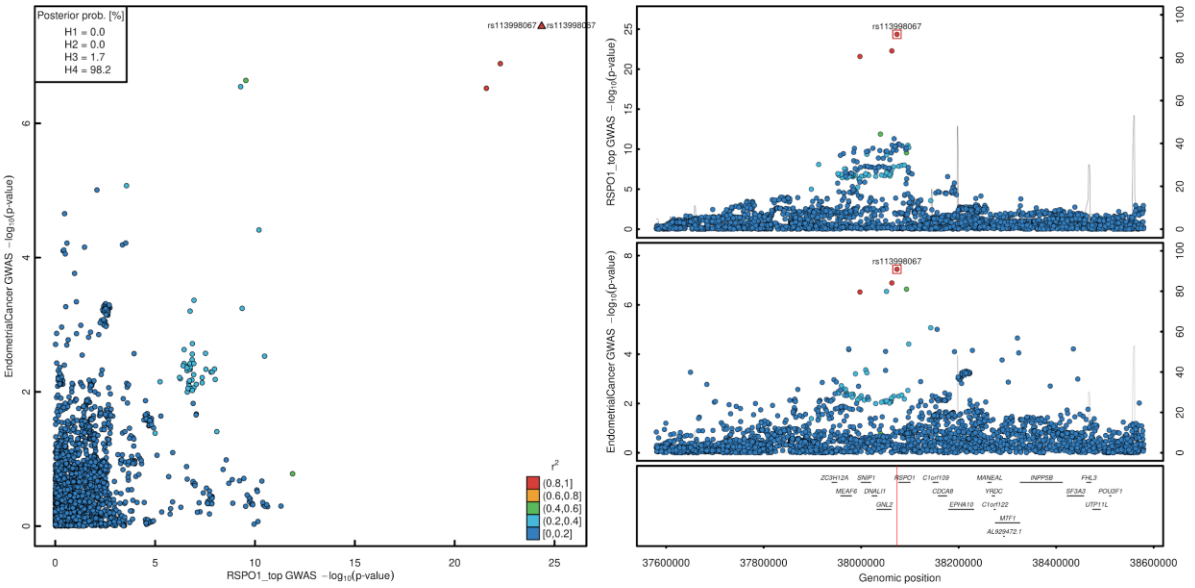


Fig. S5. Locusplot comparing association statistics for RSPO1 and endometrial cancer. The left panel displays a comparison of -log10-transformed p-values from GWAS summary statistics for genetic variants in a 500 kb region around the lead signal on chromosome 1. Coloring was done based on linkage disequilibrium with the lead variant for the protein. The right panel is a stacked locuszoom plot with annotation of protein-encoding genes underneath. Location of the lead variant is indicated by a red line.

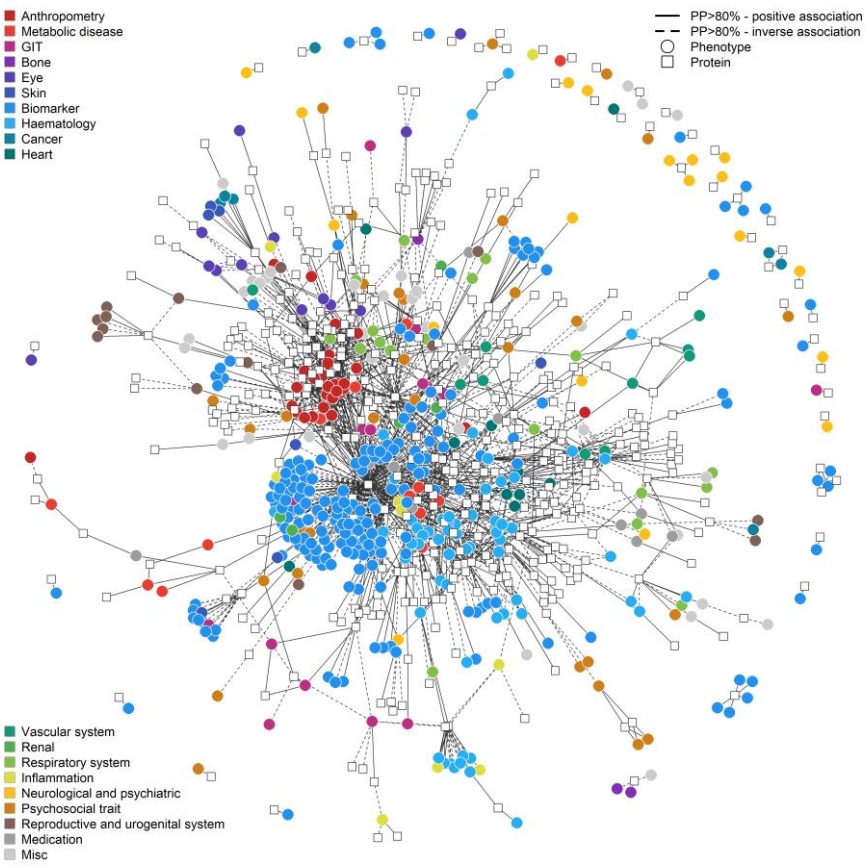


Fig. S6 Proteo-genomic map of human health. The network is composed of 412 protein targets (squares) and 506 phenotypes (circles) as nodes, which are connected (n=1,859 edges) if there is evidence of a shared genetic signal (posterior probability >80%). Only protein targets and phenotypes with at least one connection are included. Effect directions are indicated by the line type aligned with the allele associated with higher amounts of the protein target (solid – positive, dashed – inverse association with the phenotype). Colors indicate categories of phenotypes. The inset represents the entire network, including continuous and binary phenotypes, whereas the larger figure is restricted to binary phenotypes. An interactive version of the figure can be found at www.omicscience.org/apps/pgwas.

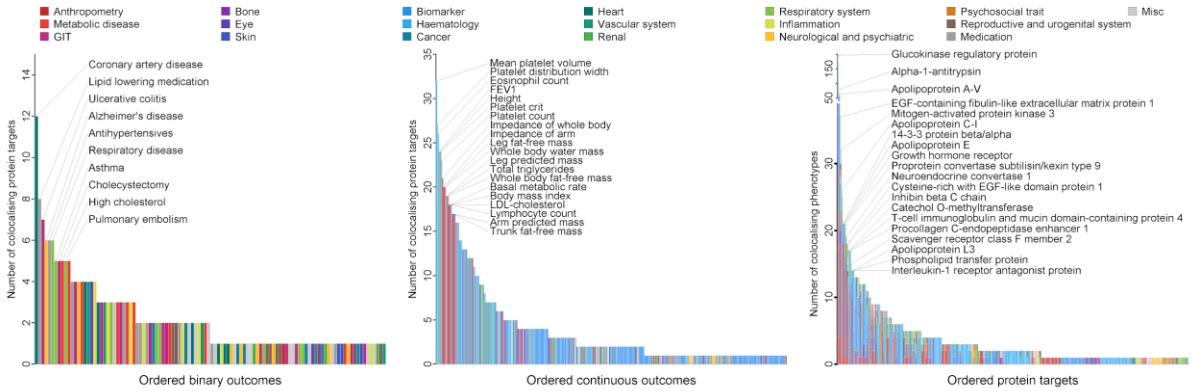


Fig. S7 Summary of cis-region-based phenome-wide colocalization analysis. The left panel represents the number of colocalizing protein targets (posterior probability >80% for a shared genetic signal) for each binary outcome, whereas the panel in the middle does the same for continuous traits. Traits were ordered by the number of colocalizing protein targets, and top traits are annotated. The right panel displays the number of colocalizing phenotypes for each protein target, and stacked bar charts were used to indicate diversity of phenotypes based on the categories indicated above the plots.

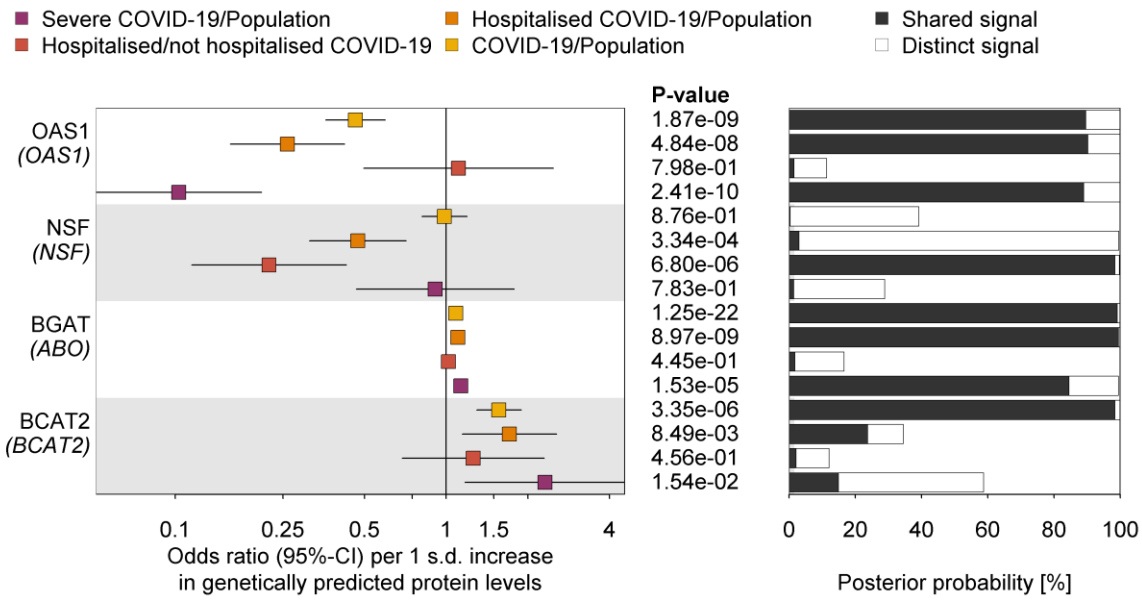


Fig. S8 Protein targets related to COVID-19. Odds ratios and 95%-CIs for the genetically predicted effect of protein levels on four different outcome definitions and control populations for COVID-19 (left), including protein targets with strong evidence for statistical colocalization for at least one definition (right). The column in the middle reports p-values.

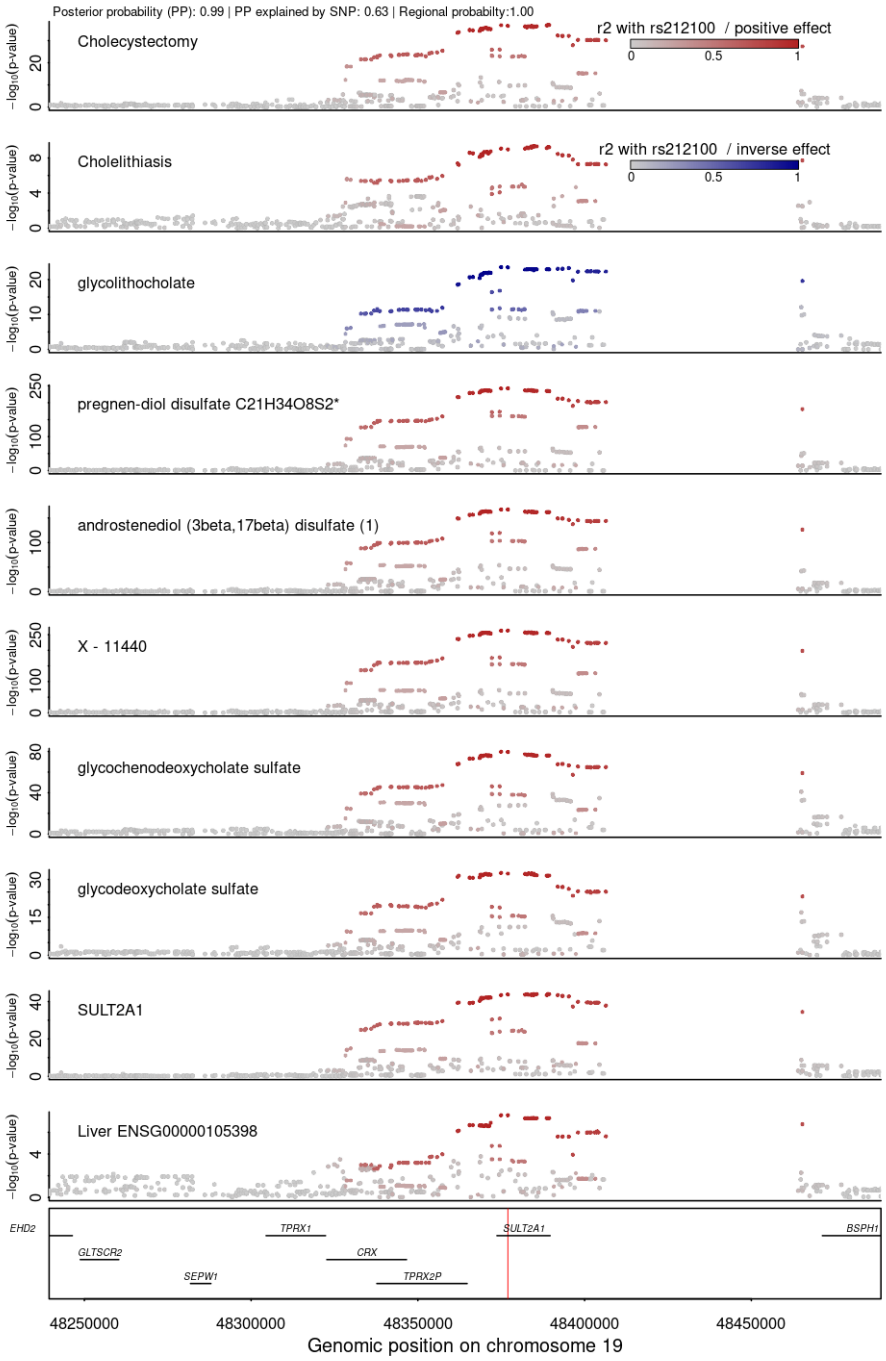


Fig. S9 Stacked regional association plots for phenotypes colocalizing with bile salt sulfotransferase (SULT2A1). Phenotypes are annotated and coloring is based on effect directions aligned to the protein-increasing allele for rs212100 (red-positively, blue-inversely). Darker colors indicate higher linkage disequilibrium with rs212100, and gray dots are below 0.1. Posterior and regional probabilities from multi-trait colocalization are given at the top of the plot.

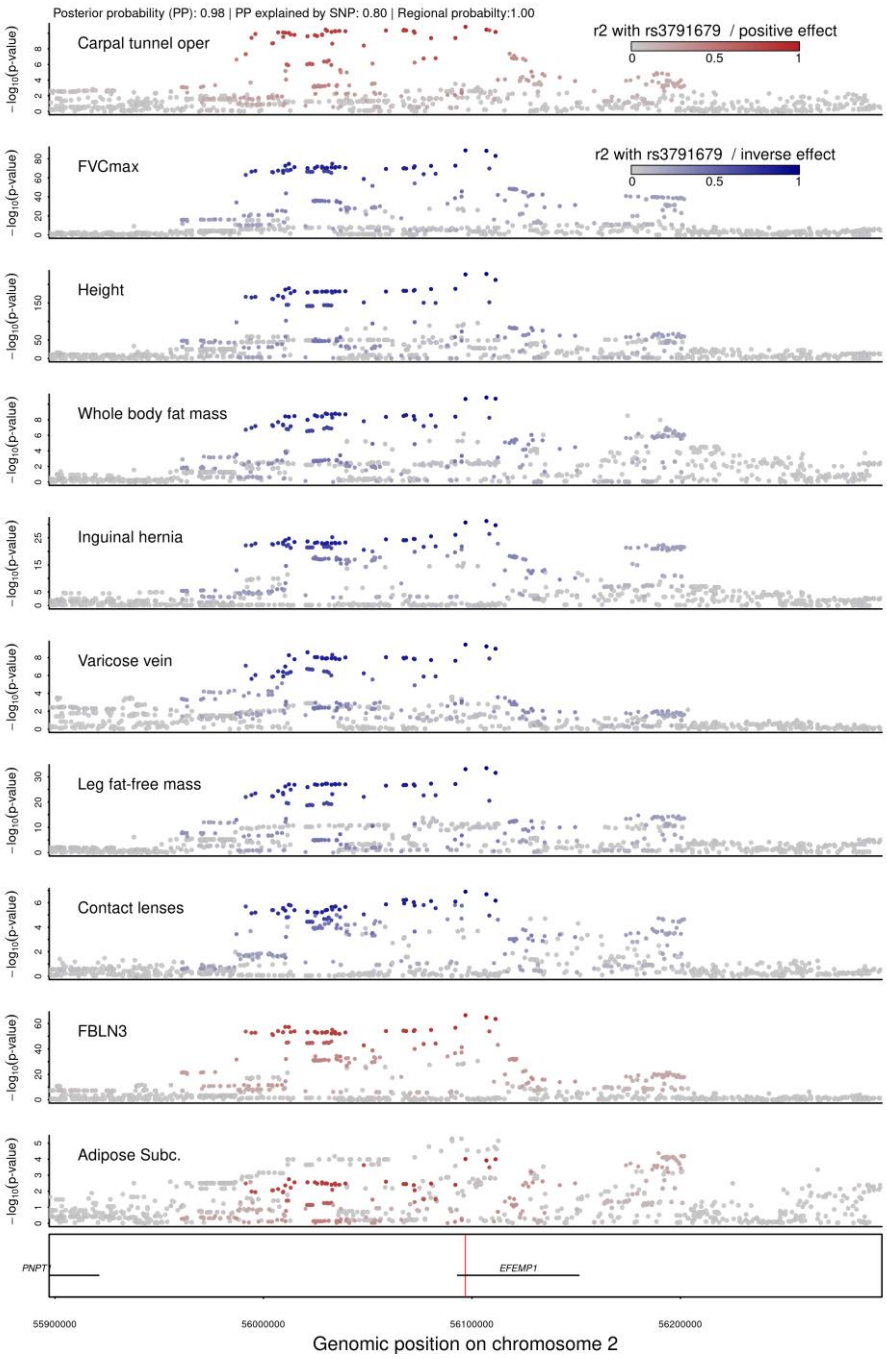


Fig. S10 Stacked regional association plots for phenotypes colocalizing with FBLN3. Phenotypes are annotated and coloring is based on effect directions aligned to the protein-increasing allele for rs3791679 (red-positively, blue-inversely). Darker colors indicated higher linkage disequilibrium with rs3791679, and gray dots are below 0.1. Posterior and regional probabilities from multi-trait colocalization are given at the top of the plot.

Table titles and legends

Table S1. Demographics of the Fenland study population.

Table S2. Summary of variant – protein target associations. The table includes all lead and secondary signals across all 2,584 identified genomic regions, including summary statistics, functional annotations, variant classification, and gene assignments.

Table S3. Explained variance for 4,030 distinct aptamers targeting 3,892 proteins with at least one pQTL. The table includes the amount of variance explained in protein abundances separately for each of the three classification criteria of pQTLs and contains further information on possible druggable targets.

Table S4. Integration of gene and splicing QTLs from the GTEx version 8 release. For each protein target the strongest eQTL and/or sQTL are listed along with colocalization priority and possible tissue specificity is indicated.

Table S5. Summary of pQTL mapping to known GWAS loci. For each mapping cis-pQTL – GWAS variant pair all curated traits are listed, and a column indicates whether the protein-encoding genes has been reported at this locus.

Table S6. Protein target – phenotype connections with strong evidence of colocalization at the protein encoding locus. The table contains all protein – phenotype connections as shown in Figure 5 with further information on association statistics for the lead cis-pQTL.

Table S7. Results from pathway enrichment analysis. For each curated phenotype with at least three associated proteins in the proteo-genomic map GO term enrichment was performed and results are presented collapsing pathways with the same gene set into one entry.

Table S8. Results from age- and sex-interaction analysis for cis-pQTLs. The table lists for 14 identified cis-pQTLs with a potential age- or sex-differential effect results from interaction testing and results within each stratum, that is, for each sex separately as well as for middle age and older individuals.

Table S9. Summary on identified druggable targets with potential for repurposing.