nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01963-z

The genetic landscape of neuro-related proteins in human plasma

Received: 21 March 2023

Accepted: 22 July 2024

Published online: 29 August 2024

Check for updates

Linda Repetto[1,2,3,4,39], Jiantao Chen[1,2,5,39], Zhijian Yang[1,2,5,6,39], Ranran Zhai[1,2,5,6,39], Paul R. H. J. Timmers[3,7], Xiao Feng[2,5], Ting Li[1,2,5], Yue Yao[2,5], Denis Maslov[8], Anna Timoshchuk[8], Fengyu Tu[1], Emma L. Twait[9], Sebastian May-Wilson[3], Marisa D. Muckian[3], Bram P. Prins[10], Grace Png[11,12], Charles Kooperberg[13], Åsa Johansson[14], Robert F. Hillary[15], Eleanor Wheeler[16], Lu Pan[6], Yazhou He[17], Sofia Klasson[18], Shahzad Ahmad[19], James E. Peters[20], Arthur Gilly[11], Maria Karaleftheri[21], Emmanouil Tsafantakis[22], Jeffrey Haessler[13], Ulf Gyllensten[14], Sarah E. Harris[23], Nicholas J. Wareham[16], Andreas Göteson[24], Cecilia Lagging[18,25], Mohammad Arfan Ikram[19], Cornelia M. van Duijn[19], Christina Jern[18,25], Mikael Landén[6,24], Claudia Langenberg[16,26,27], Ian J. Deary[23], Riccardo E. Marioni[15], Stefan Enroth[14], Alexander P. Reiner[28], George Dedoussis[29], Eleftheria Zeggini[11,30], Sodbo Sharapov[8,31], Yurii S. Aulchenko[8,32], Adam S. Butterworth[10,33,34,35,36,37], Anders Mälarstig[6,38], James F. Wilson[3,7,39], Pau Navarro[3,7,39] & Xia Shen[1,2,3,5,6,39] ✉

Understanding the genetic basis of neuro-related proteins is essential for dissecting the molecular basis of human behavioural traits and the disease aetiology of neuropsychiatric disorders. Here the SCALLOP Consortium conducted a genome-wide association meta-analysis of over 12,000 individuals for 184 neuro-related proteins in human plasma. The analysis identified 125 *cis*-regulatory protein quantitative trait loci (*cis*-pQTL) and 164 *trans*-pQTL. The mapped pQTL capture on average 50% of each protein's heritability. At the *cis*-pQTL, multiple proteins shared a genetic basis with human behavioural traits such as alcohol and food intake, smoking and educational attainment, as well as neurological conditions and psychiatric disorders such as pain, neuroticism and schizophrenia. Integrating with established drug information, the causal inference analysis validated 52 out of 66 matched combinations of protein targets and diseases or side effects with available drugs while suggesting hundreds of repurposing and new therapeutic targets.

Certain patterns of human behaviours such as cigarette smoking, alcohol consumption and a high-fat diet may elevate the risk of developing a range of complex diseases[1,2]. Neuropsychiatric disorders are among the leading causes of lifelong disability globally, affecting around 800 million people[3,4]. As of 2024, mental health remains a global crisis and priority brought to the forefront of public health discussions anew, after the impact of COVID-19 on people's lives, where stressors such as isolation, notable changes in habits, and global enhanced mortality and fear of contracting the disease have had severe consequences on mental well-being[5–7]. These conditions represent a substantial challenge for medical research owing to the high complexity of their neurobiological mechanisms and heterogeneity of symptoms, which often overlap with other neurological, psychiatric and non-psychiatric disorders[8–10].

In the past decade, genome-wide association studies (GWAS) have been successful in identifying numerous genetic variants that can partially account for variation in complex traits and diseases[11,12].

A full list of affiliations appears at the end of the paper. ✉e-mail: shenx@fudan.edu.cn

However, the effect of a genetic variant such as a single-nucleotide polymorphism (SNP) on a complex disease is usually very small and often does not provide information on the phenotype's molecular architecture. Measuring proteins may overcome this obstacle as proteins are the product of translated DNA and functional elements that bridge the genetic codes and disease outcomes. Circulating proteins in blood plasma originate from various organ tissues and cell types in the human body and have fundamental roles in different biological processes[13–15]. Thus, such proteins are often used in clinical practice as disease biomarkers. Circulating neurology-related proteins have the potential to provide insight into the pathophysiology of neurological and mental disorders and the genetic architecture of their molecular pathways, setting the basis for the improvement of diagnostic instruments and targeted therapy[16].

Protein levels are more linked to variation in cognitive function than genetic variants alone. Current studies on neurology-related proteins either focused on neurodegenerative disorders or cognitive function specifically or had a limited sample size[17–22]. In a recent study, neurology-related proteins were associated with general fluid cognitive abilities in late life, and a portion of these was observed to be mediated by brain volume, measured as a structural brain variable[20].

The field of proteomics has been rapidly expanding in recent years and produced results that have played a fundamental role in the decoding process of molecular mechanisms involved in several traits and diseases, from cardiovascular disease to general health[19,23–26]. The genomic studies of the human proteome have benefited from various high-throughput measurement techniques, such as mass spectrometry[14,27], aptamer-based assays[28] and antibody-based assays[15]. Among these, the antibody-based Proximity Extension Assay[29] has high measurement precision, especially for many functional but low-abundant proteins.

This study aims to identify genetic variants associated with 184 neurology-related blood circulating proteins via a large-scale genome-wide association meta-analysis (GWAMA) and investigate the proteins' genetic relationships with potential disease-causing behaviours, common psychiatric disorders and related comorbidities. We systematically investigate the proteins' therapeutic implications based on established drug information. We provide an atlas for the genetic architecture of these proteins as a resource for biomedical research on human behaviours and psychiatric disorders.

Results

GWAMA of 184 proteins identified 289 significant loci

In the discovery phase, we conducted a GWAMA using data from up to 12,176 individuals (mean age = 61.9, percentage females = 44.6%) for 92 proteins in the Olink Neurology panel, and up to 5,013 individuals (mean age = 49.6, percentage females = 56.1%; see Supplementary Tables 16–30 for details) for 92 proteins in the Olink Neuro-Exploratory panel, from a total of 12 participating cohorts (Supplementary Tables 17–30). Overall, we identified 289 top variants distributed across a total of 125 *cis*-pQTL and 164 *trans*-pQTL with the significance threshold of $P < 5 \times 10^{-8}$ for the *cis*-loci and $P < 5 \times 10^{-8}/184 = 2.7 \times 10^{-10}$ for the *trans*-loci (Supplementary Table 1 and Supplementary Figs. 10 and 11). Out of the 139 proteins with detected pQTL, 74 (53%) proteins had significantly associated variants both in *cis*- and *trans*-regulatory loci. The median number of primary associations per protein that we observed was 2, with the maximum number of pQTLs per protein being 6. Proteins with lower abundance tended to have weaker *cis*-pQTL association signals (Supplementary Fig. 8).

As expected, the identified *trans*-pQTL, in general, were more weakly associated than the *cis*-pQTL; nevertheless, we found that 24 proteins shared a total of 14 *trans*-pQTL. For example, well-known pleiotropic loci such as the *HLA* region and the *ABO* locus showed *trans*-regulatory effects across a number of plasma proteins (Fig. 1a). For instance, 19 proteins showed significant *trans*-pQTL at the *ABO* locus; nevertheless, the associations were not completely due to the same causal variants (Supplementary Fig. 3). The nearest coding genes for 199 mapped pQTL have *cis*-eQTL data available from the eQTLGen[30]. Among them, 185 (93%) pQTL were also expression QTL (eQTL) significantly associated with the expressions of the nearest genes ($P < 5 \times 10^{-8}$). However, regarding the underlying genetic regulation of transcripts and protein expressions, compared with *trans*-pQTL, *cis*-pQTL were more likely to colocalize with eQTL (Supplementary Figs. 1, 2 and 7). The lead variants of the *cis*-pQTL were also more centred around the transcription start sites (TSS) of the corresponding coding genes, compared with those of the *trans*-pQTL around the TSS of the nearest coding genes (Fig. 1c). The *cis*-pQTL also had stronger effects, less correlated with the minor allele frequencies (MAFs), compared with the *trans*-pQTL (Fig. 1c–d).

The fact that the *trans*-pQTL were not as colocalized as *cis*-pQTL were with eQTL could be partly due to the weaker signals of the *trans*-pQTL than those of the *cis*-pQTL. However, we hypothesized that the *trans*-pQTL may not necessarily reflect the biological regulatory mechanisms of the corresponding proteins, but rather driven by underlying features of the blood samples, owing to their influence on the immuno-reaction of the antibody-based assay. For example, the pleiotropic *trans*-pQTL across the proteins highlight major blood coagulation and clotting factors such as KLKB1 (plasma kallikrein), KNG1 (Kininogen-1) and F12 (coagulation factor XII), as well as glycosylation locus *ST3GAL4*. We thus also looked into the functional pathways and gene sets that involve the closest genes to our *trans*-pQTL using the gene set enrichment analyses (Supplementary Fig. 5). With a false discovery rate <5%, 997 significant pathways were found to be enriched for the genes of our *trans*-loci, of which 443 (44.4%) were driven or partly driven by the *HLA* genes. The top enriched pathways were clustered into inflammatory and immune responses, coagulation processes, cell-to-cell signalling and adhesion, and protein glycosylation (Supplementary Table 13). In particular, the *trans*-pQTL were found to be enriched in (1) established GWAS traits such as blood protein levels, platelet count and platelet crit; (2) GO pathways such as biological adhesion, wound healing, coagulation and glycosylation; (3) hallmark gene sets including coagulation; (4) Reactome pathways including haemostasis and clotting formation; and (5) microRNA targets and Wiki pathways for blood clotting cascade. To further justify the hypothesis that the blood coagulation factors may affect the performance of the antibody-based assay, we cross-referenced the lead variants of the mapped *trans*-pQTL in the pQTL results of the Icelandic population[31], where the proteome was measured using aptamer-based assays instead. In total, 69 *trans*-pQTL for 50 proteins overlapped between the 2 datasets, and the 8 loci whose nearest genes are involved in coagulation pathways replicate notably worse than the other *trans*-pQTL (Supplementary Fig. 6).

We assessed the overall heritabilities across the 184 analysed plasma proteins. Methods based on summary association statistics have been developed to infer heritability and genetic correlation parameters for complex traits with GWAS results; however, consistent estimates can only be obtained for genetic correlations[32–34]. Thus, we used a standard polygenic mixed model on the individual-level data collected in the ORCADES cohort to assess the narrow-sense heritability for each protein[35]. Across the analysed proteins, we found that the higher the protein's heritability, the more pQTL detected for the protein (Fig. 1e), the stronger the *cis*-pQTL effects are (Fig. 1g), and the higher amount of phenotypic variance captured by the detected pQTL (Fig. 1f). On average, the mapped pQTL together explain 49% of the proteins' heritability. This indicates that proteins as molecular phenotypes have strong major regulatory loci. Nevertheless, their genetic effects can still be widespread across the genome, having a polygenic genetic architecture.

Using data from the ORCADES cohort, we found TDGF1 (Teratocarcinoma-Derived Growth Factor 1) to have the highest

https://doi.org/10.1038/s41562-024-01963-z

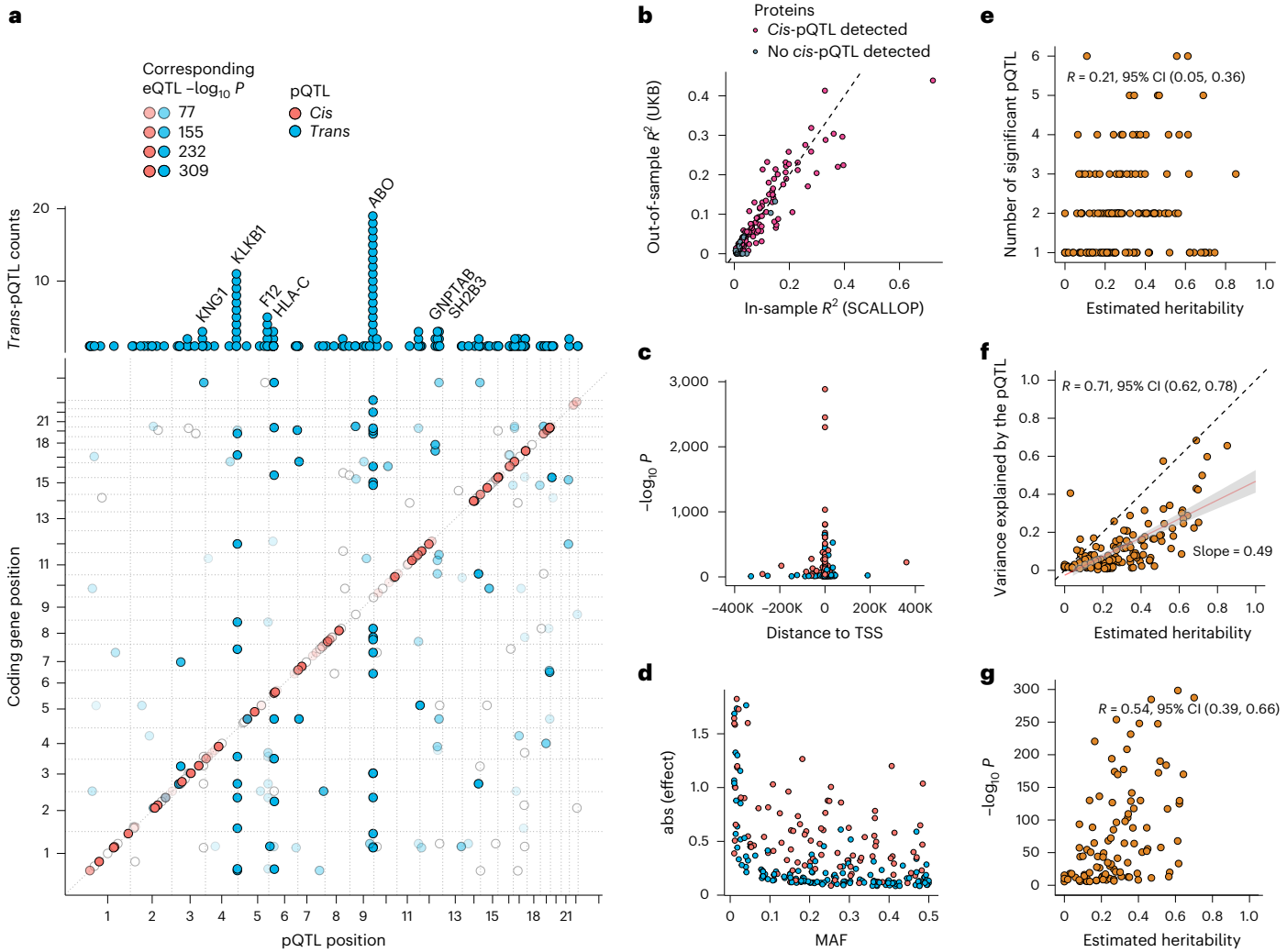

Fig. 1 | Overview of the mapped pQTL. The displayed *P* values were obtained from two-sided Wald tests in the GWAS without correction for multiple testing. **a**, Pleiotropic *trans*-pQTL counts and overlap of the mapped pQTL with existing eQTL. The upper bar plot shows the number of proteins that share *trans*-pQTL (gene annotations based on the gene closest to the *trans*-pQTL). The scatter plot shows the genomic location of significant *cis*-pQTL in red ($P < 5 \times 10^{-8}$) and significant *trans*-pQTL in blue ($P < 5 \times 10^{-8}/184$), and the shading within the dots indicates the significance of the corresponding/nearest *cis*-eQTL for the respective protein. **b**, Comparison of the phenotypic variance captured by the discovered pQTL of each protein in the SCALLOP meta-analysis (in-sample R^2) and the predictable phenotypic variance in the UKB-PPP data (out-of-sample R^2). **c**, Scatter plot of the pQTL lead variants association signals versus their distance to the TSS of the corresponding/nearest coding genes. **d**, Scatter plot of the absolute estimated genetic effects of the pQTL lead variants versus their MAFs. **e**, Number of mapped pQTL per protein versus the linear mixed model estimated heritability in the ORCADES cohort. **f**, The variance explained by the mapped pQTL summed up for each protein versus the estimated heritability. **g**, For the proteins with significant *cis*-pQTL mapped, the lead variant signal strength versus the estimated heritability of each protein. The correlation coefficients (R) and their relative confidence intervals (CI) are indicated in the plot.

heritability ($h^2 = 0.85$), followed by MDGA1 (MAM Domain-Containing Glycosylphosphatidylinositol Anchor Protein 1, $h^2 = 0.75$), CLM1 (CD300 Molecule Like Family Member F, $h^2 = 0.72$) and LAIR2 (Leukocyte-Associated Immunoglobulin-Like Receptor 2, $h^2 = 0.70$). By contrast, CTF1 (Cardiotrophin 1), EPHA10 (Ephrin Type-A Receptor 10), GSTP1 (Glutathione *S*-Transferase Pi 1), HSP90B1 (Heat Shock Protein 90 Beta Family Member 1), IFI30 (Gamma-Interferon-Inducible Lysosomal Thiol Reductase), NDRG1 (N-Myc Downstream Regulated 1) and SFRP1 (Secreted Frizzled Related Protein 1) all had an estimated h^2 value close to 0 while having at least one pQTL.

We used the PhenoScanner database[36,37] to determine whether the pQTL sentinel variants or variants in linkage disequilibrium (LD) with them ($r^2 > 0.8$) that we identified had been previously found to be significantly associated with the corresponding proteins (Supplementary Table 2). We identified 113 loci within our own results that had already been discovered in previous studies. We also checked whether the hits from the meta-analysis were significant in the individual cohorts and observed that 73 of the sentinel variants were found to be statistically significant only in the meta-analysis. We extracted the established associations between our mapped *cis*-pQTL and complex traits from the PhenoScanner database (Supplementary Table 3). At a 5% false discovery rate, 39 *cis*-pQTL showed a significant association with both complex traits and other proteins (mostly based on an aptamer-based assay). We found that the level of pleiotropy at the protein level, that is, being *trans*-pQTL for other proteins, is associated with the level of pleiotropy on the complex traits (Supplementary Fig. 4).

We performed LD pruning ($r^2 < 0.001$) to identify secondary independent associations at the *cis*-pQTL. We identified a total of 162 additional significantly associated variants across all the 125 proteins with *cis*-pQTL mapped (Supplementary Tables 7 and 8).

This meta-analysis within our SCALLOP collaborative framework is a follow-up of a previous study on the proteins from the Olink Neurology and Neuro-Exploratory panels, where data were collected

https://doi.org/10.1038/s41562-024-01963-z

Protein	Trait	OR (95% CI)	P_{MR}
Cognitive function			
ADAM22	Fluid intelligence test: conditional arithmetic	1.07 (1.02 – 1.12)	4.76×10^{-3}
CDH6	Fluid intelligence test: word interpolation	0.92 (0.88 – 0.97)	9.76×10^{-4}
DPEP2	Positional arithmetic	0.84 (0.74 – 0.96)	8.35×10^{-3}
Food/drink			
CSF3	Spread type: butter/spreadable butter	1.22 (1.12–1.32)	1.70×10^{-6}
CSF3	Cereal intake	0.93 (0.90–0.96)	1.31×10^{-5}
CTSS	Tea intake	0.97 (0.96–0.98)	5.84×10^{-10}
CTSS	Coffee intake	0.99 (0.98–0.99)	2.26×10^{-5}
CTSS	Water intake	1.02 (1.02–1.03)	1.54×10^{-9}
DEFB4A	Oily fish intake	0.96 (0.93–0.99)	4.60×10^{-3}
DEFB4A	Processed meat intake	1.10 (1.07–1.14)	7.68×10^{-9}
DEFB4A	Cheese intake	1.09 (1.06–1.13)	1.01×10^{-6}
DEFB4A	Type of meals eaten yesterday: restaurant meals	1.54 (1.21–1.98)	5.55×10^{-4}
IL32	Unsalted nuts intake	1.03 (1.01–1.05)	1.91×10^{-3}
ISLR2	Cereal intake	0.96 (0.94–0.98)	8.61×10^{-6}
LGALS8	Weekly red wine intake	0.98 (0.97–0.99)	3.37×10^{-7}
LGALS8	Alcohol usually taken with meals	0.93 (0.91–0.96)	1.73×10^{-7}
NCAN	Non–butter spread type: Flora Pro-Activ or Benecol	1.37 (1.18–1.60)	5.48×10^{-5}
ROBO2	Dark chocolate intake	0.92 (0.89–0.96)	6.16×10^{-5}
ULBP2	Side salad intake	1.07 (1.04–1.10)	2.97×10^{-5}
Lifestyle			
CDH6	Frequency of friend/family visits	0.99 (0.98–1.00)	2.07×10^{-2}
CDH6	Age first had sexual intercourse	0.98 (0.97–0.99)	2.36×10^{-4}
CDH6	None of attendance/disability/mobility allowance	0.94 (0.90–0.99)	1.03×10^{-2}
CDH6	Disability living allowance	1.09 (1.03–1.14)	1.06×10^{-3}
CTSS	Use of sun/UV protection	1.02 (1.01–1.03)	1.39×10^{-7}
DEFB4A	Usual side of head for mobile phone use: left	1.15 (1.07–1.24)	3.27×10^{-4}
DEFB4A	Usual side of head for mobile phone use: right	0.87 (0.80–0.93)	1.62×10^{-4}
DEFB4A	Prefer an 'evening' than a 'morning' person	1.06 (1.02–1.09)	9.99×10^{-4}
DEFB4A	Times of suffering sunburn in childhood	1.10 (1.06–1.13)	1.95×10^{-7}
DEFB4A	Age first had sexual intercourse	1.09 (1.06–1.13)	3.02×10^{-7}
FLRT2	Duration of walks	0.98 (0.97–0.99)	2.37×10^{-4}
NCAN	Frequency of stair climbing in last 4 weeks	0.92 (0.88–0.97)	2.48×10^{-3}
PAEP	Ever addicted to a behaviour	1.32 (1.14–1.53)	1.85×10^{-4}
WWP2	Risk taking	0.81 (0.74–0.89)	1.94×10^{-5}
Substance use			
ASAH2	Opioid dependence	0.96 (0.94–0.99)	2.48×10^{-3}
CDH17	Number of cigarettes smoked daily	0.94 (0.92–0.97)	8.92×10^{-6}
CDH6	Past tobacco smoking	0.97 (0.95–0.98)	4.03×10^{-7}
CDH6	Never smoked	0.95 (0.93–0.97)	1.41×10^{-6}
Work/education			
CDH6	Qualifications: college or university degree	0.97 (0.95–0.99)	5.14×10^{-3}
CDH6	Qualifications: A levels/AS levels or equivalent	0.97 (0.95–0.99)	9.51×10^{-3}
CDH6	Unable to work because of sickness or disability	1.07 (1.01–1.13)	1.19×10^{-2}
PRTG	Qualifications: College or university degree	0.95 (0.93–0.98)	2.91×10^{-4}
RGMA	Sometimes feeling hot at the workplace	0.92 (0.85–0.99)	2.58×10^{-2}
TPPP3	Qualifications: College or university degree	1.14 (1.07–1.22)	1.44×10^{-4}
VWC2	Work hours per week	1.06 (1.03–1.10)	2.67×10^{-4}

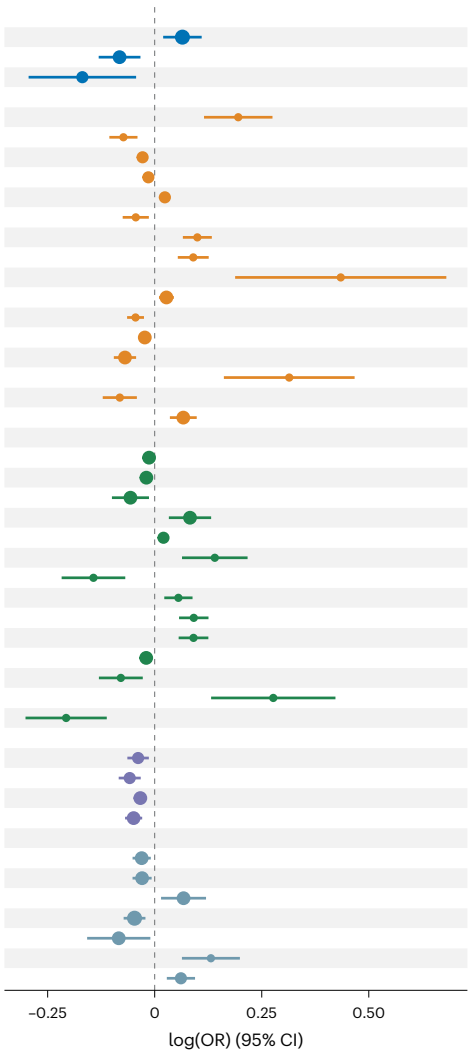

Fig. 2 | Effects of the proteins on human behavioural traits inferred by MR analyses. The forest plot shows the significant MR results (false discovery rate < 0.05) based on LD-pruned ($r^2 < 0.001$) instrumental variants within each *cis*-pQTL. IVW MR results are provided as the estimates (solid round dots) +/− half of the 95% confidence intervals (CI; the whiskers). Odds ratios (OR) are indicated in the third column with the appropriate CI. The *P* values were obtained from two-sided Wald tests in the GWAS without correction for multiple testing. The displayed protein–trait pairs all showed colocalization evidence in the standard coloc test (PP.H4 > 0.8). The numbers of instrumental variants in the *cis*-pQTL are indicated as the size of the dots. The sample sizes for deriving the GWAS summary statistics are given in Supplementary Table 10.

from the two Greek cohorts that we included in this study[38]. Our results replicated over 90% of the established loci, including the previous main discoveries of the *cis*-pQTL for CD33, GPNMB and MSR1. Furthermore, we cross-referenced the significant loci discovered in the meta-analysis with the currently available pQTL data from the UK Biobank Pharma Proteomics Project (UKB-PPP)[39]. One hundred seventy-four out of the 184 proteins in our SCALLOP analysis currently overlap with the UKB-PPP analysis (Supplementary Table 1). Among these, we reported 117 *cis*-pQTL, where 110 (94%) can be replicated in the UKB-PPP analysis ($P < 5 \times 10^{-8}$); UKB-PPP reported 22 additional *cis*-pQTL; and for 32 proteins, there were no *cis*-pQTL reported in either SCALLOP or UKB-PPP. We also reported 164 *trans*-pQTL for 88 proteins, where 84 proteins were also measured in the UKB-PPP analysis. For these 84 proteins, we mapped 155 *trans*-pQTL, where 128 (83%) can be replicated in UKB-PPP ($P < 5 \times 10^{-8}/184 = 2.7 \times 10^{-10}$).

If we consider the UKB-PPP as a gold standard for 'Olink assay detectable *cis*-pQTL', this indicates that we have a type I error rate of 1.7% and a type II error rate of 40.7%. If we lower the *cis*-pQTL discovery threshold to 5×10^{-6}, we would yield a type I error rate of 2.4% and a type II error rate of 32.6%. Because of the protein measurement consistency, the genetic effects of our discovered pQTL could be well replicated in the UKB-PPP data: the out-of-sample predictable phenotypic variances of the proteins were consistent with the in-sample phenotypic variances captured by the pQTL (Fig. 1b), especially for the proteins with *cis*-pQTL discovered.

Shared genetics between proteins and neuro-related traits

Focusing on the *cis*-pQTL regions, we investigated the shared genetic architecture between the studied proteins and other human complex traits. We collected the union of GWAS summary statistics from two sources as the outcome data: 4,085 traits from Neale's lab UKB GWAS, and 20 psychiatric or neurological disorder traits from PGC (Psychiatric Genomics Consortium) (Methods and Data Availability; Supplementary Tables 4–6). We adopted colocalization and Mendelian randomization (MR) analysis to illustrate the genetic correlations between the proteins and complex traits at the *cis*-pQTL. Potential causality might be inferred from colocalized protein–trait combinations, given the molecular biological basis of the *cis*-pQTL (Discussion).

https://doi.org/10.1038/s41562-024-01963-z

Protein	Trait	OR (95% CI)	P_{MR}
Neurological disorder			
CLEC10A	Tourette syndrome	1.16 (1.05–1.28)	2.90×10^{-3}
DPEP1	Carpal tunnel syndrome	1.06 (1.02–1.10)	6.57×10^{-3}
Psychiatric disorder			
ADAM22	Anorexia nervosa	0.86 (0.80–0.92)	1.69×10^{-3}
ADAM22	Schizophrenia	0.92 (0.89–0.96)	9.78×10^{-6}
CD302	Schizophrenia	0.90 (0.86–0.94)	6.76×10^{-6}
DEFB4A	Multiple disorders	0.80 (0.68–0.95)	9.25×10^{-3}
GFRA3	Bipolar disorder	1.43 (1.12–1.85)	4.98×10^{-3}
PRTG	Bipolar disorder	0.91 (0.85–0.98)	9.87×10^{-3}
Symptom			
CDH6	Mood swings	1.03 (1.01–1.05)	2.24×10^{-3}
CDH6	Tense/highly strung	1.04 (1.01–1.07)	7.39×10^{-3}
CDH6	Leg pain on walking	1.08 (1.03–1.12)	7.35×10^{-4}
CDH6	Back pain in last month	1.04 (1.02–1.07)	3.01×10^{-4}
CTSC	Number of depression episodes	0.96 (0.94–0.98)	1.58×10^{-4}
DEFB4A	Mood swings	1.11 (1.04–1.19)	2.66×10^{-3}
DEFB4A	Miserableness	1.21 (1.13–1.29)	3.89×10^{-8}
DEFB4A	Irritability	1.19 (1.10–1.28)	7.51×10^{-6}
DEFB4A	Sensitivity/hurt feelings	1.18 (1.10–1.26)	1.25×10^{-6}
DEFB4A	Fed-up feelings	1.14 (1.07–1.23)	1.13×10^{-4}
DEFB4A	Nervous feelings	1.26 (1.16–1.36)	1.92×10^{-8}
DEFB4A	Tense/highly strung	1.27 (1.16–1.39)	1.94×10^{-7}
DEFB4A	Worry too long after embarrassment	1.20 (1.12–1.29)	1.19×10^{-7}
DEFB4A	Neuroticism score	1.13 (1.09–1.17)	1.34×10^{-10}
DEFB4A	Guilty feelings	1.14 (1.06–1.23)	5.73×10^{-4}
DPEP2	Longest period of mania or irritability	1.11 (1.04–1.20)	2.83×10^{-3}
FLRT2	Ever self-harmed	0.86 (0.80–0.94)	4.53×10^{-4}
FRZB	Irritability	1.03 (1.01–1.05)	5.60×10^{-3}
IFNL1	Headache in last month	1.18 (1.09–1.29)	9.53×10^{-5}
NTRK3	Distress caused by unusual or psychotic experiences	0.72 (0.57–0.92)	7.19×10^{-3}

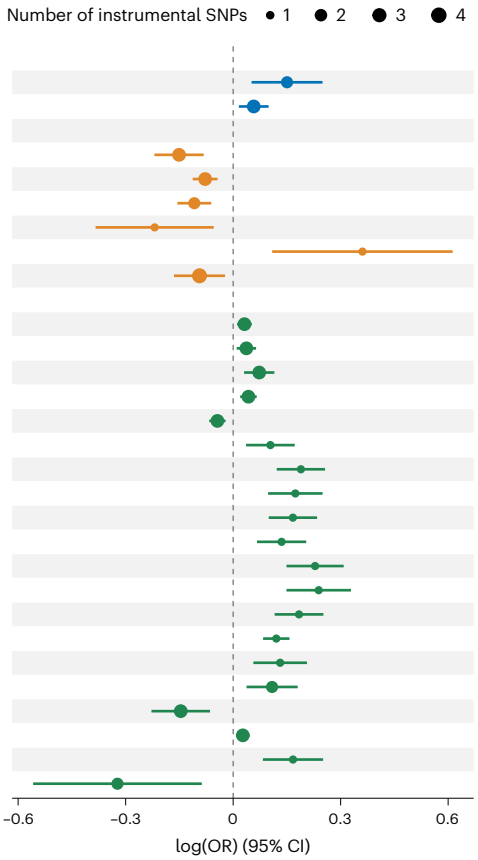

Fig. 3 | Effects of the proteins on neuro-related conditions inferred by MR analyses. The forest plot shows the significant MR results (false discovery rate < 0.05) based on LD-pruned ($r^2 < 0.001$) instrumental variants within each *cis*-pQTL. IVW MR results are provided as the estimates (solid round dots) +/− half of the 95% confidence intervals (the whiskers). The *P* values were obtained from two-sided Wald tests in the GWAS without correction for multiple testing. The displayed protein–trait pairs all showed colocalization evidence in the standard coloc test (PP.H4 > 0.8). The numbers of instrumental variants in the *cis*-pQTL are indicated as the size of the dots. The sample sizes for deriving the GWAS summary statistics are given in Supplementary Table 11.

We started by identifying the protein–trait combinations that have the pQTL and the corresponding trait association signal colocalized. Both the standard coloc method[40] (assuming a single causal variant) and the SuSiE-coloc procedure[41,42] (assuming multiple causal variants) were applied. The 291 protein–trait pairs that showed a colocalization posterior probability (PP.H4) greater than 0.8 in either of the two models were passed onto subsequent MR analysis.

We performed an inverse-variance weighted (IVW) two-sample MR analysis using the LD-pruned genetic instruments across the 125 *cis*-pQTL with colocalization evidence for the outcome phenotypes (Supplementary Table 9). For binary outcomes, to avoid the influence of the subpopulation structure, we focused on those with at least 1,000 cases. With a false discovery rate of 5%, we obtained 287 significant MR effects for 56 proteins on 201 traits (Supplementary Tables 9–11), including 43 human behavioural phenotypes (Fig. 2), 22 psychiatric or neurological conditions (Fig. 3) and 19 other disease-related traits (Fig. 4). Together with the colocalization support, these results are consistent with a potential causal role of each protein, following the assumptions of MR (Discussion).

In terms of human behaviours, for instance, Cadherin-6 (CDH6) showed a positive effect on smoking (Fig. 5a), while CDH17 showed a negative effect. PRTG and TPPP3 showed opposite effects on educational attainment. The protein Cathepsin S (CTSS) showed a positive effect on the use of sun protection; meanwhile, it showed an increasing effect on water intake and decreasing effects on tea and coffee intake (Fig. 5b). Galectin-8 (LGALS8) was a plausible marker for alcohol intake (Fig. 5d).

Regarding psychiatric and neurological conditions, for example, Dipeptidase 1 (DPEP1) showed potential risk-increasing effects on mononeuropathies of the upper limb and carpal tunnel syndrome (Fig. 5c). Besides, ADAM22 and CD302 showed protective effects on schizophrenia. The known effect of CD33 on Alzheimer's disease[38,43] could be replicated using MR-Egger regression on CD33 *cis*-pQTL and the regional associations from PGC ($P = 2.0 \times 10^{-3}$, two-tailed test).

There are protein markers that showed effects on both behavioural phenotypes and neurological disorders. For example, the protein CDH6 showed effects on behavioural traits such as smoking, word interpolation and age of first sexual intercourse, as well as neurological symptoms such as pain, tension and mood swings, where its effects on behavioural traits and neurological symptoms had different directions (Fig. 5a). ADAM22, besides its protective effects on schizophrenia and anorexia nervosa, also showed a positive effect on arithmetic skills.

In relation to other complex diseases, for example, ADAM15 showed a protective effect on infectious and parasitic diseases. CD302 both showed protective effects on hypothyroidism or myxoedema. DPEP1 showed a protective effect against hypertension. Galectin-8 (LGALS8), besides its negative effect on alcohol intake, was found to increase the risk of female genital prolapse (Fig. 5d). Neurocan core protein (NCAN) was found to be genetically associated with high cholesterol, and thus also cholesterol-lowering substitutes such as the use of Flora Pro-Activ or Benecol, since it was used more frequently than the other products, revealing its colocalized genetic associations at the *cis*-pQTL of NCAN (Fig. 5e). Besides its effects on non-butter spread use and stair climbing, the protein NCAN also showed protective effects on hypertension and diabetes. However, such genetic correlations of NCAN were likely driven by its nearby gene *TM6SF2* owing to linkage (Discussion).

https://doi.org/10.1038/s41562-024-01963-z

Protein	Trait	OR (95% CI)	P_{MR}
Other disease			
ADAM15	Infectious and parasitic diseases	0.92 (0.86–0.98)	9.98×10^{-3}
CD200R1	None of clotting disorders, respiratory issues or allergies diagnosed	1.02 (1.01–1.04)	5.50×10^{-3}
CD302	Self-reported hypothyroidism/myxoedema	0.86 (0.81–0.91)	4.34×10^{-7}
CLEC1B	Other diseases of arteries and capillaries	2.62 (1.50–4.58)	7.32×10^{-4}
DEFB4A	Self-reported hypertension	0.88 (0.81–0.94)	5.08×10^{-4}
DEFB4A	Self-reported hypothyroidism/myxoedema	1.38 (1.18–1.61)	4.31×10^{-5}
DEFB4A	None of diagnosed vascular/heart problems	1.13 (1.06–1.22)	5.11×10^{-4}
DEFB4A	Diagnosed high blood pressure	0.87 (0.81–0.94)	2.59×10^{-4}
DEFB4A	None of clotting disorders, respiratory issues or allergies diagnosed	1.17 (1.09–1.26)	1.94×10^{-5}
DEFB4A	Other arthrosis	0.74 (0.65–0.84)	5.41×10^{-6}
DEFB4A	Diseases of the musculoskeletal system and connective tissue	0.84 (0.78–0.91)	3.52×10^{-5}
DPEP1	Self-reported hypertension	0.96 (0.95–0.98)	6.19×10^{-8}
DPEP1	None of vascular/heart problems	1.03 (1.02–1.04)	1.43×10^{-5}
DPEP1	Diagnosed high blood pressure	0.97 (0.95–0.98)	1.14×10^{-6}
DPEP1	ICD10: G56 mononeuropathies of upper limb	1.08 (1.03–1.12)	7.98×10^{-4}
FCRL2	Self-reported hyperthyroidism/thyrotoxicosis	1.19 (1.05–1.35)	4.99×10^{-3}
IL12	Mouth ulcers	1.10 (1.06–1.13)	5.39×10^{-9}
LGALS8	ICD10: N81 female genital prolapse	1.09 (1.02–1.16)	8.53×10^{-3}
MATN3	Self-reported psoriasis	1.18 (1.08–1.29)	1.99×10^{-4}
NCAN	Self-reported hypertension	0.84 (0.77–0.91)	5.20×10^{-5}
NCAN	Self-reported diabetes	0.68 (0.56–0.82)	1.07×10^{-4}
NCAN	Diagnosed diabetes	0.69 (0.58–0.82)	4.40×10^{-5}
SCARB2	Self-reported high cholesterol	1.11 (1.05–1.18)	7.51×10^{-5}
SMOC1	Cataract	0.84 (0.77–0.93)	4.70×10^{-4}
SMOC1	Trigger finger	1.41 (1.12–1.78)	3.26×10^{-3}
TPPP3	Self-reported hypothyroidism/myxoedema	0.71 (0.62–0.82)	1.87×10^{-6}

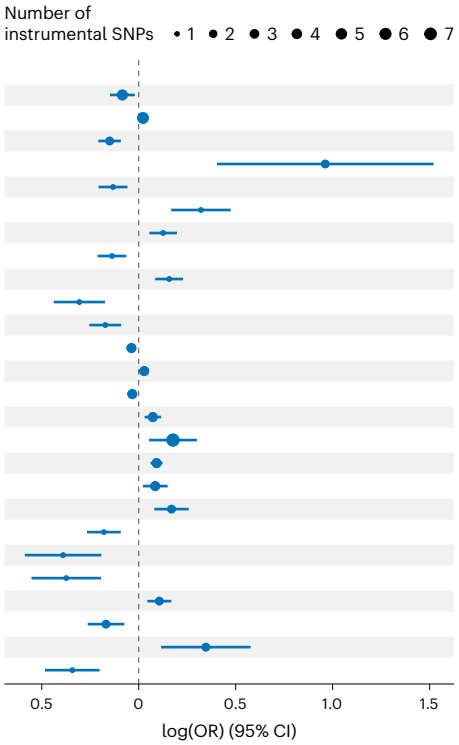

Fig. 4 | Effects of the proteins on other complex diseases inferred by MR analyses. The forest plot shows the significant MR results (false discovery rate < 0.05) based on LD-pruned ($r^2 < 0.001$) instrumental variants within each *cis*-pQTL. IVW MR results are provided as the estimates (solid round dots) +/− half of the 95% confidence intervals (the whiskers). The *P* values were obtained from two-sided Wald tests in the GWAS without correction for multiple testing. The displayed protein–trait pairs all showed colocalization evidence in the standard coloc test (PP.H4 > 0.8). The numbers of instrumental variants in the *cis*-pQTL are indicated as the size of the dots. The sample sizes for deriving the GWAS summary statistics are given in Supplementary Table 9.

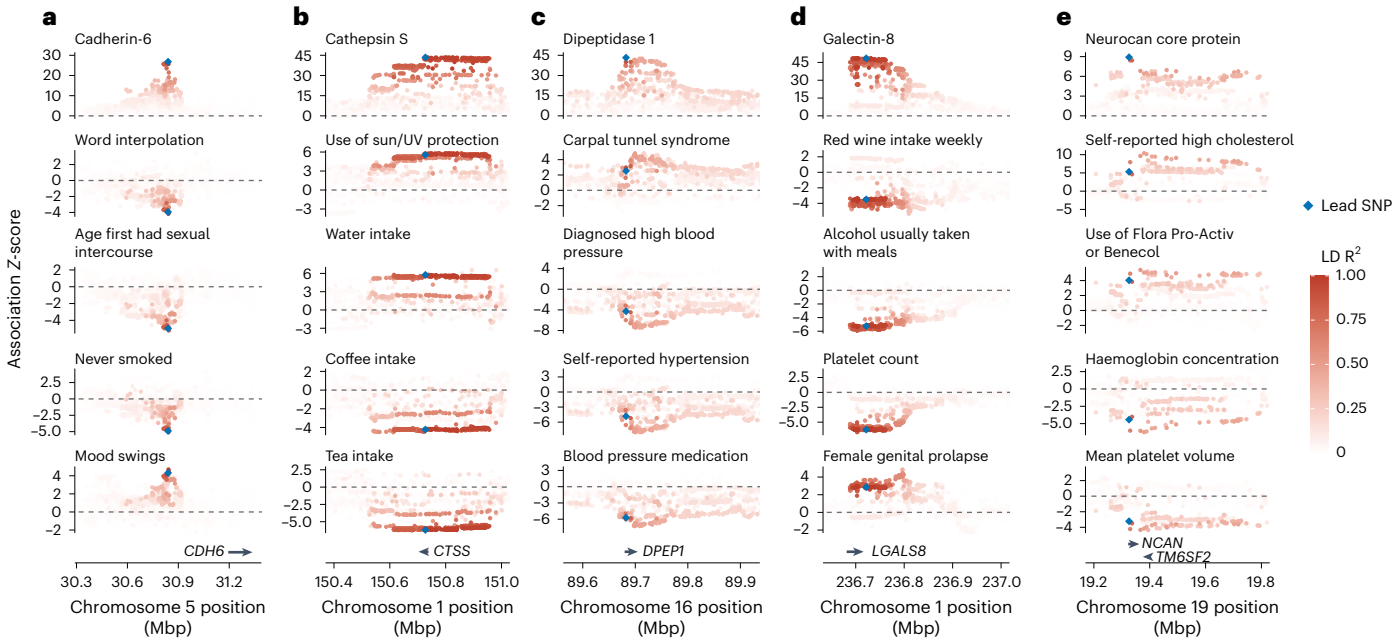

Fig. 5 | Examples of regional association patterns for colocalized *cis*-pQTL and complex traits. a–**e**, The *cis*-pQTL regions for five proteins (CDH6 (**a**), CTSS (**b**), DPEP1 (**c**), LGALS8 (**d**) and NCAN (**e**)) are visualized. Each dot represents a variant, where the lead variant for each *cis*-pQTL is marked as a blue diamond. The other variants within the region are coloured based on their LD R^2 values with the corresponding lead variants. Detailed descriptions of the complex traits analysed can be found in the supplementary tables and the original data sources.

Except for TPPP3 and hypothyroidism or myxoedema, reverse generalized summary-statistics-based MR[44] did not show evidence for reverse causality. In general, the MR estimated odds ratios at a false discovery rate (FDR) of less than 5% were found to range from 0.21 to 2.62, consistent with previous studies evaluating the MR effects of blood circulating proteins on other complex traits[15,45].

https://doi.org/10.1038/s41562-024-01963-z

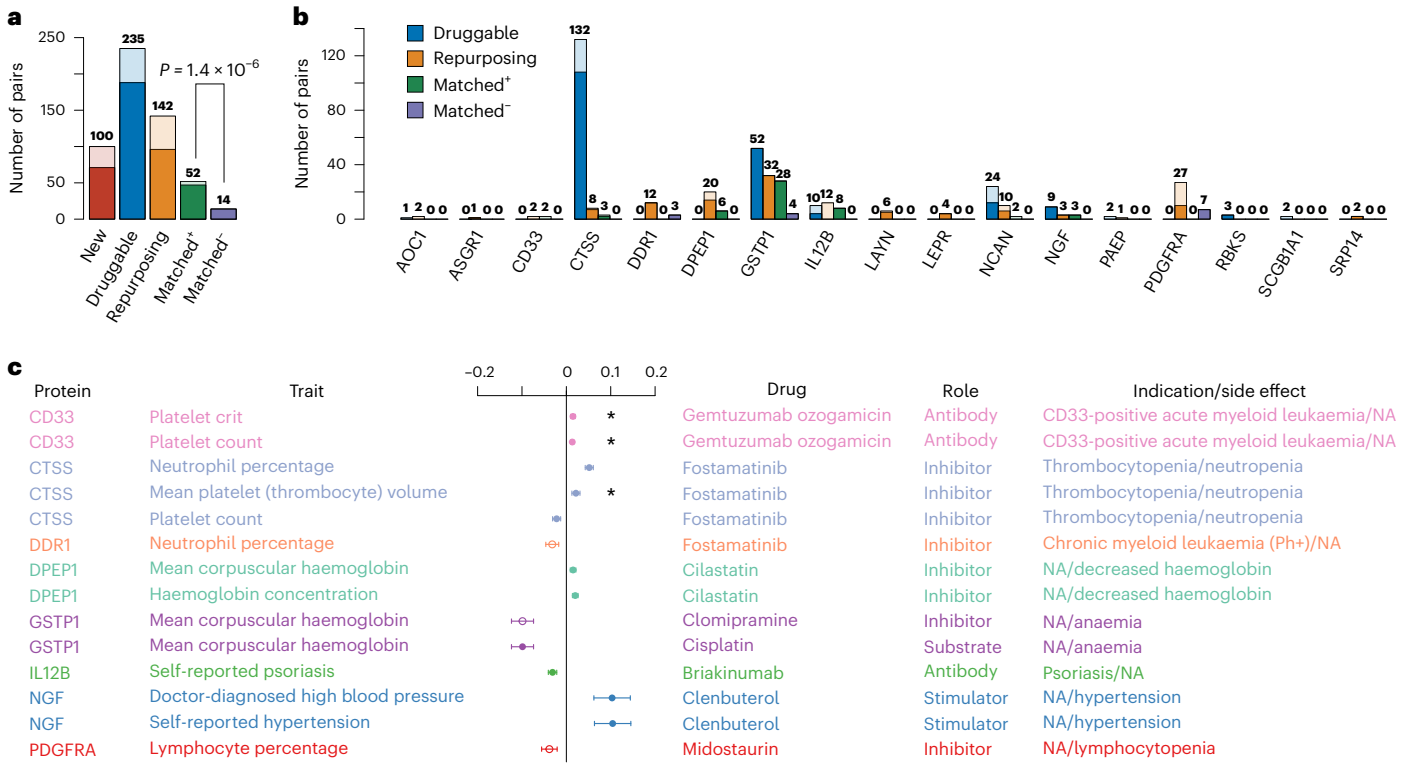

Fig. 6 | Drug targets revealed by MR analyses. The MR results with a 5% false discovery rate were considered. **a**, The number of MR-inferred pairs of proteins and traits split into five categories: new (drug) targets, druggable targets that have drugs with unclear clinical function, repurposing targets that have established drugs but for different diseases, and validated known targets where the established drugs have pharmacological effects that match the MR effect directions (Matched$^+$) and those with opposite directions (Matched$^-$). The sets of targets that have strong colocalization support (PP.H4 > 0.8) are marked in lighter colours. The *P* value was obtained from a two-sided binomial test. **b**, Numbers of different categories of drug targets per protein analysed. **c**, Summarized examples of the validated known drug targets, the description of the drugs, the indication (diseases treated)/side effects and the corresponding consistent MR estimated effects. The MR estimates with colocalization support are marked with stars. IVW MR results are provided as the estimates (solid round dots) +/− half of the 95% confidence intervals (the whiskers). NA, not applicable. The sample sizes for deriving the GWAS summary statistics are given in Supplementary Table 9.

Investigation of drug targets among the protein markers

We systematically investigated the protein markers identified by the causal inference based on colocalization and MR analysis for their therapeutic potential, as proteins are usually the direct drug targets. Among the 125 proteins with *cis*-pQTL, 53 proteins showed effects for at least one disease or health-related phenotype with FDR <5%. Using the DrugBank and Drugs.com databases, we found 91 drugs targeting 17 of the 53 proteins, resulting in 443 drug–protein–phenotype combinations (Fig. 6 and Supplementary Table 14). Thirty-nine out of the 91 drugs were documented without clear clinical indications, which resulted in 235 drug–protein–phenotype combinations that we considered 'druggable', which might be used in improving the corresponding phenotypes (Supplementary Table 15).

After matching the MR outcomes with the clinical indications or side effects of the 91 − 39 = 52 drugs, we found 142 repurposing (that is, established drugs exist for the targets but treating different diseases) and 66 matched combinations (that is, established drugs exist for the targets treating the same outcome diseases or leading to the same outcome side effects). Within 66 matched pairs of protein targets and phenotypes, 52 had the same directions between the MR-inferred drug impact and the drugs' actual pharmacological effects on the phenotypes (Fig. 6), indicating that MR is four times more likely to identify a protein marker whose effect direction aligns with a drug's pharmacological action than to identify one with an inconsistent effect direction (one-tailed binomial test $P = 1.4 \times 10^{-6}$). Five out of the 66 combinations showed strong colocalization support (coloc PP.H4 > 0.8), and they all have consistent directions between the MR estimated effects and the corresponding drugs' pharmacological effects.

For instance, gemtuzumab ozogamicin is a monoclonal anti-CD33 antibody conjugated with a calicheamicin derivative that can induce cell death upon internalization through binding with CD33 on the cell surface. One of the side effects of gemtuzumab ozogamicin is thrombocytopenia, a condition characterized by a low platelet count. This is consistent with the MR result that decreased CD33 can decrease platelet count and platelet crit as the binding process of CD33 antibody and CD33 antigen should deplete the CD33 level[46].

Taking another example, clenbuterol was used as a bronchodilator in the treatment of patients with asthma. However, it can cause long- and short-term side effects, including hypertension. From MR analysis, the beta nerve growth factor (NGF) has a positive effect on blood pressure, while clenbuterol is a stimulator of NGF. Also, it has been shown that plasma beta NGF levels were higher in patients with hypertension[47], but their causal role was never established. Our MR finding provides evidence suggesting that NGF is actively involved in the blood pressure-increasing process.

Discussion

We identified pQTL for 139 of 184 neuro-related proteins, provided insights into their molecular mechanisms and effects on complex diseases and traits, and highlighted useful therapeutic targets with established drugs. On average, we identified half of the genetic architecture underlying the concentration of these proteins. We provide a well-powered genetic landscape for these proteins with large-scale summary-level data for future research.

Although the proteins were found to have small effects individually in the MR analysis, our results indicated that for about 75% of the

identified proteins, having low levels in plasma leads to a higher chance of having poorer health conditions (Supplementary Fig. 9). These conditions include both deterioration of mental health and related non-neurological comorbidities. Such results on the neuro-related proteins are consistent with the notion that psychiatric and neurological disorders are multifactorial and not limited to the central nervous system but rather are products of interactions among multiple systems within the organism[48–51]. The intertwining of neuropsychiatric, inflammatory and cardiovascular disorders has long presented a challenge in clinical research owing to the difficulties in discerning the relationships among them[52,53]. Our results suggest that these disorders may share molecular mechanisms and pathways and provide the basis for developing new diagnostic tools and treatment strategies. We also reported a large number of drug repurposing targets, suggesting the potential use of established drugs in new clinical trials for the treatment of different symptoms and disorders.

Regarding the MR methodology, we found that the MR analysis with a single genetic instrument at the *cis*-pQTL tended to generate a stronger estimated effect (Fig. 4). This is partly due to power, as compared with multi-instrument MR, single-instrument MR tends to produce effect estimates with larger standard errors so that only the results with large effect estimates could reach statistical significance. Thus, it indicates that (1) single genetic instrument analysis may be more prone to winner's curse, that is, more likely to detect an overestimated effect on the outcome trait, and (2) using multiple independent instruments within a locus may not only improve power but also control false discoveries owing to overestimated effects in the outcome GWAS.

MR assumes that the genetic effect on the outcome is mediated through the exposure. To justify the MR direct effect assumption and infer potential causality, we strictly used only independent variants at the *cis*-pQTL as genetic instruments, and *trans*-pQTL were never considered in the MR analysis. This is based on the fundamental biology that the variants near the coding gene of a protein are most likely to directly affect the protein-coding gene expression and less likely to have other indirect actions on other phenotypes. Variants within the *cis*-pQTL thus provide strong and most likely valid genetic instruments in MR. With the colocalization between the *cis*-pQTL and the outcome phenotype, stronger causal inferences can be made owing to the high genetic correlation between the exposure protein and the outcome trait. However, it should be noted that genetic variants may regulate multiple nearby genes, including those encoding proteins not captured on our assay platform, making it challenging to rule out local pleiotropic effects. For example, although we saw that NCAN was genetically correlated with fibrosis and cirrhosis of the liver (N cases = 252), established knowledge supports the nearby gene *TM6SF2* to be causal instead of *NCAN*[54].

While MR is a robust method for establishing potential causal links between exposures and outcomes[55,56], potential pitfalls should be noted, emphasizing the need for cautious interpretation of results. MR analyses typically may be limited by unobserved confounders, nonlinear protein–outcome relationships, reverse causation and population-specific effects[57–59]. In particular, when using MR as a procedure for drug target inference, with colocalization support, the analysis shows a strong genetic link between the protein targets and the corresponding complex diseases. However, the analysis does not suggest the actionability of the targets, nor their clinical effect if targeted by certain drugs or treatments.

Furthermore, it should be noted that both MR and colocalization are statistical approaches applied to summarized data from GWAS. While they share similarities, their objectives, implementations and interpretations are different. Colocalization between exposure and outcome phenotypes is crucial for causal inference using MR because it reinforces the validity of the genetic instruments used in MR. By confirming that the same genetic variants influence both the exposure and outcome, colocalization ensures that MR analyses are based on solid genetic grounds, reducing the risk of spurious or biased results. In fact, for shared loci between the exposure and outcome, colocalization is essential or even necessary to validate causality. Specifically, if there is a positive MR result at a genetic locus without colocalization of the exposure and outcome associations, it cannot be deemed causal. Colocalization serves as a safeguard against false-positive MR results stemming from the LD structure.

The improved causal inference specificity by the colocalization analysis can also be seen from the drug target investigation. MR revealed that glutathione *S*-transferase P (GSTP1) is negatively regulating mean corpuscular haemoglobin in the blood; however, both clomipramine and busulfan have side effects of anaemia, while they have different actions on GSTP1. A similar situation was also observed with platelet-derived growth factor receptor alpha (PDGFRA). Both olaratumab and imatinib could cause lymphopenia, while they are antibody and inhibitor of PDGFRA, respectively. These controversial results indicate that, although MR is more likely to reveal potential causal effects consistent with the drugs' action directions, limitations do exist in MR analyses, because of the great complexity of pharmacological and biological processes. Nevertheless, among the matched pairs of protein targets and traits, the five pairs with strong colocalization support all showed consistent MR effects and actual drug effect directions.

The mapped *trans*-pQTL were enriched in blood clotting and coagulation pathways. For instance, a blood clotting factor *KLKB1* appeared to be a *trans*-regulatory hub for multiple proteins. We thus infer that some of the *trans*-pQTL discovered are not directly involved in the genetic mechanisms of the corresponding proteins, but rather they regulate blood characteristics that affect the performance of the antibody-based assays.

Considerable attention must be paid to the effect of coagulation factors on protein quantification methods, especially in plasma-based assays. The enrichment of *trans*-pQTLs with coagulation factors and their established links to diverse neurological conditions emphasize the need for cautious interpretation[60–62]. Previous research has demonstrated the functional relationship between psychiatric and neurological conditions, structural brain features, immune response and coagulation[60,63–65], highlighting the importance of accounting for these factors in the analysis of blood-based protein quantification data.

Similar to the effect of clotting factors on the antibody-based assay, since glycosylation could potentially impact the binding of antibodies, it is likely to reveal the *trans*-pQTL effect of the glycosylation locus *ST3GAL4* or other glycosylation-related genes[66–68]. These are important discoveries for biotechnological development in proteomics, suggesting that the features of the plasma samples and protein structure modifiers could be non-negligible factors in circulating protein quantification.

The fundamental of pQTL studies, such as this particular large-scale GWAMA by the SCALLOP Consortium, is to map the genetic basis of protein abundance (see also the SCALLOP studies of the cardiovascular[15] and inflammatory[69] proteins). Although the biology of protein functions can be complicated, the genetic coding of each protein and the effects of genetic variants on each protein are generally consistent across the human body. A large proportion of the proteins measured in plasma are not primarily synthesized by blood cells. As a result, the pQTL (particularly *cis*-pQTL) that we identify in plasma are likely to be indicative of genetic loci within the tissues or cells responsible for producing these proteins. This, in turn, offers valuable insights into the underlying intracellular processes when we assess proteins in plasma. Despite the variation in pQTL observed across different tissues or cells, a substantial level of convergence is evident, especially when examining *cis*-pQTL[70]. This suggests that, even when protein levels in plasma, brain and cerebrospinal fluid do not exhibit strong correlations, there are instances where QTL are shared. Nevertheless, the effect size of the *cis*-pQTL could vary across tissues and cell types owing to complicated biological interactions. Current proteogenomics

still lacks tissue-specific pQTL studies, which ought to be addressed in future studies.

This study substantially advances our understanding of the genetics of neuro-related proteins and provides new targets for drug discovery. The pQTL discovery and causal inference with disease outcomes can inform clinical studies to identify actionable drug targets and enable integration into multi-omics analyses. The UKB-PPP and more cohorts could provide additional insights through larger meta-analyses and replication analyses, potentially revealing secondary signals in the pQTL. The inclusion of cohorts with diverse ancestries could further elucidate pQTL alleles that are not sufficiently polymorphic in European populations, identifying distinct molecular mechanisms underlying complex diseases.

Methods

Proteins

This study focused on plasma proteins from the Olink Neurology and Olink Neuro-Exploratory panels. Circulating plasma protein levels were quantified using Proximity Extension Assay technology, consisting of pairs of oligonucleotide-labelled antibodies to bind target proteins and hybridize to have their sequence extended and amplified through polymerase chain reaction (PCR). The level of amplified DNA is then quantified by microfluidic qPCR[29].

Proteins were selected by a panel of experts to include protein biomarkers that are known to be associated with neurological disorders and conditions through existing literature. The functions of these proteins comprise axonal development, metabolism, immune response and cell-to-cell communication. The proteins have been included in their respective panel on the basis of their observed involvement in neurological conditions and disorders, as well as the general performance of the assay.

Cohorts and data collection

We obtained summary statistics from the GWAS analyses performed on the Olink Neurology proteins from ten cohorts and the Olink Neuro-Exploratory proteins from six cohorts. Cohorts comprised population-based and case-control studies. The analysis plan that was circulated to the cohorts analysts is included in Supplementary Information, and the summary statistics information for each cohort can be found in Supplementary Tables 16–30. The total sample size for the Neurology panel meta-analysis was 12,176, whereas the Neuro-Exploratory panel meta-analysis included up to 5,013 individuals. The participating cohorts used whole-genome sequencing data or imputed data using the 1000 Genomes Project (phase 1 and phase 3) or the Haplotype Reference Consortium as reference panels. An average of 14.5 million SNPs were tested per protein, and the lowest per-SNP filter imputation quality ranged from 0.4 to 0.3, depending on the cohort. Each cohort carried out quality control according to their study design, as reported in Supplementary Table 16.

Data below the Olink limit of detection is calculated based on the negative controls included in each PCR run. Data below the limit of detection was available only for some cohorts participating in the meta-analysis. As the proteins were quantified at different times across cohorts, not all studies have data on all proteins in the two Olink panels.

Genome-wide association analysis of the proteins

The normalized protein expression (NPX) values, Olink's unit of protein abundance level on a log2 scale[29], were rank-based inverse normal transformed before running the per-protein GWAS analyses. Genotypic data were the allelic dosages resulting from imputation using the Haplotype Reference Consortium or the 1000 Genomes data as reference panels. Monomorphic SNPs were excluded. The genotype–phenotype association analysis was performed using regression models adjusting for sex, age, plate number, plate column, plate row, sample time in storage, season of sample collection, population structure (when appropriate) and other study-specific covariates. The analysis was done either by a linear regression model of the normalized protein abundance (NPX values) on the genotype data of each genetic variant, where the cohort-specific covariates were included, or by a linear mixed model, where the polygenic random effects were included to correct for population structure, besides the fixed effects covariates.

Meta-analysis

The summary association statistics from each participating cohort were uploaded through a secured FTP channel to the University of Edinburgh's ECDF Eddie Mark 3 cluster. The meta-analysis was run per protein in METAL (version 2018-08-28)[71] using the IVW method. We defined *cis*-pQTL to be 500 kb upstream or downstream of the gene coding for the respective protein and set the *trans*-pQTL window to be 1 Mb around the top variants that were found outside the defined *cis*-window. A 1% MAF filter was applied to the meta-analysis summary statistics for subsequent analyses. The variants that existed in only one participating cohort were also removed before subsequent analyses. The significance threshold was set to be 5×10^{-8} for the top variants of *cis*-regulatory variants and $5 \times 10^{-8}/184 = 2.73 \times 10^{-10}$ for the variants in *trans*-regions. The meta-analysed GWAS summary statistics for the 184 proteins are publicly available (see Data availability).

Heritability analysis

We used a standard polygenic mixed model implemented in GenABEL[35] on the individual-level data collected in the ORCADES cohort to assess the narrow-sense heritability for each protein. The heritability captured by each pQTL is calculated as $2f(1-f)\hat{\beta}^2$, where f and $\hat{\beta}$ are the coding allele frequency and estimated genetic effect, respectively, assuming Hardy–Weinberg equilibrium.

Established genetic associations

We used PhenoScanner v2[36,37] to cross-reference the lead (most significant) genetic variants in the *cis*-pQTL from our meta-analysis with other phenotypes. PhenoScanner is an extensive database of over 65 billion associations from publicly available GWAS. We used the lead variants of our *cis*-loci as input without the additional option of using proxy markers. When checking the novelty of our mapped *cis*-pQTL, we consider established pQTL associations with $P < 5 \times 10^{-6}$ as known. When extracting the established complex trait associations, we set the *P*-value threshold to 1 to include all possible associations. As all these established associations had reported *P* values, *P*-value adjustment procedures can be used to compute the corresponding FDR. We used the standard `p.adjust(method = 'fdr')` function in R to calculate the corresponding FDR values. Thereafter, results with a false discovery rate of less than 0.05 were considered. We excluded the studies with non-European ancestry.

Cross-referencing and replication in other pQTL studies

For the antibody-based assay, we cross-referenced the discovered *cis*-pQTL with results from the two Greek cohorts that we included in this study[38] and those reported by the UKB-PPP[39]. We checked whether a *cis*-pQTL was also reported as genome-wide significant ($P < 5 \times 10^{-8}$) for the same protein in either one of the two pQTL studies.

For each *trans*-pQTL in UKB-PPP, we checked whether the *trans*-pQTL was reported within a ±500 kb window of the lead variant of our discovered *trans*-pQTL. Also, for the aptamer-based assay, we compared the estimated *trans*-pQTL effects in our SCALLOP study and those in the Icelandic population where the proteome was measured using the SomaScan assay[31].

Out-of-sample prediction in the UKB

Taking the independent *cis*-pQTL and *trans*-pQTL variants for each protein, we calculated the SCALLOP in-sample proportion of phenotypic variance explained as $R^2_{\text{in}} = \hat{\boldsymbol{\beta}}'_{\text{SCALLOP}}\hat{\boldsymbol{\beta}}_{\text{SCALLOP}}$, and the

out-of-sample predictable proportion of variance in the UKB-PPP data was calculated as $R^2_{out} = (\hat{\boldsymbol{\beta}}'_{SCALLOP}\hat{\boldsymbol{\beta}}_{UKB-PPP})^2/\hat{\boldsymbol{\beta}}'_{SCALLOP}\hat{\boldsymbol{\beta}}_{SCALLOP}$, where each element j in the $\hat{\boldsymbol{\beta}}$ vectors was normalized as $\hat{\beta}_j = \hat{b}_j/\sqrt{\hat{b}_j^2 + N \cdot var(\hat{b}_j)}$, and $\hat{b}_j$ is the GWAS effect estimate for SNP j in the corresponding summary association statistics.

Functional enrichment and annotation of *trans*-pQTL

We performed our gene set enrichment analyses using the GENE2FUNC in FUMA v1.3.7[72,73], which returns functional annotation to gene models for the submitted list in a biological context. We identified the genes closest to the top SNPs in our *trans*-loci using the locuszoom v0.12[74,75] database and then submitted the list of genes to the FUMA website. We selected all types of gene to use as background for this analysis, including over 57,000 genetic elements. We set the maximum FDR-adjusted *P* value for gene set association to 1.

Colocalization analysis

We used the Bayesian colocalization analysis tool `coloc` with the posterior probabilities testing the H4 colocalization hypothesis for two models: (1) testing for a single shared causal variant between the pair of traits[40]; (2) testing for multiple shared causal variants, known as a SuSiE model[42]. The tests were applied to the mapped *cis*-pQTL and the established GWAS summary statistics, as well as to the *cis*-eQTL and the mapped pQTL. For the eQTL–pQTL colocalization analysis, we adopted the v7 release of both the GTEx eQTL and eQTLGen summary-level data. For each *cis*-pQTL, we tested colocalization with the *cis*-eQTL of the corresponding coding gene in each tissue. For each *trans*-pQTL, we tested colocalization with the *cis*-eQTL of the nearest coding gene.

MR analysis

For the protein–trait pairs with strong colocalization support (PP.H4 > 0.8), we performed a two-sample MR analysis using the IVW method to evaluate effects between the proteins with genome-wide significant *cis*-pQTL and (1) 4,085 traits from Neale's lab UKB GWAS and (2) 20 psychiatric or neurological disorder traits from PGC. As the GWAS of the binary traits by Neale's lab were conducted using ordinary linear regression, we transformed the estimated genetic effects from such an observed scale to the logistic scale (that is, the log of odds ratios). As the phenotypic variance explained by the genetic variant is a very small fraction, this can be done using the estimates from the linear regression, the prevalence of the cases and the allele frequency of each variant (see formula 3.2 derived by Pirinen et al.[76]). Multiple genome-wide significant sentinel variants of our *cis*-pQTL after LD pruning ($r^2 < 0.001$) were used jointly as instrumental variables. We report the significant discoveries at a level of 5% false discovery rate, for which we also performed a reverse generalized summary-statistics-based MR from the complex trait exposures to protein outcomes.

Drug target investigation

For the protein markers from IVW MR results with a false discovery rate of less than 5%, we systematically investigated available drugs targeting these markers using the DrugBank and Drugs.com databases. We considered a drug target validated if an MR discovery between the protein marker and the trait/disease suggested the same effect direction as the drug's effect on the protein target. The protein targets that have available drugs but are not directly related to the MR-discovered outcomes were regarded as repurposing targets. The remaining MR discoveries were reported as either new (no drug available) or druggable (drugs available without clear clinical indications) targets.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The full genome-wide summary association statistics for the 184 proteins are publicly available at https://doi.org/10.7488/ds/7522; *cis*-eQTL summary-level data by eQTLGen, https://eqtlgen.org/cis-eqtls.html; GTEx data, https://gtexportal.org/home/datasets; 1000 Genomes phase 3 genotype data, https://www.cog-genomics.org/plink/2.0/resources#phase3_1kg; Neale's lab UK Biobank round 2 GWAS summary-level data, http://www.nealelab.is/uk-biobank; Psychiatric Genomics Consortium (PGC) summary-level data, https://pgc.unc.edu/for-researchers/download-results/; DrugBank, https://www.drugbank.com; and Drugs.com, https://www.drugs.com. Source data are provided with this paper.

Code availability

Software used included METAL (https://genome.sph.umich.edu/wiki/METAL_Documentation), PLINK (https://www.cog-genomics.org/plink/), GenABEL (https://cran.r-project.org/src/contrib/Archive/GenABEL/), GCTA-GSMR (https://yanglab.westlake.edu.cn/software/gsmr/), PhenoScanner (http://www.phenoscanner.medschl.cam.ac.uk), MendelianRandomization (https://cran.r-project.org/web/packages/MendelianRandomization/index.html), coloc (https://chr1swallace.github.io/coloc/index.html), locuszoom (http://locuszoom.org/) and FUMA (https://fuma.ctglab.nl).

References

1. Danaei, G. et al. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLoS Med.* **6**, 1–23 (2009).
2. Wang, X. et al. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. *BMJ* https://www.bmj.com/content/349/bmj.g4490 (2014).
3. *Mental Disorders* (WHO, 2019); https://www.who.int/news-room/fact-sheets/detail/mental-disorders
4. *Mental Health* (Ritchie, H. & Roser, M., 2020); https://ourworldindata.org/mental-health
5. Hossain, M. M. et al. Epidemiology of mental health problems in COVID-19: a review. *F1000Research* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7549174/ (2020).
6. Greenberg, N. Mental health of health-care workers in the COVID-19 era. *Nat. Rev. Nephrol.* **16**, 425–426 (2020).
7. Jones, E. A., Mitra, A. K. & Bhuiyan, A. R. Impact of COVID-19 on mental health in adolescents: a systematic review. *Int. J. Environ. Res. Public Health* **18**, 2470 (2021).
8. Bearden, C. E., Reus, V. I. & Freimer, N. B. Why genetic investigation of psychiatric disorders is so difficult. *Curr. Opin. Genet. Dev.* **14**, 280–286 (2004).
9. Sullivan, P. F. & Geschwind, D. H. Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell* **177**, 162–183 (2019).
10. Taylor, M. J. et al. Association of genetic risk factors for psychiatric disorders and traits of these disorders in a Swedish population twin sample. *JAMA Psychiatry* **76**, 280–289 (2019).
11. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
12. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
13. Chames, P., Regenmortel, M. V., Weiss, E. & Baty, D. Therapeutic antibodies: successes, limitations and hopes for the future. *Br. J. Pharmacol.* **157**, 220–233 (2009).
14. Solomon, T. et al. Identification of common and rare genetic variation associated with plasma protein levels using whole exome sequencing and mass spectrometry. *Circ. Genom. Precis. Med.* **11**, e002170 (2018).

15. Folkersen, L. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).
16. Westwood, S. et al. Plasma protein biomarkers for the prediction of CSF amyloid and tau and [18F]-flutemetamol PET scan result. *Front. Aging Neurosci.* **10**, 409 (2018).
17. Dencker, M., Björgell, O. & Hlebowicz, J. Effect of food intake on 92 neurological biomarkers in plasma. *Brain Behav.* **7**, e00747 (2017).
18. Jabbari, E. et al. Proximity extension assay testing reveals novel diagnostic biomarkers of atypical parkinsonian syndromes. *J. Neurol. Neurosurg. Psychiatry* **90**, 768–773 (2019).
19. Hillary, R. F. et al. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. *Nat. Commun.* **10**, 3160 (2019).
20. Harris, S. E. et al. Neurology-related protein biomarkers are associated with cognitive ability and brain volume in older age. *Nat. Commun.* **11**, 800 (2020).
21. Rodrigues-Amorim, D. et al. Plasma β-III tubulin, neurofilament light chain and glial fibrillary acidic protein are associated with neurodegeneration and progression in schizophrenia. *Sci. Rep.* **10**, 1–10 (2020).
22. Sandberg, J. V. et al. Proteins associated with future suicide attempts in bipolar disorder: a large-scale biomarker discovery study. *Mol. Psychiatry* **27**, 3857–3863 (2022).
23. Folkersen, L. et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* **13**, e1006706 (2017).
24. Williams, S. A. et al. Plasma protein patterns as comprehensive indicators of health. *Nat. Med.* **25**, 1851–1857 (2019).
25. Lehallier, B. et al. Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).
26. Wingo, A. P. et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nat. Genet.* **53**, 143–146 (2021).
27. Jensen, S. B. et al. Discovery of novel plasma biomarkers for future incident venous thromboembolism by untargeted synchronous precursor selection mass spectrometry proteomics. *J. Thromb. Haemost.* **16**, 1763 (2018).
28. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
29. Assarsson, E. et al. Homogenous 96-plex pea immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS ONE* **9**, e95192 (2014).
30. Võsa, U. et al. Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
31. Ferkingstad, E. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).
32. Bulik-Sullivan, B. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
33. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
34. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across human complex traits. *Nat. Genet.* **52**, 859–864 (2020).
35. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
36. Staley, J. R. et al. PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
37. Kamat, M. A. et al. PhenoScanner v2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).
38. Png, G. et al. Mapping the serum proteome to neurological diseases using whole genome sequencing. *Nat. Commun.* **12**, 7042 (2021).
39. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
40. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
41. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
42. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).
43. Wightman, D. P. et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
44. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
45. Bretherick, A. D. et al. Linking protein to phenotype with Mendelian randomization detects 38 proteins with causal roles in human diseases and traits. *PLoS Genet.* **16**, e1008785 (2020).
46. Molica, M. et al. Cd33 expression and gentuzumab ozogamicin in acute myeloid leukemia: two sides of the same coin. *Cancers* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8268215/ (2021).
47. Tomoda, F., Nitta, A., Sugimori, H., Koike, T. & Kinugawa, K. Plasma and urinary levels of nerve growth factor are elevated in primary hypertension. *Int. J. Hypertens.* **2022**, 3003269 (2022).
48. Knardahl, S. Cardiovascular psychophysiology. *Ann. Med.* **32**, 329–335 (2000).
49. Ioannidis, K., Askelund, A. D., Kievit, R. A. & Harmelen, A. L. V. The complex neurobiology of resilient functioning after childhood maltreatment. *BMC Med.* **18**, 1–16 (2020).
50. McLaughlin, K. A., Colich, N. L., Rodman, A. M. & Weissman, D. G. Mechanisms linking childhood trauma exposure and psychopathology: a transdiagnostic model of risk and resilience. *BMC Med.* **18**, 1–11 (2020).
51. Fried, E. I. & Robinaugh, D. J. Systems all the way down: embracing complexity in mental health research. *BMC Med.* **18**, 1–4 (2020).
52. Fleshner, M., Frank, M. & Maier, S. F. Danger signals and inflammasomes: stress-evoked sterile inflammation in mood disorders. *Neuropsychopharmacology* **42**, 36–45 (2016).
53. Bauer, M. E. & Teixeira, A. L. Inflammation in psychiatric disorders: what comes first? *Ann. N. Y. Acad. Sci.* **1437**, 57–67 (2019).
54. Liu, Y.-L. et al. TM6SF2 rs58542926 influences hepatic fibrosis progression in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 4309 (2014).
55. Burgess, S., Foley, C. N., Allara, E., Staley, J. R. & Howson, J. M. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat. Commun.* **11**, 376 (2020).
56. Slob, E. A. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* **44**, 313–329 (2020).
57. Smith, G. D. & Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. *Int. J. Epidemiol.* **33**, 30–42 (2004).
58. Gill, D. et al. Mendelian randomization for studying the effects of perturbing drug targets. *Wellcome Open Res.* **6**, 24 (2021).
59. Sanderson, E. et al. Mendelian randomization. *Nat. Rev. Methods Primers* **2**, 1–21 (2022).
60. Heurich, M., Föcking, M., Mongan, D., Cagney, G. & Cotter, D. R. Dysregulation of complement and coagulation pathways: emerging mechanisms in the development of psychosis. *Mol. Psychiatry* **27**, 127–140 (2022).

61. Yang, Y. et al. Altered levels of acute phase proteins in the plasma of patients with schizophrenia. *Anal. Chem.* **78**, 3571–3576 (2006).
62. Levin, Y. et al. Global proteomic profiling reveals altered proteomic signature in schizophrenia serum. *Mol. Psychiatry* **15**, 1088–1100 (2010).
63. Baumeister, D., Akhtar, R., Ciufolini, S., Pariante, C. M. & Mondelli, V. Childhood trauma and adulthood inflammation: a meta-analysis of peripheral C-reactive protein, interleukin-6 and tumour necrosis factor-α. *Mol. Psychiatry* **21**, 642–649 (2016).
64. Varatharaj, A. & Galea, I. The blood–brain barrier in systemic inflammation. *Brain Behav. Immun.* **60**, 1–12 (2017).
65. Najjar, S. et al. Neurovascular unit dysfunction and blood–brain barrier hyperpermeability contribute to schizophrenia neurobiology: a theoretical integration of clinical and experimental evidence. *Front. Psychiatry* https://pubmed.ncbi.nlm.nih.gov/28588507/ (2017).
66. Sharapov, S. Z. et al. Defining the genetic control of human blood plasma N-glycome using genome-wide association study. *Hum. Mol. Genet.* **28**, 2062 (2019).
67. Sharapov, S. Z. et al. Replication of 15 loci involved in human plasma protein *N*-glycosylation in 4802 samples from four cohorts. *Glycobiology* **31**, 82 (2021).
68. Shen, X. et al. Multivariate discovery and replication of five novel loci associated with immunoglobulin G *N*-glycosylation. *Nat. Commun.* http://www.research.ed.ac.uk/portal/files/43181419/Multivariate_discovery_and_replication_of_five_novel_loci_associated_with_Immunoglobulin_G_N_glycosylation.pdf (2017).
69. Zhao, J. H. et al. Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets. *Nat. Immunol.* **24**, 1540–1551 (2023).
70. Yang, C. et al. Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. *Nat. Neurosci.* **24**, 1302–1312 (2021).
71. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190 (2010).
72. Watanabe, K., Taskesen, E., Bochoven, A. V. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
73. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. FUMA: functional mapping and annotation of genetic associations. *Eur. Neuropsychopharmacol.* **29**, S789–S790 (2019).
74. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336 (2010).
75. Boughton, A. P. et al. LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37**, 3017–3018 (2021).
76. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* **7**, 369–390 (2013).

Acknowledgements

X.S. was in receipt of a National Key Research and Development Program grant (numbers 2022YFF1202100 and 2022YFF1202105), a National Natural Science Foundation of China (NSFC) grant (number 12171495), a Natural Science Foundation of Guangdong Province grant (number 2021A1515010866) and Swedish Research Council (Vetenskapsrådet) grants (numbers 2017-02543 and 2022-01309). P.R.H.J.T. and J.F.W. acknowledge support from the Medical Research Council Human Genetics Unit Program grant 'Quantitative Traits in Health and Disease' (U. MC_UU_00007/10). The work of D.M., A.T., S.S. and Y.S.A. was supported by the Research Program at the Moscow State University (MSU) Institute for Artificial Intelligence. The work from X.F. was supported by the China Postdoctoral Science Foundation (number 2023M740690 and 2024T170174). The work from T.L. was supported by the China Postdoctoral Science Foundation (number 2023M740696). The work from C.K. and A.P.R. was supported in part by NIH grant R01-HL136574. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper. We thank the members of the SCALLOP Consortium of genome-wide association studies for making their data available. Cohort-specific acknowledgements are given in Supplementary Information.

Author contributions

X.S., P.N. and J.F.W. initiated and coordinated the study. L.R. performed the GWAS meta-analysis. J.C. conducted the colocalization analysis. Z.Y. conducted the Mendelian randomization analysis. R.Z. performed the drug target investigations. P.R.H.J.T., X.F., T.L., F.T., E.L.T., P.N. and X.S. contributed to the analysis pipeline. Y.Y. contributed to the cross-referencing prediction analysis. D.M. and A.T. contributed to the colocalization data processing and analysis. S.S. and Y.S.A. were involved in planning and supervising the work of D.M. and A.T. S.M.-W., M.D.M., B.P.P., A.J., R.F.H., E.W., S.K., S.A., L.P., Y.H., G.P., C.K., J.E.P., U.G., S.E.H., N.J.W., C. Lagging, M.A.I., A. Gilly, A. Göteson, M.K., E.T., J.H., A.P.R., G.D., E.Z., M.L., C.M.V.D., C.J., C. Langenberg, I.J.D., R.E.M., S.E., A.S.B. and A.M. contributed to the cohort-level analysis. L.R., J.C., Z.Y., R.Z., P.N. and X.S. wrote the paper. All authors approved the submitted version of the paper.

Competing interests

P.R.H.J.T. is a salaried employee of BioAge Labs, Inc. R.E.M. has received a speaker fee from Illumina, is an advisor to the Epigenetic Clock Development Foundation and is a scientific consultant for Optima Partners. E.W. is now an employee of AstraZeneca. Y.S.A. is now an employee of GSK. A.S.B. has received grants from AstraZeneca, Bayer, Biogen, BioMarin and Sanofi. A.M. is an employee of Pfizer. X.S. is the founder of Quantix BioSciences and has received a speaker fee from Olink Proteomics. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01963-z.

Correspondence and requests for materials should be addressed to Xia Shen.

Peer review information *Nature Human Behaviour* thanks Michael Johnson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

[1]Biostatistics Group, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. [2]Center for Intelligent Medicine Research, Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University, Guangzhou, China. [3]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK. [4]Health Data Science Centre, Fondazione Human Technopole, Milan, Italy. [5]State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai, China. [6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [7]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [8]MSU Institute for Artificial Intelligence, Lomonosov Moscow State University, Moscow, Russia. [9]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht and Utrecht University, Utrecht, Netherlands. [10]BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [11]Institute of Translational Genomics, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany. [12]Technical University of Munich (TUM), TUM School of Medicine and Health, Munich, Germany. [13]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA. [14]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [15]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, UK. [16]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge, UK. [17]Department of Epidemiology and Medical Statistics, Division of Oncology, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China. [18]Institute of Biomedicine, Department of Laboratory Medicine, the Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [19]Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands. [20]Department of Immunology and Inflammation, Faculty of Medicine, Imperial College London, London, UK. [21]Echinos Medical Centre, Echinos, Greece. [22]Anogia Medical Centre, Anogia, Greece. [23]Lothian Birth Cohorts, University of Edinburgh, Edinburgh, UK. [24]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, The Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden. [25]Department of Clinical Genetics and Genomics, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden. [26]Computational Medicine, Berlin Institute of Health (BIH) at Charité—Universitätsmedizin Berlin, Berlin, Germany. [27]Precision Healthcare University Research Institute, Queen Mary University of London, London, UK. [28]Division of Public Health Sciences, Fred Hutchinson Cancer Center and Department of Epidemiology, University of Washington, Seattle, WA, USA. [29]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens, Greece. [30]Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine and Health, Munich, Germany. [31]Biostatistics Unit—Population and Medical Genomics Programme, Genomics Research Centre, Fondazione Human Technopole, Milan, Italy. [32]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia. [33]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK. [34]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK. [35]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge, UK. [36]National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge, UK. [37]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, UK. [38]Emerging Science and Innovation, Pfizer Worldwide Research, Development and Medical, Cambridge, UK. [39]These authors contributed equally: Linda Repetto, Jiantao Chen, Zhijian Yang, Ranran Zhai, James F. Wilson, Pau Navarro, Xia Shen. ✉e-mail: shenx@fudan.edu.cn

nature portfolio

Corresponding author(s): Xia Shen

Last updated by author(s): May 13, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect the data.
Data analysis	Data analysis was perdormed using the following tools: R v4.2.0 (General statistics), METAL v2018-08-28 (meta-analysis), PLINK 1.90 v2021-05-28, GenABEL 1.8-0 (heritability analysis), GCTA-GSMR v1.93.2, PhenoScanner v2 (genetic variants cross-referencing), R package MendelianRandomization v0.9.0 (Mendelian Randomization analysis), coloc v5.1.0 (colocalisation analysis), FUMA v1.3.7 (gene set enrichment analysis), locuszoom v0.12 (identification of closest genes).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The full genome-wide summary association statistics for the 184 proteins are publicly available at https://doi.org/10.7488/ds/7522;
cis-eQTL summary-level data by eQTLGen: https://eqtlgen.org/cis-eqtls.html
GTEx data: https://gtexportal.org/home/datasets;
1000 Genomes phase 3 genotype data: https://www.cog-genomics.org/plink/2.0/resources#phase3_1kg;
Neale's lab UK Biobank round2 GWAS summary-level data: http://www.nealelab.is/uk-biobank;
Psychiatric Genomics Consortium (PGC) summary-level data: https://pgc.unc.edu/for-researchers/download-results/;
DrugBank: https://www.drugbank.com.
Drugs.com: https://www.drugs.com

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Data on participants' sex was collected and used as a covariate in the GWAS model for every study that contributed their data to the meta-analysis. Sex and gender-based analyses were not performed.
Population characteristics	Covariate-relevant population characteristics, as well as information relative to the protein levels can be found in Supplementary Tables 16 to 30. Participants to the meta-analysis were all of European Ancestry. All proteins were measured with Olink's Proximity Extension Assay (PEA) technology.
Recruitment	Participants were recruited differently across studies that took part to the meta-analysis, and is described in the original studies. See the following PMID: [ORCADES: 18760389] [INTERVAL: 25230735] [NSPHS: 29615742] [LBC1936: 31320639] [Fenland: 33328453] [SAHLSIS Gothenburg: 15933254, 36418382] [RS: 2427856] [STANLEY LAH1: 21926972] [STANLEY SWE6:35697758] [HELIC POMAK: 28548082] [HELIC MANOLIS: 27989266] [WHI: 9492970, 14575938]
Ethics oversight	The study protocol of every cohort was approved by different ethics boards and committee. 1) ORCADES. Ethical approval: approved by the Research Ethics Committees in Orkney, Aberdeen (North of Scotland REC), and South East Scotland REC, NHS Lothian approved the study (reference: 12/SS/0151). 2) INTERVAL. Ethical approval: approved by the UK National Research Ethics Service approved the study (reference 11/EE/0538). 3) HELIC. 4) WHI. This study was conducted within the Belmont Report–recognized ethical guidelines. Written informed consent was obtained from all participants, and this study was approved by each institution's institutional review board (IRB). 5) The NSPHS study. Ethical approval: the study was approved by the local ethics committee at the Uppsala University (Regionala Etikprovningsnamnden, Uppsala Dnr 2005:325). 6) LBC1936. Ethics permission for the Lothian Birth Cohort 1936 protocol was obtained from the Multi-Centre Research Ethics Committee for Scotland (Wave 1: MREC/01/0/56), the Lothian Research Ethics Committee (Wave 1: LREC/2003/2/29), and the Scotland A Research Ethics Committee (Waves 2-6: 07/MRE00/58). The research was carried out in compliance with the Helsinki Declaration. 7) FENLAND. Ethical approval: the study was approved by the Cambridge Local Research Ethics Committee. 8) STANLEY. IRB approvals and study consent forms from each of the sample contributing organizations were sent to the Broad Institute before samples were sequenced and analyzed. 9) SALHSIS_Gothenburg. This study, including the procedure for obtaining consent, was approved by the IRB, i.e. the local ethics committee of Lund University. All studies were approved by the local ethics committees of the University of Gothenburg or Lund University. 10) The Rotterdam Study. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC (registration number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and Sport (Population Screening Act WBO, license number 1071272-159521-PG). The Rotterdam Study has been entered into the Netherlands National Trial Register (NTR; www.trialregister.nl) and into the WHO International Clinical Trials Registry Platform (ICTRP; www.who.int/ictrp/network/primary/en/) under shared catalogue number NTR6831.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed for this study, as there is currently no established lower bound on the pQTL effect. The meta-analysis includes multiple participating cohorts, reaching a total sample size of over 10,000 individuals. For molecular QTL analysis, the sample size is sufficient for major QTL discoveries, especially given that 1) the PEA technology quantifies the proteins abundance with high specificity and 2) even the analysis in one participating cohort can yield a number of pQTL findings.
Data exclusions	Data points that did not pass the internal QC of each cohort were excluded. No data filtering was performed on the protein readouts using PEA, in order to maximize the available sample size in the QTL analysis.
Replication	The pilot phase data analysis was performed as a discovery-replication design, where the HELIC cohorts were used for replication of the pQTL discoveries. Based on a high replication rate, we finally performed an all-cohort meta-analysis, in order to boost discovery power. Given that each cohort conducted independent analyses, the meta-analysis inherently involves replication. The discovered QTL were then cross-referenced in the UK Biobank analysis.
Randomization	No group allocation was performed in this study.
Blinding	Investigators were not blinded to group allocation during data collection and/or analysis - this is not relevant in the case of GWAS and proteomics studies, as participants are not allocated to any group.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging