Mapping the serum proteome to neurological diseases using whole genome sequencing

Grace Png, Andrei Barysenka, Linda Repetto, Pau Navarro, Xia Shen, Maik Pietzner, Eleanor Wheeler, Nicholas J. Wareham, Claudia Langenberg, Emmanouil Tsafantakis, Maria Karaleftheri, George Dedoussis, Anders Målarstig, James F. Wilson, Arthur Gilly & Eleftheria Zeggini

Despite the increasing global burden of neurological disorders, there is a lack of effective diagnostic and therapeutic biomarkers. Proteins are often dysregulated in disease and have a strong genetic component. Here, we carry out a protein quantitative trait locus analysis of 184 neurologically-relevant proteins, using whole genome sequencing data from two isolated population-based cohorts (N = 2893). In doing so, we elucidate the genetic landscape of the circulating proteome and its connection to neurological disorders. We detect 214 independently-associated variants for 107 proteins, the majority of which (76%) are cis-acting, including 114 variants that have not been previously identified. Using two-sample Mendelian randomisation, we identify causal associations between serum CD33 and Alzheimer's disease, GPNMB and Parkinson's disease, and MSR1 and schizophrenia, describing their clinical potential and highlighting drug repurposing opportunities.
Neurological disorders are the leading cause of disability worldwide, accounting for 276 million disability-adjusted life years (DALY) globally in 2016. This burden is continuously increasing with growing and ageing populations, emphasising the need for better prevention and treatment strategies. Multiple genetics and genomics efforts have established that these diseases have a substantial genetic component. Elucidating their genetic architecture can, therefore, help to forward our understanding of their aetiology by identifying causal disease mechanisms, thus opening a path towards clinical translation.

Due to their heterogeneity and overlapping clinical features, neuropsychiatric disorders such as schizophrenia and bipolar disorder are often misdiagnosed, while others with more distinct symptoms, such as Alzheimer’s disease (AD), lack effective drugs and accessible biomarkers that can detect early disease. The human serum proteome is an especially valuable resource of potential biomarkers for these highly polygenic disorders. As proteins are often dysregulated in disease, studying protein quantitative trait loci (pQTLs), which are genetic variants associated with protein expression levels, can help to bridge existing knowledge gaps. Most pharmaceutical drugs also target proteins, further increasing their actionability.

By implementing statistical methods that leverage relevant biomedical data, such as causal inference and colocalisation analysis, pQTLs can be used to determine causality and to identify disease pathways. For example, in a study focused on neurologically relevant proteins, a pQTL for serum PVR mapping to the PVR gene (cis-pQTL), was found to be causally associated with AD through Mendelian randomisation analysis. Through similar methods, a recent brain proteome-wide association (PWAS) and pQTL study identified five genes causal for AD at high confidence, of which four were novel. By validating known AD loci and identifying new causal genes, these studies demonstrate proof-of-concept.

Here, we aimed to identify biomarkers of neurological traits and enhance insight into disease pathways, by carrying out a pQTL analysis of 184 neurologically relevant serum proteins. The main advantage of serum proteins is that they are easily accessible, both as drug targets and diagnostic biomarkers. We use whole-genome sequencing (WGS) to capture the entire allele frequency spectrum in 2,893 samples from two Greek population-based cohorts, MANOLIS and Pomak. Association analysis was first carried out individually for each cohort, followed by a meta-analysis. Specifically, proteins were quantified using Olink’s proximity extension assay (PEA) and comprised established or potential markers of neurobiological processes. Using WGS, we were able to detect both rare and common pQTL variants. We then investigated the relevance of the discovered pQTLs to neurologically diseases and highlight biomarkers of high diagnostic or prognostic potential, identify drug repositioning opportunities, and describe pathways relevant to neurological traits.

Results

Protein QTL discovery. For the 184 neurologically relevant proteins analysed, we detect 214 independently-associated pQTLs \((P < 1.05 \times 10^{-10}; ‘Methods’ section) for 107 proteins from the meta-analysis, following conditional testing (Fig. 1 and Supplementary Data 1). Loci were classified into cis and trans: cis-acting pQTLs, which are defined as variants residing within 1 Mb upstream or downstream of the protein-encoding gene, are likely to regulate protein expression directly at the transcriptional level, while trans-pQTLs are likely to act through intermediaries to modulate protein levels. We observe 162 (75.7%) cis-acting pQTLs for 91 proteins, and 52 (24.3%) trans-acting pQTLs for 38 proteins. A total of 22 proteins had both cis and trans-acting pQTLs (Fig. 2b).

Sixteen proteins have only trans-pQTLs, 13 of which have pQTLs only in pleiotropic loci. We find altogether 30 variants arising at known pleiotropic loci, including those near or within \(KLLB1, ABO, F12, VTN\), and the HLA region on chromosome 6. These are loci that influence the levels of multiple proteins; the most pleiotropic being loci at \(KLLB1\) and \(ABO\), affecting 11 and 12 proteins, respectively. These have been identified in published pQTL studies and are not restricted to neurologically relevant proteins. \(ABO\) is the most extensively studied among these pleiotropic loci, and is known for its role in blood coagulation processes and determining the ABO blood types. In particular, we detect the missense variant rs8176747 affecting ADAM15, IL3RA, and KIRREL2 protein levels. rs8176747 is among the variants routinely used to determine blood group phenotype, which has been associated with multiple diseases, mainly of cardiovascular relevance. As proteins such as \(ABO\) are connected to large signalling networks, changes in their structure or expression levels could influence multiple downstream substrates, hence explaining their pleiotropy.

We identify 33 sequence variant-protein level independent associations for 15 proteins that have not been investigated for pQTLs before (Table 1). For the remaining 92 proteins, we identify 72 novel cis-pQTL variants, and 15 novel trans-pQTL variants, excluding those at known pleiotropic loci. We define novelty if no variants within 2 Mb have been previously reported in serum pQTL studies, or if associations remain significant after conditioning on established pQTLs.

Eight of the proteins we studied here have also been investigated in a pQTL study in cerebrospinal fluid (CSF)\(^{-14}\). We replicate six of these cis-pQTLs in serum: for CD33, GPNMB, LEPR, NAAA, SIGLEC-9, and TDGFI. Additionally, we find novel cis-pQTLs for CD33 and GPNMB, and trans-pQTLs for NAAA and SIGLEC-9, which had not been detected in CSF. The observed replication of CSF pQTLs indicates that the expression of these proteins in serum and CSF are governed by a shared genetic mechanism.

Of the identified independent pQTLs, 185 (86%) are common-frequency variants (minor allele frequency [MAF] > 5%), 25 (12%) are low-frequency (MAF 1–5%) and four (2%) are rare (MAF < 1%) (Fig. 2a). Eight of the low-frequency or rare pQTLs (all cis signals) have not been reported before, despite the proteins having been analysed in past studies, demonstrating the advantage of using whole-genome sequencing-based analysis to capture the full MAF spectrum.

Gene expression QTL colocalisation. Colocalisation analysis is used to test if independent association signals from two traits share the same causal variant. When comparing protein with gene expression levels, positive colocalisation is indicative of a shared regulatory mechanism, thereby acting as orthogonal validation. Through testing for colocalisation of neurological pQTLs with gene expression QTLs (eQTLs) from multiple tissues (GTEx), our results also identify disease-relevant tissues where gene expression correlates with serum protein expression. For cis-acting pQTLs, analysis was carried out between protein expression and the expression of the encoding gene, in all available tissues. Sixty-four (69%) cis-pQTLs colocalised strongly (colocalisation posterior probability 4 [CLPP4] > 0.8, ‘Methods’ section) with gene expression in at least one tissue, with 11 (12%) in whole blood, and 21 (23%) in various parts of the brain (Supplementary Data 4). This indicates that for these loci, the causal variant influences both gene and protein expression, therefore supporting transcriptional regulation as the mechanism underpinning variation in protein expression levels.

For trans-pQTLs, positive colocalisation between a pQTL and an eQTL at a distal gene increases the likelihood that the two gene
products map to the same regulatory pathway (Supplementary Note 1 and Supplementary Fig. 2). Colocalisation analysis was performed between protein traits and expression of genes within 2 Mb of the trans-acting variant. We detect 36 (75%) signals that colocalise with the expression of at least one gene in their vicinity, with three (6%) in whole blood and 30 (62%) in the brain (Supplementary Data 4). As proof-of-concept, we find known receptor-ligand pairs such as a trans signal for the KIR2DL3 (killer cell immunoglobulin-like receptor 2DL3) protein colocalising with the expression of HLA-C in multiple tissues (22 tissues; CLPP4 > 0.78). KIR2DL3 is an inhibitory receptor for HLA-C, and is responsible for preventing natural killer cells from killing healthy cells15. The analysis also enabled the identification of new protein links. For example, we observe a trans-pQTL for SMPD1 (sphingomyelin phosphodiesterase; rs10745925; MAF = 0.333; \( P = 7.75 \times 10^{-23} \); BETA = −0.2805; SE = 0.0285) that colocalises strongly with the expression of GNPTAB in the liver (CLPP4: 0.89), and moderately in

Fig. 1 pQTL signals for 107 serum proteins from Olink neurology and neuro-exploratory panels. a 3D Manhattan plot of detected pQTLs. The x axis represents each of the 107 proteins; the y axis represents the chromosome location of each signal; and the z axis represents the \(-\log_{10} p\)-values of each association signal. b Scatterplot of pQTL variant location against the location of the gene encoding the target protein. Each dot represents an independent variant. Cis-pQTLs are coloured in teal, while trans-pQTLs are in orange.

Fig. 2 Overall genetic architecture of 107 serum proteins of neurological relevance. a A total of 214 independent variants were detected. Cis-acting variants were defined as variants lying within 1 Mb upstream and downstream of the gene encoding the target protein, while trans-acting variants are variants that lie outside of this region. Most severe consequence was determined by Ensembl’s variant effect predictor (VEP). Effects more than missense included ‘stop_gained’, ‘frameshift_variant’, and ‘splice_acceptor_variant’ in our dataset; ‘Regulatory region’ variants include ‘[3/5]’primeUTR_variant’, ‘TF_binding_site_variant’, ‘splice_region_variant’, and ‘regulatory_region_variant’; while ‘Others’ comprises mostly intergenic and intronic variants. Novelty was assessed by cross-referencing published summary statistics from other pQTL studies (Supplementary Data 2). Known pleiotropic loci were not considered novel. Rare, low-frequency and common variants were defined as variants with minor allele frequency (MAF) < 1%, MAF 1–5%, and MAF > 5%, respectively. b Number of proteins for which we detected only cis-pQTLs, trans-pQTLs, or both.
other tissues (CLPP4 = 0.58 [oesophagus mucosa]; 0.57 [stomach]; 0.54 [adrenal gland]). SMPD1 is a lipid hydrolase involved in multiple cell processes; whereas GNPTAB encodes subunits of GlnAC-1-phosphotransferase, which is involved in the synthesis of mannose-6-phosphate (M6P). SMPD1 exists in two forms: secreted and lysosomal. Its lysosomal form is transported via the M6P receptor pathway, therefore supporting the observed SMPD1-GNPTAB interaction. Moreover, we find that the minor allele is associated with a decrease in circulating SMPD1 and an increase in GNPTAB expression. This could be a result of increased M6P associated with a decrease in circulating SMPD1 and an increase in lysosomal SMPD1, and likely play a distinct role in the body and abnormal levels of the secreted form have been implicated in age-related neurodegenerative conditions including Alzheimer’s disease and amyotrophic lateral sclerosis (ALS). We, therefore, identify a locus at GNPTAB that coregulates secreted SMPD1 levels and GNPTAB expression, pinpointing a possible mechanism behind SMPD1-related neurodegenerative disorders.

**Heritability.** To estimate the narrow-sense heritability of the protein traits studied, the proportion of variance explained (PVE) by all variants across the genome was calculated using GCTA-GREML for each protein. Using a single-component approach, WGS variants explained a median of 33.3% of variance in serum protein levels, with the highest observed heritability observed for CD33 (h² = 87.2%). Another three proteins had high heritability of more than 80%: TDGF1 (85.4%), VSTM1 (82.8%), and LAIR2 (82.3%). Conversely, some proteins had very low heritability estimates of h² < 5%: IKZF2 (4.9%), RNF31 (4.4%), and EPHA10 (0.001%).

We observe that for all four proteins with h² > 80%, the pQTLs colocalised with gene expression QTLs in multiple tissues, indicating regulation at the transcriptional level; therefore, the high observed h² values are likely to mirror genuine high heritability. There are, however, other non-mutually exclusive reasons that can drive very high or low estimates: (1) Variants that alter the binding specificity of the Olink antibody but not the quantity of protein may produce inaccurate heritability estimates; and (2) Known and unknown biases of single-component GREML approach, which tends to overestimate h² when causal variants are common, and underestimate h² when causal variants are rare. We explore the biological relevance of the pQTLs, we carried out colocalisation analysis with neuropsychiatric traits using data published by the Psychiatric Genomics Consortium (PGC), as well as other neurodegenerative traits, using publicly available summary statistics from recent large GWAS meta-analyses (Supplementary Data 5b). We also studied colocalisation with signals for pain-related traits that have been proven to have a neuropathic component, such as chronic back pain and osteoarthritis. A total of 15 protein–trait pairs colocalised with human disease signals, suggesting a role for the protein in mediating disease. These results are summarised in Supplementary Data 5a.
We applied two-sample Mendelian randomisation (MR) for the 107 proteins for which we detect pQTLs, and 206 neurologically relevant and behavioural traits. In contrast to colocalisation, the objective of MR is to look for causal effects of proteins on neurological phenotypes. Using both cis- and trans-acting pQTLs, fifteen proteins were found to be causal for at least one trait, and we detect significant causal effects for 25 unique protein–trait pairs (Fig. 3 and Supplementary Data 6a).

We replicate multiple known associations between protein and disease from the colocalisation and MR analyses. These include LEPR (leptin receptor) and migraine24, LTBP3 (latent-transforming growth factor beta-binding protein 3) and osteoarthritis25, FLRT2 (leucine-rich repeat transmembrane protein) with bipolar disorder26, and PLXNB127 (plexin-B1) and PLA2G1028 (group 10 secretory phospholipase A2) with schizophrenia.

The analysis also identified new protein–disease relationships. Notably, the strongest causal association was found between serum WFIKKN1 and schizophrenia (P_adj = 9.12 × 10^{-43}; WFIKKN1 (WAP, Kazal, immunoglobulin, Kunitz and NTR domain-containing protein 1) has not been associated with any neuropsychiatric disorder to date, but is highly expressed in the brain (GTEx) and regulates the activity of several growth and differentiation factors29. Similarly, we find new evidence that serum VSTM1 is causally associated with sleep apnoea (P_adj = 2.03 × 10^{-2}). VSTM1 (V-set and transmembrane domain-containing protein 1) is a cytokine that promotes the differentiation of helper T-cells (TH17), which are often implicated in autoimmune disorders that may develop secondary to sleep apnoea30,31.

The overarching aim of this study was to identify protein biomarkers that may be used in the prognosis, diagnosis, or treatment of neurological diseases. Here, we highlight various potential disease markers that are supported by multiple lines of evidence.

**GPNMB as a biomarker for Parkinson’s disease.** We identified a cis-pQTL that is associated with decreased levels of serum GPNMB (transmembrane glycoprotein NMB; rs7797870; MAF = 0.4286; P = 7.01 × 10^{-50}; BETA = −0.2109; SE = 0.0247) and colocalises with a known Parkinson’s disease (PD) locus32 (CLPP4 = 0.86) (Fig. 4b). GPNMB has been highlighted as a susceptibility gene in large PD meta-analyses32 and has been proven to be upregulated in the brains of PD patients and in mice with induced lysosomal dysfunction33. In addition to its connection to PD, we present new evidence showing that serum GPNMB shares a causal variant with GPNMB gene expression in both whole blood (CLPP4 = 0.79) and brain tissue (basal ganglia CLPP4 = 0.70; cortex CLPP4 = 0.74; anterior cingulate cortex CLPP4 = 0.83). This not only implies that GPNMB expression is regulated transcriptionally by the pQTL, but also that its expression in the blood and brain are mediated via a shared mechanism. This is supported by previous research showing that tissue GPNMB is able to shed its ectodomain and enter circulation34. The lead variant rs75801644 explained 7% of variance in antibody binding for serum GPNMB. Importantly, the identification of serum GPNMB levels as a potential marker of PD is significant as current diagnostic biomarkers are mostly found in the CSF. As serum biomarkers are much less invasive to...

---

**Fig. 3 Causal protein-disease associations identified using two-sample Mendelian randomisation.** We investigated the causal effect of serum proteins (exposure) on various neurological traits (outcome), indicated in the first two columns in the plot. PubMed IDs (PMIDs) are given where manually downloaded summary statistics were used; other IDs are those as given in MRBase (https://gwas.mrcieu.ac.uk/). The number of variants used in the analysis are given in the ‘nSNP’ column. The ‘pBH’ column contains the FDR-adjusted (Benjamini–Hochberg) P-value for each test. Protein–trait pairs with only one variant were analysed using the Wald ratio method, while those with more than one variant were analysed using the inverse variance-weighted (IVW) method. Data are represented as mean odds ratio ± SEM. *Additional signal arising from analysis using only cis-pQTLs as instrumental variables.

| Panel/Protein | Trait | nSNP | pBH | Test | Odds Ratio [95% CI] |
|---------------|-------|------|-----|------|-------------------|
| NEU CD200     | Parkinson’s disease | || | 1.30e-05 | [1.00, 1.10] |
| NEU CD16      | Spine osteoarthritis | GO consortium, 2023 | 2 | 2.03e-02 | [1.10, 1.17] |
| NEU CTSS      | Bipolar disorder | [ID:rs-808] multiple sclerosis | 1 | 1.42e-08 | [0.86, 0.97] |
| NEU FCRL2     | Rheumatoid arthritis | [ID:rs-833] | 1 | 5.09e-04 | [0.81, 0.91] |
| NEU IL19      | Ankylosing spondylitis | GO consortium, 2024 | 2 | 2.11e-02 | [1.09, 1.14] |
| NEU KNYU      | Bipolar disorder | [ID:rs-808] | 1 | 1.22e-03 | [0.85, 0.97] |
| NEU MRK1      | Schizophrenia | [ID:rs-833] | 1 | 1.53e-02 | [0.80, 0.92] |
| NEU WFIKKN1   | Schizophrenia | [ID:rs-810] | 1 | 4.86e-04 | [0.82, 0.90] |
| NFX CCLI7     | Feeling hurt | [ID:rs-808] GO consortium | 2 | 9.75e-03 | [0.97, 0.99] |
| NFX CD33      | Alzheimer’s disease | [ID:rs-297] | 2 | 4.66e-02 | [0.96, 0.98] |
| NFX DPEP1     | Non-cancer illness, self-reported osteoarthritis | [ID:rs-1448] | 2 | 2.82e-04 | [1.00, 1.00] |
| NFX FUT8      | Sleep disorder (insomniacs) | [ID:rs-808] | 2 | 2.11e-02 | [0.99, 1.00] |
| NFX KLB       | Alcohol consumption | [ID:rs-1283] | 2 | 1.63e-02 | [1.01, 1.02] |
| NFX LTRP1     | Diagnoses – main IC:10.42 Diagnose and guidance | [ID:rs-5655] | 2 | 8.58e-03 | [1.00, 1.00] |
| NFX VSTM1     | All osteoarthritis | GO consortium, 2022 | 2 | 1.22e-02 | [1.00, 1.00] |
|               | All osteoarthritis | GO consortium, 2022 | 2 | 1.22e-02 | [1.00, 1.00] |

Odds ratio per standard deviation increase in exposure

| 0.7 | 0.8 | 0.9 | 1 | 1.1 | 1.2 |

---

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-27387-1
CD33 as a biomarker for Alzheimer’s disease. Using two-sample MR, we confirm a significant causal association between serum CD33 (myeloid cell surface antigen CD33) and Alzheimer’s disease\(^{35}\) (AD; BETA = 0.0091; SE = 0.0017; inverse variance-weighted [IVW] \(P_{\text{adj}} = 3.62 \times 10^{-4}\)) (Figs. 3 and 4a). The role of CD33 in AD is further affirmed by positive colocalisation between the cis-pQTL with the causal variant rs2455069 (MAF = 0.3967; \(P = 2.03 \times 10^{-1580}\); BETA = 1.2092; SE = 0.0142), and a known AD-associated locus (CLPP4 = 0.82). CD33 is upregulated in the AD brain and is positively correlated with disease severity, while knockout mice have been shown to have reduced amyloid plaque formation\(^{36}\). Additionally, the cis-pQTL for CD33 colocalises with an eQTL for the CD33 gene in whole blood (CLPP4 = 0.95) and in the brain (cerebellar hemisphere CLPP4 = 0.62), indicating a shared regulatory pathway for gene and protein expression.

Notably, our heritability analysis revealed a very high \(h^2\) value (82.7%) for serum CD33, which is the highest proportion of variance explained observed across all analysed traits, thus reflecting high heritability. This has been similarly observed in a study showing that the most strongly AD-associated variant in CD33, rs3865444, explained more than 70% of variance in CD33 monocyte expression and was moreover unaffected by age\(^{37}\). A reverse Mendelian randomisation analysis (using AD as the exposure and serum CD33 as the outcome) confirmed that AD is causal for increased CD33. Together, these findings indicate that serum CD33 levels are a promising diagnostic marker for early AD (Supplementary Note 2).

MSR1 on the causal pathway to schizophrenia. We find that a cis-pQTL (rs150158578) associated with decreased serum MSR1 (macrophage scavenger receptor types I and II) is causal for schizophrenia, supported by evidence from colocalisation analysis (CLPP4 = 0.75) and two-sample MR (BETA = −0.2205; SE = 0.0522; Wald ratio \(P_{\text{adj}} = 1.44 \times 10^{-2}\); Fig. 3). Variants in the MSR1-encoding gene have been nominally significantly associated in a schizophrenia GWAS\(^{38}\), and have been robustly associated with AD\(^{39}\) and PD\(^{38}\).

MSR1 is an immune modulator expressed on the cell surface of macrophages. The protein plays a critical role in the clearance of infectious agents and toxic molecules, such as amyloid-beta protein\(^{40}\), damage-associated molecular patterns (DAMPs)\(^{41}\), and modified lipids, such as oxidised low-density lipoprotein (oxLDL)\(^{42}\). MSR1-mediated phagocytosis activates both pro- and anti-inflammatory responses, and has been shown to have a protective effect against multiple diseases, including bacterial and viral infections, AD, atherosclerosis and Barrett’s oesophagus (BE)\(^{43}\). Accordingly, MSR1-deficient mice have been shown to exhibit dysregulated immune response in the brain and deteriorating working memory\(^{44}\). MSR1 activation can also lead to excessive inflammation linked to sepsis and worsening the cardiac and cerebral injury. Here, we observe a causal association between decreased MSR1 expression and increased risk of schizophrenia, suggesting a protective role (Fig. 5b).

We also find colocalisation of the cis-pQTL for serum MSR1 with an eQTL for the MSR1 gene in the nucleus accumbens of the basal ganglia (CLPP4 = 0.90), aorta (CLPP4 = 0.94), tibial artery (CLPP4 = 0.90), and oesophagus (CLPP4 = 0.91) (Fig. 5c). The nucleus accumbens is central to the brain’s reward system, and is enriched in dopaminergic neurons that contribute to the pathophysiology of schizophrenia\(^{45,46}\) and other neuropsychiatric diseases\(^{47,48}\). A large comorbidity study has shown that patients with schizophrenia are more likely to suffer from coronary heart disease, cerebrovascular disease, and congestive heart failure\(^{49}\). We observed no evidence of colocalisation or causality between serum MSR1 and stroke or coronary artery disease (CAD).

To further investigate the mechanistic through which the pQTL regulates protein expression, we queried the ENCODE\(^{50}\) (https://www.encodeproject.org/) database for overlaps with cis regulatory elements. We found that, in blood cells, three variants in LD \((r^2 > 0.8)\) with rs15015857 (rs420931, rs433235, and rs59251421) reside within regulatory elements with a proximal enhancer-like signature (EH38E2612563), a promoter-like signature (EH38E2612567), and a distal enhancer-like signature (EH38E2612573), respectively. All three variants, as well as rs150158578, are also eQTLs for MSR1 gene expression in whole blood (GTEx). This suggests that the pQTL regulates MSR1 in
blood cells at the transcriptional level, possibly by altering the binding affinity of transcription factors to the promoter or an enhancer. Additionally, we note two other cis-acting rare, independent variants (rs182190568, MAF = 0.006, $P = 1.44 \times 10^{-21}$, BETA = -1.2568, SE = 0.1317; rs41341748, MAF = 0.0148, $P = 3.18 \times 10^{-38}$, BETA = -1.3351, SE = 0.1033), and a rare deletion (chr8: 16090064-16150000[bp]; MAF = 0.006; $P = 7.10 \times 10^{-23}$; BETA = -1.414, SE = 0.1436) that are significantly associated with serum MSR1 levels (Fig. 5a), illustrating the complexity of the genetic regulation of MSR1.

**Drug target evaluation.** Drug repositioning can dramatically expedite translational applications of proteomics and genomics into patient benefit. As over 95% of drugs target proteins, we sought to identify proteins included in this study that are targets of drugs that have been approved, or are in later stages of clinical trials (see ‘Methods’ section). Twenty-three of the proteins we studied in this work are targets of approved drugs. Of these, 17 proteins had pQTL signals (Supplementary Data 7). Seven of these proteins have cis-acting pQTLs that colocalise with or are causal for neurological diseases: DDR1, IL12, NEP,

---

**Fig. 5 Serum MSR1 is causally associated with schizophrenia.**

- **a** Genetic architecture of serum MSR1. Each of the three independent variants and their LD variants are represented in orange, teal, and purple, respectively; the intensity of the colours indicates the strength of linkage disequilibrium ($r^2$). A rare deletion is also indicated in purple, and is in complete LD with the independent variant rs182190568. Below the signal plot, the location of the variants respective to the gene are indicated using coloured points for the SNVs and a dotted box for the deletion.

- **b** Proposed mechanism of how decreased MSR1 may lead to neuronal damage, resulting in neuropsychiatric disease.

- **c** Association signal plots at the MSR1 locus for (clockwise) serum MSR1, schizophrenia, gene expression of MSR1 in nucleus accumbens tissue, aorta, oesophagus muscularis and tibial artery. The lead pQTL variant is denoted by a black diamond, while variants in LD are coloured according to the strength of LD with the lead variant (red [$r^2 > 0.8$]; orange [$0.5 < r^2 > 0.8$]; blue [$0.2 < r^2 > 0.5$]; grey [$r^2 < 0.2$]).
CD33, DPEP1, GPNMB, and LEPR (Supplementary Note 3). Of note is DPEP1 (dipeptidase 1), whose increased expression is causal for osteoarthritis and multisite chronic pain (MCP) (Fig. 3). DPEP1 is inhibited by the drug cilastatin, which is often used in combination with the antibiotic imipenem as an embolic agent in the treatment of serious infections. Given that DPEP1 is causally associated with osteoarthritis, cilastatin could potentially be repurposed to treat osteoarthritis. Indeed, the cilastatin/imipenem combination has been investigated as a treatment for knee osteoarthritis52,53, and has been proven to provide pain relief. Also notable is CD33, whose expression is increased in AD (Figs. 3 and 4a). CD33 has proven to be a safe target, demonstrated by the acute myeloid leukaemia (AML) drugs, gemtuzumab ozogamicin and lintuzumab. In a study investigating the repurposing of lintuzumab for reducing AD risk, the anti-CD33 drug was shown to robustly decrease cell surface expression of the protein54. We, therefore, provide further evidence genetic supporting repositioning of lintuzumab for AD treatment.

Discussion

Biomarker discovery is a process central to precision medicine, and is especially important for many neurological disorders that remain challenging to diagnose and treat. Serum proteins make ideal intermediate traits to study as they are druggable, measurable targets that are strongly linked to both causative genetic variants and medical outcomes. Having knowledge of their underlying genetic architecture and how that may correlate with diseases can also enhance our understanding of disease aetiology. We have carried out a pQTL analysis of 184 neurologically relevant serum proteins using WGS data. Altogether, we find 214 pQTLs for 107 proteins, of which 33 were for proteins that are being analysed for the first time. We detect novel pQTLs for previously studied proteins and replicate established associations of both blood and CSF pQTLs. Through downstream analysis, we highlight disease-relevant, translatable pQTLs by presenting new evidence supporting protein-disease associations; most notably, CD33 and Alzheimer’s disease, GPNMB and Parkinson’s disease, and MS1 and schizophrenia. Additionally, we observed that serum DPEP1 is causal for both osteoarthritis and multisite chronic pain (MCP). Pain is the main symptom of osteoarthritis, and osteoarthritis is the leading cause of pain and disability worldwide55. DPEP1 has been implicated in osteoarthritis through a large genome-wide association study56, and was additionally shown to be downregulated in mouse models of osteoarthritis57. These findings indicate that serum DPEP1 may serve as a valuable candidate biomarker for identifying patients with undiagnosed osteoarthritis and suffer from MCP.

Consistent with previously published pQTL studies, the majority (73.8%) of variants were intergenic and intronic; we also observed 31 (14.4%) variants in regulatory regions and 25 (11.6%) coding variants, either missense or with more severe consequences. We note a limitation of epitope-based proteomic assays, in that cis-acting protein structure-altering variants may affect epitope-binding affinity and, in turn, measured protein levels. We identified 21 proteins with cis-acting variants, which were either highly correlated ($r^2 > 0.8$) with, or were missense or more severe consequence variants themselves. Of these, 16 variants were determined by Olink to be within a possible epitope-binding site (Supplementary Data 3), including those for CD33, GPNMB, and MS1. Further errors may also be introduced due to cross-reactivity and unspecific binding. For 20 of 21 proteins with corresponding protein quantification using the SomaScan technique (an aptamer-based proteomic technology binding to varying protein sites), the correlation between Olink and SomaScan59 plasma protein measurements was evaluated in 485 individuals from the Fenland cohort60, using Spearman’s rank-based correlation (Supplementary Data 3). Notably, we observed good correlation in protein abundance between the two measurements for CD33 ($\rho = 0.60$), GPNMB ($\rho = 0.51$), and MS1 ($\rho = 0.74$). Explanations for a lack of correlation are manifold, including missing specificity of the aptamer or antibody for the selected target, the low affinity of the aptamer, targeting of different protein isoforms, a different dynamic range of the assays, as well as other technical factors as recently summarised61. Further orthogonal validation using epitope-independent assays is warranted.

We detect no pQTLs for 77 proteins and only trans-pQTLs for 16 proteins. This may be explained by other limitations including those related to epitope-binding. Firstly, only proteins in the serum were quantified. As serum contains multiple cell types originating from different tissues, pQTL detection is volatile to changes in serum composition. We note, for example, that the cis-pQTL for CD33 is also a known blood cell QTL ($r$3865444)37, highlighting how different cell-type composition and therefore, sample handling, can affect the serum proteome and drive pleiotropic signals. Secondly, the individuals included in this analysis are of European ancestry only, and variants that are absent or present in extremely low frequencies in our cohorts would not have been detected. Therefore, our findings—in both the pQTL discovery and downstream causal inference analyses—cannot be extrapolated to non-European populations. Finally, our sample size may not be adequate for the detection of rare variants of small effect sizes, again stressing the importance of larger, ethnically diverse studies.

In conclusion, we present the results of the first WGS-based pQTL analysis of neurologically relevant serum proteins to date. In addition to exploring the genetic architecture of these proteins, we show that pQTL analysis has the potential to identify disease-relevant serum biomarkers for debilitating neurological conditions. We identify opportunities for the repurposing of therapeutic targets, and deliver deeper insight into disease pathways. We recognise that an effective biomarker must be able to differentiate similarly presenting disorders to avoid misdiagnoses; hence, special attention must be given to further validation. Finally, we provide a resource that may be utilised by future studies to develop new hypotheses and advance our understanding of brain-related disorders.

Methods

Cohorts and samples. The two cohorts included in this analysis, MANOLIS and Pomak, are part of the Hellenic Isolated Cohorts (HELIC; https://www.helmholtz-muenchen.de/itg/projects-and-cohorts/HELIC/index.html). The HELIC study focuses on the genetics of complex traits, making use of characteristics of founder populations, such as increased frequency of rare variants, extended linkage disequilibrium, and reduced haplotype complexity. For MANOLIS, biological samples were collected from the mountainous Mylopomatos villages in Crete, Greece; whereas, Pomak refers to a set of mountainous villages in the North of Greece. Further phenotypic and genetic characteristics have been described in detail in previous publications62–64. The study was approved by the Harokopio University Bioethics Committee, and informed consent was obtained from all human subjects.

Sequencing and variant calling. Both MANOLIS and Pomak followed the same sequencing, alignment, and variant calling pipeline. Genomic DNA (500 ng) from 1482 MANOLIS samples and 1642 Pomak samples were sheared to a median size of 500 bp and subjected to standard Illumina paired-end DNA library construction. Adapter-ligated libraries were amplified by six cycles of PCR and subjected to DNA sequencing using the Illumina platform (Illumina) according to the manufacturer’s instructions. Basecall files for each lane were transformed into unmapped BAMs using Sbol2BAM, marking adapter contamination and decoding barcodes for removal into BAM tags. Phix control reads were mapped using BWA Backtrack and were used to remove spatial artefacts. Reads were converted to FASTQ and aligned using BWA MEM 0.7.8 to the hg38 reference (GRCh38) with deovs (HS38DH). The alignment was then merged into the master sample BAM file using
Illumina2BAM MergeAlign. PCR and optical duplicates are marked using bio-
bam tools mark_duplicates and the files are archived in CRAM format. Per-lane CRAM reads pooled on per-protein sample basis across lanes to produce library CRAMs; these were then divided into 200 chunks for parallelism. GVCWs were generated using HaplotypeCaller v.3.5 from the Genome Analysis Toolkit (GATK) for each chunk. All chunks were then merged at sample level, samples were then further combined in batches of 150 samples using GATK CombineGVCFs v.3.5. Variant calling was then performed on each batch using GATK GenotypeGVCFs v.3.5. The resulting variant callsets were then filtered across all batches into a cohort-wide VCF file using bcftools concat.

**Proteomics and QC.** Proteins from Olink’s (https://www.olink.com) Neurology and Neuro-exploratory panels were measured in the serum of 1457 MANOLIS and 1611 POMAK samples. The full list of 184 proteins is provided in Supplementary Data 8. Protein expression was quantified using Olink’s Proximity Extension Assay (PEA) technology. Briefly, each protein assay uses pairs of oligonucleotide-labelled antibody probes; when these antibody pairs bind to the target antigen, the oligonucleotides hybridise due to their proximity and are extended by DNA polymerase. These DNA barcodes are amplified by PCR and quantified using microfluoride qPCR. Protein expression levels are reported as Normalised Protein Expression (N PX) values, Olink’s relative quantification unit, which is in the Log2 scale. NPX values are derived by adjusting raw qPCR Ct values against several internal controls—an extension control, inter-control plate, and a correction factor calculated using a negative control. Additionally, the negative control determines the limit of detection (LOD) for each assay, calculated as the negative control plus three standard deviations. We included all proteins and all below-LOD NPX values in our analysis. Fifty-two and 37 MANOLIS samples, and 68 and 60 POMAK samples failed vendor QC for the Neurology and Neuro-exploratory panels, respectively, and were excluded from the analysis. Reported NPX values were then rank-based inverse normal transformed (INT) and used for the association analysis.

**Association analysis and meta-analysis.** A maximum of 1365 samples from MANOLIS and 1537 samples from POMAK were analysed for the Neurology panel; and for the Neuro-exploratory panel, a maximum of 1372 samples from MANOLIS and 1545 samples from POMAK were analysed. For each cohort, whole genome-wide association analysis with 184 proteins was performed using a linear mixed model implemented in GEMMA v.0.9465, simultaneously adjusting for covariates. The more conservative model implemented in GEMMA v.0.9465, simultaneously adjusting for covariates —an extension control, inter-plate control, and a correction factor calculated using a negative control. Additionally, the negative control determines the limit of detection (LOD) for each assay, calculated as the negative control plus three standard deviations. We included all proteins and all below-LOD NPX values in our analysis. Fifty-two and 37 MANOLIS samples, and 68 and 60 POMAK samples failed vendor QC for the Neurology and Neuro-exploratory panels, respectively, and were excluded from the analysis. Reported NPX values were then rank-based inverse normal transformed (INT) and used for the association analysis.

**Conditional analysis to identify independent variants.** Using the PeakPlotter software (https://github.com/hmgu-itt/peakplotter), we detected 171 signals. We observed several signals extending over large regions that were mistakenly broken up into multiple signals; because of this, 12 signals were excluded to give 159 sig-
individuals. Independent variants were identified using the approximate conditional and joint stepwise model selection, implemented using the “slct” option in GCTA-COJO; using a cut-off of 0.9. Before filtering, however, variants were first subjected to clumping in Plink 1.999 (www.cog-genomics.org/plink/1.9/), using a r2 threshold of 0.1 and a clumping window of 1 Mb; this reduces the number of variants input to COJO to avoid overfitting of the model. We arrived at a final set of 214 independent variants for 140 signals after filtering for minor allele count M > 1% that passed the Hardy–Weinberg equilibrium test (P > 1 × 10−5). Following per-cohort analysis, 12,392,022 variants common to the two cohorts were meta-analysed using the fixed-effects inverse variance-based method in METAL90. As no proteins displayed significant genomic inflation (0.95 < λc1 < 1.03), no genomic control was applied.

**Significance thresholds**

Single variant-based association and rare variant analysis. For single variant-based association, the significance threshold was adjusted for multiple testing by cor-
recting for the effective number of protein traits (Meff) and variants (Neff) analysed. The effective number of proteins was computed using the ratio of the effective variance to its maximum90:

\[ M_{eff} = \frac{M(1 - (M - 1) V_{se} / M)}{M} = 1 - \frac{\text{tr} (\Sigma^\top \Sigma)}{M} \]

(1)

where \( V_{se} \) is the variance of the eigenvalues of the correlation matrix. For the M = 184 Olink proteins included in this study, \( M_{eff} = 93 \) in both cohorts. The effective number of variants, or \( N_{eff} \), was determined by using the −ind and −marf options offered in Plink 1.9 to prune these variants. Specifically, variants with a minor allele count (MAC) of <10 were excluded; and parameters specified for −ind were window size of 50 kb, variant count of 5, and variance inflation factor (VIF) > 10. This method was performed separately for both cohorts, with resulting \( N_{eff} \)s of 5,078,182 and 4,144,062 in each respective cohort. The more conservative \( N_{eff} \) of 5,078,182 was considered for the calculation of the P-
value significance threshold for the meta-analysis to give a final P-value threshold of 1.05 × 10−10. The same threshold was used for the rare variant analysis.

**Novelty**

We assessed variants for novelty using a funnel approach, by first iden-
tifying (a) novel proteins, then (b) novel signals, and finally (c) novel variants. Novel proteins were defined as proteins that are being analysed for pQTLs for the first time. This was determined by comparing our proteins against protein lists from four large pQTL studies10–12, querying GWAS Catalogue for known signals, then confirmed by doing manual literature searches. Next, we determined variants associated to novel signals by checking against previously reported pQTLs (Supplementary Data 2). Signals were considered novel if no variants had been reported within 1 Mb upstream and downstream of our variants. All variants from known loci were then assessed for novelty by matching their rsIDs against pre-
viously reported variants; where no match was found, variants were conditioned on other known variants at the locus, and considered novel if the association P-value remained significant after conditioning.

**Heritability**

Heritability analysis was performed using GCTA GREML91 (https://gsnogenomics.com/software/gcta/index.html#GREML), using both the multi-
component LDMS and single-component approaches in two separate cohorts. The final meta-analysis \( h^2_{meta} \) was calculated using the following formula (provided on the GCTA website):

\[ h^2_{meta} = \frac{\sum(h_i^2 / SE_i^2)}{\sum(1 / SE_i^2)} = \sqrt{\frac{1}{\sum(1 / SE_i^2)}} \]

(2)

**GREML-LDMS.** For each cohort, the segment-based LD score was first calculated using GCTA’s −ld-score-region with the default length segment of 200 Kb. Vari-
ents were then stratified into four quartiles according to their LD scores in R, and a genetic relatedness matrix (GRM) was calculated for each group. For each protein, we then ran RELM analysis with four GRMs using default settings. RELM analysis failed to converge for 45 proteins across the two cohorts, likely due to limitations arising from a smaller sample size.

**GREML-SC.** As we were unable to obtain \( h^2 \) estimates for all proteins using GREML-LDMS, we also ran single-component GREML (GREML-SC) for all protein traits using a single GRM (also computed using GCTA). Full results may be found in Supplementary Data 9.

**Variant consequences.** We used Ensembl’s variant effect predictor72 (VEP; http://www.ensembl.org/vep) to determine the most severe consequence of each variant. To check for potential protein-altering effects, we also queried the most severe consequence of variants in LD (r2 > 0.8) with reported cis-acting variants, which were extracted using PLINK 1.9. Variants with, or in LD with variants with potentially protein-altering consequences are reported in Supplementary Data 3.

**eQTL colocalisation.** Colocalisation analysis was performed using our pQTL results and gene expression QTL (eQTL) data downloaded from the GTEx database (https://www.gtexportal.org/), using the colocal.fast function from the gtx R package (https://github.com/tobyjohnson/gtx/). The method is equivalent to coloco by Giambartolomei et al.73 and assumes only one causal variant at each associated locus. To satisfy this assumption in our pQTL data, for each independent variant, we conditioned associations on all other independent variants at the locus. For cis-
pQTLs, we tested colocalisation with an expression of the encoding gene in all available tissues. For trans-pQTLs, colocalisation was performed with all genes within 2 Mb of the causal variant for all available tissues. For all analysed genes, eQTL data within 1 Mb upstream and downstream of the causal variant was extracted.

**PheWAS colocalisation.** Using the same conditioned pQTL data from the eQTL colocalisation analysis, we performed colocalisation with psychiatric and neuro-
degenerative traits. For each analysed locus, GWAS data within 2 Mb of the causal variant was extracted. We used only publicly available summary statistics, either downloaded from the Psychiatric Genomics Consortium (PGC) website (https://www.med.unc.edu/pgc/download-results/), or as mentioned in the respective papers. A list of studies used can be found in Supplementary Data 5b. Additionally, colocalisation analysis was carried out with PhenoScanner4,27 traits of neuro-
ological relevance. The results for this are included in Supplementary Data 5a. Five different posterior probabilities are reported in the table (CLP0–CLP4), which corresponds to the five tested hypotheses explained in Giambartolomei at al.73. In particular, CLP4 indicates association with both tested traits with a shared causal variant.
Two-sample Mendelian randomisation. Two-sample MR was performed between 107 protein traits and 206 neurologically relevant phenotypes, using the TwoSampleMR R package (https://github.com/MBCIEU/TwoSampleMR). Traits available in the MRBase platform were selected based on the following: (a) Self-reported traits in UK Biobank with at least 1000 cases; (b) UK Biobank ICD10 primary and secondary traits of neurological relevance; (c) studies categorised as ‘Psychiatric/neurological’, ‘Personality’, and ‘Sleeping’; (d) other large neurologically relevant traits with more than 10,000 samples; (e) manually downloaded summary statistics (see ‘PheWAS colocalisation’ section). Independent variants with an association meta-analysis $P < 5 \times 10^{-8}$ were determined by GCTA-COJO (see ‘Peak calling and independent variants’ section) and used as instrumental variables (IV), including both cis and trans variants. All variants at pleiotropic loci, including KLKB1, FUT2, ABO, STS1GAL6, and the HLA region, were excluded from the analysis. For each protein-trait pair, pQTL summary statistics for all independent variants and their variants in LD ($r^2 > 0.8$) were first extracted, excluding those without rsIDs. This was then harmonised with the available outcome data. Where any independent variant was not available in the outcome data, an LD variant ($r^2 > 0.8$) was used as proxy instead. For protein traits with more than 1 causal variant (IV), we used the inverse variance-weighted method; otherwise, Wald ratio estimates were used. Sensitivity analysis was carried out for protein-traits pairs with more than one IV by assessing heterogeneity about the IVW estimate using Cochran’s Q tests, with $P < 0.05$ denoting significant heterogeneity. We find that none of the protein-traits pairs with an FDR-adjusted $P < 0.05$ had Cochran’s Q $P < 0.05$. The analysis was also repeated using only cis-pQTLs (Supplementary Data 6b). This resulted in an additional causal signal, LTBP3 with osteoarthritis; and the loss of three signals: ADAM23 with neurotrophin, NEP with osteoarthritis, and SIGLEC1 with osteoarthritis. We note an important caveat of our analysis, which is that when only one instrumental variable is available, a higher risk of violating the two-sample MR assumptions exists. Results from Wald ratio tests should, therefore, be interpreted cautiously and with orthogonal validation.

Drug target evaluation. Drug target evaluation was done by querying the Open Targets (https://www.targetvalidation.org/) and Drugbank (https://go.drugbank.com/) databases (Supplementary Data 7).

Ethics statement. The study was approved by the Institutional Review Board of Harokopio University and the Greek Ministry of Education, Lifelong Learning and Religious Affairs. The MAN-OLIS and Pomak studies were approved by the Harokopio University Bioethics Committee and informed consent was obtained from every participant.

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

Data availability. The MANOLIS sequencing data used in this study are available at the European Genome-Phenome Archive (EGA) under accession number EGAS00001001207. The Pomak sequencing data have not been deposited to the EGA as the data and the information derived from it are culturally and politically sensitive in the context of this religiously isolated population. We will consider requests to access the data by researchers when an alternative cohort cannot reasonably be used for their research, and will respond to such requests within 6 months. Summary statistics generated in this study are available for download in the GWAS Catalogue. Accession codes and the respective hyperlinks are provided in Supplementary Data 10.

Code availability. Analysis was performed using publicly available software as described in the ‘Methods’ section. Additional scripts may be found in our GitHub repositories (https://github.com/hmgu-ing).

Received: 15 April 2021; Accepted: 11 November 2021; Published online: 02 December 2021

References
1. Feigin, V. L. et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 18, 459–480 (2019).
2. GBD. 2015 Neurological Disorders Collaborator Group. Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet Neurol. 16, 877–897 (2017).
3. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric and Neurodevelopmental Disorders. Cell. 179, 7 (2019).
4. De Jager, P. L., Yang, H.-S. & Bennett, D. A. Deconstructing and targeting the genomic architecture of human neurodegeneration. Nat. Neurosci. 21, 1310–1317 (2018).
5. Ayano, G. et al. Misdiagnosis, detection rate, and associated factors of severe psychiatric disorders in a specialized psychiatry center in Ethiopia. Ann. Gen. Psychiatry 20, 10 (2021).
6. Hansson, O. Biomarkers for neurodegenerative diseases. Nat. Med. 27, 954–963 (2021).
7. 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).
8. Winglo, A. P. et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer’s disease pathogenesis. Nat. Genet. https://doi.org/10.1038/s41588-020-00777-z (2021).
9. Lifkesen, L. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nat Metab 2, 1135–1148 (2020).
10. Yao, C. et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat. Commun. 9, 3268 (2018).
11. Sun, B. B. et al. Genomic atlas of the human plasma proteome. Nature 558, 73–79 (2018).
12. Sutre, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat. Commun. 8, 14357 (2017).
13. Alkebils, L. et al. Validation of the accuracy of self-reported ABO blood types in the Danish Nurses’ Health Study. Asian Pac. J. Cancer Prev. 20, 789–793 (2019).
14. Sasaayama, D. et al. Genome-wide quantitative trait locus mapping of the human cerebrospinal fluid proteome. Hum. Mol. Genet. https://doi.org/10.1093/hmg/ddw366 (2016).
15. Parham, P. MHC class I molecules and KIRs in human history, health and survival. Nat. Rev. Immunol. 5, 201–214 (2005).
16. Jenkins, R. W., Canals, D. & Hannun, Y. A. Roles and regulation of secretory lysosomal acid sphingomyelinase. Cell. Signal. 21, 836–846 (2009).
17. Park, M. H., Jun, H. K. & Bae, J.-S. Potential therapeutic target for aging and age-related neurodegenerative diseases: the role of acid sphingomyelinase. Exp. Mol. Med. 52, 380–389 (2020).
18. Lee, J. K. et al. Acid sphingomyelinase modulates the autophagic process by controlling lysosomal biogenesis in Alzheimer’s disease. J. Exp. Med. 211, 1515–1570 (2014).
19. Cutler, R. G., Pedersen, W. A., Camandola, S., Rothstein, J. D. & Mattson, M. P. Evidence that accumulation of ceramides and cholesterol esters mediates oxidative stress-induced death of motor neurons in amyotrophic lateral sclerosis. Ann. Neurol. 52, 448–457 (2002).
20. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. 47, 1114–1120 (2015).
21. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nat. Genet. 50, 737–745 (2018).
22. Li, J. et al. Proportion of neuroprotective pain in the back region in chronic low back pain patients - a multicenter investigation. Scie. Rep. 8, 16537 (2018).
23. Thäkur, M., Dickenson, A. H. & Baron, R. Osteoarthrits pain: nociceptive or neuropathic? Nat. Rev. Rheumatol. 10, 374–380 (2014).
24. Pisano, C. et al. High leptin levels are associated with migraine with aura. Cephalalgia Int. J. Headache 37, 435–441 (2017).
25. Zhou, T. et al. Common variants in LTBP3 gene contributed to the risk of hip osteoarthritis in Han Chinese population. BioSci. Rep. 40, BSR20192999 (2020).
26. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association studies of bipolar disorder identifies a new susceptibility locus near ODZ4. Nature. Genet. 43, 977–983 (2011).
27. Gilbert-Latour, J. et al. Semaphorin and plexin gene expression is altered in the prefrontal cortex of schizophrenia patients with and without auditory hallucinations. Psychiatr. Res. 229, 850–857 (2015).
28. Tavares, H., Yacubian, J., Talib, L. L., Barbosa, N. R. & Gattaz, W. F. Increased phospholipase A2 activity in schizophrenia with absent response to nicotine. Schizophr. Res. 61, 1–6 (2003).
29. Sháma, G., Kondás, K., Treater, M. & Pathy, L. WFIKKKN1 and WFIKKKN2 bind growth factors TGFβ1, BMP2 and BMP4 but do not inhibit their signalling. FEBS J. 277, 5040–5050 (2010).
30. Taylor-Gjeeve, R. M., Nair, B. V. & Gjeeve, J. A. Obstructive sleep apnoea in relation to rheumatic disease. Rheumatol. Oxf. Engl. 52, 15–21 (2013).
31. Kang, J.-H. & Lin, H.-C. Obstructive sleep apnea and the risk of autoimmune diseases: a longitudinal population-based study. Sleep. Med. 13, 583–588 (2012).
32. Nalls, M. A. et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson’s disease. *Nat. Genet.* **46**, 989–993 (2014).

33. Moloney, E. B., Moskites, A., Ferrari, E. J., Isacson, O. & Hallett, P. J. The glycolyprotein GPNMB is selectively elevated in the substantia nigra of Parkinson’s disease patients and increases after lysosomal stress. *Neurobiol. Dis.* **120**, 1–11 (2018).

34. Ptaszałka, A. N. et al. ADAM10 releases a soluble form of the GPNMB/ascorbic acid extracellular domain with angiogenic properties. *PLoS ONE* **5**, e12093 (2010).

35. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies six new risk loci for Parkinson’s disease. *Nat. Genet.* **45**, 1452–1458 (2013).

36. Griscius, A. et al. Alzheimer’s disease risk gene CD33 inhibits microglial uptake of amyloid beta. *Neuron* **78**, 631–645 (2013).

37. Bradshaw, E. M. et al. CD33 Alzheimer’s disease locus: altered monocyte function and amyloid biology. *Nat. Neurosci.* **16**, 848–850 (2013).

38. Goes, F. S. et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **168**, 649–659 (2015).

39. Sherva, R. et al. Genome-wide association study of rate of cognitive decline in Alzheimer’s disease patients identifies novel genes and pathways. *Alzheimers Dement. J. Alzheimers Assoc.* **16**, 1134–1145 (2020).

40. Frenkel, D. et al. Scaral deficiency impairs clearance of soluble amyloid-β by mononuclear phagocytes and accelerates Alzheimer’s-like disease progression. *Nat. Commun.* **9**, 1370 (2018).

41. Shichita, T. et al. MAFB prevents excess inflammation after ischemic stroke by accelerating clearance of damage signals through MSR1. *Nat. Med.* **23**, 723–732 (2017).

42. Kunjathoor, V. D. et al. Scavenger receptors class A-II and CD36 are the principal receptors responsible for the uptake of modified low density lipoprotein leading to lipid loading in macrophages. *J. Biol. Chem.* **277**, 49982–49988 (2002).

43. Kelley, J. L., Ozment, T. R., Li, C., Schweitzer, J. B. & Williams, D. L. Scavenger receptor-A (CD204): a two-edged sword in health and disease. *Crit. Rev. Immunol.* **34**, 241–261 (2014).

44. Cornejo, F. et al. Scavenger Receptor-A deficiency impairs immune response of macrophages and astrocytes potentiating Alzheimer’s disease pathophysiology. *Brain Behav. Immun.* **69**, 336–350 (2018).

45. McCollum, L. A. & Roberts, R. C. Uncovering the role of the nucleus accumbens in schizophrenia: a postmortem analysis of tyrosine hydroxylase and vesicular glutamate transporters. *Schizophr. Res.* **169**, 369–373 (2015).

46. McCollum, L. A., Walker, C. K., Roche, J. K. & Roberts, R. C. Elevated excitatory input to the nucleus accumbens in schizophrenia: a postmortem ultrastructural study. *Schizophr. Bull.* **41**, 1123–1132 (2015).

47. Chaudhury, D. et al. Rapid regulation of depression-related behaviours by control of midbrain dopamine neurons. *Nature* **493**, 532–536 (2013).

48. Dujardin, K. & Sgamato, V. Neuropsychiatric disorders in Parkinson’s Disease: what do we know about the role of dopaminergic and non-dopaminergic systems? *Front. Neurosci.* **13**, 25 (2020).

49. Correll, C. U. et al. Prevalence, incidence and mortality from cardiovascular disease in patients with pooled and specific severe mental illness: a large-scale meta-analysis of 3,211,768 patients and 113,383,368 controls. *World Psychiatry* **8**, 877–892 (2017).

50. Santos, R. et al. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2017).

51. Okuno, Y., Matsumura, N. & Oguro, S. Transcatheter arterial embolization using imipenem/colistatin sodium for tendinopathy and enthesopathy refractory to nonsurgical management. *J. Vasc. Interv. Radiol.* **24**, 787–792 (2013).

52. Lee, S. H. et al. Clinical outcomes of transcatheter arterial embolisation for chronic knee pain: mild-to-moderate versus severe knee osteoarthritis. *Cardiovasc. Interv. Radiol.* **42**, 1530–1536 (2019).

53. Malik, M. et al. Genetics of CD33 in Alzheimer’s disease and acute myeloid leukemia. *Hum. Mol. Genet.* **24**, 3357–3370 (2015).

54. Hilgmann, M. et al. Health economics in the field of osteoarthritis: an expert’s consensus paper from the European Society for Clinical and Economic Aspects of Osteoporosis and Osteoarthritis (ESCEO). *Semin. Arthritis Rheum.* **43**, 303–313 (2013).

55. Tschamazidou, I. et al. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nat. Genet.* **51**, 230–236 (2019).

56. Miller, R. E., Lu, Y., Tortorella, M. D. & Malfait, A.-M. Genetically engineered mouse models reveal the importance of proteases as osteoarthritis drug targets. *Curr. Rheumatol. Rep.* **15**, 350 (2013).

57. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).

58. Gold, I. et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE* **5**, e15004 (2010).

59. Pietzner, M. et al. Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6597 (2020).

60. McCollum, L. A. et al. Cross-platform proteomics to advance genetic prioritisation strategies. http://biobasev.org/lookup/doi/10.1101/2021.03.18.435919 (2021).

61. Farmaki, A.-E. et al. The mountainous Cretan dietary patterns and their relationship with cardiovascular risk factors: the Hellenic Isolated Cohorts MANOLIS study. *Public Health Nutr.* **20**, 1063–1074 (2017).

62. Gill, A. et al. Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. *Hum. Mol. Genet.* **25**, 2360–2365 (2016).

63. Panoutsopoulou, K. et al. Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat. Commun.* **5**, 5345 (2014).

64. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for genome association studies. *Nat. Genet.* **44**, 821–824 (2012).

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

66. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **87**, 76–82 (2011).

67. Giazza, G. C. et al. Second-degree PINK1 mutation: challenging to the larger and richer data sets. *GigaScience* **4**, 7 (2015).

68. Moskvin, V. & Schmidt, K. M. On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.* **32**, 567–573 (2008).

69. Cheverud, J. M. A simple correction for multiple comparisons in interval mapping genome scans. *Hereditas* **87**, 52–58 (2001).

70. Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).

71. McLaren, W. et al. The Ensemble variant effect predictor. *Genome Biol.* **17**, 12 (2016).

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

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

74. Svedas, T. et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).

75. Hemani, G., Tilson, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).

76. Hemani, G. et al. The MR-base platform supports systematic causal inference across the human phenotype. *elife* **7**, e34408 (2018).

77. Ochoa, D. et al. Open targets platform: supporting systematic drug-target prioritisation and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).

78. Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
**Author contributions**
Sample collection and phenotyping: E.T., M.K., G.D. and E.Z. Phenotype transformation and quality control: G.P., L.R., P.N., X.S. and J.F.W. Association analysis: G.P. Downstream bioinformatics: G.P., A.G. and A.B. Somalogic comparison with Fenland cohort: M.P., E.W., N.J.W. and C.L. Software development: A.G. and A.B. Study design and supervision: E.Z. Manuscript writing: G.P. and E.Z. Manuscript editing: G.P., A.B., L.R., P.N., X.S., M.P., E.W., N.J.W., C.L., E.T., M.K., G.D., A.M., J.F.W., A.G. and E.Z.

**Competing interests**
The authors declare no competing interests.

**Additional information**
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-27387-1.

**Correspondence** and requests for materials should be addressed to Grace Png or Eleftheria Zeggini.

**Peer review information** Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021