TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Medicine and Health

Molecular Biomarkers of Complex Traits

Yunting Grace Png

Vollständiger Abdruck der von der TUM School of Medicine and Health der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Julia Höfele

Prüfende der Dissertation:

1. Prof. Dr. Eleftheria Zeggini
2. Prof. Dr. Wolfgang Wurst
3. Prof. Dr. Andreas Birkenfeld

Die Dissertation wurde am 08.05.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Medicine and Health am 07.02.2024 angenommen.

Molecular biomarkers of complex traits

Grace Png

Acknowledgements

I would like to thank my supervisor, Prof. Dr. Eleftheria Zeggini, for her guidance and support throughout this journey. I will always be grateful for the opportunities she has provided, the trust she bestowed on me, and the genuine care that she showed for my professional and personal wellbeing. I also want to thank the members of my thesis advisor committee, Dr. Anders Mälarstig and Prof. Dr. Wolfgang Wurst, for their scientific and career advice.

I am indebted to Arthur Gilly and Park Young-Chan, for all their (bio)informatics help. To my other colleagues at the Institute of Translational Genomics – with special mention to Georgia and Peter – thank you for the company, conversations (work and non-work), and beer. I am especially thankful to the Ops team for their invaluable help with administrative work and translations.

To Hao Jie, Rebecca, and my family – thank you for the love and pride you poured into me when I could not bear to do so myself. You brought light and joy to the darkest days. To my parents, who taught me to live and read, thank you for always giving me the best.

Summary

Complex diseases such as type 2 diabetes and Alzheimer's disease are growing global health burdens that negatively impact quality of life. Understanding the molecular perturbations causing these diseases is key to finding biomarkers and drug targets that enable early and effective intervention. While studies have been successful at identifying numerous disease-associated genes and proteins, discerning causal mechanisms remains a challenge.

Protein levels are intermediate phenotypes that are often dysregulated in disease; and are measurable, druggable targets. In this thesis, I show how genetic variation underlying serum protein levels (protein quantitative trait loci [pQTLs]) can help to pinpoint causal biomarkers and pathways. Using whole genome sequencing (WGS) data from two isolated Greek populations, we detect pQTLs for >400 serum proteins relevant to cardiometabolic and neurological processes. We find novel pQTLs, including some rare variants that have drifted up in frequency in the studied cohorts.

By integrating our pQTL findings with existing large-scale disease genome-wide association study (GWAS) data, we identify genes and proteins causal for cardiometabolic traits and neurological diseases through colocalisation analysis and Mendelian randomisation. These include CD33 and Alzheimer's disease; MEP1B and high-density lipoprotein levels; and MSR1 and schizophrenia.

In doing so, we validate known targets and propose potential clinical biomarkers and drug repurposing opportunities, while demonstrating the importance of isolated populations in pQTL analysis. We discuss limitations and future directions, and contribute to a growing pQTL resource that may be used by others to empower future research.

Zusammenfassung

Komplexe Krankheiten wie Typ-2-Diabetes und Alzheimer stellen weltweit eine zunehmende Belastung für die Gesundheit dar und beeinträchtigen die Lebensqualität. Kenntnis von molekularen Veränderungen, die diese Krankheiten verursachen, ist der Schlüssel für die Suche nach Biomarkern und Arzneimittelzielen, die ein frühzeitiges und wirksames Eingreifen ermöglichen. Zwar konnten in Studien zahlreiche krankheitsassoziierte Gene und Proteine identifiziert werden, doch bleibt es eine Herausforderung, die ursächlichen Mechanismen zu erkennen.

Proteinspiegel sind intermediäre Phänotypen, die bei Krankheiten oft fehlreguliert sind und messbare, medikamentöse Ziele darstellen. In dieser Dissertation zeige ich, wie genetische Variationen, die den Serumproteinspiegeln zugrunde liegen (protein quantitative trait loci [pQTLs]), dazu beitragen können, kausale Biomarker und Signalwege zu identifizieren. Anhand von Daten aus Ganzgenomsequenzierungen (WGS) aus zwei isolierten griechischen Populationen ermitteln wir pQTLs für >400 Serumproteine, die für kardiometabolische und neurologische Prozesse relevant sind. Wir finden neue pQTLs, darunter einige seltene Varianten, deren Frequenz in den untersuchten Kohorten gestiegen ist.

Durch die Integration unserer pQTL-Ergebnisse mit bestehenden groß angelegten genomweiten Assoziationsstudien (GWAS) identifizieren wir Gene und Proteine, die für kardiometabolische Merkmale und neurologische Erkrankungen kausal sind, durch Kolokalisationsanalyse und Mendelian Randomization. Dazu gehören CD33 und Alzheimer, MEP1B und High-Density-Lipoproteinspiegel sowie MSR1 und Schizophrenie.

Auf diese Weise validieren wir bekannte Zielmoleküle und schlagen potenzielle klinische Biomarker und Möglichkeiten für die Neuausrichtung von Medikamenten vor, während wir gleichzeitig die Bedeutung isolierter Populationen bei der pQTL-Analyse aufzeigen. Wir erörtern Einschränkungen und künftige Richtungen und leisten einen Beitrag zu einer wachsenden pQTL-Ressource, die von anderen genutzt werden kann, um die künftige Forschung zu unterstützen.

List of publications

First author publications contributing to this thesis:

i. **Png G**, Gerlini R, Hatzikotoulas K, Barysenka A, Rayner NW, Klarić L, Rathkolb B, Aguilar-Pimentel JA, Rozman J, Fuchs H, Gailus-Durner V, Tsafantakis E, Karaleftheri M, Dedoussis G, Pietrzik C, Wilson JF, Angelis MH, Becker-Pauly C, Gilly A, Zeggini E. **Identifying causal serum protein-cardiometabolic trait relationships using whole genome sequencing.** Hum Mol Genet. 2022 Nov 9:ddac275. doi: 10.1093/hmg/ddac275. Epub ahead of print. PMID: 36349687. (Appendix B)

ii. **Png G**, Barysenka A, Repetto L, Navarro P, Shen X, Pietzner M, Wheeler E, Wareham NJ, Langenberg C, Tsafantakis E, Karaleftheri M, Dedoussis G, Mälarstig A, Wilson JF, Gilly A, Zeggini E. **Mapping the serum proteome to neurological diseases using whole genome sequencing.** Nat Commun. 2021 Dec 2;12(1):7042. doi: 10.1038/s41467-021-27387-1. PMID: 34857772; PMCID: PMC8640022. (Appendix A)

Further publications:

i. Gilly A, Klaric L, Park YC, **Png G**, Barysenka A, Marsh JA, Tsafantakis E, Karaleftheri M, Dedoussis G, Wilson JF, Zeggini E. **Gene-based whole genome sequencing meta-analysis of 250 circulating proteins in three isolated European populations.** Mol Metab. 2022 Jul;61:101509. doi: 10.1016/j.molmet.2022.101509. Epub 2022 Apr 30. PMID: 35504531; PMCID: PMC9118462.

ii. Gilly A, Park YC, **Png G**, Barysenka A, Fischer I, Bjørnland T, Southam L, Suveges D, Neumeyer S, Rayner NW, Tsafantakis E, Karaleftheri M, Dedoussis G, Zeggini E. **Whole-genome sequencing analysis of the cardiometabolic proteome.** Nat Commun. 2020 Dec 10;11(1):6336. doi: 10.1038/s41467-020-20079-2. PMID: 33303764; PMCID: PMC7729872.

iii. **Png G**, Suveges D, Park YC, Walter K, Kundu K, Ntalla I, Tsafantakis E, Karaleftheri M, Dedoussis G, Zeggini E, Gilly A. **Population-wide copy**

number variation calling using variant call format files from 6,898 individuals. Genet Epidemiol. 2020 Jan;44(1):79-89. doi: 10.1002/gepi.22260. Epub 2019 Sep 14. PMID: 31520489; PMCID: PMC8653900.

Table of Contents

1 Introduction

1.1 Genetics of complex diseases

The study of human genetics is aimed at understanding how genetic variations can explain the differences in traits observed between people. The most common method to do this today involves testing individual variants throughout the entire human genome for associations with traits of interest – also called phenotypes – in a genome-wide association study (GWAS). Before the first landmark study in 2005[1], disease-associated genes were primarily identified through linkage analyses in high-risk families. This approach was limited, however, to traits with clear patterns of inheritance, and has thus seen most success in identifying risk genes for rare diseases caused by aberrations in only one or a few genes – also known as Mendelian traits.

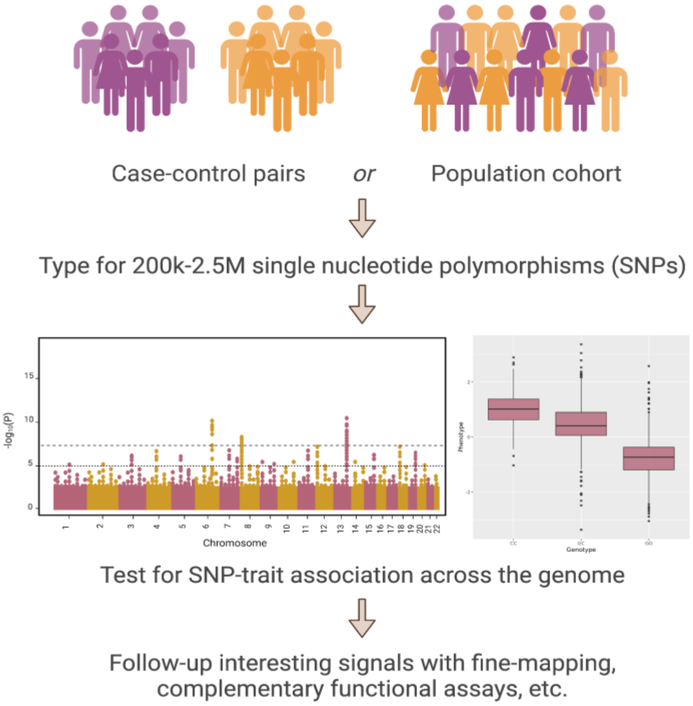

Case-control pairs
or
Population cohort
Type for 200k-2.5M single nucleotide polymorphisms (SNPs)
-log10(P)
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
Phenotype
Genotype
Test for SNP-trait association across the genome
Follow-up interesting signals with fine-mapping,
complementary functional assays, etc.

Figure 1. Basic principles of GWAS.

At the same time, linkage studies revealed a group of traits that could not be explained by the Mendelian model. In a 1994 study of Alzheimer's disease (AD)[2], the authors attempted to explain transmission patterns for late-onset AD using Mendelian models, but failed. They suggested a "more complex transmission mechanism (mixed or polygenic model)" with no singular deterministic risk gene, and called for novel methodology that could account for "oligogenic and heterogeneity models". GWAS has not only removed the need for pedigree information, but provides the power and resolution needed to capture the breadth of genetic variation associated with complex, polygenic traits like late-onset AD. For example, in one of the largest late-onset AD GWAS (>90,000 cases) to date[3], 38 associated loci implicating a range of biological processes were detected, among over 10,000 loci that have been associated with thousands of complex traits[4] in the past two decades.

1.1.1 Limitations of GWAS

The advent of GWAS soon revealed limitations regarding the method and interpretation of results. Associated variants discovered through GWAS explained a much smaller proportion of phenotypic variance than expected. This was apparent in early GWAS studies for height[5–7], which found over 40 associated loci that explained only about 5% of trait heritability – much lower than the estimated heritability of 80%[8,9].This discrepancy was hence termed the "missing heritability" problem. Sources of missing heritability have been discussed extensively since, and include overestimations of genuine heritability[10]; imprecise phenotype definitions; epistasis[11]; contributions of variants in non-coding regions, rare variants of small effect sizes, or structural variants; and inadequate sample sizes and ancestral diversity.

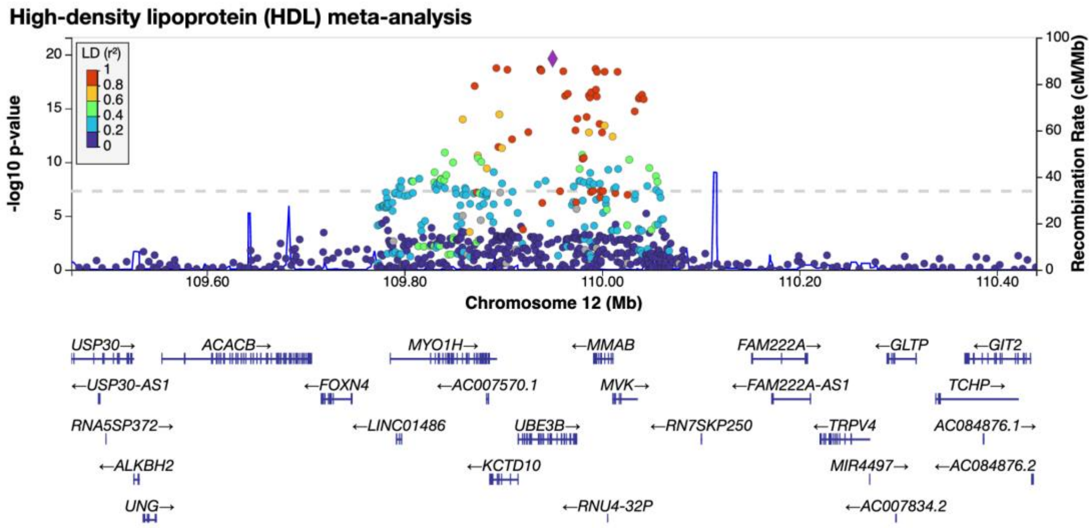


Figure 2. Regional Manhattan plot of locus on chromosome 12 associated with high-density lipoprotein levels [24097068]. The diamond represents the SNP with the lowest p-value; other SNPs are coloured according to the extent of linkage disequilibrium (r^2) with the strongest associated SNP. Nearest genes are annotated in the panel below the plot. Figure produced and downloaded from https://my.locuszoom.org/.

Downstream from variant detection, the complexity of the human genome has furthermore made it difficult to identify causal variants and causal genes. A large obstacle is linkage disequilibrium (LD), whereby alleles at different loci are more likely to be inherited together than by random chance. As a result, GWAS signals are often composed of multiple variants in LD that may span larger or smaller regions of the genome (Figure 2), obscuring the causal variant. This is even more complicated for variants in non-coding regions, which make up 99% of the human genome. Increasingly sophisticated fine-mapping tools are being developed for this purpose[12,13]. The process often includes annotating variants using deleteriousness scores that consider genomic features and protein structure, such as CADD[14] and Eigen[15]; incorporating growing information on regulatory elements (databases such as ENCODE[16]); or overlapping with quantitative trait loci.

1.2 Overcoming the limitations

1.2.1 Whole-genome sequencing

One way to find missing heritability is to detect more genetic variants, and this can be achieved by genotyping more sites. Single nucleotide polymorphism (SNP) arrays combined with imputation is the most widely-used method to genotype sites across the whole genome at decent coverage. Popular as a cost-effective solution, these arrays contain allele-specific oligonucleotide probes to which fluorescently-labelled sequences can bind to, producing a hybridisation signal that reflects the genotype at a chosen site. Modern arrays typically directly genotype between 200,000 and 2 million sites out of the ~3.3 billion bases of the human genome; and genotypes at additional sites are then imputed through LD. However, SNP arrays are predesigned based on known variants, and imputation is based on known LD patterns in studied populations. This means that while they are effective at genotyping common variation in well-studied populations, SNP arrays cannot reliably genotype rare or novel variants, especially in less studied populations.

In contrast, whole genome sequencing (WGS) aims to genotype each position in the genome directly without prior information. In addition to eliminating the ascertainment bias associated with array design[17], WGS provides coverage that is magnitudes larger, plus high-quality genotyping of ultra-rare and novel variants. Even at very low sequencing depths, Gilly et al.[18] found that WGS identified twice as many variants as a genotyping array + imputation approach in the same cohort, with a large proportion of variants only identified through WGS being rare variants. When applied in a GWAS across more than 50 phenotypes, low-depth WGS also detected twice the number of suggestive associations with almost equal sensitivity. The study shows how WGS is able to provide a more complete characterisation of underlying genetic architecture than imputation approaches, particularly at the extreme ends of the allele frequency spectrum.

1.2.2 Isolated populations

Analysing isolated populations can also address the lack of power that general population studies have to detect low-effect rare variants. Defined as subpopulations arising from a small group of individuals or founders that are geographically and/or culturally isolated due to a founder or bottleneck event (Figure 3), isolates have unique characteristics that can be leveraged to empower GWAS studies. One such population is the Icelandic population of roughly 320,000, most of whom are descended from a small group of individuals that emigrated from Scandinavia, Scotland, and Ireland approximately 1,100 years ago[19].

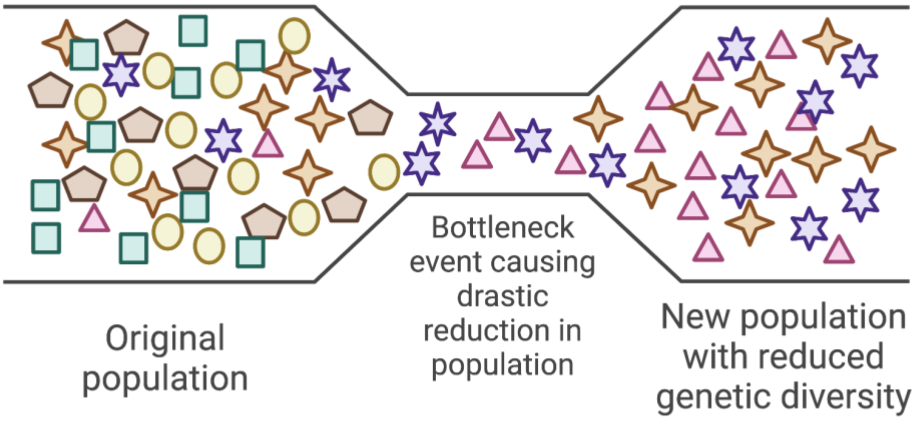

Original population
Bottleneck event causing drastic reduction in population
New population with reduced genetic diversity

Figure 3. Bottleneck effect leading to reduced genetic diversity in a new population.

An advantageous genetic characteristic of isolated populations is their reduced haplotype complexity, which is observed by stretches of LD that extend over longer distances compared to non-isolated populations. Haplotypes refer to groups of variants in LD; and higher overall levels of LD result in longer haplotypes and more haplotype sharing between individuals. This facilitates genotype imputation and increases the power of an association study by reducing heterogeneity; although it also reduces fine-mapping potential. A second advantage is the enrichment of rare disease alleles, which is a result of a combination of geographical isolation, endogamy, and stronger genetic drift. This raises certain alleles to fixation (and others to extinction) in the population (Figure 3); consequently, rare disease-causing alleles may drift up in frequency in isolated populations, enabling easier detection of a rare GWAS signal even with smaller sample sizes.

These characteristics are further complemented by reduced environmental and cultural heterogeneity (e.g., in diet, lifestyle habits, or environmental conditions), which limits potential confounding. Isolated populations have proven useful in GWAS: the cardioprotective variant in *APOC3* was first discovered in an isolated Old Order Amish population[20], and was associated with higher high-density lipoprotein (HDL) levels and lower total cholesterol levels. The variant, R19X, is a loss-of-function variant that is non-existent in the general European population but was found in 5% of the studied Amish individuals, enabling discovery. Later, R19X was also found in the isolated Greek population, MANOLIS, at a frequency of 1.9%[21]. The association with higher HDL levels was recapitulated in MANOLIS, providing evidence supporting the clinical relevance of R19X and demonstrating how population isolates can empower GWAS discovery.

1.2.3 Rare variant analysis

R19X exemplifies how functionally important rare variants contribute to trait heritability, but are often missed by GWAS studies due to insufficient power. Detecting the association in the general European population would have required an estimated sample size of 67,000[21], as compared to the 809 Old Order Amish and 1,256 MANOLIS samples in which it was detected. To attain power to test for rare variant effects, methods have been developed that allow for the effects of multiple rare variants to be tested cumulatively. Most commonly used are burden tests, kernel-based tests such as SKAT[22], and methods that unify the two approaches, such as SKAT-O[23]. The main difference between burden tests and kernel-based tests is that burden tests assume that all variants contribute to the phenotype in the same direction; whereas, kernel-based tests allow for varying effect sizes and directions. Rare variant association studies in large cohorts confirm that rare variants contribute to disease risk independently from common variation, and have implicated genes not detected in the typical single point GWAS[24–26].

1.2.3 Meta-analyses

Increasing sample size is another straightforward method to improve statistical power. The quickest way to achieve this today is through meta-analysis, which allows the aggregation of GWAS results of multiple independent studies on the same trait without the need for individual-level genotype data. Large-scale meta-analyses were pioneered by a study on type 2 diabetes (T2D)[27], where the authors combined GWAS results from three large cohorts (WTCCC, DGI, FUSION) to obtain a sample size of more than 10,000. They identified six novel loci associated with type 2 diabetes that replicated in almost 80,000 additional samples. This was a substantial increase to the ten established T2D loci at the time. In addition to increased power, meta-analysis provides replication: false signals from a single cohort are likely to be attenuated when integrated with data from other cohorts, while genuine signals are augmented. The accrual of published GWAS studies in the past two decades has made meta-analysis an increasingly convenient way to accurately and comprehensively identify disease-associated variants across the allele frequency spectrum.

1.2.4 Intermediate phenotypes

Downstream interpretation of GWAS loci is a difficult task that acts as a major obstruction blocking the path towards clinical translation. Functional variant annotation can be improved by complementing GWAS with the analysis of intermediate phenotypes. Intermediate phenotypes are known as such because they lie in between disease outcomes and their causal variants. These are typically measured, quantitative parameters that are often perturbed in disease; like body mass index, waist-hip ratio, blood pressure, and kidney function markers. With the development of multi-omic technologies, molecular traits such as methylation, gene expression, metabolite, and protein levels can now also be studied as intermediate traits. Genetic variants associated with quantitative molecular traits are known as quantitative trait loci (QTLs), and provide functional information behind disease-causing variants, shedding light on elusive pathways and interactions in disease. This thesis focuses on protein levels as an intermediate phenotype, which we elaborate on in Chapters 1.3 and 1.4.

1.3 Protein levels as an intermediate phenotype

Proteins make up a class of macromolecules essential for life, and are vital for a diverse range of biological processes as enzymes, antibodies, hormonal proteins, transport proteins, structural proteins, et cetera. This diversity is made possible due to the unique three-dimensional (3D) structures of proteins, for which hundreds of millions of configurations exist[28] . In humans, protein synthesis begins with DNA, which is transcribed to messenger RNA, followed by translation. During translation, a specific sequence of amino acids is joined together in a polypeptide chain to form the primary structure. The secondary and tertiary structures are then formed when chemical groups on the amino acids interact with each other to produce a 3D configuration. This determines the binding sites and chemical properties of a protein molecule and is therefore central to its function. Modifications to protein structure, either through changes in the encoding gene or post-translational processes, can alter protein function and disrupt normal biological processes, causing disease.

Besides changes to their structure, abnormal protein abundances are also a risk factor for disease. Numerous processes from mRNA production to protein degradation work together to regulate protein abundances at the steady state and in response to system perturbations[29]. Being dynamic and quantifiable traits makes them useful biomarkers that can be used to predict, diagnose, or monitor disease. Many proteins are already part of common blood tests: immunoglobulin tests measure antibody levels in the blood and are used to detect infections or autoimmune conditions; high levels of creatinine kinase can indicate inflammation in the heart; abnormal fasting insulin levels can indicate prediabetes; and more. As an intermediate phenotype, finding associated loci through pQTL analyses and overlapping these with existing disease GWAS through various causal inference tools can help to fine-map disease loci, reveal novel pathways, and identify clinical biomarkers and drug targets.

1.3.1 Quantifying the circulating proteome

Opportunities for biomarker discovery can be maximised by using a hypothesis-free approach that covers as much of the human proteome as possible. The human proteome refers to the entire set of proteins expressed by the human genome; this amounts to approximately 20,000 canonical proteins, corresponding to the ~20,000 genes in the genome. In reality, the proteome differs between tissues due to the different biological processes occurring in each. This thesis is focused on the circulating proteome, which is composed of immunoglobulins and proteins that have been secreted, leaked from damaged cells, or shed into the bloodstream. Because circulating proteins may originate from anywhere in the body, abnormal levels can reflect disturbed biological processes in any tissue or organ, offering a wealth of systemic information. As of 2021, ~4,395 canonical plasma proteins had been catalogued as part of the Human Plasma Proteome Project (HPPP)[30,31].

Mass-spectrometry (MS) and affinity-based assays are the two most commonly used protein quantification methods today. Quantifying the circulating proteome, however, is a challenge because of the extreme range of abundances in which the proteins are present. Only 20 proteins account for almost 99% of the blood proteome, compared to 2,500 proteins in human cells[30,32]. Typical MS-based approaches can only capture ~500 circulating proteins[33], with those that are able to capture more proteins requiring impractically extensive preparation at high costs[34,35]. Affinity-based approaches such as that offered by Olink (https://olink.com/) and Somalogic (https://somalogic.com/) are newer alternatives that are able to detect low abundance proteins in plasma and serum samples. The latest assays offered by Olink and Somalogic are able to quantify ~3,000 and ~7,000 proteins, respectively. This provides pQTL studies greater access to the circulating proteome, enabling more thorough genetic characterisation.

1.3.2 Interpreting pQTLs

Protein levels are regulated in response to changes in their environment and interact often with other molecules. This means that they may be influenced by changes in multiple genes and are complex traits themselves. In a pQTL analysis, a common initial step is to designate pQTLs as either *cis* or *trans* acting (Figure 4). *Cis*-pQTLs are variants located within or nearby (usually defined as between 1 to 2 million bases) the gene encoding the target protein; while *trans*-pQTLs are variants located outside of the defined *cis* region, including variants on entirely different chromosomes. This distinction allows us to glean different types of information on the genetic regulation of the target protein.

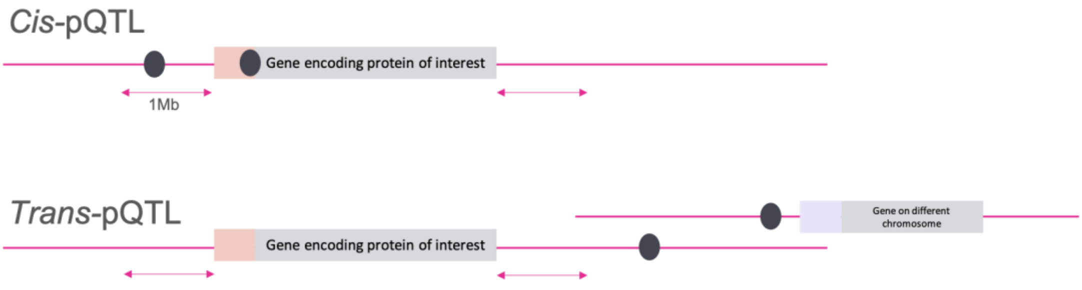

Cis-pQTL
Gene encoding protein of interest
1Mb
Trans-pQTL
Gene encoding protein of interest
Gene on different chromosome

Figure 4. Cis- and trans-pQTLs. The ovals represent pQTLs.

Due to their proximity to the encoding gene, many *cis*-pQTLs influence protein expression directly through transcriptional regulation. Evidence of this may be obtained by overlapping the *cis*-pQTL with a corresponding *cis*-eQTL (gene expression QTLs) in a colocalisation analysis (Figure 5), where positive colocalisation would indicate that *cis*-pQTL also influences mRNA levels in particular tissues.

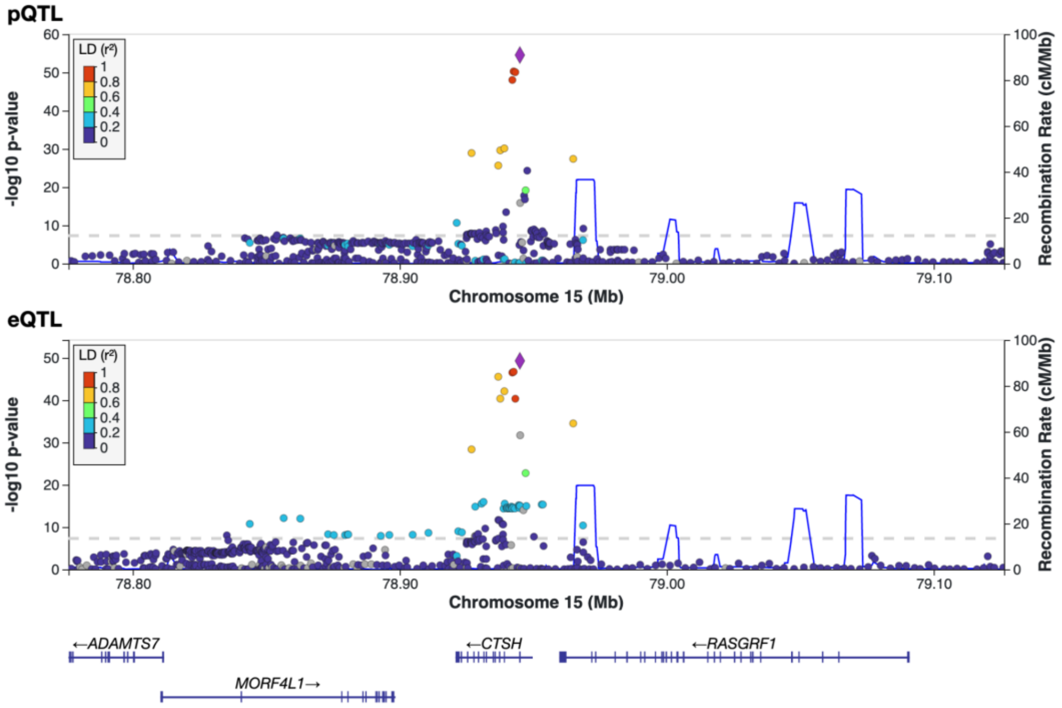


Figure 5. Visualising colocalisation between a pQTL (top panel) and an eQTL (bottom panel) using regional plots showing association signal architecture. The example shows the variant at chr15:78944951 (purple diamond), which is both a *cis*-pQTL for cathepsin H (CTSH) serum levels[36] and an *cis*-eQTL for *CTSH* gene expression in thyroid tissue (GTEx)[37].

Meanwhile, *trans*-pQTLs reflect gene or protein interactions and offer insight into molecular pathways. In a study by Folkersen et al.[38], the most likely causal genes were determined for each *trans*-pQTL using a combination of eQTL colocalisation testing, network analysis, and text mining. The authors describe a pathway comprising FOXO3, AKT1, and NGF; identifying *FOXO3* as the causal gene for the *trans*-pQTL associated with NGF levels, despite the nearest gene being *LACE1*. The three proteins are part of an apoptosis regulation pathway in neurons that has been implicated in neuropathies[39,40]:

> Reduced NGF → Inhibits Akt activity → Dephosphorylates FoxO → Activates pro-apoptotic genes

The example demonstrates the potential of *trans*-pQTL analyses to advance our current understanding of pathways relevant to disease. Additionally, existing studies have identified several pleiotropic *trans* genes that influence the levels of multiple proteins[41–46]; namely, *ABO, CFH, HLA, F12, FUT2, ST3GAL6, KLKB1,* and *VTN*. Four of these genes (*ABO, FUT2, F12, KLKB1*) are involved in blood coagulation pathways, while *CFH* and *HLA* play a role in inflammatory response.

1.3.3 Existing pQTL studies

The publication of pQTL studies picked up speed after 2016, corresponding to the time affinity-based technologies by Somalogic and Olink were made available. Before this, most pQTL studies were performed using mass spectrometry or targeted proteomic panels. From an list curated by Suhre et al.[47] (http://www.metabolomix.com/a-table-of-all-published-gwas-with-proteomics/), most studies have assayed between 80 and 1,500 proteins, in sample sizes between 100 and 5,000. Sample sizes have increased over the years – up to 50,000 – as seen most recently in a manuscript by Sun et al. in the UK Biobank cohort[48], as well as in a

2021 study by Ferkingstad et al. in more than 35,000 Icelanders[49]. Increasing sample size generally increases the pQTL discovery rate, likely owing to increased power. This was particularly so for *trans*-pQTLs, which tend to have smaller effects. Comparing two studies by Folkersen et al. that analysed roughly the same Olink proteins, 41 cis- and 38 trans-pQTLs were detected in the smaller study with 3,394 samples[38], as compared to 75 cis- and 326 trans-pQTLs in a larger meta-analysis with 21,758 samples[43]. This stresses the importance of larger sample sizes to provide a more complete understanding of protein level heritability. We additionally note that almost all studies have used genotype array data in European populations with little focus on rare (MAF < 1%) variants, highlighting the need for sequencing-based studies and greater ethnic diversity.

1.4 Connecting pQTLs to health and disease

Clinical translation is an eventual goal for most pQTL studies, and key to achieving this is the successful identification of genes, proteins, and networks that lie on the causal pathway to disease. By leveraging existing knowledge, pQTLs can be used to discover clinical biomarkers, validate drug targets, and construct disease risk scores. In this section, I will elaborate, using examples, on the various downstream methods incorporating pQTLs that accelerate clinical translation.

Identifying causal genes is a major limitation obstructing the path from GWAS to clinical translation. To overcome this, overlaps between pQTLs and disease-associated loci can help to pinpoint causal genes. This is normally done through a test of colocalisation, which evaluates if a pQTL signal and a GWAS signal share the same causal variant. *Cis*-pQTLs are particularly useful for this purpose – because we are often able to map *cis*-pQTLs to their causal genes with confidence, positive colocalisation with a GWAS signal would indicate the same causal gene for the GWAS signal. For instance, Pietzner et al.[50] identify *PRSS8* as the causal gene for a known Alzheimer's disease (AD) locus, based on colocalisation between the *cis*-pQTL for PRSS8 and the AD signal. The causal variant lies within the adjacent *KAT8* gene in a gene-rich region, and prior efforts to identify a causal gene had been unsuccessful.

Complementing colocalisation analysis is Mendelian randomisation (MR), which is used to identify proteins whose levels are causal for – not just correlated with – disease. A significant protein-disease MR association indicates that the disease is directly influenced by changes in the target protein abundance, and has been described as the equivalent of a randomised clinical trial[47] for estimating causal effects between an exposure and an outcome. In two-sample MR, analysts are allowed to use exposure or outcome GWAS data from other studies to be tested against their own; this includes data from large consortia, which can substantially increase statistical power and widen the scope of discovery. For example, Sun et al. confirmed a causal protective role of PSP-94 in prostate cancer using two-sample MR[45]. Reduced serum levels of the protein had previously been associated with the development of male prostate cancer, but it was unknown if the association was causal. Mendelian randomisation is able to resolve correlation from causation; this is an important distinction that acts as a crucial filtering step in selecting effective biomarkers and drug targets.

Both MR and colocalisation evidence are valuable for drug target discovery and prioritisation. Human genome-derived proteins make up almost 75% of known molecular drug targets[51], making the proteome a valuable source of novel therapeutics. Genetic support has been estimated to double the success rate of drug targets in clinical development[52], and even more so when causal genes are known[53]. Genes and proteins with causal links to disease may also be overlapped with drug databases to identify new indications that approved drugs may be repurposed for, which can dramatically expedite the drug development process.

Biomarker discovery can also be achieved through polygenic scores, which aggregate the effects of alleles across the genome to estimate an individual's genetic predisposition to a particular trait. Like disease, protein levels can also be genetically predicted using polygenic models trained on pQTL data. Association tests between predicted protein levels and disease occurrence in large samples can then be performed, presenting a powerful approach for biomarker discovery. This was demonstrated by a polygenic model explaining almost 20% of variance in circulating ST2 levels, that was used to predict ST2 levels in a cohort of >300,000 individuals (UK Biobank)[43]. Calculated scores were strongly associated with asthma and inflammatory bowel disease, highlighting ST2 as a biomarker for the two conditions.

The popularity of polygenic scores is founded upon its ability to identify low- or high-risk individuals within a population[54]. In the same way, protein polygenic scores can be used to identify individuals with especially high or low predicted levels of a target protein. The scores may be incorporated into existing risk models to improve predictive accuracy. For proteins that act as drug targets, they can furthermore identify patients that are likely to benefit most from the targeting drug. Growing pQTL data alongside GWAS data is, therefore, key to improving polygenic models, which will create more opportunities to further our understanding of disease and translate this knowledge into patient benefit.

1.5 Aims

For complex diseases, the traditional GWAS is often limited in the information it can convey by itself. In this thesis, we will analyse serum protein levels as an intermediate trait to seek clarification into complex disease aetiology. Using multiplexed proteomic and whole genome sequencing data in isolated populations, we aim to:

(1) Comprehensively describe the genetic regulation of serum proteins levels:

This will be achieved through genome-wide association tests to identify independently-associated *cis-* and *trans-*acting pQTLs. Using available tools, we will perform functional annotation of the detected pQTLs; characterise heritability; and carry out eQTL colocalisation tests to give further insight into our findings. Because the cohorts analysed are part of isolated populations that have been whole genome-sequenced, particular attention will also be given to rare variants that may have drifted up in frequency. The pQTLs will be made available for the public to download, contributing to the growing database of pQTLs that can be used to empower future research.

(2) Further understanding of the complex disease aetiology through pQTLs:

In this thesis, we will focus on two broad classes of complex disease: neurological and cardiometabolic. Methods that allow us to leverage existing GWAS data – namely, colocalisation tests and two-sample Mendelian randomisation – will be used to identify genes and proteins causally associated with relevant diseases. By incorporating the information acquired from these analyses with existing literature, we aim to highlight biomarkers with potential for clinical use; describe pathological pathways; and identify drug repurposing opportunities.

2 Methods

2.1 Study design

Serum proteins from five Olink panels were analysed in two phases: 184 proteins from the Neurological and Neuro-exploratory panels were first analysed with a focus on neurological disease (Chapter 3.1); and 276 proteins from the Cardiovascular II, Cardiovascular III, and Metabolism panels were analysed in the second phase, with a focus on cardiometabolic disease (Chapter 3.2). Figure 6 provides an overview of the analysis pipeline for each phase. In the rest of this chapter, I will provide further detail on the methods that were implemented at each step of the analysis.

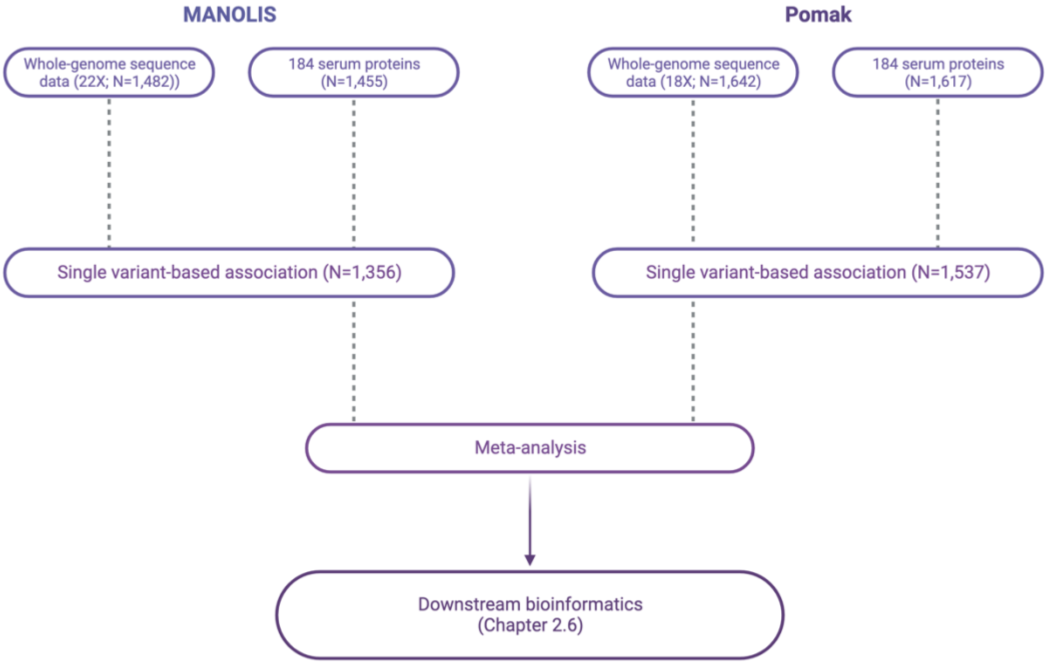

MANOLIS
Pomak
Whole-genome sequence data (22X; N=1,482))
184 serum proteins (N=1,455)
Whole-genome sequence data (18X; N=1,642)
184 serum proteins (N=1,617)
Single variant-based association (N=1,356)
Single variant-based association (N=1,537)
Meta-analysis
Downstream bioinformatics (Chapter 2.6)

Figure 6. Overview of analysis pipeline. The numbers provided are from the first phase of analysis, focusing on 184 proteins from the Neurological and Neuro-exploratory Olink panels.

2.2 Study cohorts

The HELIC (HELlenic Isolated Cohorts) study comprises two Greek cohorts, MANOLIS and Pomak. The MANOLIS (Minoan Isolates) collection contains data from individuals from the mountainous region of Mylopotamos, on the island of Crete in southern Greece[55]. Of the total population of ~5,200 (2011), 29.7% (N=1,553) participated in the study. The mean age of the participants was 61.6 years, 55.8% of whom were female. The Pomak collection, on the other hand, is made up of individuals from the Pomak villages located in the Xanthi regional unit in northeastern Greece[56]. These individuals are part of the small minority of Muslims in Greece with uncertain historical origins. Of the estimated population of 25,000, 1,702 participated in the HELIC study. The mean age was 44.9 years, with 69.3% of participants being female.

Compared to the general Greek population, MANOLIS and Pomak have genomic characteristics that establish them as genetic isolates, such as more extensive haplotype sharing, genetic drift, and enrichment of missense variants in the MANOLIS cohort[57]. For both cohorts, detailed phenotypic information have been collected, including anthropometric, biometric, biochemical and haematological blood measurements, medical history, demographic, socioeconomic, lifestyle, and dietary information[55,56]. Genome-wide association studies for these traits have also been performed[58,59], and serum measurements for 460 proteins were generated in 1,455 MANOLIS and 1,617 Pomak individuals using the Olink proximity extension assay (PEA).

2.3 Olink proteomic quantification

460 proteins in 1,455 MANOLIS and 1,617 Pomak serum samples were quantified using Olink's proximity extension assay (PEA; https://olink.com/). These proteins composed the Neurology, Neuro-exploratory, Cardiovascular II, Cardiovascular III, and Metabolism panels, which were designed by Olink based on existing literature. The PEA method involves three steps: (1) the immunoreaction step, where a pair of antibodies labelled with complementary oligonucleotides first binds specifically to a target protein. Binding of the antibody pairs brings the oligonucleotides in close proximity, causing them to hybridise. This leads to (2) the extension step, where a double-stranded barcode unique to the target protein is formed. Larger quantities of

the protein are, therefore, reflected in larger quantities of the barcode, which are quantified in (3) the amplification and detection step, using real-time quantitative polymerase chain reaction (qPCR). Olink's PEA is distinguished by its use of antibody pairs and PCR amplification, which enables highly specific protein quantification even with low sample volumes.

Relative protein quantities are reported as arbitrary Normalised Protein Expression (NPX) values, where an increase in 1 NPX means a doubling of protein concentration. NPX values are calculated based on the qPCR cycle threshold (Ct) values and various internal and sample controls, according to the following equations:

$$\mathrm{Ct}_{\mathrm{Analyte}} - \mathrm{Ct}_{\mathrm{Extension\ Control}} = \mathrm{dCt}_{\mathrm{Analyte}}$$

$$\mathrm{dCt}_{\mathrm{Analyte}} - \mathrm{dCt}_{\mathrm{Inter\text{-}plate\ Control}} = \mathrm{ddCt}_{\mathrm{Analyte}}$$

$$\mathrm{Correction\ factor} - \mathrm{ddCt}_{\mathrm{Analyte}} = \mathrm{NPX}_{\mathrm{Analyte}}$$

The extension control is used to adjust for intra-assay variability introduced during extension; inter-plate controls are used to adjust for interplate variability; and the correction factor is a value derived from negative controls to adjust background values to approximately zero. Following NPX calculations, all samples with internal controls deviating by more than 0.3 NPX from the median are then excluded as failures. Explanations for each type of control are provided in greater detail in Olink's white paper (https://www.olink.com/content/uploads/2022/04/white-paper-data-normalization-v2.1.pdf).

2.4 High-depth whole-genome sequencing and variant calling

The genomic DNA of 1,482 MANOLIS and 1,642 Pomak samples were subject to the same high-depth whole genome sequencing (WGS), alignment, and variant calling pipeline. WGS was performed using Illumina's HiSeqX platform, and reads were aligned with the hg38 reference genome (GRCh38). Variant calling and genotyping were then carried out using the HaplotypeCaller and GenotypeGVCFs tools from the Genome Analysis Toolkit (GATK) to produce a cohort-wide variant call format (VCF) file. VCF files further underwent variant-level QC using GATK's variant quality score recalibration tool (VQSR); and sample-level QC, which excluded 25 and 27 samples from MANOLIS and Pomak, respectively. Other intermediate steps and QC criteria have been described in detail in publications[36,60–62]. Altogether, 1,455 MANOLIS and 1,617 Pomak samples had both WGS and Olink data.

2.5 Association and meta-analysis

2.5.1 Single variant association analysis

Protein QTLs are identified by checking for significant associations between genotypes and target protein measurements. For quantitative phenotypes such as protein measurements, linear models are used to test for associations. The basic linear regression model, however, assumes that all observations are independent of each other. This is problematic in isolated populations such as MANOLIS and Pomak, whose genetic isolation results in high relatedness between individuals. For this reason, we use linear mixed models (LMMs) that are able to account for relatedness and population structure by incorporating a relatedness matrix. An LMM models a quantitative phenotype y as:

$$y = W\alpha + x\beta + Zu + \varepsilon$$

Where y is an $n \times 1$ vector of phenotype values and n is the number of individuals; W is an $n \times c$ matrix of covariates with fixed effects, where c is the number of covariates and α is a $c \times 1$ vector of corresponding coefficients; x is an $n \times 1$ vector of marker genotypes, β is the effect size of the marker; Z is an $n \times m$ loading matrix where m is the number of strains ($m = n$ for human studies), u is an $m \times 1$ vector of random effects and follows the multivariate normal distribution $MVN_m(0, \lambda\tau^{-1}K)$, where τ^{-1} is the variance of residual errors, λ is a scale factor that is the ratio between the two variance components, K is an $m \times m$ relatedness matrix; and ε is an $n \times 1$ vector of errors that follows the distribution $MVN_n(0, \tau^{-1}I_n)$, where I_n is an $n \times n$ identity matrix.

In the analyses presented in this dissertation, we perform association testing using the LMM algorithm implemented in GEMMA[63], which uses the full relatedness matrix K to calculate exact association statistics with high computational efficiency.

2.5.2 Meta-analysis

Meta-analysis combines the results across multiple association studies with the main objective of increasing statistical power. Using just summary statistics, the approach allows for the proper adjustment of study-specific covariates pre-meta-analysis, while offering power equivalent to that of a mega-analysis (where individual-level data across studies are first pooled before running the association tests)[64]. Methods for meta-analysis are classified into two broad families: those that operate under the "fixed effects" model, which assumes that the true underlying effects are equal across studies; or the "random effects" model, which allows for effects to vary between studies. With any model, the meta-analysis p-value may be calculated using either a p-value-based or inverse variance-based (IV) strategy. Whereas the p-value-based strategy only calculates a signed Z-score and p-value, the IV-based strategy outputs a Z-score (Z), effect size (β), standard error (SE), and p-value (P) and is more informative. It has, additionally, been shown to be more robust to between-study differences in allele frequencies[65], making it the preferred option for many researchers. Under the fixed effects model, the meta-analysis association statistics for each variant under the IV-based strategy are calculated as given[66]:

Intermediate statistics for each study i:

$$w_i = \frac{1}{se_i^2}$$

Meta-analysis statistics:

$$SE = \surd(1/\sum_i w_i)$$

$$\beta = \frac{\sum_i \beta_i \, w_i}{\sum_i w_i}$$

$$Z = \frac{\beta}{SE}$$

$$P = 2\phi(|-Z|)$$

Where w_i is the weight given to each study i, se_i is the standard error for each study i, β_i is the effect size for each study i, and β is the overall effect size.

2.5.3 Multiple testing correction

The p-value is most often used to denote significance in statistical testing. A threshold of P<0.05 is commonly applied; this means that we can reject the null hypothesis (H_0) if there is less than a 5% probability of observing a result as extreme as the observed result, under the conditions of H_0. In the case of any GWAS, H_0 states that there is no association between the observed genotype and phenotype. A p-value that falls below the decided threshold can therefore be used to reject H_0, implying a significant association between genotype and phenotype.

When carrying out multiple tests simultaneously, the p-value threshold must be adjusted to control for the type I error rate. GWAS involves the testing of millions of variants across the genome at the same time – and in the case of a large-scale pQTL analysis – with hundreds of protein measurements, amounting to a huge number of simultaneous tests. Bonferroni correction – where the p-value threshold is adjusted by dividing by the total number of tests – is a popular method to correct for multiple testing. However, in the context of a pQTL analysis, this is overly conservative as it does not take into account LD between variants, nor correlation between proteins.

Considering these two factors, we adjust only for the effective number of variants (N_{eff}) and proteins (M_{eff}) to arrive at a threshold of P < $0.5/(N_{eff} \times M_{eff})$.

N_{eff} was calculated by pruning of the variants to produce a set of variants in approximate linkage equilibrium with each other, using the "--indep" command implemented in Plink (https://www.cog-genomics.org/plink/ . M_{eff} was derived using Cheverud's formula[67], which is based on the principle that the variance ($V_{\lambda obs}$) of the eigenvalues derived from the (trait) correlation matrix is proportional to the correlation between traits; therefore, the proportional reduction in the number of traits can be calculated as the ratio of the $V_{\lambda obs}$ to the maximum variance (equal to the total number of traits) M, as $V_{\lambda obs}/M$. This produces the equation $M_{eff} = (1 - (M - 1)V_{\lambda obs}/M^2)$, where M_{eff} ranges from 1 to M.

2.5.4 Rare variant meta-analysis

Gene-based testing provides increased statistical power to detect rare variant associations, by aggregating rare variants within a gene and testing for associations on the gene level. Burden testing, kernel-based, (e.g. SKAT), and unified approaches (e.g. SKAT-O) are popular methods used for rare variant analysis in single cohorts. Until recently, however, existing meta-analysis methods had been unable to account for relatedness between subjects from multiple sources. The software package GMMAT (https://github.com/hanchenphd/GMMAT) implements a computationally efficient meta-analysis framework, SMMAT.meta[68], that provides an interface to SKAT-O and allows the specification of relatedness matrices and variant weights. Briefly, single-variant scores and covariance matrices for low-frequency to rare variants (MAF < 5%) are first computed using SMMAT, and then meta-analysed to produce a SKAT-O association p-value for every gene.

2.6 Downstream bioinformatics

With the help of existing bioinformatics tools and databases, we are able to achieve a better understanding of the genetic architecture underlying serum protein levels, and connect the proteome to health and disease using pQTLs. The following table (Table 1) is a list of databases and their purposes used in our analysis, with further details on colocalisation analysis and Mendelian randomisation in the next sections.

Table 1. Descriptions of databases used in this thesis.

Database [URL]	**Description and purpose**
Ensembl [https://www.ensembl.org/]	Ensembl is a genome browser that also incorporates information from several other databases/tools. Ensembl is particularly useful for pQTL annotation, to: 1. Map pQTLs to their nearest genes; 2. Get variant consequences and their predicted deleteriousness using the variant effect predictor tool (VEP); 3. Compare allele frequencies between populations.
GTEx [https://gtexportal.org/]	Short for genotype-tissue expression, the GTEx portal contains information on gene expression QTLs (eQTLs) in different human tissues. pQTL-eQTL colocalisation can: 1. Validate cis-pQTLs; 2. Identify causal genes for trans-pQTLs; 3. Identify specific tissues from which a pQTL may originate [Pietzner]
PhenoScanner [http://www.phenoscanner.medschl.cam.ac.uk/]	PhenoScanner is a database of summary statistics from large-scale GWAS. A command line tool is available to retrieve GWAS summary statistics that overlap queried genomic variants (e.g., a pQTL) or regions. Colocalisation between pQTLs and GWAS signals can: 1. Connect protein to disease 2. Identify causal genes for GWAS signals (see Chapter 1.4) 3. Support MR associations
STRING [https://string-db.org/]	STRING is a database of protein-protein interaction networks that aggregates information from other primary databases. STRING confidence scores may be used to prioritise causal genes or identify possible mediator genes for trans-pQTLs [10.1371/journal.pgen.1006706]
MR-Base/OpenGWAS [https://www.mrbase.org/]	Similar to PhenoScanner, MR-Base is a database of summary statistics from large-scale GWAS, developed for use in two-sample Mendelian randomisation using the connected TwoSampleMR R package (or web app). Using pQTLs as instruments, two-sample MR is able to deduce causal relationships between protein levels and disease.
Open Targets [https://www.opentargets.org/] DrugBank [https://go.drugbank.com/]	Open Targets and DrugBank are both drug databases that contain information on targets for drugs that are approved or in clinical trials. This information can be used to identify repurposing opportunities or potential side effects of drugs targeting proteins that have been causally linked to disease via two-sample MR.

2.6.1 Colocalisation analysis

A colocalisation test answers the question of whether any two independent association signals share a causal variant, which increases the probability that the two traits share a causal mechanism. A pQTL can be tested for colocalisation with a signal from any other trait, including gene expression (eQTLs) and disease (as discussed in this thesis), or other protein traits. Colocalisation differs from a simple overlap or visual comparison by also taking into account local LD patterns, thereby reducing the chances of accidental overlap (Figure 5). While several methods have been developed[69], the *coloc* method[70] is able to test for colocalisation using only the allelic effect and standard errors of local variants at the associated locus, without the need for individual-level genotype data or prior selection of variants (hence avoiding the "winner's curse" resulting in the overestimation of effect sizes).

For every locus across the two tested traits, *coloc* calculates five posterior probabilities (based on user-defined priors), each corresponding to one of five hypotheses:

- H_0: No association with either trait
- H_1: Association with trait 1 but not trait 2
- H_2: Association with trait 2 but not trait 1
- H_3: Association with both traits, but with distinct causal variants
- H_4: Association with both traits with a shared causal variant (positive colocalisation)

Briefly, for each trait, every variant within the locus is assigned a value of 0 or 1, where 1 indicates a strong association with the trait based on the given summary statistics. This produces a pair of binary vectors of (0,1) values corresponding to each trait, which are then assessed for its support for each hypothesis (Figure 7).

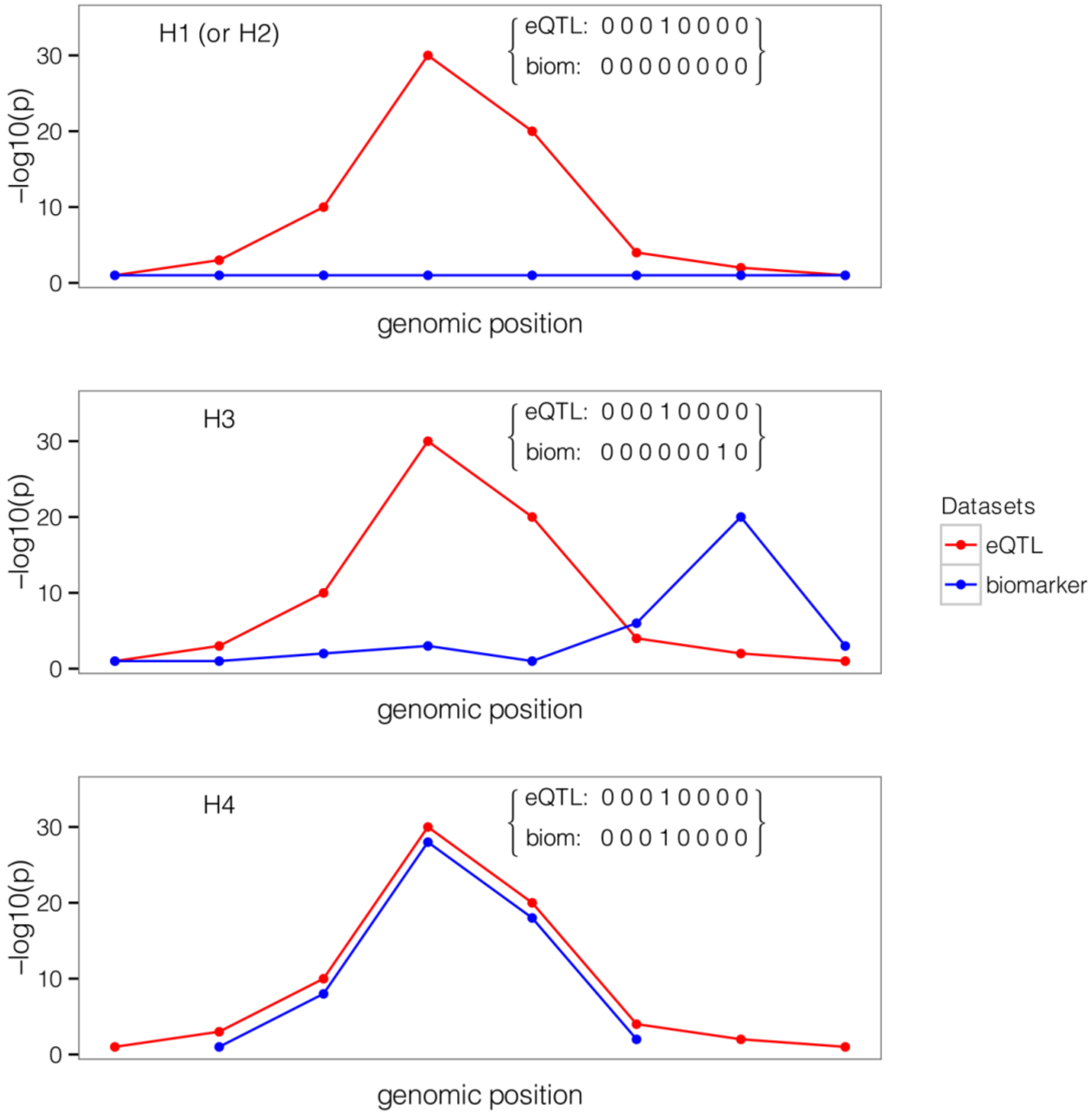


Figure 7. Example of one configuration under different hypotheses by Giambartolomei et al. (2014), used under CC BY 4.0. In this example, "eQTL" and "biomarker" refer to association signals from the two tested traits. The y-axis represents the -log10 of the association test p-values, with each point representing a variant.

The results are, however, valid only under the assumptions that (1) samples for both traits have the same ancestry; (2) for each trait, the phenotype and genotype are linearly related; (3) the causal variant is included in the set of tested variants; and (4) only one association (i.e. causal variant) is present in the region of interest. Where multiple causal variants are present (as for many pQTLs), *coloc* can be run for each independent signal by using p-values that have been conditioned on all other signals in the region.

2.6.2 Mendelian randomisation

Mendelian randomisation (MR) aims to evaluate the causal effect of an exposure on an outcome. Discerning correlation from causation is, however, difficult due to unknown confounding factors. MR overcomes this by estimating the effect of an exposure (e.g., protein level) on an outcome (e.g., disease) using proxies – known as the "instruments" – that are least likely to be related to potential confounding factors. In MR, these instruments are the genetic variants associated with the exposure, since genotypes are randomly assigned to individuals during meiosis (according to Mendel's laws).

The traditional MR (one sample MR) study requires both the exposure and outcome data to be from the same samples; but because this is often not possible, two sample MR was developed to allow for different study samples to be used for the either dataset[71]. In most cases, this increases statistical power as large datasets from consortia can be used. The power of MR relies heavily on three assumptions:

1. "Relevance": The genetic variants (the instruments) are associated with the exposure
2. "Independence": There are no confounders of the genetic variant-outcome association
3. "Exclusion restriction": The genetic variants influence the outcome only through the exposure

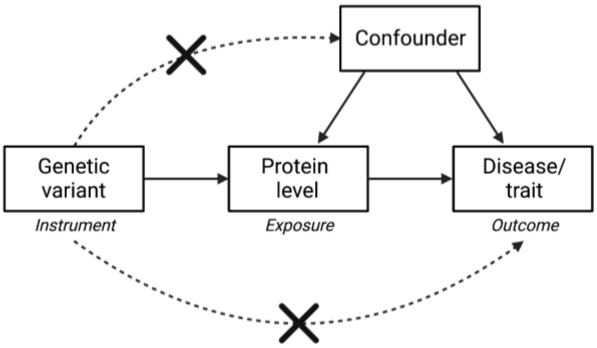


Figure 8. Representation of valid instrumental variables and potential violations of the core assumptions (dotted arrows). The effect of the genetic variant(s) on the exposure and outcome must not be mediated by confounders, which is a violation of the independence assumption (top dotted arrow). The genetic variant(s) must also not influence the outcome directly, which is a violation of the exclusion restriction (bottom dotted arrow).

Several guides have been published suggesting steps to ensure the robustness of an MR study[72,73]. In short, special care should be taken when choosing genetic instruments – multiple genetic instruments should be used where possible to ensure sufficient power, but they must not be pleiotropic. Additionally, sensitivity analyses should be performed using alternative MR methods with relaxed assumptions (e.g., MR Egger[74]); tests of heterogeneity (e.g., Cochran's Q statistic) to ensure concordant effects across instruments; or leave-one-out analyses to remove variants with dominating effects.

3 Summary of contributed publications

3.1 Mapping the serum proteome to neurological diseases using whole genome sequencing

(See Appendix A)

Background: The global burden of neurological disorders has been increasing continually over the last two decades. Despite this, few treatment options exist and diagnosis is often challenging due to the heterogeneous and overlapping nature of neurological diseases. There is an urgent demand for novel treatment strategies and biomarkers that can detect early disease; and an important source of biomarkers are circulating proteins, which are accessible, quantifiable, and actionable targets whose abundances are often perturbed in disease.

Methods: Understanding the genetic basis of circulating proteins can bridge knowledge gaps by resolving GWAS signals and uncovering causal pathways and potential clinical biomarkers. Here, we aimed to identify genetic variants influencing the levels (protein quantitative trait loci [pQTLs]) of 184 serum proteins, using whole genome sequence data from two Greek isolated population-based cohorts (N=2,893), MANOLIS and Pomak. The analysed proteins were quantified using Olink's proximity extension assay, and comprised the Neurological and Neuro-exploratory Olink panels. We integrated our findings with existing neurological disease GWAS data, using colocalisation analysis and two-sample Mendelian randomisation, to identify causal protein-disease relationships.

Results: We detected 214 independently-associated pQTLs (162 cis; 52 trans) for 107 proteins. Excluding pleiotropic loci, 87 pQTLs (40%; 72 cis; 15 trans) were independent from previously-reported pQTLs and were hence defined as novel. The remaining pQTLs have been associated in previous pQTL studies, either directly or through linkage disequilibrium (LD), validating previous findings and providing proof of concept. A large majority (70%) of pQTLs were located in intergenic regions or introns, in addition to 36 pQTLs overlapping regulatory features. Through colocalisation and two-sample Mendelian randomisation analysis, we identified 25 causal protein level-disease relationships, three of which we highlighted:

(1) Transmembrane glycoprotein NMB (GPNMB) and Parkinson's disease (PD)

We observed positive colocalisation between a cis-pQTL for GPNMB and a known locus associated with PD[75,76]. GPNMB is a glycoprotein that is involved in different cell functions, including neuroinflammation[77]. The relationship between GPNMB and PD is supported by experimental evidence showing that GPNMB is increased in both the brain[78] and plasma[79] of PD patients. Further, we observed colocalisation between the serum pQTL and eQTLs in whole blood and several brain tissues (basal ganglia, cortex, and anterior cingulate cortex), indicating shared regulatory mechanisms. Our findings suggest serum GPNMB as a potential biomarker for PD and provides genetic evidence for GPNMB as a therapeutic target for PD.

(2) Siglec-33 (CD33) and Alzheimer's disease (AD)

Through both colocalisation and two-sample MR, we observed evidence for a causal relationship between increased CD33 and AD risk. Genetic associations between variants in *CD33* and AD have been replicated in several GWAS studies[3,80–82]. CD33 is a member of the immunoglobulin superfamily that is expressed primarily on myeloid cells. The protein modulates brain microglial activation by inhibiting phagocytosis, and its role in AD is well-studied[83–85]. Drug repositioning efforts have, additionally, shown that the FDA-approved anti-CD33 drug, lintuzumab, was able to effectively downregulate cell surface expression of CD33 *in vitro*[86]. Our results provide validation and further genetic support for anti-CD33 drugs for AD treatment.

(3) Macrophage scavenger receptor I/II (MSR1) and schizophrenia

We observed an inverse relationship between MSR1 and schizophrenia risk, indicating a possible protective role for MSR1. The protein mediates the phagocytosis of toxic molecules, with known protective effects against bacterial infections, AD, and atherosclerosis[87]. MSR1 depletion in mice has also been shown to cause deteriorating working memory and dysregulated immune response, further supporting a brain protective mechanism[88]. The causes of schizophrenia are not known, and the role of MSR1 in schizophrenia has not been described. These findings provide evidence for a novel pathway in schizophrenia development, where reduced expression of MSR1 may cause the accumulation of toxins and damage signals in the brain, resulting in excess inflammation and changes in brain function that lead to schizophrenia.

Conclusion: We described the genetic architecture of 107 serum proteins relevant to neurological processes, 15 of which have not previously been investigated before. We detected both known and novel pQTLs, thereby validating previously published findings and contributing new knowledge. Using three examples – GPNMB and PD; CD33 and AD; and MSR1 and schizophrenia – we showed how pQTL studies can identify clinical biomarkers, uncover drug repurposing opportunities, and reveal novel disease pathways, respectively. All results from the colocalisation and MR analyses serve as a starting point for further experiments that will help to increase our understanding of neurological disease aetiology exponentially.

Contributions: My contributions to this work include the preparation of the proteomic data, all described analyses, and interpretation of the results. The first draft of the manuscript was written by me and revised by Prof. Dr. Eleftheria Zeggini, including all presented figures and tables.

3.2 Identifying causal serum protein-cardiometabolic trait relationships using whole genome sequencing

(See Appendix B)

Background: Cardiovascular and metabolic diseases are among the leading causes of mortality around the world today. The prevalence of diabetes, for example, has risen rapidly by 70% since 2000, and is responsible for an increasing number of deaths caused by comorbidities like kidney disease[89]. While genome- and proteome-wide association studies have been successful at identifying genetic variation linked to a wide variety of cardiometabolic diseases, pinpointing causal genes and proteins with the biggest predictive and therapeutic potential remains a challenge. Intermediate phenotypes such as protein levels provide biological information that, when integrated with GWAS data, can identify causal disease genes and proteins.

Methods: We performed a genome-wide association meta-analysis for 248 serum proteins to identify protein quantitative trait loci (pQTL) in 2,893 individuals (from the HELIC-MANOLIS and Pomak cohorts) with whole-genome sequence data. Relative abundances of serum proteins from the Cardiovascular II, Cardiovascular III, and Metabolism Olink panels were measured using Olink's proximity extension assay (PEA). This meta-analysis provides a substantial increase in power from our previous study (MANOLIS only; N=1,356)[60] by doubling the sample size. Colocalisation and two-sample MR analyses were applied to cardiometabolic traits to identify causal protein-disease relationships. In doing so, we found that the protein meprin A subunit beta (MEP1B) was causally associated with high density lipoprotein (HDL) levels, and systematically phenotyped a mouse model to better understand its biological role.

Results: We detected 301 pQTLs (215 cis; 86 trans) for 170 proteins that were present in both cohorts with concordant direction of effect. This is an 83% increase from our previous analysis in only the MANOLIS cohort, demonstrating the importance of large sample sizes to empower discovery. 58% of the pQTLs replicated in an independent cohort, ORCADES (N=950). Importantly, we detected 15 novel pQTLs that are rare in the general European population, but have drifted up in frequency in at least one of our discovery cohorts. This included a deleterious variant (rs144755357) for the matrix metalloproteinase-2 (MMP2) protein that had increased in frequency 95-fold in Pomak (MAF=1.3%; gnomAD Europeans MAF<0.01%). MMP2 has been linked to chronic airway diseases and cancer[90,91], exemplifying the advantage of isolated populations to detect high-impact rare variants relevant to disease.

Through causal inference analysis, we found 43 serum proteins to be causally associated with a cardiometabolic trait. The analysis revealed shared and distinct aetiology among lipid traits: LDL cholesterol, total cholesterol, and triglyceride levels were causally associated with the levels of seven common proteins; but HDL cholesterol was associated with a different set of proteins, suggesting distinct regulatory pathways from other lipid traits. Using novel pQTLs as instrumental variables, we found that decreased MEP1B, a protease, was associated with increased HDL cholesterol. Systematic phenotyping of an existing *Mep1b* knock-out mouse model showed increased body mass in female mice due to increased adiposity, confirming a metabolic role for MEP1B.

Conclusion: This work expands on previous work[60] and demonstrates the importance of larger sample sizes (meta-analysis) and isolated populations in pQTL discovery. Through a meta-analysis, we provided a more thorough characterisation of the genetics underlying serum protein levels and uncovered novel rare pQTLs. We highlighted proteins that are causally associated with cardiometabolic traits, including known (e.g., LDL receptor and LDL cholesterol) and novel protein-disease relationships (e.g., cathepsin H [CTSH] and delta like non-canonical Notch ligand 1 [DLK1] in diabetic kidney disease) that warrant further investigation. Finally, we showed – using a *Mep1b* knock-out mouse model – how pQTL analysis can inspire new hypotheses for downstream functional experiments and deliver novel biological insights.

Contributions: My contributions to this work include the preparation of the proteomic data, association and meta-analyses, colocalisation analysis, and interpretation of the results. The first draft of the manuscript was written by me and revised by Prof. Dr. Eleftheria Zeggini, including all presented figures and tables.

4 Discussion

The aims of this thesis were to describe the genetic basis of the human serum proteome, and through that, achieve a better understanding of the molecular aetiology of complex neurological and cardiometabolic diseases. Using comprehensive whole genome sequence data, we showed that underlying genetic variation varies greatly from protein to protein across the full allele frequency spectrum, and demonstrated the value of isolated populations in detecting rare pQTLs. By leveraging information from public databases, we identified genes, proteins, and pathways causal for disease; revealed shared aetiology between traits; and identified opportunities for drug repurposing.

4.1 Replicating published findings

Roughly half of the pQTLs we detected replicated previously-reported findings from studies that used independent cohorts and sometimes different proteomic assays (e.g., Somalogic). Across the 585 pQTLs for 290 proteins which we detect, 55% of cis-pQTLs and 27% of trans-pQTLs are located in previously-associated regions (within 2Mb). Replicated loci serve as proof of concept, which increases the probability that novel loci are genuine; while validating previous findings. This is important especially for less robust trans-acting loci, which tend to have smaller effect sizes and are more susceptible to errors due to technical and sample differences.

Conditionally independent variants, however, differed from other studies in many cases. This was most obvious in comparison with recent findings from the UK Biobank (UKB)[48]: while 88% (515) of our associated regions replicated (within 2Mb), only 42% (244) of independent variants matched directly (94) or were in LD (r2>0.8) with those reported in UKB. Possible explanations for this include: different populations, genotyping technologies, LD reference panels, and the methods used to define independence (Plink[92] clump + GCTA-COJO[93] vs Plink clump + SuSiE (UKB)[48,94]). This has implications in downstream analyses, such as MR, where the independent variants are often used as instruments. This emphasises the need for rigorous study design and validation before findings are forwarded for clinical application.

We replicate ($P<1x10^{-5}$) the following MR findings from five large pQTL studies[41–43,45,46] (data from https://www.epigraphdb.org/pqtl/[95]) (Table 2), strengthening genetic evidence for clinical translation.

Table 2. Causal circulating protein-disease associations replicated from five large pQTL studies[41–43,45,46].

Exposure (protein)	Outcome	Major function(s)
Siglec-3 (CD33)	Alzheimer's disease	Immune response (microglial activation)
Granulin (GRN)	Coronary heart disease, LDL cholesterol, HDL cholesterol, total cholesterol	Unknown, implicated in numerous functions
Intercellular adhesion molecule 2 (ICAM2)	LDL cholesterol, HDL cholesterol	Spermatogenesis, immune response
E-Selectin (SELE)	Coronary heart disease, myocardial infarction, LDL cholesterol, total cholesterol	Inflammation
P-Selectin (SELP)	LDL cholesterol, total cholesterol	Inflammation
Von Willebrand Factor (VWF)	Coronary heart disease, myocardial infarction	Blood coagulation

4.2 Rare and non-coding variant contributions

Rare genetic variants can contribute a significant proportion of heritability but are often detectable only when sample or effect sizes are large. Here, we found novel pQTLs that are rare in the general urban European population, but have drifted up in frequency in HELIC MANOLIS and/or Pomak up to 680-fold. We reported fifteen of such pQTLs for proteins from the Cardiovascular and Metabolism Olink panels, all of which were variants at newly-associated loci (Table 1 of Publication 2). These novel loci demonstrate how genetic drift in isolated populations can, even at moderate sample sizes, empower pQTL discovery.

Among the novel pQTLs in HELIC, we observed that the median effect size of rare pQTLs (MAF < 1%) was significantly higher than that of non-rare pQTLs (Wilcoxon rank sum test; *P* = $6.60x10^{-6}$). This may not be reflective of genuine genetic architecture, but rather, a lack of power to detect rare variants of smaller effect sizes. Gene-based testing can circumvent this by aggregating rare variants within a gene and testing for associations on the gene level.

In addition to single-variant analysis, gene-based rare variant meta-analysis was also performed for 250 proteins (from the Cardiovascular and Metabolism panels) from MANOLIS[60], Pomak, and an independent isolated population, ORCADES (total N = 4,422), as recently published[62]. A total of 55 signals passed stringent quality control, including a cis-signal for myeloperoxidase (MPO) that was undetectable in the single-point analysis. The study provides a rare variant analysis pipeline that supplements common-variant association analysis to give a more complete understanding of protein level genetic architecture. This serves as a starting point for larger meta-analyses, which will have greater statistical power to explain missing heritability, and are currently underway.

4.3 The importance of orthogonal validation

Antibody-based proteomic assays such as Olink's PEA are susceptible to epitope effects that can result in false pQTL signals reflecting changes in protein structure rather than protein abundance. While not necessarily biologically unimportant, such variants are not relevant to the aims of this study. This has been discussed in the publications included in this thesis, and highlights the need for pQTL validation using alternative proteomic assays. Other areas of ambiguity include the use of arbitrary thresholds for *cis/trans* annotation, and cross-reactivity with highly homologous protein isoforms.

Limitations to the methods used for downstream causal inference analysis have also been discussed in the contributed publications. In essence, two-sample MR is based on strong assumptions that are difficult to verify (see Methods). Because *trans*-pQTLs are more likely to be pleiotropic, using only *cis*-pQTLs is one way to reduce chances of horizontal pleiotropy; however, this results in fewer instruments and lower statistical power. We also observe instances where only one instrument was available, which increases the potential for false positives. In such cases, genetic colocalisation between the pQTL and the outcome GWAS signal is needed to ensure no confounding by LD has occurred. MR results need to be interpreted with caution and validated using complementary functional assays. Overall, there is a need for consistent strategy and reporting for MR studies. Guidelines (STROBE-MR[96]; https://www.strobe-mr.org/) and recommendations[97] have recently been published, that future studies should adhere to so as to ensure the quality of MR results. A public database of MR findings will also facilitate replication efforts, strengthening evidence for potential clinical targets.

4.5 Conclusion and future perspectives

Circulating proteins are accessible and dynamic markers with potential clinical applications at every step of disease progression, and thus play an indispensable role in our goal towards precision medicine. They assume a unique position in between disease genotype and phenotype; and through the convergence of genetic variation underlying protein levels (pQTLs) and diseases (GWAS signals), we show how we are able to bring findings from proteomic and GWAS studies a step closer to the clinic. This work contributes novel findings and strengthens evidence for disease-relevant proteins that warrant further translational work. All pQTL data have been shared publicly on the GWAS Catalog (https://www.ebi.ac.uk/gwas/; accession IDs provided in supplementary material of contributed publications), adding to a knowledge database that serves as a foundation for forming new hypotheses and validating existing ones.

Large national-scale projects such as the UK Biobank and the German National Cohort (NAKO; https://nako.de/) are generating sequencing and multi-omic data in hundreds of thousands of individuals that will be instrumental in future research. As proteomic technologies approach high-throughput capabilities, we will achieve more in-depth genetic characterisation of the circulating proteome. This will empower biomarker discovery, improve polygenic models, and uncover an abundance of biological insight. Future projects integrating multi-omic data in large and diverse samples, both in circulation and specific cell types/tissues will further clarify ambiguities, enabling us to harness the full potential of pQTLs. Meanwhile, efforts to recruit participants across ancestry groups and the increasing adoption of open sharing of ancestrally-diverse genetic data must be prioritised to ensure health disparities are reduced[98].

Bibliography

1. Haines, J. L. *et al.* Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science* **308**, 419–421 (2005).
2. Rao, V. S. *et al.* Multiple etiologies for Alzheimer disease are revealed by segregation analysis. *Am. J. Hum. Genet.* **55**, 991–1000 (1994).
3. 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).
4. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
5. Gudbjartsson, D. F. *et al.* Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **40**, 609–615 (2008).
6. Lettre, G. *et al.* Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.* **40**, 584–591 (2008).
7. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* **40**, 575–583 (2008).
8. Visscher, P. M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).
9. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res. Off. J. Int. Soc. Twin Stud.* **6**, 399–408 (2003).

10. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109**, 1193–1198 (2012).

11. Hemani, G., Knott, S. & Haley, C. An Evolutionary Perspective on Epistasis and the Missing Heritability. *PLoS Genet.* **9**, e1003295 (2013).

12. Broekema, R. V., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* **10**, 190221 (2020).

13. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111-119 (2015).

14. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

15. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).

16. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

17. Geibel, J. *et al.* How array design creates SNP ascertainment bias. *PLOS ONE* **16**, e0245178 (2021).

18. Gilly, 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).

19. Helgason, A., Sigureth ardóttir, S., Gulcher, J. R., Ward, R. & Stefánsson, K. mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am. J. Hum. Genet.* **66**, 999–1016 (2000).

20. Pollin, T. I. *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702–1705 (2008).

21. Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat. Commun.* **4**, 2872 (2013).

22. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).

23. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).

24. Jurgens, S. J. *et al.* Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat. Genet.* **54**, 240–250 (2022).

25. Bis, J. C. *et al.* Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Mol. Psychiatry* **25**, 1859–1875 (2020).

26. Dilliott, A. A. *et al.* Contribution of rare variant associations to neurodegenerative disease presentation. *NPJ Genomic Med.* **6**, 80 (2021).

27. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**, 638–645 (2008).

28. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

29. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).

30. Deutsch, E. W. *et al.* Advances and Utility of the Human Plasma Proteome. *J. Proteome Res.* **20**, 5241–5263 (2021).

31. Schwenk, J. M. *et al.* The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *J. Proteome Res.* **16**, 4299–4310 (2017).

32. Anderson, N. L. Counting the proteins in plasma. *Clin. Chem.* **56**, 1775–1776 (2010).

33. Bruderer, R. *et al.* Analysis of 1508 Plasma Samples by Capillary-Flow Data-Independent Acquisition Profiles Proteomics of Weight Loss and Maintenance. *Mol. Cell. Proteomics MCP* **18**, 1242–1254 (2019).

34. Keshishian, H. *et al.* Multiplexed, Quantitative Workflow for Sensitive Biomarker Discovery in Plasma Yields Novel Candidates for Early Myocardial Injury. *Mol. Cell. Proteomics* **14**, 2375–2393 (2015).

35. Keshishian, H. *et al.* Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. *Nat. Protoc.* **12**, 1683–1701 (2017).

36. Png, G. *et al.* Identifying causal serum protein–cardiometabolic trait relationships using whole genome sequencing. *Hum. Mol. Genet.* ddac275 (2022) doi:10.1093/hmg/ddac275.

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

38. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* **13**, e1006706 (2017).

39. Zareen, N., Biswas, S. C. & Greene, L. A. A feed-forward loop involving Trib3, Akt and FoxO mediates death of NGF-deprived neurons. *Cell Death Differ.* **20**, 1719–1730 (2013).

40. Du, S. & Zheng, H. Role of FoxO transcription factors in aging and age-related metabolic and neurodegenerative diseases. *Cell Biosci.* **11**, 188 (2021).

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

42. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).

43. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).

44. Pietzner, M. *et al.* Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020).

45. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

46. 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).

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

48. Sun, B. B. *et al. Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants*. http://biorxiv.org/lookup/doi/10.1101/2022.06.17.496443 (2022) doi:10.1101/2022.06.17.496443.

49. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).

50. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).

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

52. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

53. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLOS Genet.* **15**, e1008489 (2019).

54. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

55. 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).

56. Farmaki, A.-E. *et al.* A Dietary Pattern with High Sugar Content Is Associated with Cardiometabolic Risk Factors in the Pomak Population. *Nutrients* **11**, 3043 (2019).

57. 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).

58. Kuchenbaecker, K. *et al.* Insights into the genetic architecture of haematological traits from deep phenotyping and whole-genome sequencing for two Mediterranean isolated populations. *Sci. Rep.* **12**, 1131 (2022).

59. Southam, L. *et al.* Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat. Commun.* **8**, 15606 (2017).

60. Gilly, A. *et al.* Whole-genome sequencing analysis of the cardiometabolic proteome. *Nat. Commun.* **11**, 6336 (2020).

61. Png, G. *et al.* Mapping the serum proteome to neurological diseases using whole genome sequencing. *Nat. Commun.* **12**, 7042 (2021).

62. Gilly, A. *et al.* Gene-based whole genome sequencing meta-analysis of 250 circulating proteins in three isolated European populations. *Mol. Metab.* **61**, 101509 (2022).

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

64. Lin, D. Y. & Zeng, D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet. Epidemiol.* **34**, 60–66 (2010).

65. Lee, C. H., Cook, S., Lee, J. S. & Han, B. Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. *Genomics Inform.* **14**, 173–180 (2016).

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

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

68. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am. J. Hum. Genet.* **104**, 260–274 (2019).

69. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).

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

71. Pierce, B. L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).

72. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* k601 (2018) doi:10.1136/bmj.k601.

73. Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res.* **4**, 186 (2020).

74. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* **32**, 377–389 (2017).

75. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson’s disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).

76. 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).

77. Saade, M., Araujo de Souza, G., Scavone, C. & Kinoshita, P. F. The Role of GPNMB in Inflammation. *Front. Immunol.* **12**, 674739 (2021).

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

79. Diaz-Ortiz, M. E. *et al.* GPNMB confers risk for Parkinson's disease through interaction with α-synuclein. *Science* **377**, eabk0637 (2022).

80. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

81. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).

82. Alzheimer Disease Genetics Consortium (ADGC), *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

83. Griciuc, A. *et al.* Alzheimer's disease risk gene CD33 inhibits microglial uptake of amyloid beta. *Neuron* **78**, 631–643 (2013).

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

85. Griciuc, A. *et al.* TREM2 Acts Downstream of CD33 in Modulating Microglial Pathology in Alzheimer's Disease. *Neuron* **103**, 820-835.e7 (2019).

86. Malik, M. *et al.* Genetics of CD33 in Alzheimer's disease and acute myeloid leukemia. *Hum. Mol. Genet.* **24**, 3557–3570 (2015).

87. 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).

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

89. Vos, T. *et al.* Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* **396**, 1204–1222 (2020).

90. Quintero-Fabián, S. *et al.* Role of Matrix Metalloproteinases in Angiogenesis and Cancer. *Front. Oncol.* **9**, 1370 (2019).

91. Loffek, S., Schilling, O. & Franzke, C.-W. Biological role of matrix metalloproteinases: a critical balance. *Eur. Respir. J.* **38**, 191–208 (2011).

92. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

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

94. 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. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

95. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

96. Skrivankova, V. W. *et al.* Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. *JAMA* **326**, 1614 (2021).

97. Zhao, H. *et al.* Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. *Cell Genomics* **2**, 100195 (2022).

98. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

Appendices

Appendix A

Mapping the serum proteome to neurological diseases using whole genome sequencing

Available at https://doi.org/10.1038/s41467-021-27387-1.

Png G, Barysenka A, Repetto L, Navarro P, Shen X, Pietzner M, Wheeler E, Wareham NJ, Langenberg C, Tsafantakis E, Karaleftheri M, Dedoussis G, Mälarstig A, Wilson JF, Gilly A, Zeggini E. Mapping the serum proteome to neurological diseases using whole genome sequencing. Nat Commun. 2021 Dec 2;12(1):7042. doi: 10.1038/s41467-021-27387-1. PMID: 34857772; PMCID: PMC8640022.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-27387-1 OPEN

Mapping the serum proteome to neurological diseases using whole genome sequencing

Grace Png[1,2]✉, Andrei Barysenka[1], Linda Repetto[3], Pau Navarro[4], Xia Shen[3,5,6], Maik Pietzner[7], Eleanor Wheeler[7], Nicholas J. Wareham[7], Claudia Langenberg[7,8], Emmanouil Tsafantakis[9], Maria Karaleftheri[10], George Dedoussis[11], Anders Mälarstig[12,13], James F. Wilson[3,4], Arthur Gilly[1] & Eleftheria Zeggini[1,2]✉

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.

[1] Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany. [2] TUM School of Medicine, Technical University of Munich and Klinikum Rechts der Isar, Munich, Germany. [3] Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK. [4] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [5] Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University, Guangzhou, China. [6] Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden. [7] MRC Epidemiology Unit, University of Cambridge, Cambridge, UK. [8] Computational Medicine, Berlin Institute of Health (BIH), Charité University Medicine, Berlin, Germany. [9] Anogia Medical Centre, Anogia, Greece. [10] Echinos Medical Centre, Echinos, Greece. [11] Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens, Greece. [12] Department of Medicine, Karolinska Institute, Solna, Sweden. [13] Emerging Science & Innovation, Pfizer Worldwide Research, Development and Medical, Cambridge, MA, USA. ✉email: grace.png@helmholtz-muenchen.de; eleftheria.zeggini@helmholtz-muenchen.de

Neurological disorders are the leading cause of disability worldwide, accounting for 276 million disability-adjusted life years (DALY) globally in 2016[1]. This burden is continuously increasing with growing and ageing populations[2], emphasising the need for better prevention and treatment strategies. Multiple genetics and genomics efforts have established that these diseases have a substantial genetic component[3,4]. 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[5], while others with more distinct symptoms, such as Alzheimer's disease (AD), lack effective drugs and accessible biomarkers that can detect early disease[6]. 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[7], 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[8] 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 neurological 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 *KLKB1, 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 *KLKB1* and *ABO*, affecting 11 and 12 proteins, respectively. These have been identified in published pQTL studies and are not restricted to neurologically relevant proteins[9–12]. *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[13], 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 TDGF1. 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

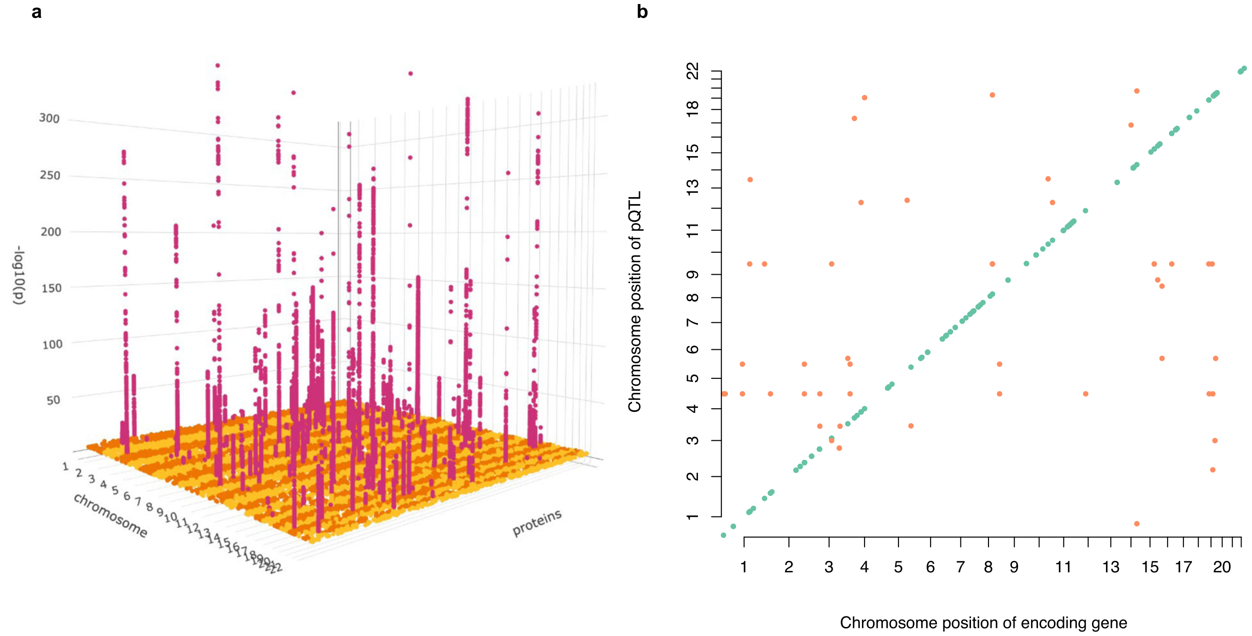


Fig. 1 pQTL signals for 107 serum proteins from Olink neurology and neuro-exploratory panels. a 3D Manhattan plot of detected pQTLs. The *x* axis represents each of the 107 proteins; the *y* axis represents the chromosome location of each signal; and the *z* axis represents the −log10 *p*-values of each 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.

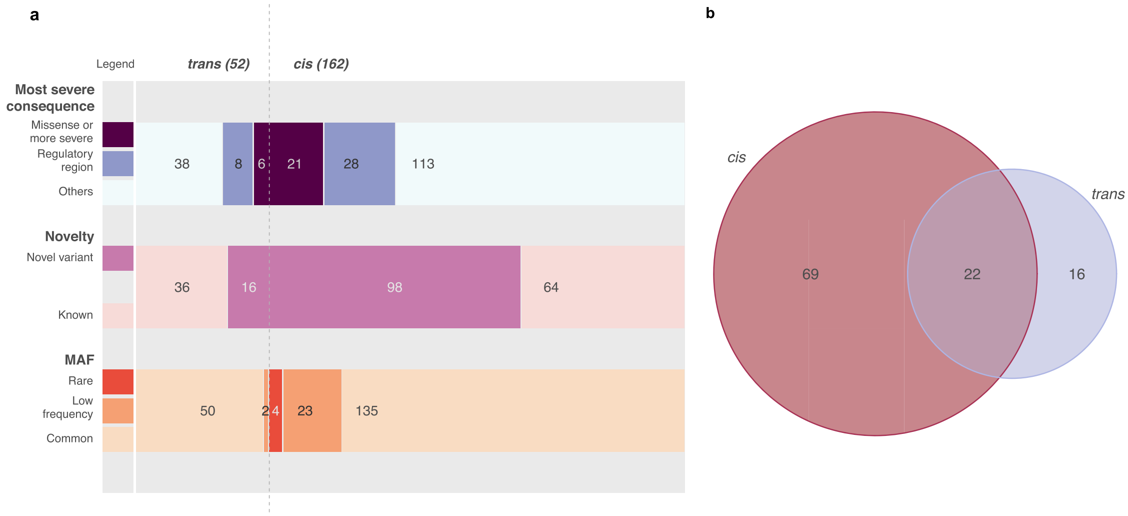


Fig. 2 Overall genetic architecture of 107 serum proteins of neurological relevance. a A total of 214 independent variants were detected. *Cis*-acting variants were defined as variants lying within 1 Mb upstream and downstream of the gene encoding the target protein, while *trans*-acting variants are 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.

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 cells[15].

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

Table 1 Independent pQTL variants for proteins that are being analysed for the first time.

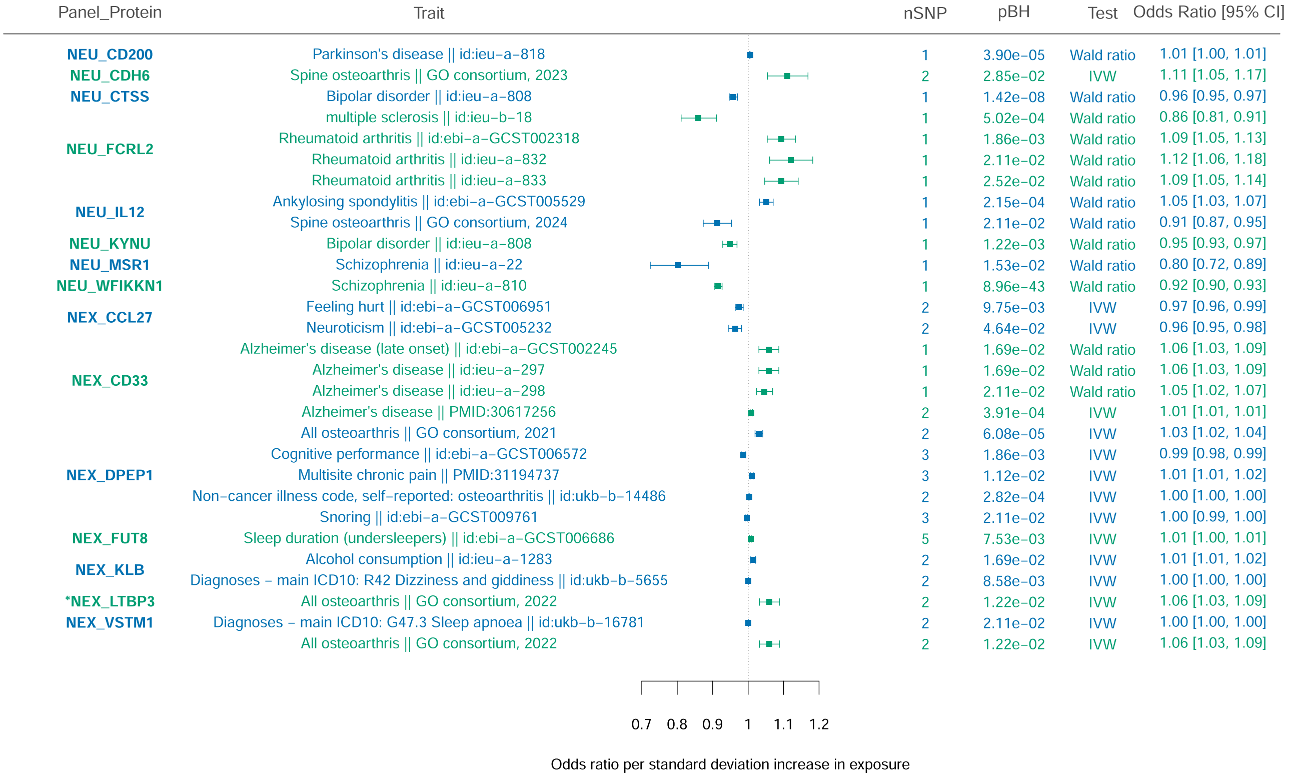
Protein	Variant	MAF	BETA	S.E.	*P*-value	rsID
ADGRB3	chr6:68956792	0.1576	0.89	0.032	8.44E−170	rs1932618
ADGRB3	chr6:68962147	0.3461	0.4947	0.0262	2.31E−79	rs3798971
ADGRB3	chr6:68968025	0.3468	0.8342	0.0225	2.83E−301	rs1953613
CD302	chr2:159745359	0.1016	−0.4303	0.0436	5.34E−23	rs5002908
CD302	chr2:159773858	0.3098	0.3731	0.0281	3.64E−40	rs1553790820
CDH17	chr9:133253728	0.0918	−0.6534	0.0462	1.70E−45	rs10793962
CDH17	chr9:133264504	0.3431	−0.3879	0.028	1.19E−43	novel
CDH17	chr19:48703205	0.4516	−0.386	0.0264	2.25E−48	rs681343
CDH17	chr8:94194571	0.4782	−0.2672	0.0276	3.61E−22	rs56129387
CDH17	chr8:94130944	0.4847	0.2889	0.0267	3.21E−27	rs1051624
GGT5	chr22:24232046	0.0064	−2.3071	0.1696	3.75E−42	rs200519116
GGT5	chr22:24235780	0.1923	−0.3614	0.0326	1.52E−28	rs6004108
GGT5	chr22:24247481	0.2015	−0.3049	0.0317	7.33E−22	rs5760275
IFI30	chr19:18172691	0.2613	0.3604	0.0295	2.10E−34	rs273266
IMPA1	chr8:81652967	0.3331	0.3338	0.0278	3.41E−33	rs2142316
KIR2DL3	chr19:54744273	0.0665	0.8024	0.0574	2.11E−44	rs10414825
KIR2DL3	chr19:54743423	0.2167	0.6973	0.0299	5.70E−120	rs11667532
KIR2DL3	chr6:31272403	0.266	0.5934	0.0307	1.71E−83	rs2524093
KLB	chr17:68883786	0.0268	−0.556	0.0849	5.79E−11	rs34931250
KLB	chr4:39431127	0.3249	−0.4173	0.0265	5.44E−56	rs2926042
KLB	chr4:39447786	0.333	0.7642	0.025	1.17E−205	rs12513342
LTBP3	chr11:65572664	0.0527	0.5989	0.058	5.49E−25	rs10896017
LTBP3	chr11:65575510	0.2504	0.253	0.0299	2.68E−17	rs67924081
NDRG1	chr5:177412889	0.2384	0.2707	0.0318	1.67E−17	rs2731674
NDRG1	chr4:186235350	0.4738	0.2847	0.0263	2.23E−27	novel
PSG1	chr19:42929524	0.02	0.7883	0.087	1.32E−19	rs146569565
PSG1	chr19:42872373	0.1525	−0.3243	0.033	7.79E−23	rs60887906
PSG1	chr19:42881078	0.192	0.8012	0.0267	5.72E−198	rs2005772
RBKS	chr2:27858572	0.009	1.9199	0.1685	4.54E−30	rs140948699
SNCG	chr10:86945549	0.2564	0.934	0.0217	3.24E−403	rs3750822
TPPP3	chr16:67267204	0.0813	−0.3312	0.0483	6.86E−12	rs7200971
VSTM1	chr19:54062922	0.1819	−0.8967	0.0309	9.59E−185	rs8111849
VSTM1	chr19:54042277	0.3968	0.8847	0.0218	4.71E−359	rs2433724

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 GlcNAc-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 tagging, which targets a disproportionate amount of the enzyme to the lysosome rather than the secretory pathway. Secreted and lysosomal SMPD1 are likely to play distinct roles in the body[16], and abnormal levels of the secreted form have been implicated in age-related neurodegenerative conditions[17] including Alzheimer's disease[18] and amyotrophic lateral sclerosis (ALS)[19]. We, therefore, identify a locus at *GNPTAB* that coregulates secreted SMPD1 levels and *GNPTAB* expression, pinpointing a possible mechanism behind SMPD1-related neuropathological 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[20] 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 ($h^2 = 87.2\%$). 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 $h^2 < 5\%$: IKZF2 (4.9%), RNF31 (4.4%), and EPHA10 (0.001%).

We observe that for all four proteins with $h^2 > 80\%$, the pQTLs colocalised with gene expression QTLs in multiple tissues, indicating regulation at the transcriptional level; therefore, the high observed h^2 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 h^2 when causal variants are common, and underestimate h^2 when causal variants are rare[21] (Supplementary Fig. 1).

Link to disease outcomes. To 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[22] and osteoarthritis[23]. 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.



Panel_Protein	Trait	nSNP	pBH	Test	Odds Ratio [95% CI]
NEU_CD200	Parkinson's disease || id:ieu−a−818	1	3.90e−05	Wald ratio	1.01 [1.00, 1.01]
NEU_CDH6	Spine osteoarthris || GO consortium, 2023	2	2.85e−02	IVW	1.11 [1.05, 1.17]
NEU_CTSS	Bipolar disorder || id:ieu−a−808	1	1.42e−08	Wald ratio	0.96 [0.95, 0.97]
NEU_FCRL2	multiple sclerosis || id:ieu−b−18	1	5.02e−04	Wald ratio	0.86 [0.81, 0.91]
	Rheumatoid arthritis || id:ebi−a−GCST002318	1	1.86e−03	Wald ratio	1.09 [1.05, 1.13]
	Rheumatoid arthritis || id:ieu−a−832	1	2.11e−02	Wald ratio	1.12 [1.06, 1.18]
	Rheumatoid arthritis || id:ieu−a−833	1	2.52e−02	Wald ratio	1.09 [1.05, 1.14]
NEU_IL12	Ankylosing spondylitis || id:ebi−a−GCST005529	1	2.15e−04	Wald ratio	1.05 [1.03, 1.07]
	Spine osteoarthris || GO consortium, 2024	1	2.11e−02	Wald ratio	0.91 [0.87, 0.95]
NEU_KYNU	Bipolar disorder || id:ieu−a−808	1	1.22e−03	Wald ratio	0.95 [0.93, 0.97]
NEU_MSR1	Schizophrenia || id:ieu−a−22	1	1.53e−02	Wald ratio	0.80 [0.72, 0.89]
NEU_WFIKKN1	Schizophrenia || id:ieu−a−810	1	8.96e−43	Wald ratio	0.92 [0.90, 0.93]
NEX_CCL27	Feeling hurt || id:ebi−a−GCST006951	2	9.75e−03	IVW	0.97 [0.96, 0.99]
	Neuroticism || id:ebi−a−GCST005232	2	4.64e−02	IVW	0.96 [0.95, 0.98]
NEX_CD33	Alzheimer's disease (late onset) || id:ebi−a−GCST002245	1	1.69e−02	Wald ratio	1.06 [1.03, 1.09]
	Alzheimer's disease || id:ieu−a−297	1	1.69e−02	Wald ratio	1.06 [1.03, 1.09]
	Alzheimer's disease || id:ieu−a−298	1	2.11e−02	Wald ratio	1.05 [1.02, 1.07]
	Alzheimer's disease || PMID:30617256	2	3.91e−04	IVW	1.01 [1.01, 1.01]
NEX_DPEP1	All osteoarthris || GO consortium, 2021	2	6.08e−05	IVW	1.03 [1.02, 1.04]
	Cognitive performance || id:ebi−a−GCST006572	3	1.86e−03	IVW	0.99 [0.98, 0.99]
	Multisite chronic pain || PMID:31194737	3	1.12e−02	IVW	1.01 [1.01, 1.02]
	Non−cancer illness code, self−reported: osteoarthritis || id:ukb−b−14486	2	2.82e−04	IVW	1.00 [1.00, 1.00]
	Snoring || id:ebi−a−GCST009761	3	2.11e−02	IVW	1.00 [0.99, 1.00]
NEX_FUT8	Sleep duration (undersleepers) || id:ebi−a−GCST006686	5	7.53e−03	IVW	1.01 [1.00, 1.01]
NEX_KLB	Alcohol consumption || id:ieu−a−1283	2	1.69e−02	IVW	1.01 [1.01, 1.02]
	Diagnoses − main ICD10: R42 Dizziness and giddiness || id:ukb−b−5655	2	8.58e−03	IVW	1.00 [1.00, 1.00]
*NEX_LTBP3	All osteoarthris || GO consortium, 2022	2	1.22e−02	IVW	1.06 [1.03, 1.09]
NEX_VSTM1	Diagnoses − main ICD10: G47.3 Sleep apnoea || id:ukb−b−16781	2	2.11e−02	IVW	1.00 [1.00, 1.00]
	All osteoarthris || GO consortium, 2022	2	1.22e−02	IVW	1.06 [1.03, 1.09]

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.

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 migraine[24], LTBP3 (latent-transforming growth factor beta-binding protein 3) and osteoarthritis[25], FLRT2 (leucine-rich repeat transmembrane protein) with bipolar disorder[26], and PLXNB1[27] (plexin-B1) and PLA2G10[28] (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 ($P_{\text{adj}} = 9.12 \times 10^{-43}$); 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 factors[29]. Similarly, we find new evidence that serum VSTM1 is causally associated with sleep apnoea ($P_{\text{adj}} = 2.03 \times 10^{-2}$). 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 apnoea[30,31].

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; $P = 7.01 \times 10^{-50}$; BETA = −0.2109; SE = 0.0247) and colocalises with a known Parkinson's disease (PD) locus[32] (CLPP4 = 0.86) (Fig. 4b). *GPNMB* has been highlighted as a susceptibility gene in large PD meta-analyses[32] and has been proven to be upregulated in the brains of PD patients and in mice with induced lysosomal dysfunction[33]. 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 circulation[34]. 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

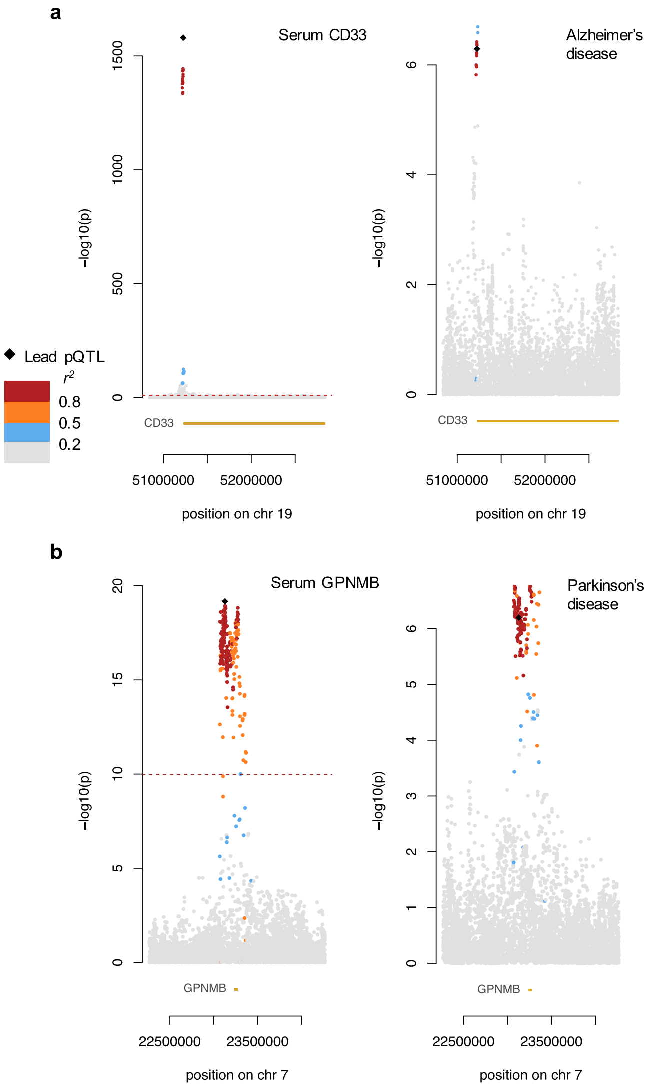


Fig. 4 Colocalisation plots. Each plot shows the association signal and the −log10 *P*-values. The lead pQTL variant is represented by a black diamond, while other points are variants that are coloured according to the extent of linkage disequilibrium with the lead variant. The location of the genes of interest are also shown in yellow at the bottom of each plot. Significance thresholds used for each respective study are shown using a dotted red line. **a** Left: Protein QTL signal for serum CD33; right: GWAS signal for Alzheimer's disease. **b** Left: Protein QTL signal for serum GPNMB; right: GWAS signal for Parkinson's disease.

measure, they are generally preferred for routine testing or monitoring disease progression. Clinical studies will be required to evaluate translational utility.

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] $P_{\mathrm{adj}} = 3.62 \times 10^{-4}$) (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 $P_{\mathrm{adj}} = 1.44 \times 10^{-2}$; 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 mechanism 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 (EH38E2612565), 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

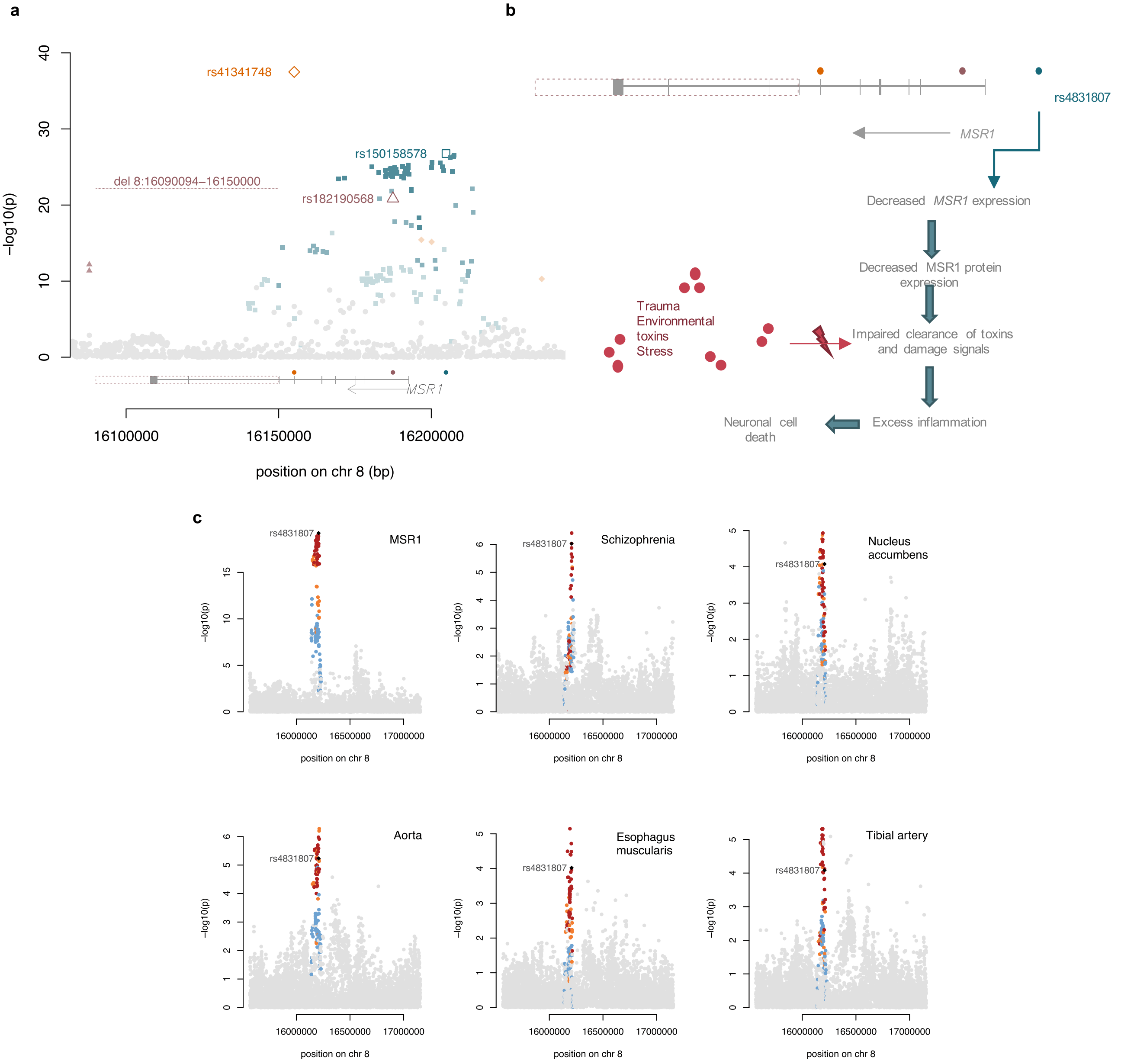


Fig. 5 Serum MSR1 is causally associated with schizophrenia. a Genetic architecture of serum MSR1. Each of the three independent variants and their LD variants are represented in orange, teal, and purple, respectively; the intensity of the colours indicates the strength of linkage disequilibrium (r^2). 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 [$r^2 < 0.2$]).

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: 16090094-16150000[b38]; 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[51], 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, 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 osteoarthritis[52,53], 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 protein[54]. We, therefore, provide further genetic evidence 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 MSR1 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 worldwide[55]. DPEP1 has been implicated in osteoarthritis through a large genome-wide association study[56], and was additionally shown to be downregulated in mouse models of osteoarthritis[57]. 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 MSR1. Further errors may also be introduced due to cross-reactivity and unspecific binding[58]. 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 SomaScan[59] plasma protein measurements was evaluated in 485 individuals from the Fenland cohort[60], 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 MSR1 ($\rho = 0.74$). 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 summarised[61]. 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 (rs3865444)[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 publications[62–64]. 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 HiSeqX platform (Illumina) according to the manufacturer's instructions. Basecall files for each lane were transformed into unmapped BAMs using Illumina2BAM, 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 decoys (HS38DH). The alignment was then merged into the master sample BAM file using

Illumina2BAM MergeAlign. PCR and optical duplicates are marked using biobambam markduplicates and the files were archived in CRAM format. Per-lane CRAMs were retrieved and reads pooled on a per-sample basis across all lanes to produce library CRAMs; these were each divided into 200 chunks for parallelism. GVCFs 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 merged 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 microfluidic qPCR. Protein expression levels are reported as Normalised Protein Expression (NPX) 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-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.

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.0.94[65], simultaneously adjusting for covariates—age, sex, season of sample collection, plate number, plate row, and plate column. An empirical relatedness matrix was also used for each cohort to account for population structure; this was calculated on an LD-pruned set of low-frequency and common variants (MAF > 1%) that passed the Hardy–Weinberg equilibrium test ($P > 1 \times 10^{-5}$). 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 METAL[66]. As no proteins displayed significant genomic inflation ($0.95 < \lambda < 1.03$), no genomic control was applied.

Conditional analysis to identify independent variants. Using the PeakPlotter software (https://github.com/hmgu-itg/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 signals. Independent variants were identified using the approximate conditional and joint stepwise model selection, implemented using the -slct option in GCTA-COJO[67], using a collinearity cut-off of 0.9. Before that, however, variants were first subjected to clumping in Plink 1.9[68] (www.cog-genomics.org/plink/1.9/), using a r^2 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 number of 214 independent variants for 140 signals after filtering for minor allele count (MAC) > 10, Hardy–Weinberg equilibrium $P > 1 \times 10^{-5}$, and replication (meta-analysis P-value < per-cohort P-value) in both cohorts.

Significance thresholds

Single variant-based association and rare variant analysis. For single variant-based association, the significance threshold was adjusted for multiple testing by correcting for the effective number of protein traits (M_{eff}) and variants (N_{eff}) analysed. The effective number of proteins was computed using the ratio of the eigenvalue variance to its maximum[69,70]:

$$M_{\text{eff}} = M(1-(M-1)V_{\lambda_{\text{obs}}}/M^2) = 1 + \frac{tr(\Sigma^T \Sigma)}{M} \quad (1)$$

where $V_{\lambda_{\text{obs}}}$ is the variance of the eigenvalues of the correlation matrix. For the $M =$ 184 Olink proteins included in the study, $M_{\text{eff}} = 93$ in both cohorts. The effective number of variants, or N_{eff}, was determined by using the --indep and --maf 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 --indep were: window size of 50 kb, variant count of 5, and variance inflation factor (VIF) of 2. This was performed separately for both the MANOLIS and Pomak 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.

Significance threshold for two-sample MR. P-values were adjusted for multiple testing by controlling for false discovery rate (FDR) using the Benjamini–Hochberg method. Results were considered significant if the FDR-adjusted P-values were below 0.05.

Novelty. We assessed variants for novelty using a funnel approach, by first identifying (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 studies[7,10–12,71], querying GWAS Catalogue for known signals, then confirmed by doing manual literature searches. Next, we determined variants belonging 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 previously 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 GREML[20] (https://cnsgenomics.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):

$$h^2_{\text{meta}} = \sum(h_i^2/\text{SE}_i^2)/\sum(1/\text{SE}_i^2), \text{SE} = \sqrt{(1/\sum(1/\text{SE}_i^2))} \quad (2)$$

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. Variants 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 REML analysis with four GRMs using default settings. REML 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 predictor[72] (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 ($r^2 > 0.8$) 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 coloc.fast function from the gtx R package (https://github.com/tobyjohnson/gtx/). The method is equivalent to coloc 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 neurodegenerative 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 PhenoScanner[74,75] traits of neurological relevance. The results for this are included in Supplementary Data 5a. Five different posterior probabilities are reported in the table (CLPP0-CLPP4), which corresponds to the five tested hypotheses explained in Giambartolomei et al.[73]. In particular, CLPP4 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[76] (https://github.com/MRCIEU/TwoSampleMR). Traits available in the MRBase[77] 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, ST3GAL6*, 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–trait 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–trait 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 neuroticism, 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[78] (https://www.targetvalidation.org/) and Drugbank[79] (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-itg).

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 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 specialized psychiatry centers 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. Wingo, 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/s41588-020-00773-z (2021).
9. Folkersen, 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. Suhre, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).
13. Alkebsi, L. et al. Validation of the accuracy of self-reported ABO blood types in the Japan Nurses' Health Study. *Asian Pac. J. Cancer Prev.* **20**, 789–793 (2019).
14. Sasayama, D. et al. Genome-wide quantitative trait loci 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 and lysosomal acid sphingomyelinase. *Cell. Signal.* **21**, 836–846 (2009).
17. Park, M. H., Jin, 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**, 1551–1570 (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 neuropathic pain in the back region in chronic low back pain patients -a multicenter investigation. *Sci. Rep.* **8**, 16537 (2018).
23. Thakur, M., Dickenson, A. H. & Baron, R. Osteoarthritis pain: nociceptive or neuropathic? *Nat. Rev. Rheumatol.* **10**, 374–380 (2014).
24. Pisanu, C. et al. High leptin levels are associated with migraine with aura. *Cephalalgia Int. J. Headache* **37**, 435–441 (2017).
25. Zhao, 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 analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43**, 977–983 (2011).
27. Gilabert-Juan, J. et al. Semaphorin and plexin gene expression is altered in the prefrontal cortex of schizophrenia patients with and without auditory hallucinations. *Psychiatry 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 niacin. *Schizophr. Res.* **61**, 1–6 (2003).
29. Szláma, G., Kondás, K., Trexler, M. & Patthy, L. WFIKKN1 and WFIKKN2 bind growth factors TGFβ1, BMP2 and BMP4 but do not inhibit their signalling activity. *FEBS J.* **277**, 5040–5050 (2010).
30. Taylor-Gjevre, R. M., Nair, B. V. & Gjevre, 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 glycoprotein 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. Rose, A. A. N. et al. ADAM10 releases a soluble form of the GPNMB/osteoactivin extracellular domain with angiogenic properties. *PLoS ONE* **5**, e12093 (2010).
35. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
36. Griciuc, A. et al. Alzheimer's disease risk gene CD33 inhibits microglial uptake of amyloid beta. *Neuron* **78**, 631–643 (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. Scara1 deficiency impairs clearance of soluble amyloid-β by mononuclear phagocytes and accelerates Alzheimer's-like disease progression. *Nat. Commun.* **4**, 2030 (2013).
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. V. et al. Scavenger receptors class A-I/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 microglia 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. & Sgambato, V. Neuropsychiatric disorders in Parkinson's Disease: what do we know about the role of dopaminergic and non-dopaminergic systems? *Front. Neurosci.* **14**, 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 Off. J. World Psychiatr. Assoc.* **16**, 163–180 (2017).
50. Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
51. Santos, R. et al. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2017).
52. Okuno, Y., Matsumura, N. & Oguro, S. Transcatheter arterial embolization using imipenem/cilastatin sodium for tendinopathy and enthesopathy refractory to nonsurgical management. *J. Vasc. Interv. Radiol.* **24**, 787–792 (2013).
53. 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).
54. Malik, M. et al. Genetics of CD33 in Alzheimer's disease and acute myeloid leukemia. *Hum. Mol. Genet.* **24**, 3557–3570 (2015).
55. Hiligsmann, 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).
56. Tachmazidou, I. et al. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nat. Genet.* **51**, 230–236 (2019).
57. 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).
58. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).
59. Gold, L. et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE* **5**, e15004 (2010).
60. Pietzner, M. et al. Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020).
61. Pietzner, M. et al. Cross-platform proteomics to advance genetic prioritisation strategies. http://biorxiv.org/lookup/doi/10.1101/2021.03.18.435919 (2021).
62. 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).
63. Gilly, 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).
64. 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).
65. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
66. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
67. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
68. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
69. Moskvina, V. & Schmidt, K. M. On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.* **32**, 567–573 (2008).
70. Cheverud, J. M. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* **87**, 52–58 (2001).
71. Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).
72. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
73. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
74. Staley, J. R. et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
75. Kamat, M. A. et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).
76. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).
77. Hemani, G. et al. The MR-base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
78. Ochoa, D. et al. Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).
79. Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).

Acknowledgements

HELIC-MANOLIS study: We thank the residents of the Pomak and Mylopotamos villages for taking part. This work was funded by the Wellcome Trust [098051] and the European Research Council [ERC-2011-StG 280559- SEPI]. TEENAGE study: The TEENAGE study has been supported by the Wellcome Trust (098051), European Union (European Social Fund—ESF) and Greek national funds through the Education and Lifelong Learning Operational Program of the National Strategic Reference Framework (NSRF)—Research Funding Program: Heracleitus II, Investing in knowledge society through the European Social Fund. The GATK3 programme was made available through the generosity of the Medical and Population Genetics Program at the Broad Institute, Inc. We thank the Human Genetics DNA Pipelines and Human Genetics Informatics departments at the Wellcome Sanger Institute for performing sequencing and variant calling. This study has been conducted using the UK Biobank Resource (project ID 10205). J.F.W. and P.N. acknowledge support from the MRC Human Genetics Unit programme grant, 'Quantitative traits in health and disease' (U. MC_UU_00007/10). The Fenland Study (10.22025/2017.10.101.00001) and C.L., E.W. and N.J.W. are funded by the Medical Research Council (MC_UU_12015/1). We are grateful to all the volunteers and to the General Practitioners and practice staff for assistance with recruitment. We thank the Fenland Study Investigators, Fenland Study Co-ordination team and the Epidemiology Field, Data and Laboratory teams. We further acknowledge support for genomics and metabolomics from the Medical Research Council (MC_PC_13046). Proteomic measurements were supported and governed by a collaboration agreement between the University of Cambridge and Somalogic.

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.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix B

Identifying causal serum protein-cardiometabolic trait relationships using whole genome sequencing

Available at https://doi.org/10.1093/hmg/ddac275.

Png G, Gerlini R, Hatzikotoulas K, Barysenka A, Rayner NW, Klarić L, Rathkolb B, Aguilar-Pimentel JA, Rozman J, Fuchs H, Gailus-Durner V, Tsafantakis E, Karaleftheri M, Dedoussis G, Pietrzik C, Wilson JF, Angelis MH, Becker-Pauly C, Gilly A, Zeggini E. Identifying causal serum protein-cardiometabolic trait relationships using whole genome sequencing. Hum Mol Genet. 2022 Nov 9:ddac275. doi: 10.1093/hmg/ddac275. Epub ahead of print. PMID: 36349687.

OXFORD

Human Molecular Genetics, 2022, Vol. 00, 00, 1–10
https://doi.org/10.1093/hmg/ddac275
Advance access publication date 9 November 2022
Original Article

Identifying causal serum protein–cardiometabolic trait relationships using whole genome sequencing

Grace Png[1,2], Raffaele Gerlini[3,4], Konstantinos Hatzikotoulas[1], Andrei Barysenka[1], N. William Rayner[1], Lucija Klarić[5], Birgit Rathkolb[3,4,6], Juan A. Aguilar-Pimentel[3], Jan Rozman[3,4,7], Helmut Fuchs[3], Valerie Gailus-Durner[3], Emmanouil Tsafantakis[8], Maria Karaleftheri[9], George Dedoussis[10], Claus Pietrzik[11], James F. Wilson[5,12], Martin Hrabe de Angelis[3,4,13], Christoph Becker-Pauly[14], Arthur Gilly[1] and Eleftheria Zeggini[1,15,*]

[1]Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg 85764, Germany
[2]Technical University of Munich (TUM), School of Medicine, Munich 80333, Germany
[3]Institute of Experimental Genetics, German Mouse Clinic, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg 85764, Germany
[4]German Center for Diabetes Research (DZD), Neuherberg 40225, Germany
[5]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH8 9QN, UK
[6]Institute of Molecular Animal Breeding and Biotechnology, Gene Center, Ludwig-Maximilians University Munich, Munich 80539, Germany
[7]Institute of Molecular Genetics of the Czech Academy of Sciences, Czech Centre for Phenogenomics, Vestec 25250, Czech Republic
[8]Anogia Medical Centre, Anogia 74150, Greece
[9]Echinos Medical Centre, Echinos 67300, Greece
[10]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens 17671, Greece
[11]Institute for Pathobiochemistry, University Medical Center of the Johannes Gutenberg University Mainz, Mainz 55122, Germany
[12]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh EH8 9QN, UK
[13]Chair of Experimental Genetics, TUM School of Life Sciences, Technical University of Munich, Freising 80333, Germany
[14]Institute of Biochemistry, Unit for Degradomics of the Protease Web, University of Kiel, Kiel 24118, Germany
[15]Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine, Munich 80333, Germany
*To whom correspondence should be addressed: Institute of Translational Genomics, Helmholtz Zentrum München, Ingolstaedter Landstr. 1, D-85764 Neuherberg, Germany. Tel: +49 89 3187 49728; Email: eleftheria.zeggini@helmholtz-muenchen.de

Abstract

Cardiometabolic diseases, such as type 2 diabetes and cardiovascular disease, have a high public health burden. Understanding the genetically determined regulation of proteins that are dysregulated in disease can help to dissect the complex biology underpinning them. Here, we perform a protein quantitative trait locus (pQTL) analysis of 248 serum proteins relevant to cardiometabolic processes in 2893 individuals. Meta-analyzing whole-genome sequencing (WGS) data from two Greek cohorts, MANOLIS ($n = 1356$; 22.5× WGS) and Pomak ($n = 1537$; 18.4× WGS), we detect 301 independently associated pQTL variants for 170 proteins, including 12 rare variants (minor allele frequency < 1%). We additionally find 15 pQTL variants that are rare in non-Finnish European populations but have drifted up in the frequency in the discovery cohorts here. We identify proteins causally associated with cardiometabolic traits, including *Mep1b* for high-density lipoprotein (HDL) levels, and describe a knock-out (KO) *Mep1b* mouse model. Our findings furnish insights into the genetic architecture of the serum proteome, identify new protein–disease relationships and demonstrate the importance of isolated populations in pQTL analysis.

Introduction

Cardiovascular and metabolic disorders, such as hypertension, hyperlipidaemia, coronary artery disease (CAD) and type 2 diabetes (T2D), impose a heavy and increasing health burden (1,2). Significant progress has been made in disentangling the complex and overlapping genetic aetiology of these diseases through genome-wide association studies (GWAS), which have successfully identified multiple genetic variants associated with disease risk. At the same time, multiplex proteomic assays have enabled the identification of disease-associated proteins (3–5).

However, statistical association with disease does not always mean that the gene or protein plays a causal role. This can be elucidated by coupling genetics with proteomics to identify genetic variants associated with protein levels, known as protein quantitative trait loci (pQTLs). By complementing pQTL analysis with causal inference approaches such as two-sample Mendelian randomization (MR), non-spurious protein–disease relationships and, therefore, disease pathways, genetic variants, and proteins of clinical relevance can be identified (6–12).

We have previously (10) assessed the genetic architecture of 257 serum protein levels in a Greek isolated cohort, MANOLIS, through which we found 164 independently associated pQTLs for 109 proteins, and demonstrated the value of genetically predicted protein levels in clinical risk models. Here, we substantially increase power by doubling the sample size, meta-analyzing whole genome sequencing data from MANOLIS with an additional isolated

Received: June 28, 2022. **Revised:** October 24, 2022. **Accepted:** October 27, 2022

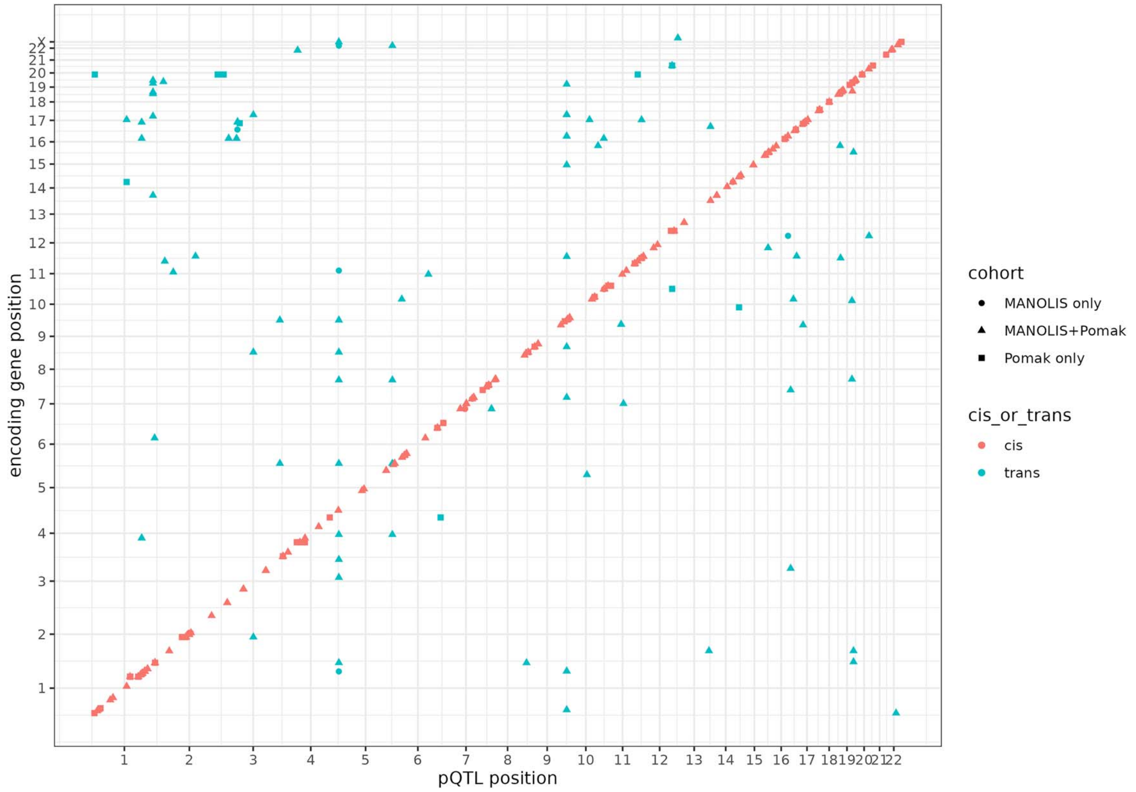


Figure 1. Chromosomal location of *cis*- (red) and *trans*-pQTLs (blue) plotted against the chromosomal location of the gene encoding the proteins of interest. Cis-pQTLs were defined as variants lying within 1 Mb of the start of the gene encoding the target protein.

population cohort, Pomak. We find 301 independent pQTLs for 170 proteins and describe pQTLs that are driven up in frequency in either discovery cohort, illustrating the value of population isolates in the discovery of protein-associated variation. We further highlight previously undetected causal protein–disease associations using genetic colocalization analysis and two-sample MR.

Results

Genetic architecture of 170 proteins

We detect 301 independently associated pQTLs ($P < 7.45 \times 10^{-11}$) for 170 proteins (Supplementary Material, Table S2) that are present in both cohorts with a consistent direction of effect. Of these, 133 variants belong to loci that were not detected previously in MANOLIS only (10). All protein targets had between one and eight independently associated variants (Supplementary Material, Fig. S1), highlighting the varying complexity of protein level genetic architecture. Additional evidence for replication was sought in a protein level dataset of plasma samples obtained from up to 950 individuals (Methods) from the ORCADES study (13), an isolated population from the Orkney islands in the Northern Isles of Scotland. In sum, 177 (58.8%) pQTLs replicated (Methods) in this independent cohort (Supplementary Material, Table S2).

Detected pQTLs were categorized into *cis*- and *trans*-pQTLs according to their distance to the target protein-encoding gene (Methods); we found 215 *cis*-acting pQTLs for 138 proteins, and 86 *trans*-pQTLs for 63 proteins (Fig. 1). In sum, 31 proteins had both *cis*- and *trans*-pQTLs. By mapping *trans*-pQTLs to their nearest gene, we determined 42 *trans*-pQTLs located in known pleiotropic genes; namely, *ABO, CFH, HLA, F12, FUT2, ST3GAL6* and *KLKB1*. Four of these genes (*ABO, FUT2, F12, KLKB1*) are involved in blood coagulation pathways, whereas *CFH* and *HLA* are closely related to inflammatory response.

Protein QTLs that act in *trans* are also useful for identifying unknown molecular interactions. As proof of principle, we detect an intronic *trans*-pQTL for C-C motif chemokine ligand 3 (CCL3) located within the encoding gene for C-C motif chemokine receptor 3, *CCR3*. CCL3 is a known agonist of CCR3 that may contribute to the aggregation of eosinophils to inflammation sites (14). Mapping *trans*-pQTLs to their causal genes, however, remains a challenge as causal genes are often not the closest ones (8,12) (Supplementary Material, Note 1).

The majority of pQTLs are common variants (minor allele frequency [MAF] > 5%). We find 12 rare (MAF < 1%) pQTLs and 42 low-frequency pQTLs (1% < MAF < 5%). Using Ensembl's variant effect predictor (VEP), we find altogether 36 (12%) pQTLs that have a most severe consequence of missense, whereas two variants for PRSS27 (*trans*-pQTL) and IL17D (*cis*-pQTL), respectively, are stop-gain variants. The PRSS27-associated variant acts in *trans* and is located within the pleiotropic gene, *FUT2*. The *cis*-pQTL for IL17D and five other missense variants are all rare and were previously undetected in MANOLIS, showing how larger sample sizes provide increased power to detect rare associated variants of severe consequences.

Excluding *trans*-pQTLs located within pleiotropic genes, we find 35 pQTLs (11.6%) in regions that have not been reported in other large-scale pQTL analyses (Supplementary Material, Table S3), comprising 22 *cis*-pQTLs for 18 proteins, and 13 *trans*-pQTLs for 12 proteins. As isolated populations often contain private, rare variants that have drifted up in frequency because of founder effects (15), we additionally interrogate 69 pQTLs that are present in only one discovery cohort, of which 7 replicate in ORCADES (10%) and 28 (40.5%) have not been previously reported (Supplementary Material, Table S2). In sum, 15 novel pQTLs are rare (MAF < 1%) in non-Finnish Europeans (gnomAD) but have drifted up in frequency in one or both of our discovery

Table 1. Novel and previously unreported pQTLs that have drifted up in frequency in MANOLIS and/or Pomak. The gnomAD-NFE MAF column contains the minor allele frequencies (MAF) of each variant in non-Finnish Europeans (NFE) from the Genome Aggregation Database (gnomAD). MAFs (gnomAD and 1000 Genomes) of all other detected variants are reported in Supplementary Material, Table S4. The most severe consequences were obtained using Ensembl's variant effect predictor (VEP). An expanded table containing the genotype counts, Hardy–Weinberg equilibrium test *P*-values, and the full VEP results are in Supplementary Material, Tables S5A and B. Abbreviations: Chr, chromosome; Pos, position; HELIC, Hellenic isolated cohorts; MAF, minor allele frequency; NFE, non-Finnish Europeans

Protein	Chr	Pos	rsID	Cohorts	*cis/trans*	HELIC MAF	gnomAD-NFE MAF	Most severe consequence
SUMF2	7	71973324	rs568788425	MANOLIS	*cis*	0.80%	0.04%	Intron
CD1C	1	158292108	rs201448758	MANOLIS+Pomak	*cis*	1.21%	0.01%	Missense
ENO2	12	6862641	rs184861396	MANOLIS+Pomak	*cis*	0.45%	0.20%	Intron
ITGB7	12	53519700	rs541150953	MANOLIS+Pomak	*cis*	1.53%	0.18%	Intron
ACP6	1	121470180	rs114127018	Pomak	*cis*	0.90%	0.01%	Intergenic
APLP1	19	35871901	rs767668877	Pomak	*cis*	1.00%	0.00%	Missense
CD93	1	3888781	rs912070506	Pomak	*trans*	0.20%	0.01%	Intergenic
CD93	2	207672303	rs942471010	Pomak	*trans*	0.40%	0.01%	Intergenic
CD93	2	227266736	rs1396628045	Pomak	*trans*	0.40%	0.01%	Non-transcript exon
IGFBP7	4	67658568	rs539585543	Pomak	*cis*	0.70%	0.02%	Intron
IL1RL2	2	89009162	rs543843028	Pomak	*cis*	2.00%	0.13%	Intergenic
KYAT1	9	126833282	rs746374838	Pomak	*cis*	0.60%	0.00%	Missense
MMP2	16	55496937	rs144755357	Pomak	*cis*	1.30%	0.01%	Missense
PSGL1	12	97893711	rs185338771	Pomak	*cis*	0.40%	0.00%	Intergenic
VSIG2	11	124706898	rs959226701	Pomak	*cis*	0.60%	0.15%	Intergenic

cohorts by at least 2.25-fold (Table 1; Supplementary Material, Fig. S2; Supplementary Material, Table S4), including four missense variants. None of the 15 variants were present in the replication cohort, and proxies in linkage disequilibrium (LD) failed to replicate. In particular, a *cis*-pQTL for 72 kDa type IV collagenase (MMP2; rs144755357) that has drifted up 95-fold in Pomak is predicted to be deleterious by SIFT and PolyPhen-2 (Supplementary Material, Table S5). The MMP2-increasing variant causes a p.Arg495Gln substitution within the hemopexin C domain, which binds the inhibitor TIMP-2 (16) (Supplementary Material, Fig. S3). We therefore demonstrate the importance of including isolated populations in pQTL association studies as they may contribute to high-impact variants otherwise undetectable in cosmopolitan populations.

Identifying proteins associated with cardiometabolic traits

To identify causal relationships between serum proteins and cardiometabolic traits, we applied two-sample Mendelian randomization and colocalization analysis using GWAS summary statistics of complex traits. We defined cardiometabolic traits as follows: all lipid traits; glycaemic traits; diabetes; kidney disease and measures of kidney function; all heart conditions; hypertension; and body-mass index (BMI) (Methods). We find 43 serum proteins that are associated with at least one cardiometabolic trait (Supplementary Material, Table S6 and S7).

Of these, 18 proteins show strong evidence of causal association (≥2 instrumental variables, using the inverse variance-weighted [IVW] method) with at least one cardiometabolic trait (Fig. 2). Of note are the TYRO3 (tyrosine-protein kinase receptor), DLK1 (protein delta homologue 1) and CTSH (cathepsin H) proteins, which are significantly associated with diabetic kidney disease (DKD). Increased TYRO3 and CTSH levels are associated with an increased risk of DKD in individuals with type 1 or 2 diabetes, and reduced DLK1 levels are associated with an increased risk of DKD in individuals with T2D. Whereas CTSH and DLK1 have not been associated with kidney disease (Supplementary Material, Note 2), studies have shown increased *TYRO3* mRNA expression (17) and increased circulating and urinary TYRO3 levels (18) in patients with DKD, further supporting a causal role. We also note that TYRO3 is targeted by an approved drug for rheumatoid arthritis, fostamatinib, highlighting an opportunity for the repurposing of fostamatinib to treat DKD. We elaborate on other previously unreported examples in Supplementary Material, Note 2.

The MR analysis further validates known protein–disease links, showing causal associations between increased serum LDLR (low-density lipoprotein [LDL] receptor) protein and decreased LDL, total cholesterol and risk of coronary heart disease (19). We also replicate a previously reported finding showing that LRIG1 (leucine-rich repeats and immunoglobulin-like domains 1) lies on the causal path for atrial fibrillation, T2D and self-reported hypercholesterolemia (10).

For two proteins, sulfatase modifying factor 2 (SUMF2; Supplementary Material, Note 2) and meprin A subunit beta (*Mep1b*), we observe association with cardiometabolic traits using novel replicating pQTLs as instrumental variables. We find that decreased serum *Mep1b* is causally associated with increased HDL levels (Wald ratio $P_{\mathrm{FDR}} = 3.38 \times 10^{-2}$; beta = −0.008; SE = 0.002). The intronic *cis*-pQTL, rs680321, is robustly associated with serum *Mep1b* (MAF = 0.37; beta = −1.07; SE = 0.026; $P = 2.50 \times 10^{-372}$; Supplementary Material, Note 3). Two other independently associated *Mep1b cis*-pQTLs are private to Pomak (rs763953724, rs1410442909); both variants are non-existent in non-Finnish Europeans and lie upstream of the *Mep1b* gene.

To better understand the potential metabolic role played by *Mep1b*, we systematically phenotyped an existing *Mep1b* KO mouse model at the German Mouse Clinic. Monitoring body weight from age 9 to 19 weeks revealed that *Mep1b* depletion in the mouse impacts on the body mass of females, which were heavier as a result of increased adiposity (Supplementary Material, Note 4; Supplementary Material, Figs S4–S6; Supplementary Material, Table S8). This sex-specific effect was not observed for the *cis*-pQTL, rs680321 (sex heterogeneity *P* = 0.086; Supplementary Material, Fig. S7 and Supplementary Material, Table S9).

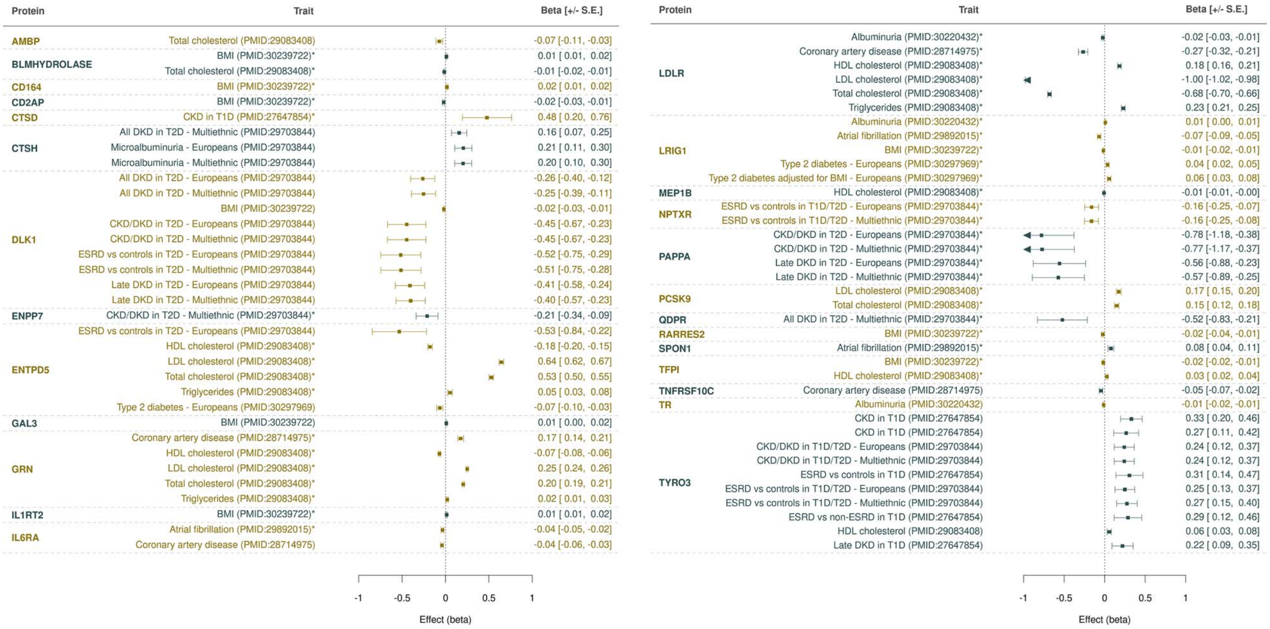


Figure 2. Two-sample Mendelian randomization between proteins (exposure) and cardiometabolic traits (outcome), using only downloaded summary statistics. Points represent the effect size (beta) and direction of each causal association, with errors bars representing ±SE. Arrows indicate beta coefficients that are below −1. Actual beta and SE values are given to the right of each plot. Traits marked with an asterisk (∗) indicate that a Wald ratio test was performed; otherwise, the inverse-variance weighted method was used. Full MR results with MRBase traits are given in Supplementary Material, Table S6.

Discussion

The relationship between *Mep1b* and cholesterol or adiposity remains largely unexplored. *Mep1b* is a metalloprotease that is involved in post-translational proteolysis of numerous targets (20,21) in mammals. Closely related to meprin *α* (MEP1A), both proteins have been implicated in inflammatory disorders, Alzheimer's disease, kidney disease and cancer (20). Several substrates of *Mep1b* have also been linked to cholesterol levels, such as dipeptidyl peptidase 4 (DPP4) and amyloid precursor protein (APP) (22,23). Results from our MR analysis and mouse phenotyping support a direct role of *Mep1b* in influencing adiposity, which is a risk factor for a multitude of complex diseases, including those previously linked to *Mep1b*. Given its involvement in complex networks, however, further experiments will be needed to identify specific pathways.

Our causal inference analysis additionally revealed cardiometabolic traits that are associated with multiple shared proteins (Supplementary Material, Figs S8–S10). LDL cholesterol, total cholesterol and triglyceride levels were all causally associated with the serum levels of seven proteins: GRN, LDLR, SUMF2, KIM1, ENTPD5, CHI3L1 and FGF21. HDL cholesterol was associated with four of the same proteins (GRN, LDLR, SUMF2, ENTPD5), but additionally with eight other proteins (TYRO3, HBEGF, SPON1, SCF, TIMP4, TFPI, MEP1B, ANGPTL1, AXL) that were not significantly associated with LDL or total cholesterol, suggesting a complex and distinct underlying proteomic landscape. This demonstrates the potential of such analyses to furnish insights into molecular similarities and differences between similarly presenting diseases or disease subtypes in future studies, facilitating efforts for more precise diagnosis and treatment.

In this work, we detect 133 new pQTLs, 40% of which are *trans*-pQTLs for 48 proteins, including the *CCR3*-*CCL3* receptor-ligand interaction. We were able to reproduce 92% of the 164 independent pQTLs reported previously (10), including 12 variants exclusive to MANOLIS. The remaining 13 pQTLs (12 *cis*, 1 *trans*) were not reproduced because of either the exclusion of the protein from meta-analysis (QC failure) or a loss of significance. Overall, 59% of our pQTLs replicated in an independent cohort. There are several possible explanations for lack of replication, including insufficient statistical power because of the smaller sample size of the replication cohorts, a lack of proxies for private variants, and differences in cell type and protein composition between serum (MANOLIS and Pomak) and plasma (ORCADES) (Supplementary Material, Fig. S11).

Population isolates have special population genetics characteristics that can boost the discovery of rare variant associations. Here, we identify 15 rare pQTLs that have drifted up in frequency in one or both cohorts. Whole genome sequencing enables access to the analysis of rare variants through gene-based burden testing. We have recently described (24) five rare variant burden pQTLs in MANOLIS, Pomak and ORCADES that are independent of the single point signals reported in this work. Projects with larger sample sizes will further increase power and are currently underway.

We recognize several limitations to this work. First, as Olink's immunoassay relies on the binding of antibodies to target antigens, genetic variation can alter binding sites and, therefore, the affinity of the antibody probes to the target protein. This may result in association signals that reflect altered protein structure rather than changes in protein abundance. For 25 proteins with protein-altering variants (based on Ensembl VEP classification [Methods]), we checked for such effects through a comparison of proteomic data by Olink versus an aptamer-based assay by Somalogic (with different antigen binding sites) in an independent cohort, Fenland (12). We observed good correlation (Spearman correlation>0.5) for 13 (59%) of 22 proteins that were measured using both technologies (Supplementary Material, Table S10), suggesting genuine pQTL signals. Other than altered antibody binding as a result of protein structure changes, weak correlations may be explained by different technical and protein characteristics, as recently investigated (25). Orthogonal validation is therefore necessary for accurate downstream biological interpretation.

Secondly, the validity of the two-sample MR results relies on the assumptions that the genetic instruments (pQTLs) influence the outcome (cardiometabolic trait) only through the exposure (protein level) and are not associated with confounders (Methods). Moreover, we note that the GWAS summary statistics used in this analysis were not derived from WGS-based studies, and therefore several of our instruments were not found in these datasets and could not be used. As we only assess causality unidirectionally, future studies will benefit from bidirectional analyses using larger, sequence-based exposure and outcome GWAS datasets that can produce a greater number of reliable instruments and provide validation. Finally, all individuals in the discovery and replication cohorts are of European descent. Larger, ethnically diverse sample sizes are needed to fully characterize the genetic architecture of the serum proteome.

Materials and Methods

Sequencing and variant calling

The two cohorts were sequenced in an identical way. Genomic DNA (500 ng) from 1482 and 1642 samples for MANOLIS and Pomak, respectively, was 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 HiSeqX platform (Illumina) according to manufacturer's instructions.

Basecall files for each lane were transformed into unmapped BAMs using Illumina2BAM, marking adaptor 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 decoys (HS38DH). The alignment was then merged into the master sample BAM file using Illumina2BAM MergeAlign. PCR and optical duplicates are marked using biobambam markduplicates and the files were archived in CRAM format.

Per-lane CRAMs were retrieved and reads pooled on a per-sample basis across all lanes to produce library CRAMs; these were each divided in 200 chunks for parallelism. GVCFs were generated using HaplotypeCaller v.3.5 from the Genome Analysis Toolkit (GATK) (26) 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 merged across all batches into a cohort-wide VCF file using bcftools concat.

Variant and sample quality control

Variant-level QC was performed using the Variant Quality Score Recalibration tool from the GATK v. 3.5–0-g36282e4 (26), using a tranche threshold of 99.4% for SNPs, which provided an estimate false positive rate of 6% and a true positive rate of 95%. For INDELs, we used the recommended threshold of 1%. For sample-level QC, we made extensive use of genotyping array datasets in overlapping samples, which provided sample matching information for 1386 and 1511 samples in MANOLIS and Pomak, respectively. In MANOLIS, a total of 25 individuals were excluded ($n = 1457$) based on sex checks, low concordance (<0.8) with chip data, duplicate checks, average depth (<10×), missingness (>0.5%) and contamination (Freemix or CHIPMIX score from the verifyBamID suite^{32} > 5%). This number was 27 for the Pomak cohort. In the case of sample duplicates, the sample with highest quality metrics (depth, freemix and chipmix score) was kept.

Proteomics

The serum levels of 275 unique from three Olink (https://www.olink.com/) panels—Cardiovascular II, Cardiovascular III and Metabolism—were measured using Olink's proximity extension assay (PEA) technology (Supplementary Material, Table S1). Briefly, for each assay, the binding of a unique pair of oligonucleotide-labelled antibody probes to the protein of interest results in the hybridization of the complementary oligonucleotides, which triggers extension by DNA polymerase. DNA barcodes unique to each protein are then amplified and quantified using microfluidic real-time qPCR. Measurements were given in a natural logarithmic scale in Normalized Protein eXpression (NPX) levels, a relative quantification unit. NPX is derived by first adjusting the qPCR Ct values by an extension control, followed by an inter-plate control and a correction factor predetermined by a negative control signal. This is followed by intensity normalization, where values for each assay are centred around its median across plates to adjust for inter-plate technical variation. Further details on the internal and external controls used can be found at http://www.olink.com. Additionally, a lower limit of detection (LOD) value is determined for each protein based on the negative control signal plus three standard deviations. In this study, NPX values that fall below the LOD were set to missing.

We adjusted all phenotypes using a linear regression for age, age squared, sex, plate number and per-sample mean NPX value across all assays, followed by inverse-normal transformation of the residuals. We also adjusted for the season, given the observed annual variability of some circulating protein levels. Given the dry Mediterranean climate of Crete, we define the season of collection as hot summer or mild winter. Plate effects are partially offset by the median-centring implemented by Olink. MANOLIS and Pomak samples were plated in the order of sample collection, which results in plate and season information to be largely correlated.

In MANOLIS, we excluded 13 protein measurements across all panels with missingness or below-LOD proportion greater than 40%. BNP was measured across all three panels and was excluded because of high missingness in all three. In sum, 26, 2 and 14 samples failed vendor QC and were excluded from Cardiovascular II, III and Metabolism, respectively. Also, 42 samples were excluded because of missing age. In Pomak, we excluded 15 proteins and 49, 6 and 13 samples in Cardiovascular II, III and Metabolism. No samples were excluded because of missing covariates. Seven proteins in MANOLIS and five in Pomak were further excluded because of failing QC in the other cohort. A total of 255 proteins were included in the final single-point analysis (Supplementary Material, Table S1).

Single-point association and meta-analysis

We carry out single-point association using the linear mixed model implemented in GEMMA v.0.94 (27). We use an empirical relatedness matrix calculated on an LD-pruned set of low-frequency and common variants (MAF > 1%) that pass the Hardy–Weinberg equilibrium test ($P < 1 \times 10^{-5}$). We further filter out variants with missingness higher than 1% and MAC < 10. Following single-point association, a further seven proteins (GDF15, TFF3, TINAGL1, LOX1, SRC, CTSL1, IDUA) were excluded because of having a genomic control $\lambda_{GC} < 0.97$ or $\lambda_{GC} > 1.05$ after association in either cohort.

GEMMA truncates alleles to a single character. In order to enable unambiguous meta-analysis of indels, we updated alleles in summary statistics by matching it to the VCF. More precisely, we join both files by chromosome and position, and match the alleles by frequency for biallelics. For multiallelics, we compute the difference in allele frequency between the GEMMA output MAF, which is based on samples with non-missing phenotypes, and the AF fields of each allele in the VCF, and use the alleles with the lowest difference.

We use the 25 March 2011 release of METAL (28) for meta-analysis of 248 proteins using inverse-variant based weighting. Full summary statistics are available for download from the GWAS Catalogue (https://www.ebi.ac.uk/gwas/); accession IDs are provided in Supplementary Table 14.

Signal extraction and conditional analysis

Using a *P*-value threshold of 1 × 10^{-6}, 495 signals were extracted using the peakit.py routine of PeakPlotter commit 545191d6db51d87f2b549351e5cda19aaf50330e (https://github.com/hmgu-itg/peakplotter), after filtering out index variants with a minor allele count (MAC) of <10 or do not pass the Hardy–Weinberg equilibrium test. PeakPlotter is based on a combination of distance-based and LD-based pruning; specifically, the software sorts variants passing the significance threshold by increasing the *P*-value, then for each variant, computes SNPs in LD greater than $r^2 = 0.2$, removes them and moves on to the next variant. Variants selected in this way located within <2 Mb of each other are then grouped together, and the index variant is set to the variant with the lowest *P*-value. Each index variant defines a signal, and we use locus and signal interchangeably in this article. A total of 380 index variants passing the study-wide significance threshold of $P < 7.45 \times 10^{-11}$ were extracted. We then extracted independent SNV at each associated locus using an approximate conditional and joint stepwise model selection analysis as implemented in GCTA-COJO[34], using merged cross-cohort genotypes for LD calculation. To avoid overfitting when too many predictors are included in the model, we perform LD-based clumping using Plink v.1.9 (29) (www.cog-genomics.org/plink/1.9/), based on an r^2 value of 0.1 and a window of 1 Mb before the GCTA-COJO analysis (30). The extended LD present within population isolates can cause very large peaks to be broken up into several signals. We identified and manually investigated 44 regions where multiple peaks were present in close proximity of each other, reducing the number of independent signals to 257 and the number of conditionally independent variants to 370 (301 present in both cohorts).

Sex-specific meta-analysis

To look for sex-specific pQTLs, we investigated the heterogeneity between males and females for all 370 conditionally independent pQTLs present in at least one cohort. Single-point association analyses for males and females in both discovery cohorts were first run separately for each pQTL using GEMMA v.0.94 (27), using the same methods as described for the main single point analysis. With the output files, we then performed a sex-specific meta-analysis using the GWAMA v2.2.2 software (31,32) by specifying the —sex option. None of the 370 pQTLs show significant sex heterogeneity using a Bonferroni-corrected *P*-value significance threshold ($P < 1.35 \times 10^{-4}$) (Supplementary Material, Table S9).

Defining *cis*- and *trans*-pQTLs

We define *cis*-pQTLs as variants that lie within 1 Mb upstream or downstream of the encoding gene, whereas *trans*-pQTLs are all variants lying outside of this region.

Comparison of Olink and Somalogic proteomic data in Fenland

Cis-acting protein-altering variants may result in false-positive associations because of epitope effects. We note that 26 *cis*-acting variants for 25 proteins have a potentially protein-truncating effect (IMPACT of MODERATE or HIGH according to Ensembl VEP). Comparison of Olink measurements with an alternative assay, Somalogic, in the Fenland (12,25) cohort (https://www.omicscience.org/apps/pgwas/) showed good correlation between the two measurements for 13 out of 22 proteins (with *cis*-pQTLs) with both Olink and Somalogic proteomic data (Supplementary Material, Table S10).

Significance threshold

We based our significance threshold on the effective number of variants and traits analyzed. We excluded variants with MAC < 10 from the MANOLIS cohort, then performed LD-pruning using Plink v.1.9 (29) using the parameter—indep 50 5 2. This yielded an $N_{\mathrm{eff}} = 5\,078\,182$ unique variants for MANOLIS. As computing a similar value for the meta-analysis would have required a computationally intensive merging of genotypes across cohorts and handling of cohort-specific variants, we note that the Pomak estimate is similar and that the majority of variants in the meta-analysis will be common to both cohorts, with a further portion of cohort-specific variants likely in LD with common ones. We therefore use the MANOLIS N_{eff} in our analysis. For M_{eff}, the effective number of phenotypes, we compute the ratio of the eigenvalues of the phenotype correlation matrix to its maximum and obtain 132. The resulting *P*-value threshold is 7.45 × 10^{-11}.

Replication

Replication was performed in the ORCADES isolated cohort from the Orkney archipelago in the Northern Isles of Scotland (13). In sum, 1348 samples were sequenced using the same WGS protocol as described for MANOLIS and Pomak. An identical phenotype transformation was performed on 275 proteins from the CVDII, III and META Olink panels in 995 samples. Because of quality control, between 928 and 950 samples overlapped between the WGS and Olink datasets. All 255 proteins analyzed in MANOLIS and Pomak were also found in the ORCADES dataset. Association was performed using GCTA v.1.93.0 beta using the MLMA algorithm (33). In ORCADES, using common LD-pruned variants for calculating the relatedness matrix was not sufficient, as persistent inflation was present. We assumed this was because of a different relatedness structure being expressed in rare variants, and we therefore included all sequence variants in the relatedness calculation, using five partitions of the autosomal genome. Following this, inflation was controlled. We sought replication for each of the 370 independent variants identified by COJO that are present in at least one cohort, using a Bonferroni threshold of 0.05/371 = 1.35 × 10^{-4}.184 variants replicated in this way.

Novelty

Previous associations with identical proteins was of particular interest as it determines novelty of our findings. To assess whether a protein had been previously studied, we examined protein lists and summary statistics from 33 large published proteomics GWAS (Supplementary Material, Table S3). To determine the novelty of genetic *cis*- and *trans*-association with proteins in our study, we first determined previously reported variants within a 2 Mb window around the association peaks. We used GEMMA (27) to perform association analysis using previously reported independent variants as covariates. The variants were declared novel if either there were no known signals in the 2 Mb window, or the associations were still study-wide significant (*P*-value threshold: 7.45×10^{-11}) after conditioning. For *trans* associations, we further annotated signals depending on whether they fell within highly pleiotropic genes that were associated with more than 1 protein in the current study and had evidence of additional associations in the literature (*KLKB1, ABO, APOE, FUT2, F12, VTN, CFH, HLA*), or whether they were independent of any *cis* signals in the vicinity. After this procedure, 42 *cis*-associated variants for 30 proteins were either not within 1 Mb or independent of a signal reported in previous proteomics GWAS. In sum, 37 *trans*-associated variants for 34 proteins were both novel and independent from *cis* loci. Only 15 of these were not located within highly pleiotropic genes. For all loci annotated as provisionally novel using the above method, we queried the GWAS Catalogue (34) (https://www.ebi.ac.uk/gwas/home)in a 2 Mb window through the Ensembl (35) REST API, as well as our PhenoScanner results. As proteomics GWAS signals are often designated generically in Ensembl, we additionally performed direct queries to the GWAS catalogue REST API when phenotype descriptions were not specific enough. We manually investigated the list of signals in search of variants associated with the protein trait of interest. When such a variant was found, conditional analysis was performed and the novelty status was updated accordingly. Novelty of each independent variant is annotated in Supplementary Material, Table S2.

Variant consequences

Consequence was evaluated using Ensembl VEP (35,36) for each variant with respect to any transcript of the *cis* gene for *cis*-associated variants and to the mapped gene for *trans*-associated variants. For *trans* associations, variants were manually mapped to any gene in a 1 Mb window coding for known ligands or interactants when they were not contained within gene boundaries. In sum, 38 replicating independent variants were protein-altering variants with a most severe consequence equal to or more severe than missense (https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html) according to Ensembl VEP. For every variant, we extracted tagging SNVs at $r^2 > 0.8$ using PLINK; however, none of these tagging variants had a more severe consequence on the target gene than the independent variant. Similarly, we overlapped all independent variants with regulatory features using the Ensembl REST API. 21 variants in 19 loci overlapped with a regulatory feature. Variant consequences are annotated in Supplementary Material, Table S2.

Gene expression QTL colocalization

We perform colocalization testing with eQTL data from the GTEx database (37) (https://gtexportal.org/home/). First, to account for multiple independent variants at the same locus, for every signal, regions are extended 1 Mb either side of every independent variant, and associations are conditioned on every other variant in the peak using GCTA-COJO; the results are used as input for the colocalization analysis. For *cis* signals, expression information for the *cis* gene is extracted from the GTEx database over the same region. For *trans* signals, expression information is restricted to all genes located within a 2 Mb region surrounding the variant. Then, for every variant/gene pair, we perform colocalization testing using the fast.coloc function from the gtx R package (https://github.com/tobyjohnson/gtx).We use the commonly chosen value of 0.8 as a posterior threshold to declare colocalization (38), and default values of 1×10^{-4}, with a standard deviation of 1, for the prior probability of a variant to be causal for either trait, and 1×10^{-5}, with a standard deviation of 1, for the prior probability of a variant to be causal for both traits. In sum, 77 (35%) independent *cis* variants colocalize with an expression quantitative trait locus for the *cis* gene. In addition, we find that 61 (73%) *trans*-pQTL variants colocalize with eQTLs for at least one gene in their vicinity (±1 Mb), in any tissue (Supplementary Material, Table S11; Supplementary Material, Figs S12–S13; Supplementary Material, Note 1).

PheWAS colocalization

We use the PhenoScanner python command line tool (39,40) (https://github.com/phenoscanner/phenoscannerpy) to query 1 Mb upstream and downstream of every lead variant in each signal. We only considered previous associations with a reported *P*-value of $0 < P < 5 \times 10^{-8}$. Using the PhenoScanner associations, we then perform colocalization testing using the same input pQTL data and methods that were used for the eQTL colocalization analysis. We additionally perform colocalization testing using downloaded summary statistics for atrial fibrillation, T2D, Alzheimer's disease, albuminuria, BMI, waist-hip ratio, estimated glomerular filtration rate, diabetic kidney disease and lipid levels. References to each study and full pheWAS colocalization results are presented in Supplementary Material, Table S7.

Drug target evaluation

For evaluating whether associated genes were drug targets, we used the OpenTargets (41) and DrugBank (42) databases. We accessed OpenTargets using the OpenTarget API. We converted the DrugBank XML file to flat files using the dbparser R package, and performed gene name matching using the USCS Gene Info database (https://genome.ucsc.edu/), downloaded May 6, 2019.35 of the proteins for which a signal was detected at study-wide significance were targeted by drugs according to OpenTargets. This was true for 70 proteins when queried against the DrugBank database (Supplementary Material, Table S12). In sum, 29 proteins are targeted by drugs according to both OpenTargets and DrugBank databases.

Mouse phenotype evaluation

We use the Ensembl (35) REST API to extract mouse orthologs for all of the 170 genes that encode proteins for which genetic associations were found in our study. According to the IMPC (43) API (https://www.mousephenotype.org/), KO experiments for 36 of these orthologs were associated with 70 unique phenotypes, with a *P*-value smaller than 1×10^{-4} (Supplementary Material, Table S13).

Two-sample MR

We extracted variants characterized as independent signals by GCTA-COJO (30) on a protein-by-protein basis across all *cis*- and *trans*-loci, and excluded novel variants without an rsID. For each remaining variant, we then extracted their pQTL summary statistics. When a variant was not present in the outcome GWAS summary statistics, we considered pQTL summary statistics for tagging positions with $r^2 > 0.8$. All such records were then merged by protein and carried over to MR analysis using the MRBase R package (44), where they were merged with the exposure datasets by rsID. MR was performed for 105 proteins on a set of 261 medically relevant traits available in MRBase. We defined cardiometabolic traits as: all lipid traits; glycaemic traits; diabetes; kidney disease and measures of kidney function; all heart conditions; hypertension; and BMI. These are annotated in Supplementary Material, Table S6. As all of our instruments involved a small number of variants (≤10), we used the inverse-variance weighted method, except for single-instrument analyses where we use the Wald ratio test, which consists of dividing the instrument-outcome by the instrument-exposure regression coefficient. All *P*-values were adjusted for multiple testing using the Benjamini–Hochberg method, using the adjusted $P < 0.05$ as the threshold for significant association.

An important caveat of our overlap-maximizing approach is that we did not require overlapping variants to be lead variants in the outcome trait GWAS. This could potentially lead to false-positives for single-instrument tests if the variant is located at the shoulders of an association peak in the outcome trait GWAS. The future availability of population-scale association studies with WGS or WES will greatly enhance the variant overlap compared with GWAS, and hence increase the power of MR analyses in proteomics. In addition to summary statistics available in MRBase, we also leveraged summary statistics manually downloaded from recent large association studies for: albuminuria, diabetic kidney disease, atrial fibrillation, BMI, CAD, lipid levels, T2D. PMID references for these studies are provided in Supplementary Material, Table S6.

Mep1b mouse model

Mep1b −/− (C57BL/6 N) mouse model is described in our previous study (45). The targeted mutation leads to the disruption of the catalytic centre in exon7 of the wild-type allele.

Mouse phenotyping

Mice were maintained in IVC cages with water and standard mouse chow according to the directive 2010/63/EU, German laws and GMC housing conditions (https://www.mouseclinic.de). All tests were approved by the responsible authority of the district government of Upper Bavaria.

In total, 18 mutant mice (9 males, 9 females) and wild-type control littermates (10 males, 10 females) underwent a systematic, comprehensive phenotyping screen by the German Mouse Clinic at the Helmholtz Zentrum Muenchen (https://www.mouseclinic.de) as previously described (46–49). This screen started at the age of 8 and 9 weeks for male and females respectively and covered multiple parameters in the areas of behaviour, cardiovascular function, clinical chemistry, dysmorphology, energy metabolism, eye analysis and vision, haematology, immunology, neurology, allergy and pathology.

Body weight

Body weight was measured at different time-points at a range of 8–19 weeks.

Body composition analysis

Body composition was analyzed at 13 and 18 weeks. Lean tissue and body fat in live mice without anaesthesia were measured by the whole-body composition analyzer (Bruker MiniSpec LF 50) based on Time Domain Nuclear Magnetic Resonance.

Blood collection

Blood samples were collected under isoflurane anaesthesia by retrobulbar puncture after overnight food withdrawal at 11–12 weeks of age and as a final blood withdrawal from ad libitum fed animals at 19–20 weeks. Blood samples for clinical chemistry analyses were collected in Li-heparin-coated tubes and stored at room temperature for one to three hours until centrifugation (4500 × g, 10 min) and separation of plasma aliquots for further analyses.

Clinical chemistry

The clinical chemistry analyses of circulating biochemical parameters in blood was performed using a clinical chemistry analyzer (AU480 autoanalyzer, Beckman Coulter, Krefeld, Germany). Fasting plasma lipid and glucose levels at 11–12 weeks of age and a broad set of parameters from fed animals at 19–20 weeks were measured using the respective kits provided by Beckman Coulter, including various enzyme activities as well as plasma concentrations of specific substrates and electrolytes in *ad libitum* fed mice (50).

Statistics

Data generated by the German Mouse Clinic were analyzed using R (Version 3.2.3). Tests for genotype effects were made by Wilcoxon rank sum test, linear models, or ANOVA depending on the assumed distribution of the parameter and the questions addressed to the data. A *P*-value <0.05 has been used as level of significance; a correction for multiple testing has not been performed. Figures were prepared using GraphPad Prism version 7.00 for Windows (GraphPad Software, La Jolla, California, USA).

Acknowledgements

We would like to thank Professor Claudia Langenberg for her valuable input. HELIC-MANOLIS and HELIC-Pomak studies: We thank the residents of the Pomak and Mylopotamos villages for taking part in this study. The GATK3 programme was made available through the generosity of the Medical and Population Genetics programme at the Broad Institute, Inc. We thank the Human Genetics DNA Pipelines and Human Genetics Informatics departments at the Wellcome Sanger Institute for performing sequencing and variant calling. This study has been conducted using the UK Biobank Resource (application number 10205). The Orkney Complex Disease Study (ORCADES): DNA extractions were performed at the Edinburgh Clinical Research Facility, University of Edinburgh. We would like to acknowledge the invaluable contributions of the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney.

Conflict of Interest statement. The authors declare no conflicts of interest.

Funding

The Wellcome Trust (098051) and the European Research Council (ERC-2011-StG 280559- SEPI); the Chief Scientist Office of the Scottish Government (CZB/4/276, CZB/4/710, a Royal Society URF to J.F.W.); Arthritis Research UK; the European Union Framework Programme 6 EUROSPAN project (contract no. LSHG-CT-2006-018947); the MRC Human Genetics Unit programme grant, ‘Quantitative traits in health and disease’ (U. MC_UU_00007/10 to J.F.W.); the National Productivity Investment Fund (MR/R026408/1 to L.K.); the German Federal Ministry of Education and Research (Infrafrontier Grant 01KX1012 to M.H.d.A.) and the German Center for Diabetes Research (DZD) (to M.H.d.A.); the Deutsche Forschungsgemeinschaft (DFG) (grant number 125440785 SFB 877), Projects A9 (to C.B.-P.), A15 (to C.B.-P. and C.U.P.).

References

1. Roth, G.A., Mensah, G.A., Johnson, C.O., Addolorato, G., Ammirati, E., Baddour, L.M., Barengo, N.C., Beaton, A.Z., Benjamin, E.J.,

Benziger, C.P. *et al.* (2020) Global burden of cardiovascular diseases and risk factors, 1990–2019. *J. Am. Coll. Cardiol.*, **76**, 2982–3021.

2. Lin, X., Xu, Y., Pan, X., Xu, J., Ding, Y., Sun, X., Song, X., Ren, Y. and Shan, P.-F. (2020) Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. *Sci. Rep.*, **10**, 14790.
3. Ferreira, J.P., Sharma, A., Mehta, C., Bakris, G., Rossignol, P., White, W.B. and Zannad, F. (2021) Multi-proteomic approach to predict specific cardiovascular events in patients with diabetes and myocardial infarction: findings from the EXAMINE trial. *Clin. Res. Cardiol.*, **110**, 1006–1019.
4. Feldreich, T., Nowak, C., Fall, T., Carlsson, A.C., Carrero, J.-J., Ripsweden, J., Qureshi, A.R., Heimbürger, O., Barany, P., Stenvinkel, P. *et al.* (2019) Circulating proteins as predictors of cardiovascular mortality in end-stage renal disease. *J. Nephrol.*, **32**, 111–119.
5. Cauwenberghs, N., Sabovčik, F., Magnus, A., Haddad, F. and Kuznetsova, T. (2021) Proteomic profiling for detection of early-stage heart failure in the community. *ESC Heart Fail.*, **8**, 2928–2939.
6. Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P. *et al.* (2018) Genomic atlas of the human plasma proteome. *Nature*, **558**, 73–79.
7. Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., Sun, B.B., Laser, A., Maranville, J.C., Wu, H. *et al.* (2018) Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.*, **9**, 3268.
8. Folkersen, L., Gustafsson, S., Wang, Q., Hansen, D.H., Hedman, Å.K., Schork, A., Page, K., Zhernakova, D.V., Wu, Y., Peters, J. *et al.* (2020) Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab.*, **2**, 1135–1148.
9. Emilsson, V., Ilkov, M., Lamb, J.R., Finkel, N., Gudmundsson, E.F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S.R., Aspelund, T. *et al.* (2018) Co-regulatory networks of human serum proteins link genetics to disease. *Science*, **361**, 769–773.
10. Gilly, A., Park, Y.-C., Png, G., Barysenka, A., Fischer, I., Bjørnland, T., Southam, L., Suveges, D., Neumeyer, S., Rayner, N.W. *et al.* (2020) Whole-genome sequencing analysis of the cardiometabolic proteome. *Nat. Commun.*, **11**, 6336.
11. Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., DeLisle, R.K. *et al.* (2017) Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.*, **8**, 14357.
12. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M.A., Oerton, E., Cook, J., Stewart, I.D., Kerrison, N.D. *et al.* (2021) Mapping the proteo-genomic convergence of human diseases. *Science*, **374**, eabj1541.
13. McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A. *et al.* (2008) Runs of homozygosity in European populations. *Am. J. Hum. Genet.*, **83**, 359–372.
14. Combadiere, C., Ahuja, S.K. and Murphy, P.M. (1995) Cloning and functional expression of a human eosinophil CC chemokine receptor. *J. Biol. Chem.*, **270**, 16491–16494.
15. Panoutsopoulou, K., Hatzikotoulas, K., Xifara, D.K., Colonna, V., Farmaki, A.-E., Ritchie, G.R.S., Southam, L., Gilly, A., Tachmazidou, I., Fatumo, S. *et al.* (2014) Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat. Commun.*, **5**, 5345.
16. Howard, E.W. and Banda, M.J. (1991) Binding of tissue inhibitor of metalloproteinases 2 to two distinct sites on human 72-kDa gelatinase. Identification of a stabilization site. *J. Biol. Chem.*, **266**, 17972–17977.
17. Zhong, F., Chen, Z., Zhang, L., Xie, Y., Nair, V., Ju, W., Kretzler, M., Nelson, R.G., Li, Z., Chen, H. *et al.* (2018) Tyro3 is a podocyte protective factor in glomerular disease. *JCI Insight*, **3**, 123482.
18. Ochodnicky, P., Lattenist, L., Ahdi, M., Kers, J., Uil, M., Claessen, N., Leemans, J.C., Florquin, S., Meijers, J.C.M., Gerdes, V.E.A. *et al.* (2017) Increased circulating and urinary levels of soluble TAM receptors in diabetic nephropathy. *Am. J. Pathol.*, **187**, 1971–1983.
19. Brown, M.S. and Goldstein, J.L. (1984) How LDL receptors influence cholesterol and atherosclerosis. *Sci. Am.*, **251**, 58–66.
20. Broder, C. and Becker-Pauly, C. (2013) The metalloproteases meprin α and meprin β: unique enzymes in inflammation, neurodegeneration, cancer and fibrosis. *Biochem. J.*, **450**, 253–264.
21. Jefferson, T., Auf dem Keller, U., Bellac, C., Metz, V.V., Broder, C., Hedrich, J., Ohler, A., Maier, W., Magdolen, V., Sterchi, E. *et al.* (2013) The substrate degradome of meprin metalloproteases reveals an unexpected proteolytic link between meprin β and ADAM10. *Cell. Mol. Life Sci.*, **70**, 309–333.
22. Monami, M., Lamanna, C., Desideri, C.M. and Mannucci, E. (2012) DPP-4 inhibitors and lipids: systematic review and meta-analysis. *Adv. Ther.*, **29**, 14–25.
23. Pierrot, N., Tyteca, D., D'auria, L., Dewachter, I., Gailly, P., Hendrickx, A., Tasiaux, B., Haylani, L.E., Muls, N., Nkuli, F. *et al.* (2013) Amyloid precursor protein controls cholesterol turnover needed for neuronal activity. *EMBO Mol. Med.*, **5**, 608–625.
24. Gilly, A., Klaric, L., Park, Y.-C., Png, G., Barysenka, A., Marsh, J.A., Tsafantakis, E., Karaleftheri, M., Dedoussis, G., Wilson, J.F. *et al.* (2022) Gene-based whole genome sequencing meta-analysis of 250 circulating proteins in three isolated European populations. *Mol. Metab.*, **61**, 101509.
25. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Kerrison, N.D., Oerton, E., Koprulu, M., Luan, J., Hingorani, A.D., Williams, S.A., Wareham, N.J. *et al.* (2021) Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.*, **12**, 6822.
26. de Auwera, G.A.V. and O'Connor, B.D. (2020) *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*, 1st edn. O'Reilly, Beijing, Boston, Farnham, Sebastopol, Tokyo.
27. Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.
28. Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinform. Oxf. Engl.*, **26**, 2190–2191.
29. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.
30. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
31. Mägi, R. and Morris, A.P. (2010) GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics*, **11**, 288.
32. Magi, R., Lindgren, C.M. and Morris, A.P. (2010) Meta-analysis of sex-specific genome-wide association studies. *Genet. Epidemiol.*, **34**, 846–853.
33. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. and Price, A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.
34. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

35. Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
36. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P. and Cunningham, F. (2016) The Ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
37. Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., Buia, S.A., Compton, C.C., DeLuca, D.S., Peter-Demchok, J., Gelfand, E.T. *et al.* (2015) A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv. Biobank.*, **13**, 311–319.
38. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.
39. Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J. *et al.* (2016) PhenoScanner: a database of human genotype-phenotype associations. *Bioinform. Oxf. Engl.*, **32**, 3207–3209.
40. Kamat, M.A., Blackshaw, J.A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A.S. and Staley, J.R. (2019) PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinform. Oxf. Engl.*, **35**, 4851–4853.
41. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M. *et al.* (2021) Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.*, **49**, D1302–D1310.
42. Wishart, D.S. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
43. The International Mouse Phenotyping Consortium, Dickinson, M.E., Flenniken, A.M., Ji, X., Teboul, L., Wong, M.D., White, J.K., Meehan, T.F., Weninger, W.J., Westerberg, H. *et al.* (2016) High-throughput discovery of novel developmental phenotypes. *Nature*, **537**, 508–514.
44. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R. *et al.* (2018) The MR-Base platform supports systematic causal inference across the human phenome. *eLife*, **7**, e34408.
45. Norman, L.P., Jiang, W., Han, X., Saunders, T.L. and Bond, J.S. (2003) Targeted disruption of the meprin beta gene in mice leads to underrepresentation of knockout mice and changes in renal gene expression profiles. *Mol. Cell. Biol.*, **23**, 1221–1230.
46. Gailus-Durner, V., Fuchs, H., Becker, L., Bolle, I., Brielmeier, M., Calzada-Wack, J., Elvert, R., Ehrhardt, N., Dalke, C., Franz, T.J. *et al.* (2005) Introducing the German Mouse Clinic: open access platform for standardized phenotyping. *Nat. Methods*, **2**, 403–404.
47. Fuchs, H., Gailus-Durner, V., Adler, T., Pimentel, J.A.A., Becker, L., Bolle, I., Brielmeier, M., Calzada-Wack, J., Dalke, C., Ehrhardt, N. *et al.* (2009) The German Mouse Clinic: a platform for systemic phenotype analysis of mouse models. *Curr. Pharm. Biotechnol.*, **10**, 236–243.
48. Fuchs, H., Aguilar-Pimentel, J.A., Amarie, O.V., Becker, L., Calzada-Wack, J., Cho, Y.-L., Garrett, L., Hölter, S.M., Irmler, M., Kistler, M. *et al.* (2018) Understanding gene functions and disease mechanisms: phenotyping pipelines in the German Mouse Clinic. *Behav. Brain Res.*, **352**, 187–196.
49. Fuchs, H., Gailus-Durner, V., Adler, T., Aguilar-Pimentel, J.A., Becker, L., Calzada-Wack, J., Da Silva-Buttkus, P., Neff, F., Götz, A., Hans, W. *et al.* (2011) Mouse phenotyping. *Methods*, **53**, 120–135.
50. Rathkolb, B., Hans, W., Prehn, C., Fuchs, H., Gailus-Durner, V., Aigner, B., Adamski, J., Wolf, E. and Hrabě de Angelis, M. (2013) Clinical chemistry and other laboratory tests on mouse plasma or serum. *Curr. Protoc. Mouse Biol.*, **3**, 69–100.

Appendix C

The two articles included in Appendices A and B are Open Access articles published under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited. Proof of permission from the publishers have been provided via RightsLink from the Copyright Clearance Center (https://www.copyright.com/solutions-rightslink-permissions/), and can be found in the following pages.

SPRINGER NATURE

Mapping the serum proteome to neurological diseases using whole genome sequencing

Author: Grace Png et al

Publication: Nature Communications

Publisher: Springer Nature

Date: Dec 2, 2021

Creative Commons

You are not required to obtain permission to reuse this article.
To request permission for a type of use not listed, please contact Springer Nature

Identifying causal serum protein–cardiometabolic trait relationships using whole genome sequencing

Author: Png, Grace; Gerlini, Raffaele

Publication: Human Molecular Genetics

Publisher: Oxford University Press

Date: 2022-11-09

Creative Commons