Ancestry diversity in the genetic determinants of the human plasma proteome and associated new drug targets

Saredo Said[1], Alfred Pozarickij[1], Kuang Lin[1], Sam Morris[1], Christiana Kartsonaki[1,2], Neil Wright[1], Hannah Fry[1,2], Yiping Chen[1,2], Huaidong Du[1,2], Derrick Bennett[1,2], Daniel Avery[1,2], Dan Valle Schmidt[1,2], Liming Li[3,4,5], Jun Lv[3,4,5], Canqing Yu[3,4,5], Dianjianyi Sun[3,4,5], Pei Pei[4], Junshi Chen[6], Michael Hill[1], Richard Peto[1], Rory Collins[1], Robert Clarke[1], Iona Y Millwood[1,2], Zhengming Chen[1,2], Robin G Walters[1,2], on behalf of China Kadoorie Biobank Collaborative Group[‡]

1 Clinical Trial Service Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
2 Medical Research Council Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
3 Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China
4 Peking University Center for Public Health and Epidemic Preparedness and Response, Beijing, China
5 Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China
6 China National Center for Food Safety Risk Assessment, Beijing, China

[‡] *Members of the CKB Collaborative Group are shown in the Appendix*

Address for correspondence:

Prof Robin Walters
NDPH
University of Oxford
Old Road Campus
Roosevelt Drive
Oxford OX3 7LF
robin.walters@ndph.ox.ac.uk

Prof Zhengming Chen
NDPH
University of Oxford
Old Road Campus
Roosevelt Drive
Oxford OX3 7LF
zhengming.chen@ndph.ox.ac.uk

Word count: abstract 150, main text 6,233
(5 figures, 1 table and 26 eTables/eFigures)

CKB Research Tracker No.: 2022-0031

Summary (word 149 / max 150)

The proteome is fundamental to human biology and disease but little is known about ancestral diversity of its genetic determinants. In GWAS of plasma levels of 1,451 proteins in 3,974 Chinese adults, we identified pQTLs for 1,082 proteins, including 743 with at least one *cis*-pQTL. Fine-mapping defined credible sets for 3,336 independent pQTLs, of which 31% did not overlap with corresponding analyses in European adults. We assessed 777 sentinel *cis*-pQTLs in phenome-wide MR analyses using GWAS Catalog and identified Bonferroni-significant associations for 22 protein-disease pairs. Among 10 protein-disease pairs identified from East Asian-specific GWAS, four had evidence of colocalisation. Evaluation of current drug development confirmed indications for one protein target, identified potential repurposing for seven, and discovered nine potential novel targets, including GP2 for Type-2-diabetes. The findings demonstrate the importance of extending genome-wide plasma proteomic analyses to non-European ancestry populations to identify potential novel drug targets for major diseases.

Introduction

Proteins mediate the effects of genes in determining body structure and function and play a key role in human health and disease[1,2]. These include proteins that are actively secreted such as classic plasma proteins (e.g. albumin secreted by the liver or lipoproteins by gastrointestinal organs), hormones, immunoglobulins, membrane proteins, exogenous proteins from infectious organisms released into circulation and proteins present in plasma due to cellular leakage[1]. Most existing drug targets are proteins [3,4], and of the protein genes with experimental evidence of disease involvement ~1,800 are yet to be evaluated as potential drug targets in experimental studies[5,6].

Previous quantification of protein levels has mostly been limited to low-throughput proteomic assays, including antibody-based ELISA or western blot assays[7]. Advances in high-throughput proteomics assays (e.g. Olink, SomaScan or mass-spectrometry) now enable measurement of plasma levels of several thousand proteins simultaneuosly[7–9]. Application of these novel assays in clinical and population studies afford new research opportunities for hypothesis testing and gene-protein discovery[10].

Integrated analyses of plasma proteomics with genetic data and disease outcomes in population-based biobanks have great potential to enhance our understanding of the role of proteins in health and disease[11,12]. Discovery of the genetic determinants of protein levels[13,14] is important for elucidation of the causal relevance of individual proteins for disease risk, and to identify potential drug targets for treatment or prevention of such diseases[15]. Plasma proteomic analyses can also improve risk prediction[16] and enhance early detection and diagnosis of specific diseases. Thus, large-scale proteomics can help to extend the scope of discoveries from candidate-driven to hypothesis-free investigations in diverse populations[13,17].

Identification of genetic variants associated with altered protein levels (protein quantitative trait loci, pQTLs) can provide instruments for use in Mendelian randomisation (MR) analyses to assess the causal relevance of individual proteins for particular diseases or traits (e.g. adiposity, blood pressure)[18–20], thereby identifying novel or repurposing drug targets[15]. To date, analysis of the genetic architecture of the plasma proteome has been largely restricted to European ancestry (EUR) populations[13,21], including the UK Biobank Pharma Proteomics Project (UKB-PPP)[22,23]. The GWAS of 2,923 proteins assayed in ~50,000 UK Biobank (UKB) participants has expanded the catalogue of pQTLs for the plasma proteome measured by the Olink Explore platform[22,23] . However, little is known about the extent to which these differ in other ancestry populations[13,21], for the discovery cohort in UKB-PPP was confined to EUR individuals, with replication in a small number of non-EUR populations, including 148 individuals of East Asian ancestry (EAS)[22,23]. To date, no well-powered comprehensive genetic analyses of the Olink Explore platform have yet been reported in an EAS population.

The present study examined the genetic architecture of the plasma proteome in ~4,000 adults from the China Kadoorie Biobank (CKB). The aims of this study were to: (i) conduct GWAS analyses of 1,451 proteins measured by the Olink Explore panel; (ii) compare the genetic architecture of the proteome between EAS and EUR populations; (iii) assess the causal relevance of plasma proteins for a range of diseases; and (iv) explore the potential of specific proteins as drug targets.

Results

Figure 1 summarises the study design, analytic approaches and main findings. Details of the population characteristics are provided in **Table S1**. The distributions of all protein measurements are shown in **Figure S1**. Overall twelve proteins were excluded from analysis due to overt bi-modal distributions (**Table S2**).

GWAS pQTL discovery

Overall, GWAS analyses of 1,451 proteins measured in 3,974 CKB participants identified 2,091 autosomal pQTLs, comprising 2,872 conditionally independent associations at genome-wide significance (i.e. $P \leq 5 \times 10^{-8}$) (**Figure 1, Table S4, S5**). At least one pQTL was identified for 1,082 (75%) unique proteins and among them 743 (69%) proteins had at least one *cis*-pQTL (defined as those with their sentinel variant within ±500Kbp of the protein structural gene, **Table S3**), with *cis*-pQTLs accounting for 37.3% (779/2,091) of all pQTLs. When applying a more stringent significance threshold to account for multiple testing across 1,451 proteins ($P \leq 5 \times 10^{-8}/1{,}451 = 3.45 \times 10^{-11}$), 1,174 pQTLs for 805 proteins were identified comprising 1,683 conditionally independent associations, of which 1,154 were *cis*-pQTL.

Figure 2A shows the associations of all autosomal conditional pQTL sentinel variants (**Table S4**) with measured proteins (including suggestive associations at P-value$<5 \times 10^{-6}$), with many loci having associations across the proteome. The per-allele effect estimates for the sentinel variant at each conditionally independent pQTL (**Figure 2B**) ranged from –2.40 to 3.02 (median –0.14) SDs in protein level, with larger mean absolute effect sizes at *cis* (mean=0.09) than at *trans*-pQTLs (mean=0.04). Minor alleles were associated with lower protein levels for 53% of *cis*- and 48% of *trans*-pQTLs.

For the 1,082 proteins with at least one pQTL, heritability estimates based on all identified pQTLs (P-value<$5x10^{-8}$) ranged from 0.75% for MTSN to 72% for PDGFRB (**Figure 2C**, **Table S6**). Estimates of *cis*-heritability ranged from 0.77% (MITD1) to 72% (PDGFRB). Of the 341 proteins with no *cis*-pQTLs, *trans*-heritability ranged from 0.75% (MSTN) to 48% (for ICAM2). Heritability estimates in CKB were strongly correlated with those in UKB for both overall (r = 0.75) and *cis*-heritability (r = 0.74) (**Figure S2**).

Among the 1,082 proteins with pQTLs, there was a median of two independent pQTLs per protein (range 1 to 19), with 31% having a single pQTL and 12% having ≥5 pQTLs (**Figure 2D**). The largest number of independent pQTLs for a single protein was 19 (KIR3DL1), followed by 14 (ASAH2, CD177 and SFTPA2), and 12 (CCL8, IL6R and LILRB2) (**Figure 2D).**

Fine-mapping

Fine-mapping at the 2,091 pQTLs identified 3,336 credible sets (CSs) for the association signals for 1,080 proteins (**Tables S7**-**S8**), with at least one additional independent signal being identified for 378 proteins by this method compared to step-wise conditional analysis. CSs were not identified for pQTLs for two proteins (TNFRSF1B and TP53), for which SuSiE fine-mapping did not converge. CSs contained a median of 6 variants (mean 19.8, range 1 to 770), with 789 CSs (24%) consisting of a single variant. Overall, *cis*-pQTLs had a better resolution than *trans*-pQTLs (*cis*: median 5, mean 15.4 variants per CS; *trans*: median 8, mean 25.1), *and* were more likely to have a single variant CS (*cis*: N=431, 12.9%; *trans*: N=358, 10.7%). For pQTLs which met the Bonferroni-adjusted threshold for pQTL discovery (P-value<$3.45x10^{-11}$), there was a median of 7 variants per CS (mean 20.3), and again *cis*-pQTLs were more likely to have a single variant CS (*cis*: N=248, 7.4%; *trans*: N=192, 5.8%).

To identify pQTLs common to multiple proteins, CSs from all proteins were compared and those with one or more overlapping variants were merged into a single locus. This identified 212 loci associated with at least 2 different proteins, of which 17 were associated with ≥10 proteins (**Figure 2E**, **Table S9**). The locus at chr9:133227766-133279871, corresponding almost precisely to the *ABO* blood group-determining locus, was associated with levels of 133 proteins. The next most pleiotropic locus, associated with 53 proteins, was chr6:29444646-31088094 within the HLA region, followed by chr19:48642611-48711017 (*NTN5*/*FUT2*, 30 proteins), chr8:105534496-105566749 (*ZFPM2*, 25), and chr7:80547085-80582402 (*CD36*, 23).

To investigate the functional characteristics of likely causal variants, the lead variants for each CS (i.e. with the highest posterior inclusion probability, PIP) (see **Supplementary Table ST8**) were categorised according to their functional impact and location with respect to coding and transcript sequences, and according to whether they affected the gene encoding the target (i.e. assayed) protein, a gene for a different protein within the *cis* region, or at a *trans*-pQTL (**Figure 3A**).

For *cis*-pQTLs, 999 lead variants were within the transcript for the target protein, although these were predominantly in non-coding regions (**Figure 3A**). Among those located in the coding region, 304 were non-synonymous (i.e. affecting the amino-acid sequence of the translated protein, e.g. stop-gain, missense, insertion/deletion). For *trans*-pQTLs 892 lead variants were located in non-coding regions of gene transcripts.

Comparison with European ancestry pQTLs

Based on comparisons with reported pQTLs from the UKB-PPP [22], two proteins had a *cis*-pQTL in CKB but not in UKB-PPP: CDH1 (rs28372783) and HARS1 (rs117579809). Furthermore, the GWAS sentinel variants for a further 94 CKB *cis*-pQTLs were within the

structural gene whereas the *cis*-pQTL lead variants for those proteins reported in UKB-PPP lay outside the structural gene (**Table S10**). Additionally, for 31% (1018/3336) of all CSs identified in CKB there was no overlap with CSs for that protein from UKB-PPP, comprising 16% (284/1820) of *cis*-pQTLs CSs and 48% (734/1516) of *trans*-pQTL CSs (**Figure 3B**). In particular, there were 152 proteins for which none of the CSs for *cis*-pQTLs in CKB had any overlap with corresponding CSs from UKB-PPP.

Phenome-wide scan and Mendelian randomisation analyses

EAS and multi-ancestry (including EAS) studies in the GWAS Catalog were searched for associations with 777 unique *cis*-pQTL sentinel variants (where a protein had >1 *cis*-pQTL, we selected the lead SNP for each locus) (**Table S11)**; where association statistics were not available in the GWAS Catalog for a variant, the best available proxy (LD r^2>0.8) was used. All *cis*-pQTLs had F-statistics >10 and were potentially strong genetic instruments, including all proxy *cis*-pQTLs (**Table S10**). At the default search threshold of P-value<$5x10^{-6}$, a total of 176 protein-phenotype associations were identified (**Table S12)**. Of these, 132 met a Bonferroni-adjusted threshold of P-value<$6.40x10^{-10}$ (0.05 / 777 SNPs * 100544 traits tested), comprising 25 protein-disease (**Figure 4**) and 106 protein-quantitative trait pairs (**Figure S3**) (35 and 141, respectively, at $5x10^{-6}$). Alternatively, 164 associations met genome-wide significance (P-value <$5x10^{-8}$), with 118 surpassing a Bonferroni-adjusted threshold of P-value <$6.44x10^{-11}$ ($5x10^{-8}$ / 777 variants tested).

We further assessed 22 of the 25 protein-disease pairs for which effect size estimates and effect alleles were available, using a two-sample MR Wald ratio approach (**Figure 5, Table S12**). Among the associations derived from EAS-specific studies, TNFSF13 showed the strongest adverse effect on disease risk (IgA nephropathy: OR=1.69, 95%CI 1.44-1.98 per 1 SD higher concentration), followed by GP2 (type-2-diabetes (T2D): 1.43, 1.59-1.28),

CD40 (Kawasaki disease: 1.39, 1.27-1.51), and FGF5 (hypertension: 1.35, 1.27-1.42). Conversely, there were strong protective effects of higher levels of HSPA1A (Takayasu arteritis: 0.11, 0.06-0.22) and TNFRSF10A (age-related macular degeneration: 0.47, 0.39-0.58; and central serous retinopathy: 0.49, 0.41-0.59), with moderate protective effects for a further three proteins (CD40, OBP2B, MSMB). Importantly, the inferred causal effects for HSPA1A and GP2 were based on CKB *cis*-pQTLs that were distinct from those previously reported in Europeans – none of the CSs for these pQTLs overlapped with those reported by UKB-PPP. Furthermore, AGER and ERBB2 *cis*-pQTLs were within the structural gene in CKB but not for the corresponding sentinel *cis*-pQTLs in UKB-PPP.

Multi-ancestry studies that included data from EAS samples identified further apparent causal associations of protein levels with disease risk, including AGER, ESAM, UMOD, IL10RB, NID2, FGF5 and SPON1. Similarly, indications of causal protective effects on disease risk were observed for ERBB2, SUSD2, UMOD and LRIG1. However, these were not investigated further, since they could potentially be driven by signals from non-EAS samples, and all were dependent on pQTLs with CS-overlap with pQTLs from EUR. Indeed, two of these latter protein-disease pairs (ESAM-schizophrenia and MSMB-prostate cancer) were significantly associated in a recent MR-PheWAS analysis in EUR[20]. (**Table S14**) – i.e. 20 of the protein-disease associations identified in CKB were not identified as significant in EUR[20]. Furthermore, two of these associations (AGER-COPD and ERBB2-asthma) involved lead variants within the corresponding structural gene for CKB, whereas the *cis*-pQTL in UKB-PPP was not.

Colocalisation

For EAS specific PheWAS-MR associations, we accessed six available disease summary statistics and used these in colocalisation analyses[24] (**Figure 5**, **Tables S14-S15**, and

Figures S4-S9). There was strong evidence for colocalisation between FGF5 and hypertension (posteriori probability H4=0.999), GP2 and T2D (H4=0.995), and TNFRSF10A and central serous retinopathy (H4=0.995). Although MSMB and prostate cancer did not colocalise in our original analysis (H4=1.31 $x10^{-5}$), exclusion of a single lead SNP associated with prostate cancer (see **Figure S7**) revealed a shared signal (H4=0.961); similarly, colocalisation of separate association signals identified in a SuSiE fine-mapping framework also identified a shared association signal between MSMB and prostate cancer. There was no evidence for colocalisation between CD40 and chronic hepatitis B infection (H4=0.016) nor OBP2B and gastric cancer (H4=2.45$x10^{-4}$).

Potential drug target and repurposing opportunities

For the 17 unique disease-associated proteins identified above, we searched the Therapeutic Target Database[25] and DrugBank[26] to identify whether there are existing drugs targeting them, and their corresponding disease indications (**Table 1**). We identified eight proteins targeted by a total of 24 relevant drugs either approved, under investigation in clinical trials, or reported in the literature. In particular, corresponding to our identification of MSMB as being protective for prostate cancer, Tigapotide (a synthetic 15-mer peptide derived from MSMB) is currently being evaluated for treatment of prostate cancer in a phase-1 clinical trial. By contrast, for seven proteins that are existing drug targets, we identified potential causal associations with diseases that are distinct from the current intended indications. These include HSPA1A, for which we found a very strong causal relationship with Takayasu arteritis but which is the target of six different drugs being evaluated for treatment of six different conditions, none with a direct relationship to Takayasu arteritis. Importantly, for nine proteins identified in the present study as having a

causal relationship to disease, including GP2, we identified no current drugs targeting them.

Additional PheWAS in CKB population subset

For the two proteins whose inferred causal effects were based on CKB *cis*-pQTLs, no unrelated associations were identified in our phenome-wide screen of GWAS Catalog. To further investigate whether these proteins might have potential causal associations with other diseases for which EAS GWAS data were not available, we conducted additional analyses in a population representative subset of CKB, testing for association of the corresponding sentinel *cis*-pQTL variants (SNP dosages aligned to allele increasing measured protein levels) with phecodes (aggregate ICD10 codes). After correction for multiple testing, we identified no associations which might indicate potential contra-indications for their use as drug targets (**Table S16**). The strongest association for GP2 was the expected association with increased risk of T2D (OR=1.12 [1.04-1.20], P-value=2.66×10^{-3}), while for HSPA1A the lead associated phenotype was increased risk of hypopotassemia (OR=1.75 [1.18-2.59], P-value=5.36×10^{-3}), in the opposite direction to the effect of HSPA1A on Takayasu arteritis.

Discussion

To date, no well-powered, comprehensive, genetic analyses of the human proteome have been reported in an EAS population[13,27]. In this study use of high-throughput proteomic assays has enabled detailed characterisation of the genetic architecture of plasma proteins in Han-Chinese adults and hypothesis-free analyses of proteome-phenome relationships, leading to identification of potential protein drug targets. We identified 2,872 independent pQTLs for 1,082 proteins at genome-wide significance, and fine-mapped likely causal *cis*-variants for 743 proteins. MR-PheWAS analyses identified 22 protein-disease causal associations, of which 20 have not previously been reported in EUR populations.

In the recently-reported GWAS of >50K UKB participants (94% White) using the same Olink Explore proteomic panel, pQTLs were identified for 94% (1,377/1,463) of the proteins, with 82% of proteins having at least one *cis*-pQTL[22]. Notably, despite an approximately ~10-fold smaller discovery sample size, we identified pQTLs at genome-wide significance for 1,082 out of 1,463 proteins (74%). Even at the more stringent Bonferroni-adjusted threshold used in UKB-PPP, we identified 1,668 pQTLs for 798 (54%). Furthermore, we identified *cis*-pQTLs for HARS1 and CDH1 whereas UKB-PPP did not. Systematic comparison of our study with UKB-PPP, by testing for overlap between corresponding sets of CSs for each protein, found that 31% of CKB CSs had no variants in common with the corresponding associations in UKB-PPP, demonstrating substantial ancestry differences in the genetic architecture of the proteome. Thus, extending population diversity in proteomics studies has substantial potential to increase the range and variety of genetic instruments that can be used for analysis of the relevance of proteins in disease outcomes. This should also help to reduce health disparities between

populations, by enhancing our understanding of disease aetiology across ancestries, with a positive impact on the effectiveness of genomics-guided therapeutic interventions [28,29].

In line with the findings in UKB-PPP [22], global assessment across all 2,872 association signals in CKB found only minor differences in the patterns of effect sizes of *cis*- and *trans*-pQTL minor alleles, with the minor alleles of lead variants being approximately equally likely to be associated with lower or higher measured protein levels. This is consistent with the hypothesis that only a small proportion of pQTL lead variants are predicted to alter protein coding, with the large majority being predicted to modulate protein abundance through regulatory effects. Thus, even for *cis*-pQTLs very few of the observed associations are likely to reflect disruption of the Olink assay rather than genuine effects on protein level, so that MR analyses using these variants are unlikely to be confounded by assay artefacts. Conversely, 9% of *trans*-pQTLs were associated with more than one of the assayed proteins, emphasising the risk of artefactual findings due to pleiotropy if these loci are included in MR instruments.

Comparison of pQTLs identified in CKB and UKB-PPP showed that while the overall genetic architecture of the proteome appeared similar, for instance with substantial pleiotropy at the same loci including *ABO*, *FUT2*, and within the HLA region, details of individual associations varied substantially. In addition to identifying *cis*-pQTLs for two proteins for which none were found in UKB-PPP, for a further 94 proteins we identified lead variants within the relevant structural gene that were not identified in UKB-PPP. Furthermore, 31% of all credible sets for CKB pQTLs did not overlap at all with corresponding credible sets in UKB-PPP, with 152 proteins having entirely distinct *cis*-pQTLs. We note that this is likely to represent an underestimate of the proportion of EAS-specific causal variants, since overlap between CKB and UKB-PPP CSs can occur even

for distinct causal variants. Thus, extending proteomic analyses to individuals of East Asian ancestry has great potential to enhance detection of causal associations of proteins with disease, for example where different causal variants influence circulating levels of protein through effects on protein expression in different tissues (e.g. due to secretion from those tissues or release of protein fragments from the cell surface). Indeed, in our search for GWAS results which at least included a proportion of EAS individuals, we identified a total of 176 protein-phenotype associations, of which 19 were based on associations with EAS-specific *cis*-pQTLs. Furthermore, 20 out of 22 protein-disease associations were not identified in a recent proteomics PheWAS using data from EUR populations [20]. For 6 diseases where full EAS-specific GWAS summary statistics were available, colocalisation analysis provided evidence of a causal role for 4 of the associated proteins.

The validity of this approach to the identification of potential drug targets is supported by our finding of a causal protective association between MSMB (microseminoprotein beta protein) and prostate cancer, which replicated previous findings in EUR[20] and reflects an existing drug target-disease indication. MSMB is synthesised in prostate epithelial cells and secreted into seminal fluid, and levels of MSMB protein have been consistently found to be lower in prostate cancer tissues compared with benign prostate tissues[30]. The drug Tigapotide, a synthetic 15-mer peptide derived from the MSMB protein sequence, is currently being evaluated in a phase-1 clinical trial to test whether the drug inhibits metastatic protein MMP-9 in prostate cancer patients can improve prognosis. While the association at the *MSMB* locus with prostate cancer risk is well established [30,31], our finding of colocalisation between proteomics and prostate cancer association signals provides further supporting evidence for continued drug development targeting MSMB for treatment of prostate cancer. It is of note that colocalisation was only identified for individual association signals and/or after excluding the primary prostate cancer signal, perhaps

reflecting a separate association with prostate cancer at the locus in a tissue that does not directly contribute to plasma levels of this protein, but which possibly affects levels in other, disease-relevant tissues.

Among the 20 protein-disease pairs not previously reported, two involved *cis*-pQTLs distinct to EAS (i.e. HSPA1A-Takayasu arteritis and GP2-T2D), and two for which *cis*-pQTLs within the relevant structural gene were only identified in EAS (i.e., AGER-COPD and ERBB2-asthma). None of these are indications for which there were drugs existing or in development which target the corresponding protein, and so represent new opportunities for drug development or repurposing. Such genetic support for protein drug targets of interest has been found to predict improved progression and approval of new drug developments by more than 2-fold[32], with two-thirds of new FDA approved drugs in 2021 supported by genetic evidence[33].

The two novel findings driven by EAS-specific *cis*-pQTLs exemplify how proteomic MR-PheWAS can add to and complement existing knowledge, to strengthen the case for development or reassessment of drugs targeting specific proteins, and how ancestry diversity is potentially important in this process. First, we identified Heat shock protein (HSP) family A member 1A (HSPA1A, also known as HSP70-1) as potentially being protective for Takayasu arteritis (also known as pulseless disease), a rare disease characterised by chronic granulomatous vasculitis affecting large arteries, chiefly the aorta and its major branches. Vascular inflammation in this disease can cause stenosis, occlusion or aneurysm formation and involves an autoimmune-mediated process leading to inflammation and stenosis[34]. HSPs are expressed across a broad range of cell and tissue types[35], and are involved in a range of autoimmune and inflammatory conditions[36,37], with important roles in cellular stress responses, protein folding,

maintaining protein-homeostasis and immune regulation[36,38]. Elevated HSPA1A has been identified as a biomarker of adverse vascular disease outcomes following cardiac events[39], and there is evidence suggesting it is protective against atherosclerosis and vascular risk[40]. Given the systemic nature of Takayasu arteritis and its association with immune dysfunction, and since variants within the chr6 MHC region surrounding the *HSPA1A* gene are associated with Takayasu arteritis in Chinese adults[41], a protective effect for HSPA1A is plausible, but remains to be established. With several drugs currently being evaluated in randomized trials, each of which is an analogue or activator of HSPA1A, there is an opportunity for drug repurposing to explore their potential for treating Takayasu arteritis. Notably, although it has been reported in other age groups and ancestries, Takayasu arteritis commonly presents in East and South Asian women before the age of 40[34] and, therefore, the identification of this possible drug target indication would have been unlikely in the absence of ancestry-diverse studies.

Conversely, while a relationship between GP2 (Glycoprotein 2, a trans-membrane protein) and T2D is well-established, the mechanistic basis for this remains obscure. Association at the *GP2* locus was identified first with risk of T2D in Japanese[42,43] and subsequently with gestational diabetes in a Chinese-Han population[44], but has not been found in EUR populations – consistent with the GP2 *cis*-pQTL identified in CKB being distinct from that in EUR. GP2 is implicated in pancreatic biology, being identified as a marker of pancreatic progenitor cells, in particular beta-cells[45], and with variants altering the GP2 protein being associated with both pancreatitis[46], and pancreatic cancer[47]. Nevertheless, no direct link between GP2 protein levels and T2D has been established. A complicating factor is that GP2 is primarily expressed not only in pancreatic tissue[35] but also in the gut, particularly in small intestine M cells[48], which play an important role in the immune response by sampling antigens from the gut lumen and presenting them to immune cells[48]. Gut microbiota

composition[49], intestinal permeability[50], and immune responses[51], have all been suggested as potentially contributing to T2D development. Our identification of a causal relationship between GP2 and T2D, strengthened by evidence from colocalisation that this reflects a shared genetic signal adds to the evidence that GP2 may be a potential drug target for prevention or treatment of T2D. Since no association between the GP2 locus and T2D has been found in EUR, elucidation of the precise relationship between GP2 protein levels and T2D, and of any potential therapeutic implications, will likely need to include studies in EAS populations, once again emphasising the importance of performing proteomics studies in different ancestral populations.

Despite having proteomics measures in ~4,000 participants, we were still able to identify a large number of pQTLs, highlighting the extensive genetic contribution to plasma levels of these proteins. Nevertheless, UKB-PPP found that increasing sample size would be likely to lead to identification of many additional pQTLs, particularly *trans*-pQTLs. Thus, an increase in the sample size for EAS, and other ancestries, would enable discovery of many additional pQTLs and corresponding associations with disease[22]. In addition, many proteins not included in the Olink platform can be interrogated using other methods, as shown in a recent report of many potential protein-disease associations based on mass-spectrometry proteomics of 304 proteins in 2,958 Han Chinese participants[27].

One important limitation of this and other proteomics studies is that analyses based on plasma protein levels do not necessarily reflect levels of the protein in disease-relevant tissue. Furthermore, functional alterations to a protein (e.g. arising from missense mutations) which do not affect the abundance or conformation of a protein might remain undetected, potentially masking associations between protein and diseases. Conversely, missense variants can result in detection of artefactual *cis*-pQTLs, if such variants affect

assay binding[52], which may disrupt MR approaches such that causal inference remains valid but the direction of effect may be uncertain[20]. However, since the majority of variants responsible for *cis*-pQTLs lie outside the coding regions the number of proteins affected will likely to be very small. Moreover, while initial protein-disease associations mainly used *cis*-pQTLs, additional analyses using *trans*-pQTLs can help to replicate initial MR findings and to further characterise upstream biological pathways. A further limitation of this study is that, while the large number of pQTLs identified enabled extensive exploratory hypothesis-generating investigations across the phenome, the number of phenotypes we were able to analyse was limited by the availability of suitable GWAS studies with EAS ancestry. This further highlights the growing need for further large-scale GWAS in non-EUR populations.

In conclusion, the present study is the largest proteome GWAS in an EAS population in terms of both number of proteins assayed and sample size[13]. Our analyses highlighted distinct ancestral differences in the genetic architecture of plasma proteomics and identified causal association for 22 protein-disease pairs which warrant further investigation. The findings demonstrate the importance of extending genome-wide plasma proteomic analyses to non-European ancestry populations to identify potential novel biomarkers and drug targets for major diseases.

Acknowledgments

The chief acknowledgment is to the participants, the project staff, and the China CDC and its regional offices for assisting with the fieldwork. We thank Judith Mackay in Hong Kong; Yu Wang, Gonghuan Yang, Zhengfu Qiang, Lin Feng, Maigeng Zhou, Wenhua Zhao, and Yan Zhang in China CDC; Lingzhi Kong, Xiucheng Yu, and Kun Li in the Chinese Ministry of Health; and Sarah Clark, Martin Radley, and Mike Hill in the CTSU, Oxford, for assisting with the planning, conduct and organization of the study.

Funding

The CKB baseline survey and the first re-survey were supported by the Kadoorie Charitable Foundation in Hong Kong. The long-term follow-up and subsequent resurveys have been supported by Wellcome grants to Oxford University (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z) and grants from the National Natural Science Foundation of China (82192901, 82192904, 82192900) and from the National Key Research and Development Program of China (2016YFC0900500).The UK Medical Research Council (MC_UU_00017/1, MC_UU_12026/2, MC_U137686851), Cancer Research UK (C16077/A29186, C500/A16896) and the British Heart Foundation (CH/1996001/9454), provide core funding to the Clinical Trial Service Unit and Epidemiological Studies Unit at Oxford University for the project. The proteomic assays were supported by BHF (18/23/33512), Novo Nordisk and Olink. DNA extraction and genotyping were supported by GlaxoSmithKline and the UK Medical Research Council (MC-PC-13049, MC-PC-14135). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR

Oxford Biomedical Research Centre; the views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Author contributions

SS, RGW & ZC designed the study, SS analysed the data and drafted the manuscript. AP and SM further analysed or annotated the data. RGW, IM, RC and ZM supervised the study. KL, CK, NW, DA, HF, YC, HD, DVS, PP, JL, CY, DS, JC, LL, and DB collected/ prepared the data. All authors including RP, RC and MH reviewed the manuscript.

Declaration of interests

The authors declare no competing interests.

Data availability

Full summary statistics data are available at [URL accessible on publication]. Other data presented in this study are included in this publication supplementary information.

License

This research was funded in whole, or in part, by the Wellcome Trust [212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z]. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Methods

Study population and design

The CKB study design and population have been previously reported[53–55]. Briefly, 512,891 Chinese adults aged 30-79 years were recruited from the general population in 5 urban and 5 rural areas in 2004-2008. Questionnaire data, physical measurements and blood samples were collected by trained health workers in local assessment centres. Study participant characteristics are reported in **Table S1**.

The present study design was a case-subcohort within the prospective CKB study. This included 1,951 incident cases of IHD (ICD-10: I21, I20-I25]) accrued during a 12-year follow-up prior to 1st January 2019 and 2,026 sub-cohort participants. The IHD cases comprised incident IHD cases that had GWAS data and no prior history of CVD (i.e. IHD, stroke, transient ischaemic attack or rheumatic heart disease) and no use of statins at enrolment. The sub-cohort was randomly-selected from a population subset of 69,353 participants with GWAS data[55], who were genetically unrelated, and had no reported history of CVD at baseline.

Proteomic assays

Relative levels of 1,463 unique plasma proteins were measured in 3,977 participants using the Olink Explore panel (Olink Inc, Uppsala, Sweden)[21], expressed as NPX units on a $\log_2$ scale. Protein values below the lower limit of detection were included in analyses while those with QC or assay warnings were excluded. Data for three outlier samples were excluded based on PCA of the protein levels (|Z-score|>5 for any of the first 10 PCs), giving a sample size of 3,974 participants. Only autosomal pQTLs were included in the present report and 1,445 out of 1,463 proteins are autosomal encoded.

Histograms of protein levels were visually assessed for multi-modal distribution, which identified 12 proteins with bi-modal distribution that were subsequently excluded from further analyses (**Table S2**). Protein measures were analysed as rank inverse normal transformed residuals following linear regression on sex, age, age^2, and region (10-level categorical variable).

Genomic data

Genotyping of CKB participants has been previously described[55]. This study used new imputation, as follows: genotypes of 531,565 variants passing QC in all 100,706 genotyped samples were converted to genome build 38 using CrossMap v0.6.1[56], and were checked for consistency by reversing the process ("liftUnder"). Variants not mapped, mapped to different chromosomes, or not mapped back to the same locations after liftUnder, were excluded. The remaining 531,542 variants were prephased using SHAPEIT v4.2 (SHAPEIT v2.904 for chromosome X)[57]. They were then imputed using (i) the TOPMed imputation server[58] and (ii) the Westlake Biobank for Chinese imputation server[59]; the two sets of imputed data were then merged, for each variant retaining the imputed genotypes with the higher imputation INFO score. Variants with INFO<0.3 or MAF=0 were excluded, giving a final imputed dataset of 17,933,159 variants with MAF>0.001.

Genome-wide association analyses

Genome-wide association analyses were performed for the rank inverse normal transformed protein levels using BOLT-LMM v.2.3.4[60], with 11 PCs and genotyping array version as covariates (other covariates were adjusted for in the transformation of protein measures described above). For 216 proteins for which BOLT-LMM failed to run due to insufficient estimated heritability, we instead used SNPTEST v.2.5.6[61] with 11 PCs and

genotyping array as covariates. Post GWAS filtering removed SNPs with effective minor allele count <20 (MAF * INFO score * 2 * N <20).

pQTL loci were identified by LD clumping around association signals reaching genome-wide significance (P-value $<5 \times 10^{-8}$), using PLINK v1.9[62] to identify variants in linkage disequilibrium (LD $r^2>0.05$, P-value >0.05) within 5Mbp of the sentinel variant (or within 20Mbp for association signals near the HLA region, chr6:21744977-39074734[55]). Loci identified by clumping were extended by ±10kbp, and overlapping loci were merged. Conditionally independent associations signals within each merged locus were identified using GCTA-COJO[63,] with stepwise model selection and threshold of P-value $<5 \times 10^{-8}$. Association signals were categorised as *cis*-pQTLs if the sentinel variant was within 500Kbp of the structural gene for the assayed protein (**Table S3**). All other associations were categorised as *trans*-pQTLs. We defined more stringent statistical significance using a conservative multiple test-corrected threshold of P-value $<3.45\times10^{-11}$ (genome-wide significance threshold adjusted for number of unique proteins tested: $5\times10^{-8}/1{,}451$).

Since the study design was a nested case-cohort study, we applied the following methods to determine whether case status may have biased the GWAS results. Using genotype data extracted for all our conditionally independent variants associated across the proteome we tested 4 approaches, which demonstrated consistent findings with and without case-ascertainment: (i) Logistic regression to assess if case status was associated with the variant genotype adjusting for covariates (age, age^2, sex, region, array, 11 PCs); (ii) tests using the R anova() function for improved model fit by addition of case ascertainment to a linear regression model for protein measurement associated with variant genotype and covariates age, age^2, sex, region, array version, and 11 PCs; (iii) comparison of effect size (betas) and P-values from the same two linear regression

models (**Figure S10**); and lastly (iv) using our in-house GWAS of MI we examined whether the reported variants from the current study were significantly associated with MI.

Fine-mapping

Fine-mapping of each associated (merged) locus as defined by clumping was performed using Sum of Single Effects mode (SuSiE)[64] (version0.12.16) to obtain CS's of likely causal variants using parameters 'L=10, max_iter= 100,000, min_abs_corr=0.1, refine = TRUE', and using an internal LD reference consisting of 40,000 unrelated CKB participants. Each independent CS was defined/annotated according to the SNP with the largest posterior inclusion probability (PIP).

Identification of newly reported pQTLs and assessing signal overlap

Loci and CSs for CKB *cis*-pQTLs were compared to those reported in the preliminary UKB-PPP preprint[22]. Presence or absence of a *cis*-pQTL for a protein in UKB-PPP used their definition a window +/- 1Mbp around a sentinel variant. Assessment of whether CKB pQTL loci were novel compared UKB-PPP were according to whether or not the full set of CSs at a locus had one or more variants in common with the equivalent set of CSs reported by UKB-PPP for that protein.

Heritability, variance explained and F-statistic

We calculated heritability attributable to each association according to the formula:

(1) $h^2 = 2 \times MAF \times (1 - MAF) \times \text{effect size}^2$

Where MAF is minor allele frequency and effect size (beta) is the SNP effect size in SD units. Heritability for a protein was determined as the sum across all single sentinel SNPs

at each associated locus for that protein. *Cis*- and *trans*-heritability were partitioned according to the corresponding pQTLs.

For SNPs tested in MR, we performed linear regression of each protein measurement with a reduced model with covariates only (age, age^2, sex and region, 11 PCs, array version) and a full model including its corresponding sentinel *cis*-SNP genotype dosage with adjustment plus covariates. We then used the R anova() function to calculate the F-statistic and percentage of variance explained by the variants (estimated as the partial r^2, derived as the proportional change in error sum of squares (SSE), i.e. (reduced - full model) / reduced model).

Functional annotation of pQTLs

Variant annotation of pQTLs was conducted according to Ensemble VEP 107[65]. We selected the most severe annotation for all variants across all available transcripts using BCFtools 'split-vep -s worst'[66]. We collapsed these annotations into 4 groups; non-synonymous (non-synonymous coding transcript variants e.g. stop-gain, missense), synonymous (coding transcript variant), non-coding (non-coding transcript variant e.g. splice region, intron) and intergenic (non-coding, non-transcript variant).

Phenome-wide pQTL to trait associations

GWAS Catalog v1.0.2 (29/07/2023)[67] was used to search for associations with sentinel variants for each *cis*-pQTL. We selected only those studies conducted in EAS populations or which included EAS population in meta-analyses. If there were multiple distinct *cis*-loci for a protein, we tested sentinel SNP for each locus. If a *cis*-pQTL sentinel variant did not have a reported trait in the phenome scan we searched all available proxy SNPs as identified by PLINK LD statistic reports command '--ld-snp-list' , r^2 threshold=0.8, LD-

window 500kb, using the internal CKB LD reference of 40,000 unrelated individuals. We selected results for one proxy SNP per *cis*-pQTL in the following order of priority: (i) if EAS trait was reported; (ii) by highest r^2; (iii) smaller P-value for the association with the outcome. We applied an initial threshold of $5x10^{-6}$ to the downloaded results across all GWAS traits for all *cis*-pQTLs. Grouping results based on chromosome position, repeated protein-SNP-trait entries were identified, and where available we selected associations from EAS-only studies, and only report multi-ancestry analyses that include EAS where EAS-only associations were not found, selecting the most recent study entry. Where the remaining duplicate entries still matched the above criteria we selected the entry with the smaller reported P-value. We applied Bonferroni thresholds to correct (0.05/N SNPs tested* traits searched) for multiple testing.

Mendelian randomisation of cis-pQTLs

We utilised the sentinel *cis*-pQTL variants as instrumental variables for protein levels as exposure with outcome identified through GWAS Catalog using two-sample MR to compute the Wald ratio statistic. The Wald ratio was defined as a change in the outcome due to a 1SD increase in the exposure, calculated as the ratio of the regression coefficient of the gene-outcome association to that for the gene-exposure association[68,69]. Outcome effect sizes reported as odd ratios (ORs) effect were first recalculated as a beta estimate. Exposure pQTLs were set to the protein-increasing allele and harmonised with outcome estimates. We focused on unique protein-trait pairs and excluded associations with outcome traits without reported estimates. Where there was an estimate reported but no effect allele, a manual look up was performed and where the effect allele was not specified, the reported association was excluded. We further excluded associations where the outcome effect allele did not match either effect or alternative allele for the exposure.

The largest published proteome MR-PheWAS in EUR (https://www.epigraphdb.org/pqtl/site)[20] was searched for the identified protein-disease pairs to identify which pairs were replicated and which were newly-reported with P-value<0.05. We also searched the latest EAS proteome-wide MR-PheWAS[27]

Colocalisation

Upon identifying publicly accessible summary statistics for the EAS specific outcomes identified in PheWAS-MR (**Table S15**), we retrieved and converted them to build GRCh38 using UCSC-liftOver[70]. Using the R package coloc[24] we evaluated shared genetic variants within a genomic locus (as defined by LD clumping) to assess support for hypotheses: H0 implying no association with either trait, H1 denoting association with trait 1 but not trait 2, H2 indicating association with trait 2 but not trait 1, H3 suggesting association with both trait 1 and trait 2 as two independent single nucleotide polymorphisms (SNPs), and H4 indicating association with both trait 1 and trait 2, encompassing a shared SNP. This framework enables us to interpret evidence of H4 as indicative of colocalisation between the two traits. We further analysed MSMB-prostate cancer using the R package susieR[71,72] which employs SuSiE framework to fine-map genetic signals for colocalisation optimising analysis when multiple casual SNPs exist.

Drug targets

Proteins with identified disease associations were checked for evidence of druggability using publicly-available drug databases: Therapeutic Target Database[25] (https://db.idrblab.net/ttd/, latest version update: 13/07/23), Drug Bank[26] (https://go.drugbank.com/, latest version 5.1.10: 01/04/23).

Additional PheWAS in CKB population subset

For proteins with no *cis*-pQTL overlap with UKB-PPP CS we examined all CKB phecodes (aggregated ICD10 codes of similar traits[73]) with minimum of 100 cases to identify any further disease associations. The SNP genotype dosages were aligned to the protein-increasing allele and used for logistic regression with adjustments for age, age^2, sex, region and PC1-11 in the models. We applied Bonferroni threshold of 0.05/N phecodes tested to determine significance.

References

1. Anderson, N. L. & Anderson, N. G. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. *Molecular & Cellular Proteomics* **1**, 845–867 (2002).

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

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

4. Ghadermarzi, S., Li, X., Li, M. & Kurgan, L. Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins. *Front Genet* **10**, 1075 (2019).

5. Uhlén, M. *et al.* The human secretome. *Sci Signal* **12**, (2019).

6. The human proteome in druggable - The Human Protein Atlas. https://www.proteinatlas.org/humanproteome/tissue/druggable#potential_drug_targets.

7. Alharbi, R. A. Proteomics approach and techniques in identification of reliable biomarkers for diseases. *Saudi J Biol Sci* **27**, 968–974 (2020).

8. Zhong, W. *et al.* Next generation plasma proteome profiling to monitor health and disease. *Nature Communications 2021 12:1* **12**, 1–12 (2021).

9. Aslam, B., Basit, M., Nisar, M. A., Khurshid, M. & Rasool, M. H. Proteomics: Technologies and Their Applications. *J Chromatogr Sci* **55**, 182–196 (2017).

10. Katz, D. H. *et al.* Proteomic profiling platforms head to head: Leveraging genetics and clinical traits to compare aptamer- And antibody-based methods. *Sci Adv* **8**, 5164 (2022).

11. Corlin, L. *et al.* Proteomic Signatures of Lifestyle Risk Factors for Cardiovascular Disease: A Cross-Sectional Analysis of the Plasma Proteome in the Framingham Heart Study. *J Am Heart Assoc* **10**, 1–16 (2021).

12. Arnett, D. K. & Claas, S. A. Omics of Blood Pressure and Hypertension. *Circ Res* **122**, 1409–1419 (2018).

13. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nature Reviews Genetics 2020 22:1* **22**, 19–37 (2020).

14. Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biology 2020 21:1* **21**, 1–22 (2020).

15. Chakravarti, B., Mallik, B. & Chakravarti, D. N. Proteomics and systems biology: application in drug discovery and development. *Methods Mol Biol* **662**, 3–28 (2010).

16. Nurmohamed, N. S. *et al.* Targeted proteomics improves cardiovascular risk prediction in secondary prevention. *Eur Heart J* **43**, 1569 (2022).

17. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

18. Bretherick, A. D. *et al.* Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. *PLoS Genet* **16**, e1008785 (2020).

19. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications 2017 8:1* **8**, 1–14 (2017).

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

21. Wik, L. *et al.* Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol Cell Proteomics* **20**, 100168 (2021).

22. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv* **20**, 2022.06.17.496443 (2022).

23. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature 2023 622:7982* **622**, 329–338 (2023).

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

25. Zhou, Y. *et al.* TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res* (2023) doi:10.1093/NAR/GKAD751.

26. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074–D1082 (2018).

27. Xu, F. *et al.* Genome-wide genotype-serum proteome mapping provides insights into the cross-ancestry differences in cardiometabolic disease susceptibility. *Nature Communications 2023 14:1* **14**, 1–12 (2023).

28. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).

29. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nat Rev Genet* **19**, 175–185 (2018).

30. Whitaker, H. C. *et al.* The rs10993994 risk allele for prostate cancer results in clinically relevant changes in microseminoprotein-beta expression in tissue and urine. *PLoS One* **5**, (2010).

31. Wang, X. *et al.* Validation of prostate cancer risk variants rs10993994 and rs7098889 by CRISPR/Cas9 mediated genome editing. *Gene* **768**, (2021).

32. Id, E. A. K., 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. (2019) doi:10.1371/journal.pgen.1008489.

33. Rusina, P. V. *et al.* Genetic support for FDA-approved drugs over the past decade. *Nat Rev Drug Discov* (2023) doi:10.1038/D41573-023-00158-X.

34. Espinoza, J. L., Ai, S. & Matsumura, I. New Insights on the Pathogenesis of Takayasu Arteritis: Revisiting the Microbial Theory. *Pathogens* **7**, (2018).

35. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, (2015).

36. Spierings, J. & van Eden, W. Heat shock proteins and their immunomodulatory role in inflammatory arthritis. *Rheumatology* **56**, 198–208 (2017).

37. Mišunová, M. *et al.* Molecular markers of systemic autoimmune disorders: the expression of MHC-located HSP70 genes is significantly associated with autoimmunity development. *Clin Exp Rheumatol* **35**, 33–42 (2016).

38. Daugaard, M., Rohde, M. & Jäättelä, M. The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. *FEBS Lett* **581**, 3702–3710 (2007).

39. Jenei, Z. M. *et al.* Persistently elevated extracellular HSP70 (HSPA1A) level as an independent prognostic marker in post-cardiac-arrest patients. *Cell Stress Chaperones* **18**, 447 (2013).

40. Dulin, E., García-Barreno, P. & Guisasola, M. C. Genetic variations of HSPA1A, the heat shock protein levels, and risk of atherosclerosis. *Cell Stress Chaperones* **17**, 507 (2012).

41. Liu, L. *et al.* Whole Exome Sequencing Revealed Variants That Predict Pulmonary Artery Involvement in Patients with Takayasu Arteritis. *J Inflamm Res* **15**, 4817–4831 (2022).

42. Suzuki, K. *et al.* Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nature Genetics 2019 51:3* **51**, 379–386 (2019).

43. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics 2021 53:10* **53**, 1415–1424 (2021).

44. Zhang, T. *et al.* Common Variants in NUS1 and GP2 Genes Contributed to the Risk of Gestational Diabetes Mellitus. *Front Endocrinol (Lausanne)* **12**, 1 (2021).

45. Aghazadeh, Y. *et al.* GP2-enriched pancreatic progenitors give rise to functional beta cells in vivo and eliminate the risk of teratoma formation. *Stem Cell Reports* **17**, 964–978 (2022).

46. Boulling, A., Le Gac, G., Dujardin, G., Chen, J. M. & Férec, C. The c.1275A>G putative chronic pancreatitis-associated synonymous polymorphism in the glycoprotein 2 (GP2) gene decreases exon 9 inclusion. *Mol Genet Metab* **99**, 319–324 (2010).

47. Lin, Y. *et al.* Genome-wide association meta-analysis identifies GP2 gene risk variants for pancreatic cancer. *Nature Communications 2020 11:1* **11**, 1–12 (2020).

48. Ohno, H. & Hase, K. Glycoprotein 2 (GP2): Grabbing the FimH+ bacteria into M cells for mucosal immunity. *Gut Microbes* **1**, 407 (2010).

49. Cunningham, A. L., Stephens, J. W. & Harris, D. A. Gut microbiota influence in type 2 diabetes mellitus (T2DM). *Gut Pathogens 2021 13:1* **13**, 1–13 (2021).

50. Cox, A. J. *et al.* Increased intestinal permeability as a risk factor for type 2 diabetes. *Diabetes Metab* **43**, 163–166 (2017).

51. Riedel, S., Pheiffer, C., Johnson, R., Louw, J. & Muller, C. J. F. Intestinal Barrier Function and Immune Homeostasis Are Missing Links in Obesity and Type 2 Diabetes Development. *Front Endocrinol (Lausanne)* **12**, 833544 (2022).

52. Solomon, T. *et al.* Identification of Common and Rare Genetic Variation Associated With Plasma Protein Levels Using Whole-Exome Sequencing and Mass Spectrometry. *Circ Genom Precis Med* **11**, e002170 (2018).

53. Chen, Z. *et al.* Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *Int J Epidemiol* **34**, 1243–1249 (2005).

54. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* **40**, 1652–1666 (2011).

55. Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genomics* **3**, 100361 (2023).

56. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

57. Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nature Communications 2019 10:1* **10**, 1–10 (2019).

58. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).

59. Cong, P. K. *et al.* Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nature Communications 2022 13:1* **13**, 1–15 (2022).

60. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics 2015 47:3* **47**, 284–290 (2015).

61. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics 2007 39:7* **39**, 906–913 (2007).

62. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559 (2007).

63. J, Y. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369–375 (2012).

64. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Series B Stat Methodol* **82**, 1273–1300 (2020).

65. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 1–14 (2016).

66. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).

67. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012 (2019).

68. Pagoni, P., Dimou, N. L., Murphy, N. & Stergiakouli, E. Using Mendelian randomisation to assess causality in observational studies. *Evid Based Ment Health* **22**, 67–71 (2019).

69. Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian randomization. *https://doi.org/10.1177/0962280215597579* **26**, 2333–2355 (2015).

70. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* **51**, D1188–D1195 (2023).

71. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet* **17**, (2021).

72. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet* **16**, (2020).

73. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, (2019).

Table 1. Protein drug targets and selected main indications for all disease associated proteins

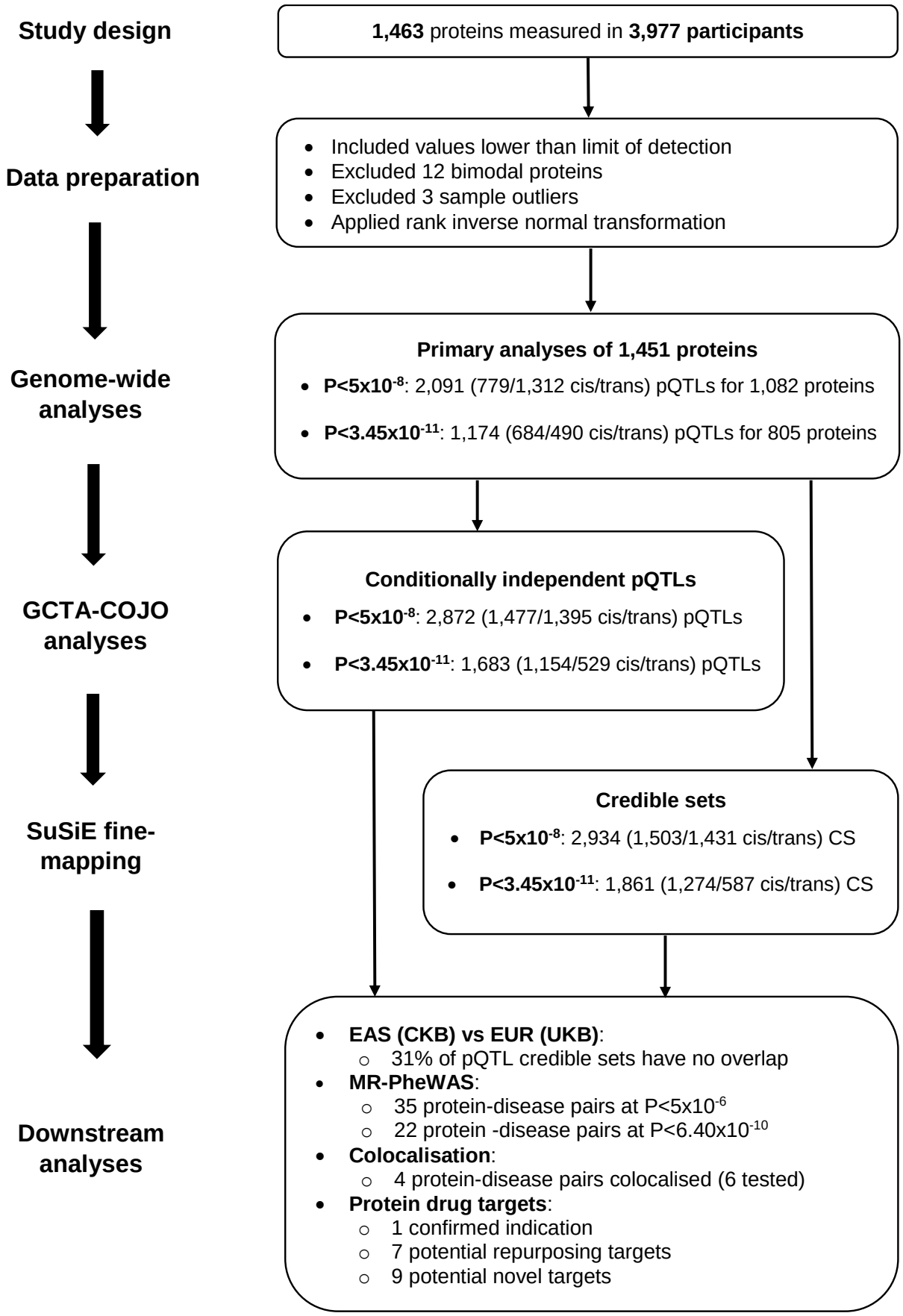
Protein	MR indication	Existing indication	Drug name¶	Trial phase
a) Confirmed drug indication				
MSMB	Prostate cancer	Prostate cancer	Tigapotide	I
b) Potential drug repurposing				
# AGER	Chronic obstructive pulmonary disease	Dementia	PF-4494700	III
		Diabetic nephropathy	Pyridoxamine	III
		Hypotension	Alagebrium chloride	II/III
		Alzheimer's disease	TTP-448	II
CD40	Chronic hepatitis B infection; Kawasaki disease	Lupus	BI 655064	II
		Transplant rejection	ASKP-1240	II
		Melanoma	APX005M	II
		Rheumatoid arthritis	CFZ533	II
		Chron's disease	ABBV-323	I
		Cancer	CDX-1140	I
# ERBB2	Asthma	HER2+ breast cancer	Tucatinib	approved
		Breast cancer	Lapatinib	approved
		Prostate cancer	Masoprocol	approved
		Non-small-cell lung cancer	Dacomitinib	approved
* HSPA1A	Takayasu arteritis	Diabetic neuropathy	AEG-33773	II
		Leukaemia	AG-858	II
		Herpes simplex virus infection	AG-707	I
		diabetic foot ulcer	BRX-005	I
		Cancer	Enkastim-ev	I
		Metastatic malignant neoplasm	H-103	I
IL10RB	COVID-19 hospitalisation	Inflammation	VT-310	literature reported
TNFRSF10A	Central serous retinopathy; age-related macular degeneration	Non-small-cell lung cancer	Mapatumumab	II
TNFSF13	IgA nephropathy	Lymphoma	Atacicept	III
c) Potential novel targets				
ESAM	Schizophrenia	-	-	-
FGF5	Hypertension; myocardial infarction	-	-	-
* GP2	Type 2 diabetes	-	-	-
LRIG1	Atrial fibrillation	-	-	-
NID2	Colon polyp	-	-	-
OBP2B	Gastric cancer	-	-	-
SPON1	Atrial fibrillation, coronary artery disease	-	-	-
SUSD2	Coronary artery disease	-	-	-
UMOD	Urolithiasis; chronic renal failure; kidney stones	-	-	-

¶ Drug search was conducted using TTD | Customized Search (db.idrblab.net), Targets | DrugBank Online

Protein with *cis*-SNP within gene in CKB and not in UKB.

* Protein with no overlapping UKB credible sets.

Figure 1. Flow diagram of study design, analytic approaches and key findings.



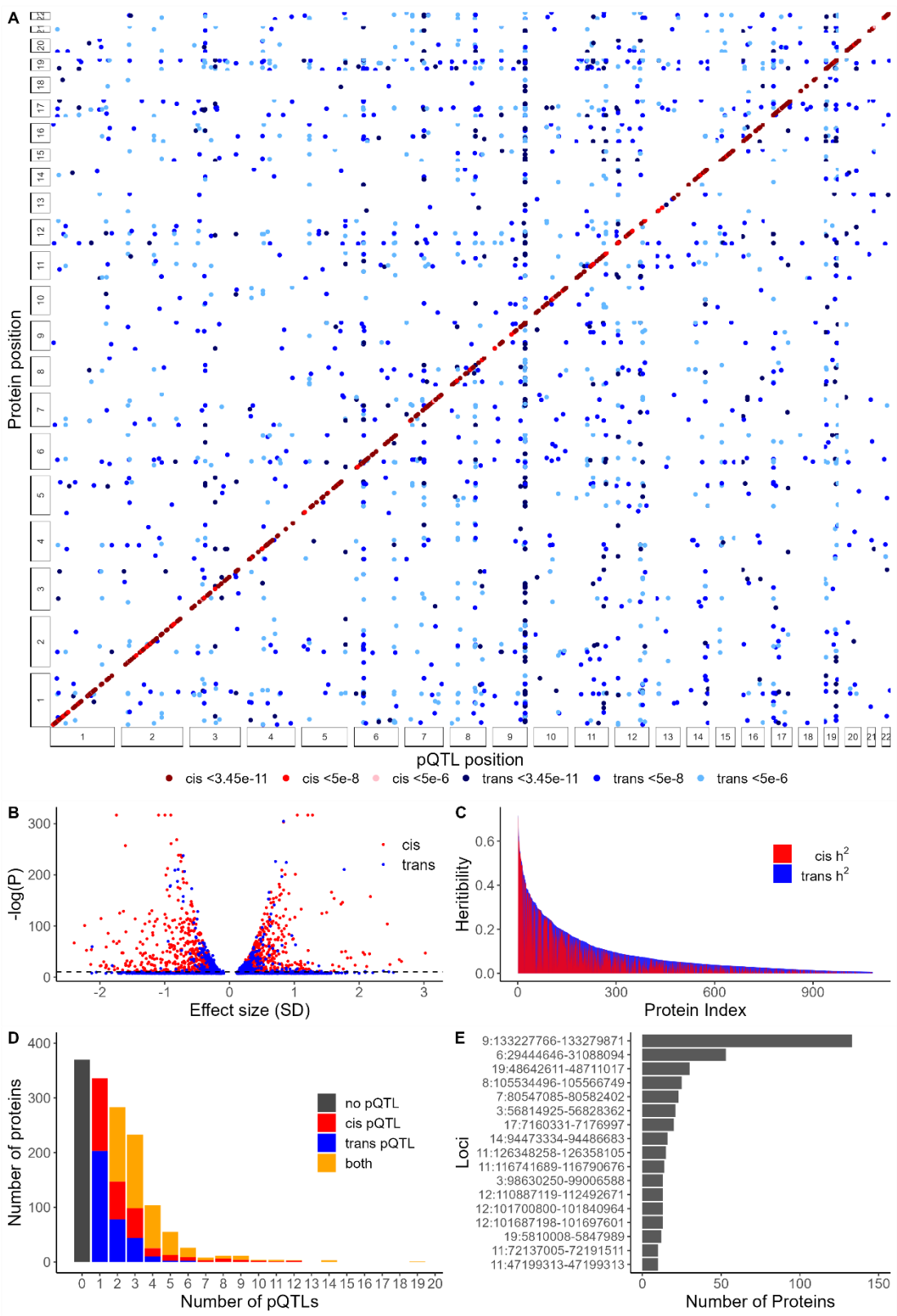


Figure 2 Summary of key GWAS findings on (A) Association of pQTLs across all assayed proteins. Associations for all conditionally independent pQTLs across all proteins are shown, according to the respective chromosomal locations. Colour intensity denotes association at suggestive significance P-value <5x10^{-6}, genome-wide significance P-value <5x10^{-8}, and genome- and proteome-wide Bonferroni adjusted significance P-value

$<3.45 \times 10^{-11}$. (**B) pQTL effect estimates**. Plot shows strength of association for all conditionally-independent pQTLs signals, relative to minor allele effect size. **(C) Heritability of assayed protein level**. *cis* and *trans* heritability is shown for all proteins with at least one pQTL, as determined by the effect size and allele frequency of the sentinel variant at each pQTL, ordered according to total estimated heritability. **(D) Multiple pQTLs**. Frequency distribution of proteins with different numbers of independent pQTL associations, according to whether their pQTLs are *cis* only, *trans* only, or both. **(E) pQTL pleiotropy**. Number of protein associations at a given locus, for all loci associated with at least 10 proteins at P-value $<5 \times 10^{-8}$. In all panels, *cis*- and *trans*-pQTLs are shown in red and blue, respectively.

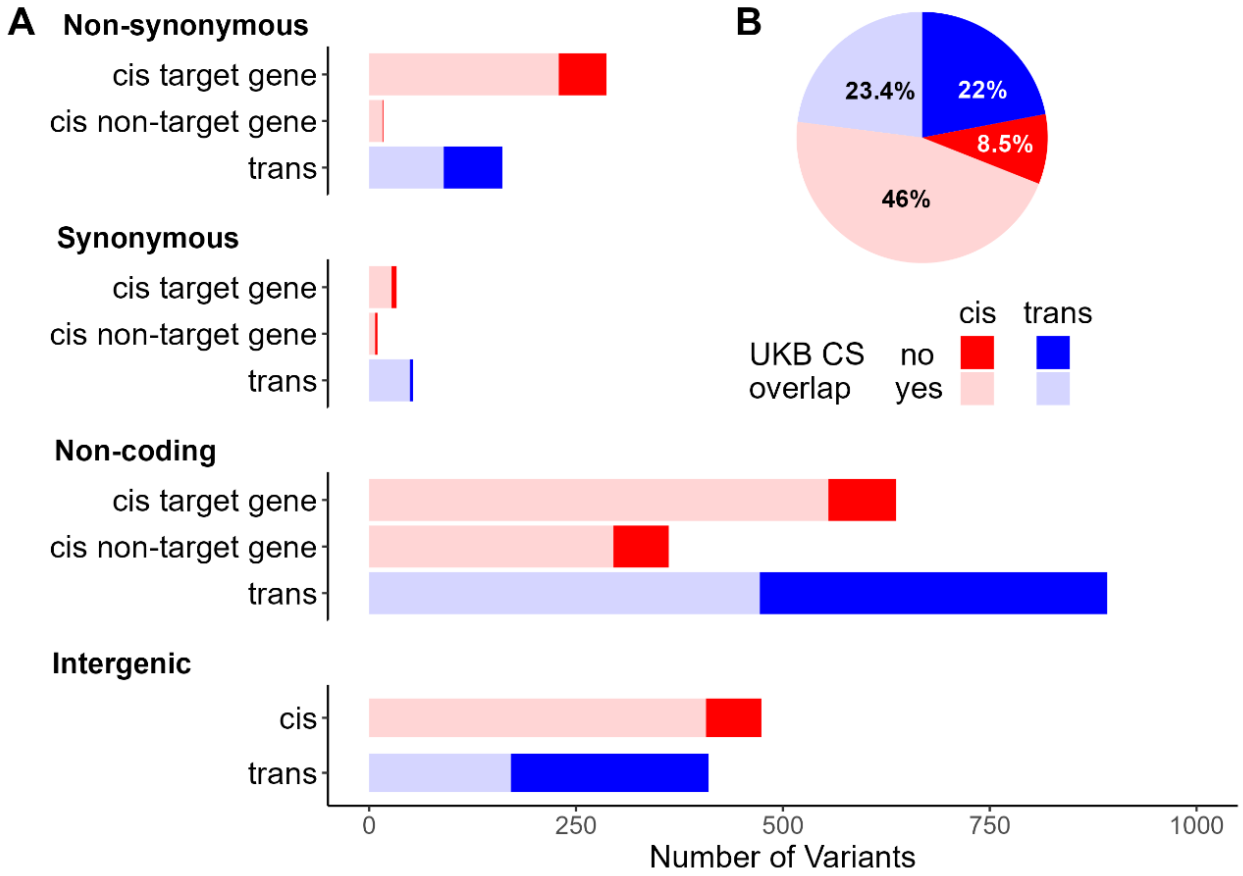


Figure 3. Description of fine-mapped pQTLs on (A) Functional annotation of *cis* and *trans* fine-mapped variants grouped into non-synonymous, synonymous, non-coding and intergenic. The *cis*-pQTLs were categorised as within the target gene transcript (i.e. for assayed protein), non-target gene transcript or intergenic. The colours indicate whether the corresponding credible set (CS) overlapped with any CS for the same protein as reported in UKB-PPP. **(B) EAS-specific pQTLs**. Proportion of *cis* and *trans*-pQTLs with a CS overlap with a corresponding UKB-PPP EUR CS.

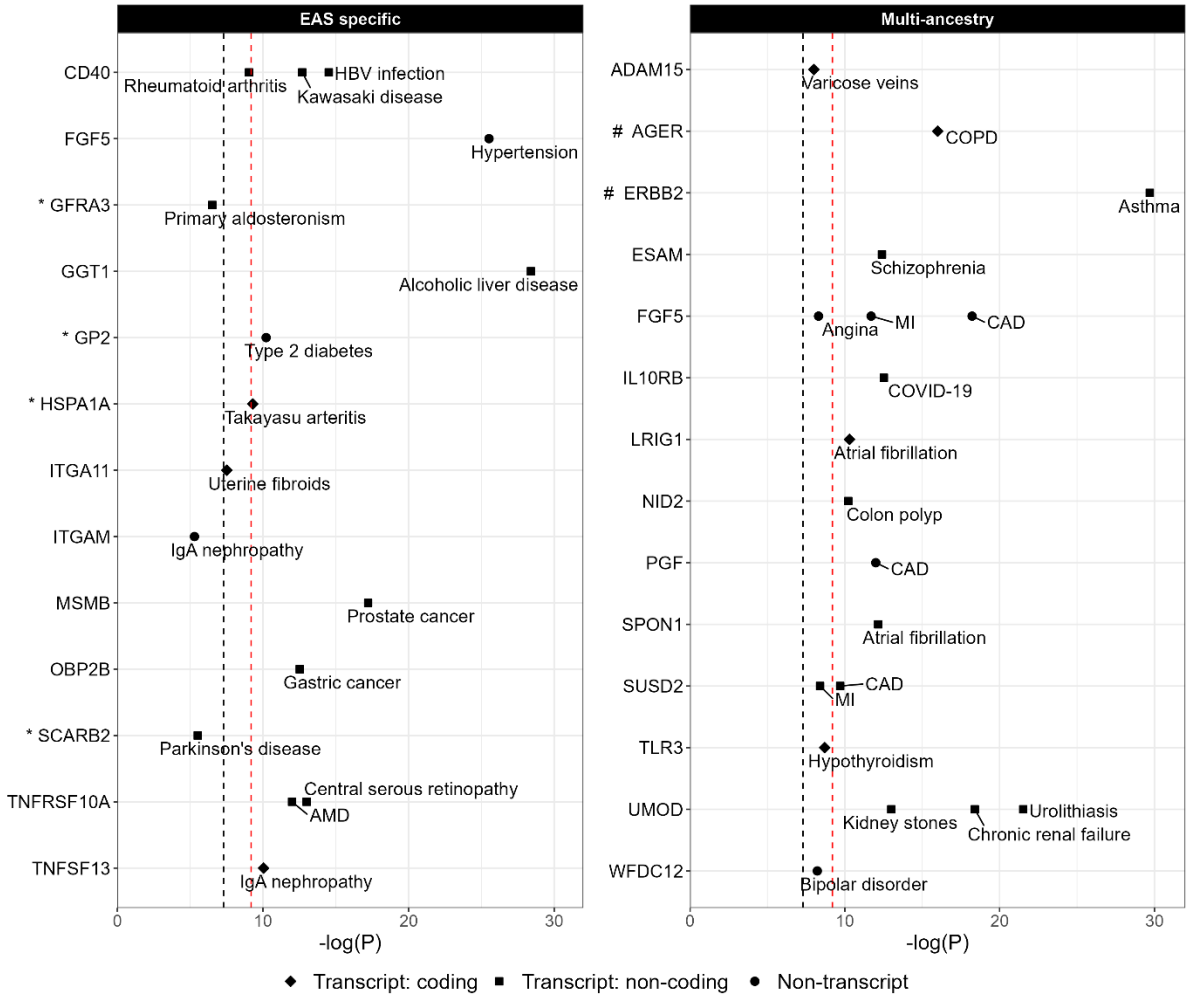


Figure 4. C*is*-pQTL-disease associations in phenome-wide scan of EAS and multi-ancestry studies in GWAS Catalog. Dotted lines denote genome-wide significance (black, P-value =5x10^{-8}) and a Bonferroni-corrected significance threshold (red, P-value =6.40x10^{-10}). Asterisks denote *cis*-pQTL loci for which no credible set had an overlap with corresponding credible sets for the same protein in EUR. Hashtag denotes where *cis*-pQTL is in protein gene and UKB lead *cis*-pQTL not in protein gene. Abbreviations: AMD=age related macular degeneration; HBV=chronic hepatitis B viral infection; CAD=coronary artery disease; COPD=chronic obstructive pulmonary disease; MI=myocardial infarction.

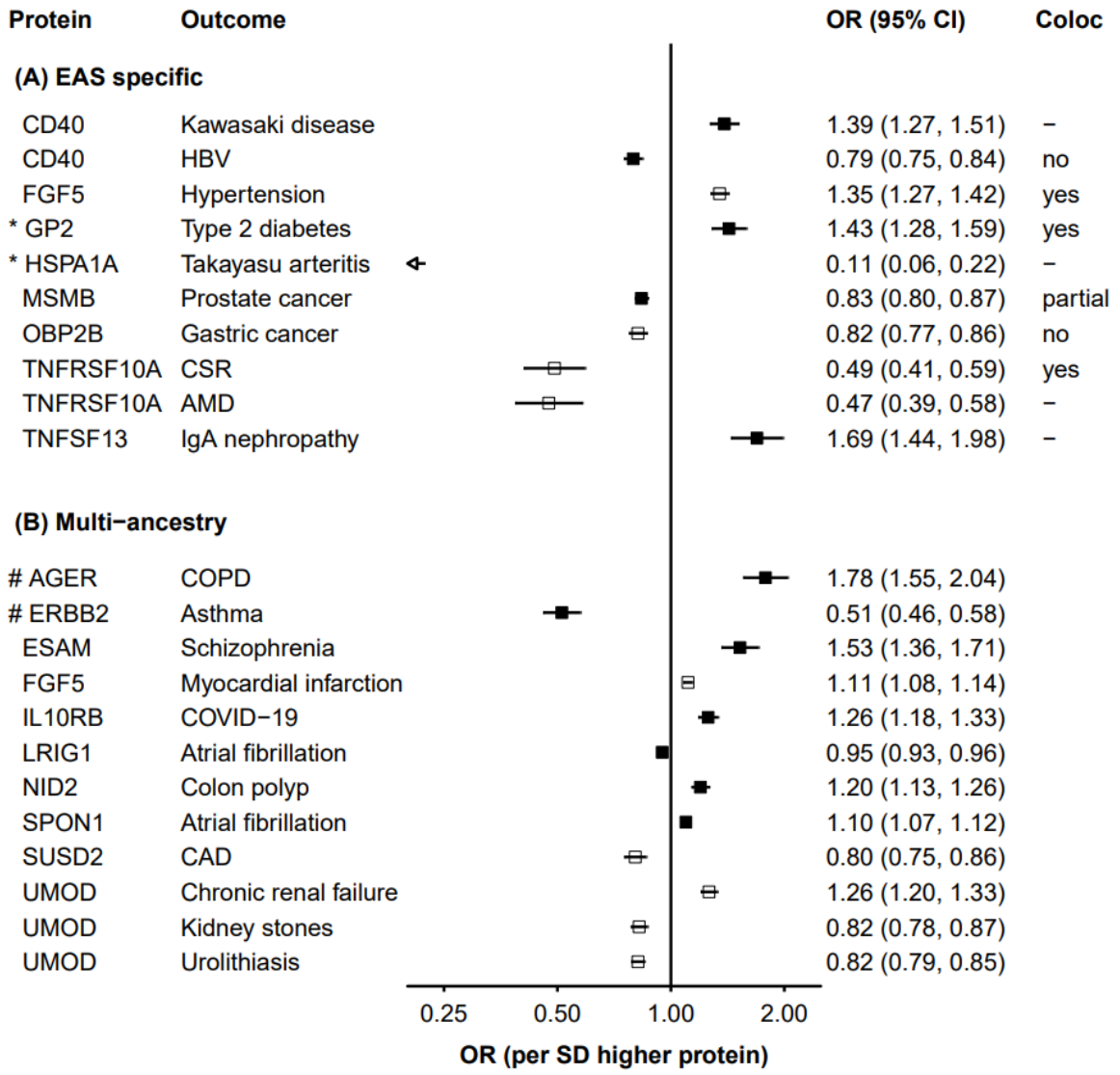


Figure 5. Two sample MR results of significant protein-disease phenotype pairs. Effect estimates per 1 SD increase in protein level are shown for disease associations identified in EAS-specific or multi-ancestry GWAS, for lead variants within (filled shape) or outside (open shape) the protein structural gene locus. Asterisks denotes *cis*-pQTL loci for which no credible set had an overlap with corresponding credible sets in EUR. Hashtag denotes where *cis*-pQTL is in protein gene and UKB lead *cis*-pQTL not in protein gene, hyphen denotes no disease summary statistics obtained. Abbreviations: CSR=central serous retinopathy; AMD=age related macular degeneration; HBV=chronic hepatitis b viral infection; CAD=coronary artery disease; COPD=chronic obstructive pulmonary disease.