Large-scale imputation models for multi-ancestry proteome-wide association analysis

Running title: Multi-ancestry PWAS models

Chong Wu[1]*, Zichen Zhang[1], Xiaochen Yang[2], and Bingxin Zhao[3-6]*

[1]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

[2]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.

[3]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA.

[4]Applied Mathematics and Computational Science Graduate Group, University of Pennsylvania, Philadelphia, PA 19104, USA.

[5]Center for AI and Data Science for Integrated Diagnostics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

[6]Penn Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

**Corresponding authors:*

Chong Wu (cwu18@mdanderson.org) and Bingxin Zhao (bxzhao@upenn.edu)

Abstract

Proteome-wide association studies (PWAS) decode the intricate proteomic landscape of biological mechanisms for complex diseases. Traditional PWAS model training relies heavily on individual-level reference proteomes, thereby restricting its capacity to harness the emerging summary-level protein quantitative trait loci (pQTL) data in the public domain. Here we introduced a novel framework to train PWAS models directly from pQTL summary statistics. By leveraging extensive pQTL data from the UK Biobank, deCODE, and ARIC studies, we applied our approach to train large-scale European PWAS models (total *n* = 88,838 subjects). Furthermore, we developed PWAS models tailored for Asian and African ancestries by integrating multi-ancestry summary and individual-level data resources (total *n* = 914 for Asian and 3,042 for African ancestries). We validated the performance of our PWAS models through a systematic multi-ancestry analysis of over 700 phenotypes across five major genetic data resources. Our results bridge the gap between genomics and proteomics for drug discovery, highlighting novel protein-phenotype links and their transferability across diverse ancestries. The developed PWAS models and data resources are freely available at www.gcbhub.org.

Keywords: Biobank Japan; FinnGen; GBMI; IEU OpenGWAS; PWAS; MVP.

Human proteome provides unique insights into human biology and disease[1-3]. Large-scale data resources and computational approaches have been developed to link genomics and proteomics, identifying protein quantitative trait loci (pQTL) where genetic variations affect protein levels[4-14]. To identify potential therapeutic targets and empower drug discovery, it is of great interest to integrate the genomic profile of proteins with the genetic architecture of complex traits depicted in genome-wide association studies (GWAS)[15,16]. Proteome-wide association studies (PWAS) have emerged as a promising approach to identify proteome-phenotype connections using genetically imputed proteomic models and GWAS summary statistics[17-19]. PWAS has been applied to understand the pathological mechanisms of multiple complex traits and diseases, including stroke[20], Alzheimer's disease[19,21], prostate cancer[22], depression[23,24], post-traumatic stress disorder (PTSD)[25,26], neuropsychiatric and substance use disorders[27,28], and multiple sclerosis[29], aiding in the mapping of genetic causal pathways and identifying druggable targets.

Due to data privacy restrictions and logistical challenges, the genetic community has endorsed the practice of sharing summary-level pQTL data rather than individual-level proteomic and genetic data. This poses a challenge for traditional PWAS models, which require access to individual-level reference proteomes for training. Consequently, existing PWAS models are typically trained on a relatively small reference proteomic dataset, often encompassing just a few hundred participants[17-19]. Furthermore, the constraints of small sample sizes (and consequently reduced statistical power) in PWAS training have restricted the potential of large-scale PWAS applications, particularly when scrutinizing massive phenotypes across curated GWAS databases, given the heavy multiple testing burden. As a result, aside from a few above-noted examples, PWAS has not been widely applied for studying complex traits and diseases.

To bridge the gap between proteomic data sharing and PWAS model training, here we introduced a novel framework to train large-scale PWAS models with only pQTL summary statistics. We applied this pipeline to pQTL summary-level data from the UK Biobank[4] (UKB, n = 46,066 participants), deCODE genetics[7] (n =

35,559), and Atherosclerosis Risk in Communities study[8] (ARIC, *n* = 7,213 for European and 1,871 for African ancestries). Additionally, we used the super learner approach[30] to develop PWAS models tailored for Asian and African ancestries, utilizing publicly available individual-level data from the UKB study (*n* = 914 for Asian and 1,171 for African ancestries). To validate the performance of these PWAS models, we conducted a systematic multi-ancestry PWAS analysis for over 700 phenotypes by using GWAS summary statistics from the Million Veteran Program[31] (MVP), Global Biobank Meta-analysis Initiative[32] (GBMI), FinnGen[33], IEU OpenGWAS[34], and Biobank Japan[35] (BBJ) projects. The developed models and data resources are freely accessible at www.gcbhub.org. **Figure 1A** provides a schematic representation of the model training and data analysis framework.

RESULTS

Development of PWAS models using BLISS

We introduced Biomarker expression Level Imputation using Summary-level Statistics (BLISS), a novel method designed for developing protein imputation models using summary-level pQTL data. Briefly, BLISS used *cis*-single nucleotide polymorphisms (SNPs, within 1 megabase of the transcription start and end sites) as predictors. Using a penalized regression framework, protein imputation models were trained on summary-level pQTL data alongside a linkage disequilibrium (LD) reference panel, such as the 1000 Genomes project[36]. For model parameter tuning, we selected parameters by maximizing the predictive R-squared (R^2) in a summary data-only pseudo-validation setup[37]. Notably, BLISS overcame the need for individual-level pQTL data for parameter tuning by generating independent pseudo-training and validation datasets. This was achieved through sampling marginal association statistics from a subset of individuals, conditional on the full summary-level pQTL data (Methods).

We applied BLISS to develop PWAS models for plasma proteins from two platforms: i) SomaScan version 4, an aptamer-based multiplex protein assay; and ii) Olink Explore 3072, an antibody-based proximity extension assay. Specifically,

for European ancestry, we trained the model using pQTL summary-level data of 35,559 Icelanders from the deCODE[7] project measured via the SomaScan platform, 7,213 European American individuals from the ARIC[8] project also measured via the SomaScan, and 46,066 white British individuals from the UKB pharma proteomics project[4] (UKB-PPP) measured by the Olink platform. For African ancestry, our PWAS models were trained using data from 1,871 African American ARIC individuals measured via the SomaScan. To satisfy the relevance assumption in subsequent association tests, we focused on proteins with a *cis*-heritability estimate greater than 0.01[8,19], including 759 proteins for UKB, 1,201 for deCODE, and 1,209 for ARIC studies. Of these models, the median predictive R^2 values were 0.058, 0.030, 0.066, and 0.080 for UKB, deCODE, ARIC (European), and ARIC (African), respectively (**Table S1**).

To evaluate the reliability of the predictive R^2 derived from BLISS, we assessed the individual-level predictive R^2 for the 759 UKB proteins for which publicly available individual-level data exists. The individual-level predictive R^2 was calculated in an independent UKB testing dataset consisting of 2,961 White non-British individuals. We found that the summary-level predictive R^2 values from the BLISS closely aligned with those computed from the individual-level testing data (**Fig. 1B**, correlation = 0.979), indicating that BLISS-derived R^2 were reliable estimates for predictive performance. We also investigated the impact of LD clumping parameters during model construction. We clumped the *cis*-SNPs to remove the highly correlated SNP pairs, increasing the signal-to-noise ratio. The predictive performance of BLISS remained robust, regardless of whether highly correlated SNPs (with clumping r^2 = 0.9) or relatively independent SNPs (with clumping r^2 = 0.05) were used (**Fig. S1**). Thus, we chose the clumping r^2 = 0.2 in our developed PWAS models to reduce computational time and enhance model stability in subsequent association testing without compromising predictive accuracy.

Furthermore, we used the publicly accessible individual-level data from the UKB study to build non-European PWAS models with the super learner approach[30].

Super learner is a robust statistical ensemble algorithm that has been applied in GWAS applications such as cross-ancestry polygenic risk scores[38,39]. Here we applied super learner to combine standard ancestry-specific PWAS imputation models, which were built through Elastic-net using individual-level data, with BLISS-trained European PWAS models built on large-scale summary-level pQTL data. Details of model training steps were provided in the Methods section. We found that the predictive performance of super learner models was generally higher compared to using standard PWAS models alone in both Asian and African ancestries, especially for proteins that had low prediction accuracy in standard models (**Figs. 1C-D** and **S2**). For the Asian standard models with a predictive R^2 below 5%, the median predictive R^2 for standard and super learner models was 0.62% and 1.48%, respectively (Wilcoxon rank test, $P = 3.13\times10^{-11}$). The super learner method increased the number of models with an R^2 exceeding 1% from 486 to 577. For African models, a similar improvement was observed, albeit more modest, potentially attributed to the closer genetic relationship between Europeans and Asians as compared to between Europeans and Africans. Specifically, among the African standard models with a predictive R^2 under 5%, the median predictive R2 was 1.10% for super learner models, whereas it was 0.72% for standard models (Wilcoxon rank test, $P = 4.13\times10^{-4}$).

BLISS performance in simulation

We conducted simulations to evaluate the performance of BLISS against traditional PWAS models trained on individual-level proteomic data. Specifically, we assessed the predictive R^2 of BLISS in an independent test dataset under various *cis*-protein expression heritability (h_g^2) settings. For example, when $h_g^2 = 0.05$, the average predictive R^2 of BLISS models was 0.0466, 0.0461, and 0.0392 for sample sizes 45,000, 7,000, and 1,000, respectively. These results were comparable with that of standard PWAS models trained on individual-level data, which yielded predictive R^2 of 0.0498, 0.0487, and 0.0402 for the same sample sizes (**Fig. 2A**). Importantly, as the sample size increased, the predictive R^2 improved, particularly in scenarios of low protein expression heritability, highlighting the importance of having large sample sizes to construct powerful

PWAS models, particularly for proteins with low heritability. Consistent with the pattern observed in the UKB real proteomic data, we found that the summary-level predictive R^2 of BLISS was closely aligned with the individual-level predictive R^2 calculated in the testing dataset (correlation coefficient = 0.979, **Fig. 2B**).

Next, we evaluated the statistical power to detect protein-phenotype associations using GWAS summary data for simulated complex traits. We compared the power of BLISS and standard PWAS models trained on varying training sample sizes (1,000, 7,500, and 45,000). Using real UKB genotype data, we simulated a causal SNP-protein-phenotype model with realistic effect sizes and set the phenotypic heritability for each gene between 0.001 and 0.01. We then evaluated the pattern of the empirical statistical power of BLISS models in different simulation scenarios (**Fig. 2C**). Across a wide range of phenotype genetic architectures, we also found that BLISS models consistently had a similar performance to that of standard PWAS models (**Figs. 2D-2E**). Particularly, BLISS with larger training sample sizes (say, 45,000) may outperform standard PWAS trained on smaller sample sizes (say, 1,000). In summary, these results underscore the benefits of the summary data-based model training with BLISS, which avoids the challenges of accessing large-scale, individual-level training proteomic data. Additionally, under the null hypothesis where there was no association between protein expression and phenotype, BLISS models yielded well-controlled empirical type 1 error rates (**Fig. S3**).

We further investigated the impact of sample overlap between the datasets used for PWAS model training and those used for generating GWAS summary statistics. Specifically, we tested scenarios 0%, 50%, and 100% sample overlap between the PWAS model training data and GWAS. Our analyses revealed that the degree of sample overlap had a negligible impact on the results, both in terms of statistical power and type 1 error rate (**Fig. S3**). These findings indicate the robustness of our approach regardless of the sample overlap between PWAS model building and GWAS datasets. We also conducted several additional simulations with different settings and observed similar patterns **(Figs. S4-5)**.

Transferable PWAS results between African and European cohorts

We carried out PWAS on 42 phenotypes for which GWAS summary statistics were available from both European and African ancestries. Most of these phenotypes were obtained from the MVP[31] database. We also considered data on Alzheimer's disease, which had GWAS summary statistics for both European and African ancestries[40,41]. At a significance threshold that controlled the false discovery rate (FDR) at the 5% level across all models and phenotypes, we identified 2,561 significant associations (P range = [4.32×10^{-90}, 6.20×10^{-4}]). The European phenotypes accounted for 2,152 significant associations, while the African phenotypes contributed 409 (**Fig. 3A** and **Table S2**). Among these, 75 unique protein-phenotype association pairs were significant in both ancestries (**Table S3**), and their Z-scores exhibited a high degree of consistency between the two ancestries, with a correlation coefficient of 0.785. Furthermore, the majority (68 out of 75) of these associations displayed consistent directions of effects (**Fig. 3B**). These consistent associations encompassed 57 genes and spanned 14 phenotypes across six categories: lipids, cardiovascular outcomes, anthropometric measurements, metabolic conditions, renal traits, and Alzheimer's disease. A comprehensive overview of the PWAS results is available on our website www.gcbhub.org. Below we highlight several findings from different phenotype categories (**Fig. 3C**).

For Alzheimer's disease, our multi-ancestry PWAS pinpointed associations with *TREM2*, a gene previously highlighted in GWAS[42,43]. *TREM2* played a pivotal role in microglial biology associated with Alzheimer's disease[44]. *TREM2* was upregulated on microglia around amyloid plaques in Alzheimer's disease, and its heterozygous mutations, which regulated innate immunity[45], have been linked to an elevated Alzheimer's risk due to decreased microglial phagocytosis of these plaques[46,47]. For cardiovascular phenotypes such as venous thromboembolism[48] (VTE), varicose veins[49], and the cardiovascular ideal health score (IHS)[50], we identified *ERBB4, HP, ABO, F11, PROC, PROS1*, and *THBD*. Of these, *PROC, PROS1,* and *THBD* were in the protein C anticoagulant pathway, a central system

governing blood coagulation and thrombosis regulation[51]. *ERBB4* encoded a tyrosine kinase receptor involved in the neuregulin-I/ErbB signaling pathway, which was essential for managing cardiac development, repair, and overall cardiovascular activities[52]. *HP* was an established cardiovascular risk factor in individuals with diabetes mellitus, primarily due to its involvement with hemoglobin and oxidative damage[53]. *ABO* determined the ABO blood group structures and had correlations with various cardiovascular diseases[54]. *F11*, encoding factor XI, has been studied for its potential cardiovascular benefits and the decreased risk of VTE associated with factor XI deficiency[55].

For non-alcoholic fatty liver disease (NAFLD), we found *IL1RN*. This gene encoded a cytokine active in the interleukin-1 pathway, a central mechanism in initiating innate immune responses. The interleukin-1 cytokines prominently contributed to liver inflammation, and their inhibition could potentially stop NAFLD progression[56]. In type 2 diabetes (T2D), we highlighted *MSR1* and *TREML2*. *MSR1* was responsible for encoding a cell surface receptor that played a role in adipose tissue inflammation and had connections to the pathogenesis of obesity and insulin resistance-linked T2D[57]. *TREML2* belonged to a gene cluster located on human chromosome 6p21.1, encoding the TREM family of receptors[58]. These TREM proteins modulated the inflammatory response and were potentially linked to obesity-triggered chronic inflammation and insulin resistance[59]. Regarding the estimated glomerular filtration rate (eGFR), we identified *UMOD*. *UMOD*, known for its largest genetic effect in GWAS of eGFR and chronic kidney disease, specifically encoded uromodulin, the most abundant protein expressed in the kidney[60]. Additionally, our analysis discovered 22 genes linked with lipids across both ancestries and 25 genes associated with anthropometric traits including height and body mass index. These findings uncover the protein-phenotype associations shared between European and African ancestries.

Consistent PWAS findings across reference proteomic datasets

We further analyzed the consistency of associations pinpointed by PWAS models trained on distinct reference proteomic datasets. Among the 393 proteins that

overlapped in the UKB, deCODE, and ARIC European ancestry models, 711 significant associations were identified ($P < 6.20\times10^{-4}$). Of these, 64.1% were significant in at least two of the PWAS cohorts, and 37.1% were significant across all three, yielding 88 unique protein-phenotype associations (**Table S4**). The *Z*-scores of these 88 associations demonstrated high consistency across the three cohorts, with correlations ranging from 0.932 to 0.977. Impressively, over 97.7% of these associations presented consistent effect directions (**Figs. 4A-C**). When evaluating overlapping genes between any two of the three PWAS cohorts, the percentage of pairs significant in both cohorts ranged from 46.6% to 54.9% (**Figs. S6A-C**). For example, deCODE and ARIC models shared 1,197 genes, accounting for 1,610 of the 2,152 European PWAS associations. Of these, 53.2% were significant in both cohorts, resulting in 428 unique shared protein-phenotype associations. The *Z*-scores of these 428 associations had a strong alignment, with a correlation coefficient of 0.882 (**Fig. S6A**). Below we highlight some phenotype-gene associations that were significant across all three PWAS cohorts (**Fig. 4D**)

Beyond the consistently identified *TREM2* in both European and African PWAS, we pinpointed an additional set of six genes consistently associated with Alzheimer's disease across all three European PWAS models, namely, *LILRB1, CD55, BCAM, SIRPA, EPHB4,* and *GRN*. Notably, many of these genes were tied to the immune system. *CD55*, expressed in both neurons and glial cells[61], played a role in Alzheimer's disease progression by influencing both the complement[62] and inflammatory systems[63]. *GRN* encoded the protein progranulin, a microglial protein linked to Alzheimer's disease, acting as a mediator of neuroinflammation and exacerbating tau-related brain pathology[64]. *BCAM* and *SIRPA* both encoded glycoproteins of the immunoglobulin superfamily. *BCAM* was located within the *APOE* 19q13.3 region, a well-known locus associated with Alzheimer's disease risk[65]. Moreover, recent research has implicated the immune pathway gene *LILRB1* in the pathogenesis of Alzheimer's disease[66]. In addition, our results pointed to significant associations of *CTSF, CTSV,* and *CD14* with PTSD and related symptom severity measures[67]. *CTSF* and *CTSV* encoded cathepsins, whose functional disruptions have consistently been linked to psychiatric disorder

pathophysiology. These cathepsins played roles in memory and influence depression and anxiety behaviors. Importantly, they regulated neuronal homeostasis and the plasticity of neurons in response to acute stress or trauma[68]. Moreover, studies have highlighted *CD14* as a central mediator in the cognitive behavioral reductions induced by social stressors[69].

We identified 12 genes associated with seven cardiovascular phenotypes. Specifically, for coronary artery disease[70] and peripheral artery disease[71], the implicated genes included *IL6R, PCSK9, C2,* and *MMP12*. *PCSK9* had a key role in regulating low-density lipoprotein cholesterol and emerged as a therapeutic target for various cardiovascular diseases (CVDs)[72,73]. Both *IL6R* and *MMP12* were also associated with multiple CVDs. *IL6R*, a consistently identified genetic risk variant for CVDs such as coronary artery disease[74] and atrial fibrillation[75], controlled IL-6 signaling, presenting a target to lower the risk of coronary heart disease[76]. *MMP12* encoded one member of the matrix metalloproteinases (MMP) family and has attracted great attention in CVD research due to the MMPs' pivotal roles in biological processes including atherosclerosis and myocardial infarction remodeling[77]. Furthermore, *IL6R* and *PCSK9* showed associations with other cardiovascular phenotypes, such as VTE, varicose veins, angiographic burden[70], and IHS. These phenotypes also expanded the list with eight additional cardiovascular-associated genes: *GP6, NPPB, OBP2B, SPARCL1, ENTPD6, FABP2, RSPO3,* and *TNFSF12,* many of which were already recognized in CVD research. For example, *FABP2*, part of the fatty acid-binding protein family, facilitated multiple lipid-centric biological and metabolic pathways that intersected with prominent chronic diseases, including CVDs[78]. *GP6* had a central role in both arterial and venous thrombus formation and has been pinpointed as a genetic risk factor for venous thrombosis[79]. *NPPB* encoded the B-type natriuretic peptide, a molecule that was widely used as both a risk biomarker and therapeutic agent in CDVs[80]. *RSPO3*, previously linked to varicose veins in GWAS[35], encoded an enhancer of WNT signaling and was highly expressed in vasculature with an important role in its remodeling[81].

We pinpointed six genes linked with T2D: *MLN, NCAN, NELL1, PAM, AGER,* and *BST1*. *AGER* encoded a receptor for advanced glycation end-products, the engagement of which can promote a proinflammatory state, impacting various biological pathways and diseases, including diabetes[82]. *NELL1* and *PAM* have been associated with T2D from GWAS[83,84]. Decreased PAM function negatively affected insulin in response to glucose through molecular mechanisms related to pancreatic β-cells[85]. For NAFLD, our findings highlighted *APOH, FCRLB, IL1RN,* and *RSPO3*. *RSPO3* was involved in the Rspos-LGR4 signaling, which modulated hepatic cholesterol synthesis via the AMPKα-SREBP2 pathway, suggesting its potential function in managing NAFLD[86]. For eGFR, we identified associations with *ANGPTL3, DPEP1, IDI2, NCAN,* and *TIMP4*. In mouse studies, *ANGPTL3* has shown its influence on kidney architecture and functionality, implying its potential benefits in slowing chronic kidney disease progression[87]. *DPEP1* was in a recognized genetic risk locus for eGFR and kidney function[88]. This gene has been implicated in kidney disease development and ferroptosis by mediating cellular iron trafficking[89]. Additionally, we found 14 genes linked to lipids and 17 genes associated with anthropometric measures.

Comparing the results of UKB and ARIC African PWAS models, we found 13 consistently significant associations, exhibiting a high correlation of *Z*-scores at 0.97 (**Fig. S6D** and **Table S4**). Interestingly, three of these associations were not identified in European PWAS models. For example, both UKB and ARIC African PWAS models identified an association between *SPON1* and NAFLD. *SPON1* was involved in cell adhesion and signaling, potentially holding an important role in hepatic fibrogenesis and steatosis in NAFLD[90]. In summary, the PWAS-derived novel protein-phenotype associations both reinforce existing literature through a proteomic lens and unveil new potential therapeutic targets for various diseases. Our results also provide compelling evidence for the robustness and consistency of associations identified by different PWAS models trained on independent reference proteomic datasets.

Multi-ancestry PWAS analysis for GBMI database

We performed an expanded multi-ancestry PWAS analysis targeting seven diseases within the GBMI[32] database, which had GWAS summary statistics across European, African, and Asian ancestries. These diseases included asthma, chronic obstructive pulmonary disease (COPD), gout, heart failure, idiopathic pulmonary fibrosis, stroke, and VTE. For stroke, we additionally incorporated data from its subtypes, obtained from a recent multi-ancestry GWAS[20]. At the 5% FDR threshold, we identified 286 significant associations (P range = [2.18×10^{-58}, 1.89×10^{-4}]). Specifically, for European, African, and Asian phenotypes, we observed 247, 14, and 25 associations, respectively (**Fig. S7A** and **Table S5**). These associations encompassed 135 genes linked to seven diseases in European ancestry, 12 genes for six diseases in African ancestry, and 10 genes for five diseases in Asian ancestry. Asthma and stroke constituted the largest proportion of findings across each ancestry group. For example, we found consistent associations between *MMP12* and stroke (including its subtypes, $P < 2.49\times10^{-5}$) across the three European PWAS models. *MMP12* has consistently been linked to both ischemic stroke and large-artery stroke[91], likely due to its role in elastin degradation and macrophage invasion in plaque inflammation[92].

We identified eight disease-gene associations that were significant in more than one ancestry (**Fig. S7B** and **Table S6**). For asthma, three genes emerged in both European and Asian ancestries, including *IL1RL1, AGER,* and *C2*. *IL1RL1* encoded the receptor for the interleukin-33 (IL-33) cytokine, and the IL-33/IL1RL1 pathway played a critical role in asthma pathogenesis[93]. *AGER* encoded RAGE, which had the highest expression in the lung, was linked to several pulmonary diseases, including asthma. The related RAGE-signaling pathway has been shown to be hyperactive across various asthma-associated pathologic mechanisms[94]. *C2* encoded a component of the complement system and was involved in both inflammation and airway tissue remodeling processes in asthma[95]. For COPD, we identified *PILRB*, a gene previously discovered in GWAS[96]. Additionally, we identified *EFNA1* for gout, which was prioritized to be a prospective drug target in a recent study[97].

For both European and African ancestries, we found *IL2RB* for asthma, *SCARA5* for stroke, and *ABO* for VTE. The association between VTE and *ABO* has been observed in the above MVP data analysis. *IL2RB* was a susceptibility gene to asthma identified in GWAS and was differentially expressed during lung development[98]. The protein abundance of *SCARA5* was recently linked to stroke and its subtypes in Mendelian randomization (MR) studies[99,100]. This association might be related to *SCARA5*'s role in encoding a scavenger receptor for the delivery of ferritin-bound iron, as both decreased and heightened iron levels can elevate the risk of stroke[101].

Applying to FinnGen and IEU OpenGWAS databases

We further assessed the performance of European PWAS models using GWAS summary statistics from 278 clinical outcomes in the FinnGen[33] study, with an average of 31,800 cases per disease. At the 5% FDR significance level across all models and traits ($P < 3.02\times10^{-4}$), we identified 5,119 significant associations across 1,360 models (340 from UKB, 532 from deCODE, and 508 from ARIC) and 273 diseases. These associations covered over 12 major disease domains, as detailed in **Figure 5A** and **Table S7**. Diseases related to the circulatory system (I9), such as hypertension, constituted the largest part of significant disease-gene associations. The proportion of significant associations remained consistent across different PWAS models (UKB, deCODE, and ARIC) (**Fig. 5B**). Of the 1,360 models that showed significant associations with at least one disease, there were no noticeable trends linking their prediction accuracy to the number of significant associations (**Fig. S8A**). This indicates that the patterns of associations we identified are not driven by model prediction accuracy.

Out of the 5,119 PWAS signals, 2,017 were contributed by 393 genes that were present in all three models (UKB, deCODE, and ARIC). Of these associations, 63.1% were significant in at least two of the cohorts, and 30.5% achieved significance across all three. These associations that were significant across all three cohorts amounted to 205 unique protein-disease pairs. The *Z*-scores of these 205 pairs demonstrated notable consistency across all three cohorts, with

correlation coefficients ranging from 0.703 to 0.913. Moreover, most (> 90.7%) of these pairs exhibited consistent directional trends (**Figs. 5C** and **S8B-C**). Considering the genes that overlap between any two of the three cohorts, the percentage of pairs significant in both cohorts ranged from 41.6% to 51.6%. For example, deCODE and ARIC models shared 1,195 genes, which accounted for 3,820 of the 5,119 PWAS associations. Among these, 46.1% of the pairs were significant in both cohorts, leading to 881 unique overlapping protein-disease associations. The *Z*-scores for these 881 associations were strongly consistent, with a correlation of 0.839 (**Fig. 5D**).

Here we showcase some findings from these 205 protein-disease associations (encompassing 73 genes and 115 diseases) that were significant across all three cohorts (**Table S8**). Consistent with our MVP data analysis, *PCSK9, IL6R,* and *FABP2* were associated with multiple cardiovascular outcomes. Particularly, *PCSK9*, a pivotal target in CVD treatment, displayed the highest number of significant associations (16 in total) with circulatory system diseases, including heart failure and coronary heart disease. We also found links between *MPO* with hypertension and statin medication use. Recognized as a pivotal inflammatory enzyme associated with conditions like hypertension and heart failure, MPO and its downstream inflammatory pathways have been considered as potential therapeutic targets of CVDs for two decades[102,103]. Our analysis also unveiled consistent correlations between *RNASET2* and *DPEP1* with basal cell carcinoma and malignant skin neoplasm. *RNASET2* encoded ribonuclease T2, a known tumor suppressor[104] involved in cellular stress responses[105]. Previous GWAS and expression QTL studies have linked this gene with basal cell carcinoma[106]. Similarly, *DPEP1*, which managed leukotriene activity, has shown involvement in various cancers[107].

We next examined 309 phenotypes of European ancestry from the MRC IEU OpenGWAS database[34] (average n = 202,612). At the FDR 5% level ($P < 3.56\times10^{-4}$), we identified 6,848 significant associations. These were distributed across 1,963 PWAS models: 459 from UKB, 770 from deCODE, and 734 from ARIC,

covering 290 distinct phenotypes (**Fig. 6A** and **Table S9**). Among the 393 overlapping genes common to UKB, deCODE, and ARIC models, we identified 2,441 protein-phenotype association pairs that were significant across all three cohorts. Similar to our FinnGen analysis, 66.7% of these pairs were significant in at least two cohorts, while 37.1% were significant across all three, resulting in 302 unique protein-phenotype pairs (**Table S10**). The *Z*-scores of these pairs were highly consistent across the three cohorts, with correlation coefficients ranging from 0.928 to 0.974. Furthermore, over 97.4% of these associations displayed consistent directional trends (**Figs. 6B-D**). When examining the genes that overlap between any two of the three cohorts, the proportion of pairs that were significant in both cohorts ranged from 48.2% to 54.6% with a high consistency in *Z*-scores (**Figs. S9A-C**).

We highlighted eight genes associated with six cancer traits across all three PWAS model cohorts, including *CTSF, RSPO3, CRTAM, LAYN, CHRDL2, DPEP1, NFASC,* and *NAAA*. Many of these genes have been associated with cancer pathophysiology. Importantly, four genes (*DPEP1, RSPO3, CHRDL2,* and *CTSF*) were also significantly associated with cancer or neoplasm traits in our FinnGen analysis across all three PWAS model cohorts (*P* range= [4.95×10^{-32}, 1.99×10^{-4}]). *CTSF* was a member of the cysteine cathepsin protease family, which was among the most intensively studied protease families in cancer research[108]. Numerous studies have reported abnormal expression of cathepsin family members in a variety of cancers[109], and their regulatory roles in cancer have been detailed through extensive mouse model studies[110]. RSPO3, recognized as an oncogenic protein, was implicated in cancer progression in multiple body systems and influenced cancer clinical features[111]. The causal oncogenic capacity of RSPO3 has been demonstrated in mouse studies, underscoring its potential role in the Wnt/β-catenin pathway and its viability as a diagnostic or therapeutic target[112]. *CHRDL2* was overexpressed in colorectal cancer and was believed to promote cell proliferation and tumorigenesis[113]. *CRTAM* played important roles in enhancing tumor-associated immune cell infiltration[114]. LAYN, a cell surface receptor for hyaluronan, was instrumental in regulating cell spreading and migration, playing a

role in the suppressive activity of cancer cells[115]. *NAAA* encoded a lysosomal enzyme, playing an important role in inflammatory response[116]. *NAAA* has dysregulated expression patterns, which were correlated with tumor mutational burden and microsatellite instability in multiple cancers[117]. In summary, the application of our PWAS models to various complex traits and diseases suggests their consistent and reliable performance. Our findings offer novel insights into the associations between proteins and phenotypes, illuminating valuable biological connections and potential therapeutic targets for a range of complex traits and diseases.

Conducting Asian PWAS using the Biobank Japan data

We examined the performance of our Asian PWAS models using GWAS summary statistics for 107 phenotypes from the BBJ[35] study (*n* cases > 2,000 for binary traits). We applied standard models trained on UKB Asian reference proteomic data (referred to as ASN) as well as super learner models (referred to as ASN+EUR). At an FDR of 5% across all models and traits ($P < 8.55\times10^{-5}$), we found 159 significant associations (**Fig. S10A** and **Table S11**) that encompassed a wide range of disease categories, such as genitourinary system, cancer, circulatory system, and nervous system. The ASN+EUR and ASN models had 89 and 70 significant associations, respectively. There were 57 protein-phenotype pairs that were significant in both sets of models.

We present the PWAS findings for cancer traits from the BBJ study and highlight their intersections with those from the FinnGen and IEU analyses of European datasets. In total, we identified 12 genes linked to three cancer traits in the BBJ. Of these, six were found significant in our European IEU analysis, and three of these six were also significant in FinnGen (**Fig. S10B** and **Table S12**). Several of these genes have established roles in cancer biology. For example, *AGER* encoded RAGE, a cell surface receptor belonging to the immunoglobulin superfamily. Both RAGE and its ligands were associated with immune-inflammatory responses and metastasis of multiple cancers[118]. *BTN3A2*, a mediator of immune activation, played an important role in immune infiltration and

various cancer progressions[119]. *EFNA1* produces Ephrin-A1, the functional ligand for the EphA2 receptor, both presented in numerous cancers, with EFNA1 having critical roles in modulating cancer angiogenesis and invasive properties[120]. *FCRL1* was a member of the Fc receptor-like family, playing pivotal roles in immune function and tumor progression[121]. It regulated B-cell responses and was implicated in B-cell malignancies[122]. These results suggest the shared proteomic biomarkers and therapeutic cancer targets across Asian and European populations.

Concordance with MR analysis and drug-target genes

We compared our PWAS findings with results from MR[15] and Bayesian colocalization analyses[123]. We focused on European PWAS models due to the large sample size of their pQTL summary data. Overall, the pQTL MR results suggest that a substantial majority (ranging between 73.9% to 79.9%) of our PWAS findings discussed in the above sections were underpinned by causal evidence. To illustrate, out of 88 consistently identified PWAS associations across the three different proteomic datasets (making a total of 264 associations) in the MVP analysis, 84 demonstrated significant causal implications in at least one proteomic dataset (P range = [5.16×10^{-192}, 3.96×10^{-2}], when controlling the FDR at the 5% level). Furthermore, 51 of these 88 associations were MR-significant across all the cohorts, and 25 were significant in two of the three datasets. In addition, 33.3% of these associations exhibited strong evidence for colocalization, characterized by a posterior probability favoring a shared causal variant hypothesis (PPH4) greater than 0.8[123,124] (**Fig. 4D** and **Table S4**). The proportion of associations with colocalization PPH4 > 0.8 ranged from 21.0% to 33.3% across different GWAS databases in our PWAS application. We have integrated these pQTL MR and colocalization results into the tables and figures in the above sections that detailed our PWAS results.

We used GREP[125] to evaluate the drug-target enrichment of PWAS genes identified in our analyses of the MVP, GBMI, FinnGen, IEU OpenGWAS, and BBJ databases. We considered seven primary disease categories, each containing

over 200 PWAS-significant genes, including cancer (and neoplasms, CD2 & ICD-O-3), endocrine, nutritional, and metabolic (E4), nervous system (G6), circulatory system (I9), respiratory system (J10), digestive system (K11), as well as musculoskeletal and connective (M13). For each disease category, we tested its enrichment in drug-target genes linked to its corresponding International Classification of Diseases (ICD)-10 groups. At a nominal significance level ($P < 0.05$), 22 out of 82 tests displayed significant enrichment. Of these, 16 passed the FDR 5% significance level (**Table S13**). The majority of the enrichment was found between cancer-associated PWAS-significant genes and drug targets for cancers or benign neoplasms. As an example, of the 263 unique PWAS genes linked to cancer phenotypes, 24 were known drug targets for malignant tumors of male genital organs (C60-C63, odds ratio = 2.32, Fisher's exact test $P = 1.76\times10^{-3}$). Beyond cancer, the 479 PWAS-significant genes for circulatory system diseases exhibited enrichment for drug-target genes associated with various cardiovascular outcomes. The most substantial enrichment was observed for conditions of veins, lymphatic vessels, and lymph nodes (I80-89, 14 drug-target genes, odds ratio = 2.27, Fisher's exact test $P = 6.90\times10^{-3}$). These results suggest the promising potential of PWAS-driven drug discovery.

Discussion

In this study, we developed a statistical framework for training PWAS models using summary-level pQTL data and super learner. We applied these approaches across extensive proteomic data resources from the UKB, deCODE, and ARIC studies, leading to a set of multi-ancestry PWAS models spanning European, Asian, and African ancestries. To evaluate the reliability of our models, we applied them to massive GWAS summary statistics sourced from the MVP, GBMI, FinnGen, IEU OpenGWAS, and BBJ projects. All developed PWAS models and the full set of our analysis results have been made freely available at www.gcbhub.org.

In extensive real data applications of our PWAS models, we uncovered biological insights into the proteomic architecture of various complex traits and diseases. Most of these protein-phenotype associations were novel within the context of

PWAS analysis. This novelty largely resulted from the limited training sample sizes of existing PWAS models and/or the absence of large-scale applications to integrate biobank-scale GWAS databases. However, our PWAS findings were not isolated in the broader scientific context. Many of our results aligned with findings from previous non-PWAS studies, reinforcing the established understanding of disease pathology. In previous sections, we have showcased our results related to CVDs, cancer traits, and Alzheimer's disease. For example, we pinpointed a consistent set of genes linked to cancers across diverse GWS databases, including FinnGen, IEU OpenGWAS, and BBJ. Many of these genes have well-defined roles in the pathophysiology of cancer and have been linked to specific clinical features. We further observed enrichments with known cancer drug-target genes. For Alzheimer's disease, we discovered substantial associations with genes related to the immune system, emphasizing the crucial involvement of the immune system and its cellular components in Alzheimer's neuropathology. In CVDs and lipids, our findings aligned with multiple signals identified in earlier GWAS, prioritizing pivotal genes for potential therapeutic interventions. We anticipate that our PWAS models will continue to shed light on many more discoveries in future applications with powerful multi-ancestry genetic data resources.

When training our PWAS models, we used the 1000 Genomes reference panel[36] data to account for the LD patterns. However, recent studies indicate improved performance when using internal LD data for other genetic analyses, such as fine mapping[126] and genetic risk prediction[127]. Moreover, certain proteomic cohorts, such as deCODE, may have fine-grained population structures not fully represented by the 1000 Genomes or other genomic reference panels. Therefore, we suggest the sharing of the cohort-specific LD matrices along with the pQTL summary statistics in future genetic studies of proteomics. Given the expected growth in the availability of proteomic data resources, especially pQTL summary-level data, we plan to continuously enhance our trained PWAS models (say, every six months), ensuring our online resources remain most updated. For example,

once the second batch of UKB proteomics data is released, we will train and update our PWAS models accordingly.

METHODS

Methods are available in the **Methods** section.

Note: One supplementary pdf file and one supplementary table zip file are available.

ACKNOWLEDGEMENTS

We thank the individuals represented in the genetic and proteomic studies for their participation and the research teams for their work in collecting, processing, and disseminating these datasets for analysis. We gratefully acknowledge all the studies and databases that made pQTL and GWAS summary data available. This research has been conducted using the UK Biobank resource (application number 76139), subject to a data transfer agreement. We would like to thank the research computing groups at the University of Texas MD Anderson Cancer Center, Purdue University, and the Wharton School of the University of Pennsylvania for providing computational resources and support that have contributed to these research results.

AUTHOR CONTRIBUTIONS

B.Z. and C.W. designed the study. C.W., Z.Z., X.Y., and B.Z. developed the models, analyzed the data, and interpreted the results. Z.Z. and C.W. performed the simulation study. Z.Z. developed the online data resources. B.Z. and C.W. wrote the manuscript with feedback from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular systems biology* **15**, e8503 (2019).

2. Jiang, L. *et al.* A quantitative proteome map of the human body. *Cell* **183**, 269-283. e19 (2020).
3. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
4. Sun, B.B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *BioRxiv*, 2022.06. 17.496443 (2022).
5. Dhindsa, R.S. *et al.* Influences of rare protein-coding genetic variants on the human plasma proteome in 50,829 UK Biobank participants. *bioRxiv*, 2022.10. 09.511476 (2022).
6. Sun, B.B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).
7. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nature genetics* **53**, 1712-1721 (2021).
8. Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nature Genetics* **54**, 593-602 (2022).
9. Xu, F. *et al.* Genome-wide genotype-serum proteome mapping provides insights into the cross-ancestry differences in cardiometabolic disease susceptibility. *Nature Communications* **14**, 896 (2023).
10. Gudjonsson, A. *et al.* A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nature communications* **13**, 480 (2022).
11. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
12. Koprulu, M. *et al.* Proteogenomic links to human metabolic diseases. *Nature Metabolism* **5**, 516-528 (2023).
13. Lotta, L.A. *et al.* A cross-platform approach identifies genetic regulators of human metabolism and health. *Nature Genetics* **53**, 54-64 (2021).
14. Png, G. *et al.* Mapping the serum proteome to neurological diseases using whole genome sequencing. *Nature communications* **12**, 7042 (2021).
15. Zhao, H. *et al.* Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. *Cell Genomics* **2**(2022).

16. Wu, Y. *et al.* Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large fraction of GWAS signals shared with molecular phenotypes. *Cell Genomics* (2023).
17. Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome biology* **21**, 1-22 (2020).
18. Hu, T. *et al.* Omnibus proteome-wide association study (PWAS-O) identified 43 risk genes for Alzheimer's disease dementia. *medRxiv*, 2022.12. 25.22283936 (2022).
19. Wingo, A.P. *et al.* Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nature genetics* **53**, 143-146 (2021).
20. Mishra, A. *et al.* Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature* **611**, 115-123 (2022).
21. Ou, Y.-N. *et al.* Identification of novel drug targets for Alzheimer's disease by integrating genetics and proteomes from brain and blood. *Molecular psychiatry* **26**, 6065-6073 (2021).
22. Zhong, H. *et al.* Identification of blood protein biomarkers associated with prostate cancer risk using genetic prediction models: analysis of over 140,000 subjects. *Human Molecular Genetics*, ddad139 (2023).
23. Wingo, T.S. *et al.* Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nature neuroscience* **24**, 810-817 (2021).
24. Deng, Y.-T. *et al.* Identifying causal genes for depression via integration of the proteome and transcriptome from brain and blood. *Molecular Psychiatry* **27**, 2849-2857 (2022).
25. Wingo, T.S. *et al.* Integrating human brain proteomes with genome-wide association data implicates novel proteins in post-traumatic stress disorder. *Molecular psychiatry* **27**, 3075-3084 (2022).
26. Zhang, Z. *et al.* Brain proteome-wide association study identifies candidate genes that regulate protein abundance associated with post-traumatic stress disorder. *Genes* **13**, 1341 (2022).

27. Gedik, H. *et al.* Identifying potential risk genes and pathways for neuropsychiatric and substance use disorders using intermediate molecular mediator information. *Frontiers in Genetics* **14**, 1191264 (2023).
28. Sathyanarayanan, A. *et al.* Multi-omics data integration methods and their applications in psychiatric disorders. *European Neuropsychopharmacology* **69**, 26-46 (2023).
29. Jia, T., Ma, Y., Qin, F., Han, F. & Zhang, C. Brain proteome-wide association study linking-genes in multiple sclerosis pathogenesis. *Annals of Clinical and Translational Neurology* **10**, 58-69 (2023).
30. Van der Laan, M.J., Polley, E.C. & Hubbard, A.E. Super learner. *Statistical applications in genetics and molecular biology* **6**(2007).
31. Gaziano, J.M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology* **70**, 214-223 (2016).
32. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).
33. Kurki, M.I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023).
34. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *BioRxiv* (2020).
35. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nature genetics* **53**, 1415-1424 (2021).
36. Consortium, G.P. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
37. Zhao, Z. *et al.* Optimizing and benchmarking polygenic risk scores with GWAS summary statistics. *bioRxiv*, 2022.10. 26.513833 (2022).
38. Zhang, H. *et al.* Novel methods for multi-ancestry polygenic prediction and their evaluations in 5.1 million individuals of diverse ancestry. *bioRxiv*, 2022.03. 24.485519 (2022).
39. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nature genetics* **54**, 573-580 (2022).
40. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature genetics* **54**, 412-436 (2022).

41. Kunkle, B.W. *et al.* Novel Alzheimer disease risk loci and pathways in African American individuals using the African genome resources panel: a meta-analysis. *JAMA neurology* **78**, 102-113 (2021).
42. Sherva, R. *et al.* African ancestry GWAS of dementia in a large military cohort identifies significant risk loci. *Molecular Psychiatry* **28**, 1293-1302 (2023).
43. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics* **45**, 1452-1458 (2013).
44. Ulland, T.K. & Colonna, M. TREM2—a key player in microglial biology and Alzheimer disease. *Nature reviews neurology* **14**, 667-675 (2018).
45. Li, J.-T. & Zhang, Y. TREM2 regulates innate immunity in Alzheimer's disease. *Journal of neuroinflammation* **15**, 1-7 (2018).
46. Brown, G.C. & George-Hyslop, S. Does soluble TREM2 protect against Alzheimer's disease? *Frontiers in Aging Neuroscience* **13**, 834697 (2022).
47. Gratuze, M., Leyns, C.E. & Holtzman, D.M. New insights into the role of TREM2 in Alzheimer's disease. *Molecular neurodegeneration* **13**, 1-16 (2018).
48. Klarin, D. *et al.* Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nature genetics* **51**, 1574-1579 (2019).
49. Levin, M.G. *et al.* Genetics of varicose veins reveals polygenic architecture and genetic overlap with arterial and venous disease. *Nature Cardiovascular Research* **2**, 44-57 (2023).
50. Huang, R.D. *et al.* Genome-wide and phenome-wide analysis of ideal cardiovascular health in the VA Million Veteran Program. *Plos one* **17**, e0267900 (2022).
51. Dahlbäck, B.r. & Villoutreix, B.O. Regulation of blood coagulation by the protein C anticoagulant pathway: novel insights into structure-function relationships and molecular recognition. *Arteriosclerosis, thrombosis, and vascular biology* **25**, 1311-1320 (2005).
52. Pentassuglia, L. & Sawyer, D.B. The role of Neuregulin-1β/ErbB signaling in the heart. *Experimental cell research* **315**, 627-637 (2009).

53. Adams, J.N. *et al.* Genetic analysis of haptoglobin polymorphisms with cardiovascular disease and type 2 diabetes in the Diabetes Heart Study. *Cardiovascular diabetology* **12**, 1-9 (2013).
54. Zhang, H., Mooney, C.J. & Reilly, M.P. ABO blood groups and cardiovascular diseases. *International journal of vascular medicine* **2012**(2012).
55. Preis, M. *et al.* Factor XI deficiency is associated with lower risk for cardiovascular and venous thromboembolism events. *Blood, The Journal of the American Society of Hematology* **129**, 1210-1215 (2017).
56. Mirea, A.-M., Tack, C.J., Chavakis, T., Joosten, L.A. & Toonen, E.J. IL-1 family cytokine pathways underlying NAFLD: towards new treatment strategies. *Trends in molecular medicine* **24**, 458-471 (2018).
57. Nance, S.A., Muir, L., Delproprosto, J. & Lumeng, C.N. MSR1 is not required for obesity-associated inflammation and insulin resistance in mice. *Scientific Reports* **13**, 2651 (2023).
58. Colonna, M. The biology of TREM receptors. *Nature Reviews Immunology*, 1-15 (2023).
59. Subramanian, S., Pallati, P.K., Sharma, P., Agrawal, D.K. & Nandipati, K.C. Significant association of TREM-1 with HMGB1, TLRs and RAGE in the pathogenesis of insulin resistance in obese diabetic populations. *American journal of translational research* **9**, 3224 (2017).
60. Devuyst, O., Bochud, M. & Olinger, E. UMOD and the architecture of kidney disease. *Pflügers Archiv-European Journal of Physiology* **474**, 771-781 (2022).
61. Helgadottir, H.T. *et al.* Somatic mutation that affects transcription factor binding upstream of CD55 in the temporal cortex of a late-onset Alzheimer disease patient. *Human Molecular Genetics* **28**, 2675-2685 (2019).
62. Kolev, M.V., Ruseva, M.M., Harris, C.L., Morgan, B.P. & Donev, R.M. Implication of complement system and its regulators in Alzheimer's disease. *Current neuropharmacology* **7**, 1-8 (2009).
63. Dho, S.H., Lim, J.C. & Kim, L.K. Beyond the role of CD55 as a complement component. *Immune network* **18**(2018).
64. Vardarajan, B.N. *et al.* Progranulin mutations in clinical and neuropathological Alzheimer's disease. *Alzheimer's & dementia* **18**, 2458-2467 (2022).

65. Kulminski, A.M. *et al.* Apolipoprotein E region molecular signatures of Alzheimer's disease. *Aging Cell* **17**, e12779 (2018).
66. Reddy, J.S. *et al.* Polygenic risk score analysis identifies deleterious protein-coding variants in novel immune pathway genes ATP8B4, FCGR1A, and LILRB1 that associate with Alzheimer's disease. *medRxiv*, 2022.07. 12.22277557 (2022).
67. Stein, M.B. *et al.* Genome-wide association analyses of post-traumatic stress disorder and its symptom subdomains in the Million Veteran Program. *Nature genetics* **53**, 174-184 (2021).
68. Niemeyer, C., Matosin, N., Kaul, D., Philipsen, A. & Gassen, N.C. The role of cathepsins in memory functions and the pathophysiology of psychiatric disorders. *Frontiers in Psychiatry* **11**, 718 (2020).
69. Jaggers, R.M. *et al.* Stressor-induced reduction in cognitive behavior is associated with impaired colonic mucus layer integrity and is dependent upon the LPS-binding protein receptor CD14. *Journal of Inflammation Research*, 1617-1635 (2022).
70. Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nature medicine* **28**, 1679-1692 (2022).
71. Klarin, D. *et al.* Genome-wide association study of peripheral artery disease in the Million Veteran Program. *Nature medicine* **25**, 1274-1279 (2019).
72. Zhang, P.-Y. PCSK9 as a therapeutic target for cardiovascular disease. *Experimental and Therapeutic Medicine* **13**, 810-814 (2017).
73. Sobati, S. *et al.* PCSK9: a key target for the treatment of cardiovascular disease (CVD). *Advanced Pharmaceutical Bulletin* **10**, 502 (2020).
74. Nikpay, M. *et al.* A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature genetics* **47**, 1121 (2015).
75. Roselli, C. *et al.* Multi-ethnic genome-wide association study for atrial fibrillation. *Nature genetics* **50**, 1225-1233 (2018).
76. Consortium, I.-R.M.R.A. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *The Lancet* **379**, 1214-1224 (2012).

77. Bräuninger, H. *et al.* Matrix metalloproteinases in coronary artery disease and myocardial infarction. *Basic Research in Cardiology* **118**, 1-17 (2023).
78. Furuhashi, M. & Hotamisligil, G.S. Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets. *Nature reviews Drug discovery* **7**, 489-503 (2008).
79. Perrella, G., Nagy, M., Watson, S.P. & Heemskerk, J.W. Platelet GPVI (glycoprotein VI) and thrombotic complications in the venous system. *Arteriosclerosis, thrombosis, and vascular biology* **41**, 2681-2692 (2021).
80. Man, J.C. *et al.* Genetic dissection of a super enhancer controlling the Nppa-Nppb cluster in the heart. *Circulation Research* **128**, 115-129 (2021).
81. Scholz, B. *et al.* Endothelial RSPO3 controls vascular stability and pruning through non-canonical WNT/Ca2+/NFAT signaling. *Developmental cell* **36**, 79-93 (2016).
82. Serveaux-Dancer, M. *et al.* Pathological implications of receptor for advanced glycation end-product (AGER) gene polymorphism. *Disease markers* **2019**(2019).
83. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nature genetics* **52**, 680-691 (2020).
84. Mahajan, A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature genetics* **50**, 559-571 (2018).
85. Thomsen, S.K. *et al.* Type 2 diabetes risk alleles in PAM impact insulin release from human pancreatic β-cells. *Nature genetics* **50**, 1122-1131 (2018).
86. Liu, S., Gao, Y., Zhang, L., Yin, Y. & Zhang, W. Rspo1/Rspo3-LGR4 signaling inhibits hepatic cholesterol synthesis through AMPKα-SREBP2 pathway. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **34**, 14946 (2020).
87. Dai, R. *et al.* Angiopoietin-like-3 knockout protects against glomerulosclerosis in murine adriamycin-induced nephropathy by attenuating podocyte loss. *BMC nephrology* **20**, 1-11 (2019).
88. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics* **51**, 957-972 (2019).

89. Guan, Y. *et al.* A single genetic locus controls both expression of DPEP1/CHMP1A and kidney disease development via ferroptosis. *Nature communications* **12**, 5078 (2021).
90. Gao, R. *et al.* Comprehensive analysis of endoplasmic reticulum-related and secretome gene expression profiles in the progression of non-alcoholic fatty liver disease. *Frontiers in Endocrinology* **13**, 967016 (2022).
91. Cárcel-Márquez, J. *et al.* Causal effect of MMP-1 (matrix metalloproteinase-1), MMP-8, and MMP-12 levels on ischemic stroke: a Mendelian randomization study. (Am Heart Assoc, 2021).
92. Mahdessian, H. *et al.* Integrative studies implicate matrix metalloproteinase-12 as a culprit gene for large-artery atherosclerotic stroke. *Journal of internal medicine* **282**, 429-444 (2017).
93. Jayalatha, A.S., Hesse, L., Ketelaar, M., Koppelman, G. & Nawijn, M. The central role of IL-33/IL-1RL1 pathway in asthma: From pathogenesis to intervention. *Pharmacology & therapeutics* **225**, 107847 (2021).
94. Perkins, T.N., Donnell, M.L. & Oury, T.D. The axis of the receptor for advanced glycation endproducts in asthma and allergic airway disease. *Allergy* **76**, 1350-1366 (2021).
95. Khan, M.A., Assiri, A.M. & Broering, D.C. Complement mediators: key regulators of airway tissue remodeling in asthma. *Journal of translational medicine* **13**, 1-9 (2015).
96. Wain, L.V. *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nature genetics* **49**, 416-425 (2017).
97. Namba, S., Konuma, T., Wu, K.-H., Zhou, W. & Okada, Y. A practical guideline of genomics-driven drug discovery in the era of global biobank meta-analysis. *Cell Genomics* **2**(2022).
98. Melén, E. *et al.* Expression analysis of asthma candidate genes during human and murine lung development. *Respiratory research* **12**, 1-10 (2011).
99. Chong, M. *et al.* Novel drug targets for ischemic stroke identified through Mendelian randomization analysis of the blood proteome. *Circulation* **140**, 819-830 (2019).

100. Liu, H. *et al.* Identification of Potential targets for Stroke and Its Subtypes by Integrating Proteomes and Genetics from Human Plasma. *medRxiv*, 2023.08.29.23294808 (2023).
101. Gill, D., Monori, G., Tzoulaki, I. & Dehghan, A. Iron status and risk of stroke: a Mendelian randomization study. *Stroke* **49**, 2815-2821 (2018).
102. Piek, A. *et al.* Pharmacological myeloperoxidase (MPO) inhibition in an obese/hypertensive mouse model attenuates obesity and liver damage, but not cardiac remodeling. *Scientific Reports* **9**, 18765 (2019).
103. Nicholls, S.J. & Hazen, S.L. Myeloperoxidase and cardiovascular disease. *Arteriosclerosis, thrombosis, and vascular biology* **25**, 1102-1111 (2005).
104. Acquati, F. *et al.* Microenvironmental control of malignancy exerted by RNASET2, a widely conserved extracellular RNase. *Proceedings of the National Academy of Sciences* **108**, 1104-1109 (2011).
105. Lualdi, M. *et al.* Pleiotropic modes of action in tumor cells of RNASET2, an evolutionary highly conserved extracellular RNase. *Oncotarget* **6**, 7851 (2015).
106. Chahal, H.S. *et al.* Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. *Nature communications* **7**, 12510 (2016).
107. Park, S.Y. *et al.* Dehydropeptidase 1 promotes metastasis through regulation of E-cadherin expression in colon cancer. *Oncotarget* **7**, 9501 (2016).
108. Olson, O.C. & Joyce, J.A. Cysteine cathepsin proteases: regulators of cancer progression and therapeutic response. *Nature Reviews Cancer* **15**, 712-729 (2015).
109. Song, L. *et al.* Expression signature, prognosis value and immune characteristics of cathepsin F in non-small cell lung cancer identified by bioinformatics assessment. *BMC Pulmonary Medicine* **21**, 1-17 (2021).
110. Akkari, L. *et al.* Combined deletion of cathepsin protease family members reveals compensatory mechanisms in cancer. *Genes & development* **30**, 220-232 (2016).
111. Ter Steege, E.J. & Bakker, E.R. The role of R-spondin proteins in cancer biology. *Oncogene* **40**, 6469-6478 (2021).
112. Ter Steege, E.J. *et al.* R-spondin-3 is an oncogenic driver of poorly differentiated invasive breast cancer. *The Journal of Pathology* **258**, 289-299 (2022).

113. Sun, J. *et al.* Overexpression of colorectal cancer oncogene CHRDL2 predicts a poor prognosis. *Oncotarget* **8**, 11489 (2017).
114. Zheng, S. *et al.* CRTAM promotes antitumor immune response in triple negative breast cancer by enhancing CD8+ T cell infiltration. (2023).
115. Pan, J.-h. *et al.* LAYN is a prognostic biomarker and correlated with immune infiltrates in gastric and colon cancers. *Frontiers in immunology* **10**, 6 (2019).
116. Xie, X. *et al.* Genetic blockade of NAAA cell-specifically regulates fatty acid ethanolamides (FAEs) metabolism and inflammatory responses. *Frontiers in Pharmacology* **12**, 817603 (2022).
117. Huang, D. *et al.* Insights into the prognostic value and immunological role of NAAA in pan-cancer. *Frontiers in Immunology* **12**, 812713 (2022).
118. Chen, M.-C. *et al.* RAGE acts as an oncogenic role and promotes the metastasis of human lung cancer. *Cell death & disease* **11**, 265 (2020).
119. Cai, P. *et al.* BTN3A2 serves as a prognostic marker and favors immune infiltration in triple-negative breast cancer. *Journal of Cellular Biochemistry* **121**, 2643-2654 (2020).
120. Chu, L.-Y. *et al.* EFNA1 in gastrointestinal cancer: Expression, regulation and clinical significance. *World Journal of Gastrointestinal Oncology* **14**, 973 (2022).
121. Liang, X., Du, L. & Fan, Y. The potential of FCRL genes as targets for cancer treatment: insights from bioinformatics and immunology. *Aging (Albany NY)* **15**, 4926 (2023).
122. Yousefi, Z. & Eskandari, N. Prognostic significance of Fc receptor-like 1 in patients with chronic lymphocytic leukemia, hairy cell leukemia, and various B-cell non-Hodgkin's lymphoma. *Leukemia research reports* **12**, 100181 (2019).
123. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* **10**, e1004383 (2014).
124. Kibinge, N.K., Relton, C.L., Gaunt, T.R. & Richardson, T.G. Characterizing the causal pathway for genetic variants associated with neurological phenotypes using human brain-derived proteome data. *The American Journal of Human Genetics* **106**, 885-892 (2020).

125. Sakaue, S. & Okada, Y. GREP: genome for REPositioning drugs. *Bioinformatics* **35**, 3821-3823 (2019).
126. Benner, C. *et al.* Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *The American Journal of Human Genetics* **101**, 539-551 (2017).
127. Zhao, B., Zheng, S. & Zhu, H. On block-wise and reference panel-based estimators for genetic data prediction in high dimensions. *arXiv preprint arXiv:2203.12003* (2022).
128. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47**, 284-290 (2015).
129. Kowalski, M.H. *et al.* Use of> 100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS genetics* **15**, e1008500 (2019).
130. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529 (2009).
131. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols* **7**, 500-507 (2012).
132. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nature communications* **8**, 15452 (2017).
133. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
134. Zhao, B. *et al.* Common genetic variation influencing human white matter microstructure. *Science* **372**, eabf3736 (2021).
135. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics* **51**, 1749-1755 (2019).
136. Zhang, Z., Bae, Y.E., Bradley, J.R., Wu, L. & Wu, C. SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification. *Nature communications* **13**, 6336 (2022).

137. Wen, X. & Stephens, M. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics* **4**, 1158 (2010).
138. Lloyd-Jones, L.R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature communications* **10**, 5086 (2019).
139. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091 (2015).
140. Xue, H., Shen, X. & Pan, W. Causal Inference in Transcriptome-Wide Association Studies with Invalid Instruments and GWAS Summary Data. *Journal of the American Statistical Association*, 1-27 (2023).
141. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 1273-1300 (2020).
142. Burgess, S., Thompson, S.G. & Collaboration, C.C.G. Avoiding bias from weak instruments in Mendelian randomization studies. *International journal of epidemiology* **40**, 755-764 (2011).
143. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS genetics* **13**, e1007081 (2017).
144. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine* **27**, 1133-1163 (2008).

METHODS

pQTL data resources.

We used pQTL summary statistics from the UKB, deCODE, and ARIC studies. Detailed descriptions of the deCODE and ARIC pQTL summary statistics can be found in Ferkingstad, et al. [7] and Zhang, et al. [8], respectively. Briefly, in the deCODE cohort, the average participant age was 55 years and 57% were women. The deCODE pQTL study included 35,559 Icelanders who had both genetic and proteomic data available. The plasma proteome data were measured using the

SomaScan version 4 assay (SomaLogic). Imputed variants with minor allele frequency (MAF) > 1% and imputation information > 0.9 were used in the deCODE pQTL study. The rank-inverse normal transformed levels were used in the linear mixed model implemented in BOLT-LMM[128], while adjusting for age and sex effects. The ARIC pQTL study, also using the SomaScan version 4 assay, comprised data from 7,213 European American participants and 1,871 African American participants. Genotypic data were collected by the Affmetrix 6.0 DNA microarray and subsequently imputed to the TOPMed reference panel[129,130] (Freeze 5b). For pQTL analysis, the log-transformed relative abundance values were residualized to adjust for probabilistic estimation of expression residual factors[131], sex, age, study site, and ten genetic principal components (PCs). The residuals were rank-inverse normalized and pQTL analysis was performed using QTLtools[132]. For both deCODE and ARIC pQTL data, we followed the pipeline outlined in Zhang, et al. [8], focusing on 4,483 unique proteins or protein complexes encoded by 4,435 autosomal genes passed quality controls (QCs), including excluding aptamers that mapped to multiple gene targets and without any SNPs in the *cis*-region.

We derived the UKB pQTL summary statistics from 46,066 white British participants in the UKB-PPP project. We used the imputed genotyping data[133] and any imputed variants with a QC r^2-value below 0.8 were excluded. Subsequent to this, we applied standard QC procedures[134], which included i) omitting subjects with over 10% missing genotypes; ii) discarding variants with an MAF under 0.01; iii) removing variants with a missing genotype rate exceeding 10%; and iv) excluding variants that did not pass the Hardy-Weinberg equilibrium test at a threshold of 1×10^{-7} level. For every protein, outliers were excluded, which were defined as values exceeding five times the median absolute deviation from the median. We used linear mixed-effect models with fastGWA[135], accounting for effects of covariates including age, sex, genotype array, visiting center ID, age squared, the interaction between age and sex, the interaction between age squared and sex, and the top 40 genetic PCs[133]. We considered *cis*-SNPs within each gene region and extended our consideration to a 1-megabase window on

either side. Additionally, we used three smaller UKB proteomic datasets for the purpose of assessing prediction accuracy or generating non-European super learner PWAS models. These datasets included 2,923 white non-British subjects, 1,171 African ancestry subjects, and 914 Asian ancestry subjects. To avoid potential biases, we randomly eliminated one from any pair of related individuals within each of the three small datasets and excluded any relations of the white British subjects.

Biomarker expression Levels Imputation using Summary Statistics (BLISS). To train the protein expression imputation model using pQTL summary statistics, we introduced a new method called BLISS. Consider the following linear regression model for protein expression:

$$X = \sum_{j=1}^{p} w_j G_j + \varepsilon,$$

where X is the protein expression levels, $G = (G_1, \cdots, G_p)'$ is the standardized genotype matrix of p *cis*-SNPs around the gene, $w = (w_1, \cdots, w_p)'$ is the effect size to be estimated, and ε is random noise with a mean of zero. The objective function $f(w)$ for a penalized regression to estimate w is:

$$f(w) = \frac{(X - Gw)'(X - Gw)}{N} + J_\lambda(w) = \frac{X'X}{N} + w'\frac{G'G}{N}w - 2w'\frac{G'X}{N} + J_\lambda(w),$$

where $J_\lambda(w)$ is a penalty term such as Elastic-net, Lasso, or MCP, and N is the sample size of the pQTL study. Following SUMMIT[136], we generated solutions for a candidate set of tuning parameters. Specifically, $G'X/N$ represented the correlation between cis-SNPs and protein expression levels, which can be estimated by $Z_j/\sqrt{N_j - 1 + Z_j^2}$, where Z_j and N_j are the *Z* score (estimated effect size divided by its standard error) and sample size for SNP j, respectively. The term $G'G/N$ was the LD correlation matrix, which can be estimated by an LD reference panel such as the 1000 Genomes project[36] of matched ancestry. We used the shrinkage estimator and shrunk the off-diagonal entries toward zero based on their genetic distance[136-138]. To ensure a unique solution upon

optimization, a small L_2 penalty term $\theta w'w$ was added to the objective function. Then the objective function was optimized through a coordinate descent algorithm.

To generate independent pseudo-training and validation datasets for selecting tuning parameters, the key idea was to sample marginal association statistics of pQTL data for a subset of individuals conditional on the complete summary-level pQTL data. Specifically, we generated the pseudo-training data $G'_{(tr)}X_{(tr)}$ with sample size n by a conditional distribution: $G'_{(tr)}X_{(tr)}|G'X \sim N(\frac{(N-n)G'X}{N}, \frac{(N-n)}{N}\Sigma)$. Here $\cdot|\cdot$ denotes the conditional distribution and Σ is the observed covariance matrix, which can be estimated by using minor allele, estimated effect sizes, and standard errors provided by pQTL summary data[37]. We then calculated the estimated effect sizes and their standard errors for pseudo-training data. After generating the pseudo-training data $G'_{(tr)}X_{(tr)}$, we obtained the pseudo-validation data $G'_{(v)}X_{(v)}$ by $G'X - G'_{(tr)}X_{(tr)}$. Subsequently, we calculated the summary-level predictive R^2, which was the squared Pearson correlation coefficient between genetically predicted and directly measured protein expression levels, using the pseudo-validation data[37]. We repeated the above procedure five times and selected the parameters that maximized the average predictive R^2 across these five replicates. We then trained protein prediction models using all data and the selected tuning parameters. These models were used in the subsequent analysis.

Super learner integration.

Using the super learner approach[30], we developed non-European PWAS models using the individual-level UKB data of African and Asian ancestries. For illustrations, we used Asian ancestry as an example; the procedures for African ancestry were similar. First, following the steps in PrediXcan[139], we built the protein imputation model for each protein by Elastic-net using *cis*-SNPs (within 1-megabase of the transcription start and end site) as predictors. The tuning parameters for Elastic-net were selected based on a nested five-fold cross-validation procedure, and the final model was refitted using all data and the selected tuning parameters. Second, recognizing that the PWAS models for

Europeans were built on much larger sample sizes and could potentially improve the prediction accuracy of Asian models, we applied super learner to combine the standard (Elastic-net) Asian models and BLISS-based European models, resulting in super learner Asian PWAS models. We used non-negative least squares (NNLS) to produce a weighted sum of predictions from standard and BLISS models, where the weights were learned from nested five-fold cross-validation. In this nested setup, the whole dataset was divided into five independent validation datasets and corresponding training datasets. An Asian model was trained and tuned on each training set and the super learner weights were subsequently constructed by minimizing the NNLS of the weighted sum of predictions from both standard and BLISS models on the validation datasets. The average weights across all five folds were used to construct the final super learner Asian PWAS model. Following PrediXcan[139], we evaluated the model performance using predictive R^2 within the nested five-fold cross-validation to avoid the upward bias commonly seen in standard cross-validation.

PWAS model training and evaluation.

To satisfy the relevance assumption, we focused on 1,201 SomaScan proteins measured in deCODE, 1,209 SomaScan proteins in ARIC, and 759 Olink proteins measured in UKB. These proteins were selected based on a *cis*-heritability threshold greater than 0.01[8,19], estimated using the REML algorithm in GCTA[4,8]. For genotype data, we excluded SNPs with MAF < 1%, those that were non-biallelic or had ambiguous strand mapping (i.e., no A/T or C/G SNPs), and those did not present in the LD reference panel (i.e., the 1000 Genomes[36]). Additionally, SNPs were clumped if they were $r^2 > 0.2$ with the index variant to improve the signal-to-noise ratio (window size = 10 megabases, which performed whole-gene clumping). We applied BLISS to train PWAS models using summary-level pQTL data from deCODE, UKB (European), and ARIC (European and African separately). We applied a 1-megabase window on either side for both deCODE and UKB models. For ARIC, we used a 500-kilobase window due to the available pQTL summary data. We focused on models where the number of SNPs with non-zero weights was greater than 15, since having a sufficiently large number of these

SNPs was crucial to ensure reliable asymptotic statistical properties during the second stage of association tests. Additionally, we applied the super learner framework to train PWAS models using individual-level UKB data of African and Asian populations, alongside BLISS-trained UKB European models.

The performance of the imputation models was evaluated by predictive R^2, the squared Pearson correlation coefficient between genetically predicted and directly measured protein expression levels, which was provided by both BLISS and super learner. To assess the reliability of the summary-level predictive R^2 values in BLISS, we calculated the individual-level predictive R^2 using an independent testing dataset comprising 2,923 white non-British individuals from the UKB study. Specifically, we used BLISS weights derived from our UKB pQTL analysis from white British individuals to predict protein expression levels in the 2,923 white non-British UKB individuals. The individual-level predictive R^2 values were compared with those calculated from BLISS. Similarly, we also assessed the impact of clumping r^2 using both summary-level and individual-level predictive R^2. We summarized the PWAS models and related information in **Table S1**.

PWAS association testing.

When only summary-level GWAS data were available, standard PWAS applied a burden-type *Z* test to examine the association between the genetically predicted protein expression levels and phenotype[8,19]. Under the null hypothesis, the test statistics was given by $T = Z'\widehat{w}/\sqrt{\widehat{w}'V\widehat{w}} \sim N(0,1)$, where Z is the vector of z-scores from GWAS, $\widehat{w}$ is the estimated weights from PWAS model, and V is the LD matrix of corresponding SNPs estimated from a population reference panel (such as the 1000 Genomes project[36]). It was worth mentioning that one assumption underlying this *Z* test was that the LD matrix estimated from the reference panel precisely matched that from the GWAS data. This alignment may be crucial as any discrepancy could potentially lead to an inflated type 1 error rate[140]. To reduce the potential negative effects of such LD covariance matrix discrepancy, in our PWAS association test, we estimated the effect size ($\hat{\gamma}$) and its variance between the protein and phenotype by[140]

$$\hat{\gamma} = \frac{\widehat{w}'Z/\sqrt{n_s}}{\sigma_r} \quad \text{and} \quad \widehat{Var}(\hat{\gamma}) = \left(\frac{1}{n_s} + \frac{1}{n_r}\right)\hat{\gamma}^2 + \frac{\zeta^2}{n_s \sigma_r} \ ,$$

where n_s is the sample size of the GWAS data, $\sigma_r = (G_r\widehat{w})'(G_r\widehat{w})/n_r$, and $\zeta^2 = 1 - 2\,\widehat{w}'Z\hat{\gamma}/\sqrt{n_s} + \sigma_r\,\hat{\gamma}^2$, G_r is the standardized (mean 0 and variance 1) genotype matrix of the population reference panel, and n_r is its sample size.

Simulations.

We used the UKB imputed genotype data[133] to conduct simulations, focusing on evaluating the protein prediction accuracy, testing power, and type 1 error rate of different methods. Specifically, we randomly selected three separate sets of unrelated white British individuals: 45,000 individuals as a training dataset for PWAS model building, 10,000 individuals as an independent testing dataset for evaluating predictive R^2, and 50,000 individuals for generating GWAS summary statistics. For our main simulations, the randomly chosen one protein (VIM), which contained 323 *cis*-SNPs that have been used in our model-building step in UKB.

Protein expression levels (X) and phenotype values (Y) were simulated by $X = Gw + \varepsilon_x$ and $Y = \gamma X + \varepsilon_y$, respectively. Here G is the genotype matrix, w is the vector of effect sizes in the PWAS model, the scalar γ is the association coefficient between protein expression and phenotype, and ε_x and ε_y are vectors of normally distributed noise to achieve pre-specified protein expression heritability and phenotypic heritability. We generated pQTL summary data by running a marginal association scan between protein expression and each SNP in the PWAS training data. Additionally, GWAS summary statistics were generated by regressing the phenotype on each SNP in the GWAS data. We considered three different PWAS training sample sizes—1,000, 7,000, and 45,000—to investigate the impact of sample size on model prediction accuracy. Given that biobank data have been widely meta-analyzed in modern GWAS, we considered three settings where 0%, 50%, and 100% of individuals in the PWAS training dataset were also part of the GWAS dataset.

We considered multiple scenarios that varied the number of causal SNPs (1, 3, and 5), represented as non-zero elements in w that were sampled from a standard normal distribution. We also examined different levels of protein expression heritability (0.01, 0.05, 0.1) and phenotypic heritability (0.001, 0.003, 0.005, 0.007, and 0.01). For each scenario, we generated 500 PWAS training and testing datasets to evaluate the predictive R^2. For each PWAS training dataset, we produced 10 replicated GWAS datasets for evaluating empirical statistical power. The empirical statistical power was calculated as the proportion of 5,000 repeated simulations with a P value less than 0.005. Furthermore, we produced 300 replicated GWAS datasets for each PWAS training dataset under the null (where γ=0), resulting in 1.2 million replicates across all settings considered. The type 1 error rates were evaluated using the quantile-quantile (QQ) plot with a 95% confidence interval.

PWAS analysis with five GWAS databases.

We applied our PWAS models to over 700 complex traits and diseases, most of which were from five major genetic data resources, including the MVP[31], GBMI[32], FinnGen[33], IEU OpenGWAS[34], and BBJ[35] studies. **Table S14** summarized the information of these GWAS summary statistics. Briefly, the MVP GWAS summary statistics were obtained from dbGaP (study accession: phs001672.v3.p1). We focused on 41 MVP phenotypes for which GWAS summary statistics were available for both European and African ancestries. In addition, we included Alzheimer's disease, for which we also had GWAS summary statistics for both European and African ancestries[40,41]. The GBMI GWAS summary statistics were downloaded from https://www.globalbiobankmeta.org/resources and we focused on seven diseases with available GWAS summary statistics across European, African, and Asian ancestries. For stroke, we additionally incorporated data of its various subtypes, obtained from a recent multi-ancestry GWAS[20]. FinnGen GWAS summary statistics were accessed from https://www.finngen.fi/en/access_results (version R9). We focused on 278 clinical outcomes defined by FinnGen, most of which had more than 10,000 cases. We also incorporated three diseases with fewer than 10,000 cases, namely Alzheimer's disease (n cases = 9,301), aortic

aneurysm (*n* cases = 7,395), and chronic kidney disease (*n* cases = 9,073). For the IEU OpenGWAS database (https://gwas.mrcieu.ac.uk/datasets), we first downloaded all European phenotypes in the 'ebi-' and 'ieu-' categories that had more than 5×10^6 genetic variants and a GWAS sample size exceeding 10,000. Furthermore, we included cancer-related traits with a GWAS sample size greater than 3,000. After manually reviewing each downloaded GWAS dataset, we excluded 31 phenotypes that posed potential issues (e.g., missing an autosome), leading to a finalized list of 309 traits. For the BBJ database, we downloaded GWAS statistics for 220 Asian phenotypes[35] (https://www.ebi.ac.uk/gwas/publications/34594039). We set a threshold of *n* cases > 2,000 for binary traits, resulting in a selection of 107 phenotypes. In all analyses, we only considered protein-phenotype pairs where the number of SNPs in the test exceeded 10.

Post-PWAS analyses.

We conducted an MR analysis in line with the methodology outlined by Zhao, et al. [15]. We focused on European PWAS findings due to the substantial sample sizes of their pQTL data. Specifically, we examined the European PWAS transferable (significant in both European and African datasets) and consistent (across the three European PWAS cohorts) results from the MVP study, significant European associations from the GBMI study, and consistently significant associations from both the FinnGen and IEU OpenGWAS studies. We selected a set of conditionally independent *cis*-pQTLs as the MR instrumental variables. We first restricted our analysis to the common set of variants that were shared by pQTLs and GWAS. To avoid the potential issue of collinearity, we removed pQTLs strongly correlated with index genetic variants (r^2 > 0.6) by applying LD clumping. We then applied SuSiE[141] to fine-map candidate instrumental variables. To increase the reliability of MR analysis, we further filtered the candidate instrumental variables. First, we estimated the instrument strength using F-statistics and excluded pQTLs with F-statistics < 15 from the candidate set to meet the relevance assumption and avoid potential weak instrument bias[142]. Second, we applied the Steiger filter[143] approach to exclude instrumental variables with potential reverse causality. After selecting

the instrumental variables, we applied either the Wald ratio (when only one instrumental variable was available) or inverse variance weighting[144] to test the causal link between protein and phenotype.

We also performed Bayesian colocalization analysis to estimate the posterior probability that the protein and phenotype shared the same causal variant, using summary-level *cis*-pQTLs and GWAS data. Specifically, we used the coloc[123] package (v5.2.2), with its default setups in the coloc function, to estimate the posterior probability of both protein and phenotype being influenced by the same causal variant (i.e., the PPH4). We chose PPH4 > 0.8[123,124] as the threshold. Proteins satisfying this threshold would suggest a shared causal variant for the *cis*-pQTLs and GWAS associations.

We assessed the enrichment overlap between PWAS-significant genes and known drug-target genes for each primary ICD-10 disease category. First, we obtained a list of genes significant for diseases from any of our PWAS analyses across the five GWAS databases (MVP, GBMI, FinnGen, IEU OpenGWAS, and BBJ). Subsequently, we focused on seven disease categories with more than 200 PWAS-significant genes, including cancer (and neoplasms, CD2 & ICD-O-3), endocrine, nutritional, and metabolic (E4), nervous system (G6), circulatory system (I9), respiratory system (J10), digestive system (K11), as well as musculoskeletal and connective (M13). For each disease category, we examined the enrichment with drug-target genes of corresponding ICD-10 categories, as defined by GREP[125]. For example, for PWAS-significant genes of the circulatory system (I9) diseases, we tested enrichment using Fisher's exact tests for several disease groups with an ICD-10 code starting with "I".

Code availability

Our PWAS analysis pipeline can be accessed at https://github.com/gcb-hub/BLISS. All software and tools utilized in this study are publicly accessible. The code for model development will be available on Zenodo.

Data availability

The PWAS data resources and trained models are available at www.gcbhub.org. GWAS summary statistics used in our analysis can be downloaded or requested via their individual project websites. The individual-level data used in this study can be accessed from the UK Biobank study at https://www.ukbiobank.ac.uk/.

Figure legends

Fig. 1 Study design and model performance.

(A) Overview of the PWAS model training framework and data analysis pipeline. **(B)** Illustrates the similarity between the summary and individual-level predictive R^2 values. Each dot represents one protein. **(C)** Compares the predictive R^2 of the super learner models (ASN+EUR) and standard PWAS model (ASN). Each dot represents one protein. The horizontal and vertical dotted lines represent predictive R^2 = 1%. **(D)** Displays the count of proteins with predictive R^2 values within each respective bin.

Fig. 2 Simulation results.

(A) Predictive R^2 of BLISS and standard PWAS models across various *cis*-protein expression heritability (0.01, 0.05, and 0.1) and sample sizes (1,000, 7,000, and 45,000). **(B)** Illustrates the similarity between the summary and individual-level predictive R^2 values for BLISS (sample size n = 7,000). **(C)** Shows the pattern of statistical power in identifying associations for BLISS PWAS models. **(D)** Compares the statistical power of BLISS and standard PWAS models across different expression heritability (sample size n = 7,000). **(E)** Compares the statistical power of BLISS and standard PWAS models across different sample sizes (phenotypic heritability = 0.005). In all these results, the number of causal SNPs is three.

Fig. 3 Transferability between African and European PWAS results.

(A) Demonstrates the significant PWAS associations at the FDR 5% level ($P < 6.20\times10^{-4}$). The horizontal dotted line represents the more stringent Bonferroni-significant level. We label the categories of phenotypes with different colors. **(B)** Illustrates the consistency of *Z*-scores for associations that were significant in both African (AFR) and European (EUR) ancestries. Each dot represents a protein-phenotype pair that was found significant in both ancestries. Protein-phenotype pairs with inconsistent directions are marked in grey, while those with consistent directions are color-coded according to their categories. **(C)** Summarizes the PWAS associations that were significant in both African and European ancestries. We also listed genes previously identified in GWAS or those that passed either the colocalization or Mendelian randomization tests.

Fig. 4 Consistent PWAS findings across reference proteomic datasets.

(A)-(C) Illustrate the consistency of European PWAS association *Z*-scores between the UKB and deCODE models (in **A**), between the ARIC and UKB models (in **B**), and between the ARIC and deCODE models (in **C**). Each dot represents a protein-phenotype pair that was found significant by all three model sets. protein-Phenotype pairs with inconsistent directions are marked in grey, while those with consistent directions are color-coded according to their categories. **(D)** Summarizes the PWAS associations that were significant in all three European model sets (UKB, deCODE, and ARIC) and/or two African model sets (UKB and ARIC). We also listed genes previously identified in GWAS or those that passed either the colocalization or Mendelian randomization tests.

Fig. 5 FinnGen data analysis.

(A) Demonstrates the significant PWAS associations across UKB, deCODE, and ARIC models at the FDR 5% level ($P < 3.02\times10^{-4}$). The horizontal dotted line represents the more stringent Bonferroni-significant level. We label the categories of FinnGen diseases with different colors. **(B)** Displays the proportion of PWAS signals in each disease category across the UKB, deCODE, and ARIC models. **(C)**

& (D) Illustrate the consistency of association *Z*-scores between the UKB and deCODE models (in **C**) and between the ARIC and deCODE models (in **D**). In **C**, each dot represents a protein-disease pair that was found significant by all three model sets. In **D**, each dot represents a protein-disease pair that was found significant by both ARIC and deCODE models. Protein-disease pairs with inconsistent directions are marked in grey, while those with consistent directions are color-coded according to their categories.

Fig. 6 MRC IEU OpenGWAS data analysis.

(A) Demonstrates the significant PWAS associations across UKB, deCODE, and ARIC models after controlling multiple testing at the FDR 5% level ($P < 3.56\times10^{-4}$). The horizontal dotted line represents the more stringent Bonferroni-significant level. We label the categories of phenotypes with different colors. **(B)-(D)** Illustrate the consistency of association *Z*-scores between the UKB and deCODE models (in **B**), ARIC and UKB models (in **C**), as well as ARIC and deCODE models (in **D**). Each dot represents a protein-phenotype pair that was found significant by all three model sets. Protein-phenotype pairs with inconsistent directions are marked in grey, while those with consistent directions are color-coded according to their categories.

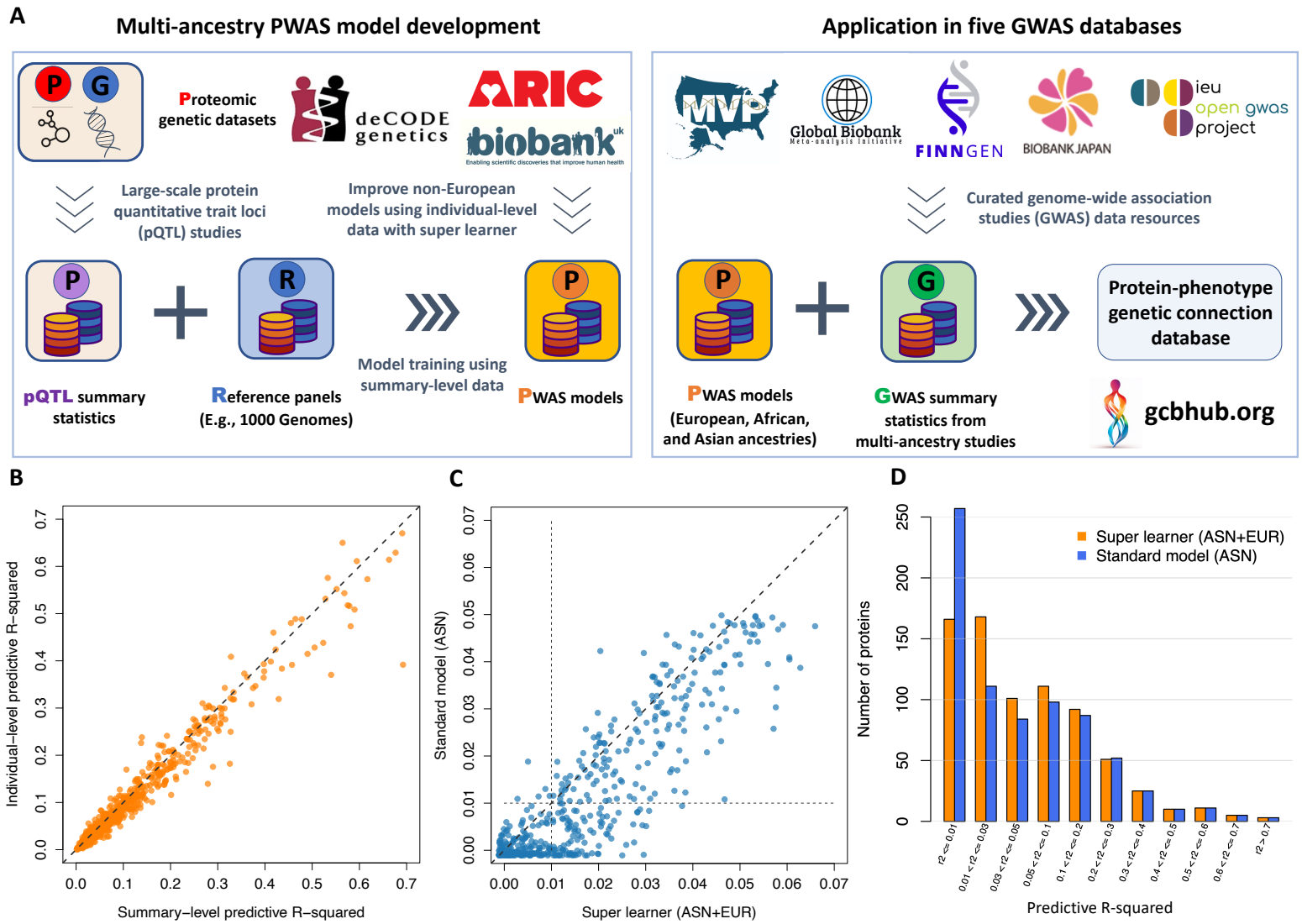

A
Multi-ancestry PWAS model development
Proteomic genetic datasets
deCODE genetics
ARIC
biobank uk
Large-scale protein quantitative trait loci (pQTL) studies
Improve non-European models using individual-level data with super learner
pQTL summary statistics
Reference panels (E.g., 1000 Genomes)
Model training using summary-level data
PWAS models
Application in five GWAS databases
MVP
Global Biobank Meta-analysis Initiative
FINNGEN
BIOBANK JAPAN
ieu open gwas project
Curated genome-wide association studies (GWAS) data resources
Protein-phenotype genetic connection database
PWAS models (European, African, and Asian ancestries)
GWAS summary statistics from multi-ancestry studies
gcbhub.org
B
Individual−level predictive R−squared
Summary−level predictive R−squared
C
Standard model (ASN)
Super learner (ASN+EUR)
D
Super learner (ASN+EUR)
Standard model (ASN)
Number of proteins
Predictive R-squared

Figure 1

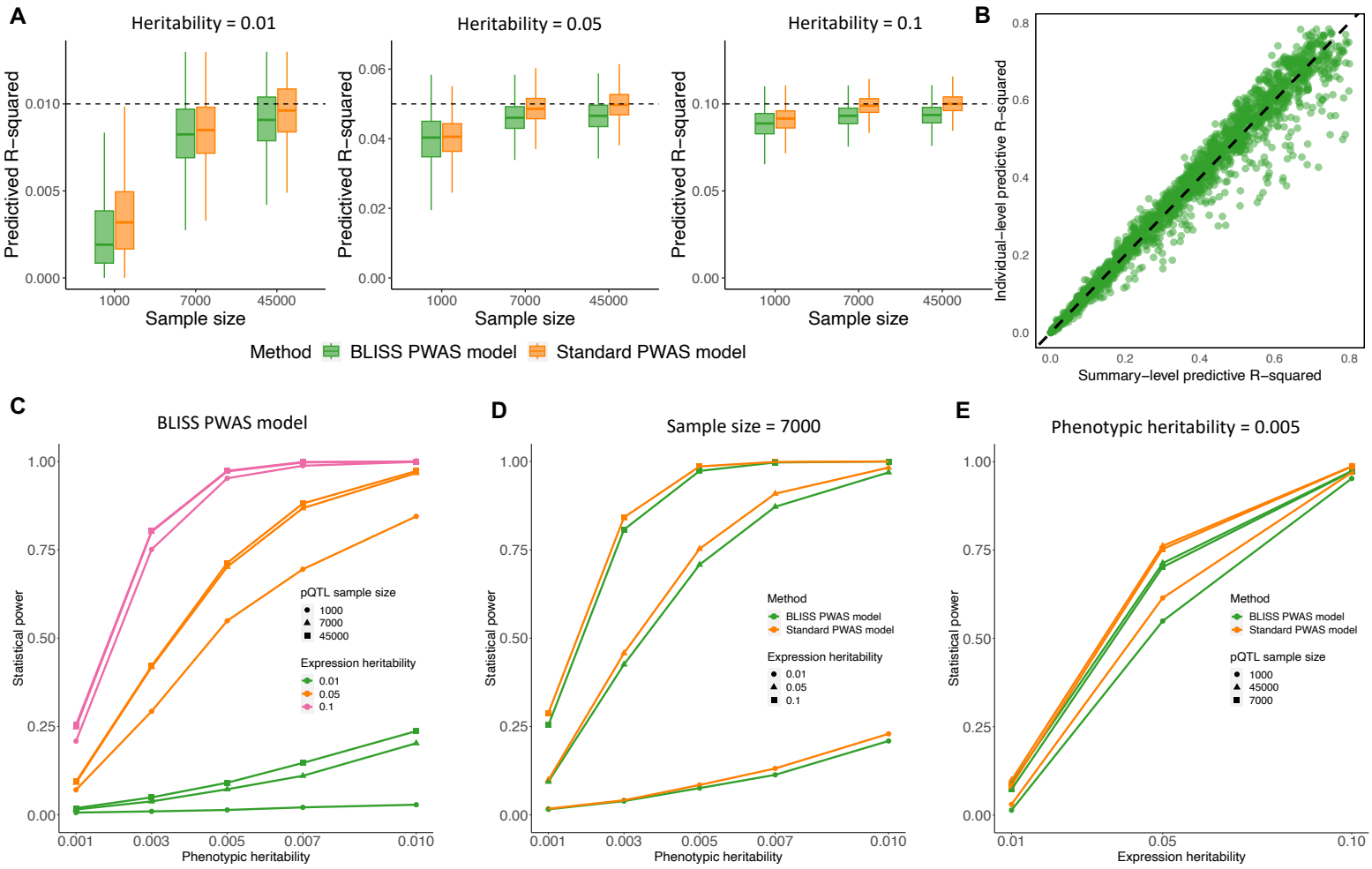
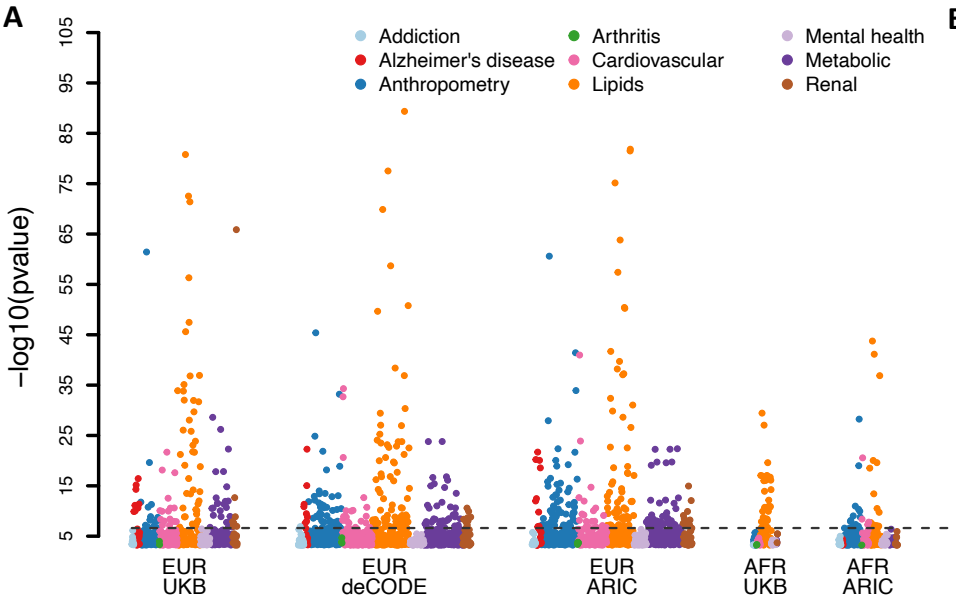
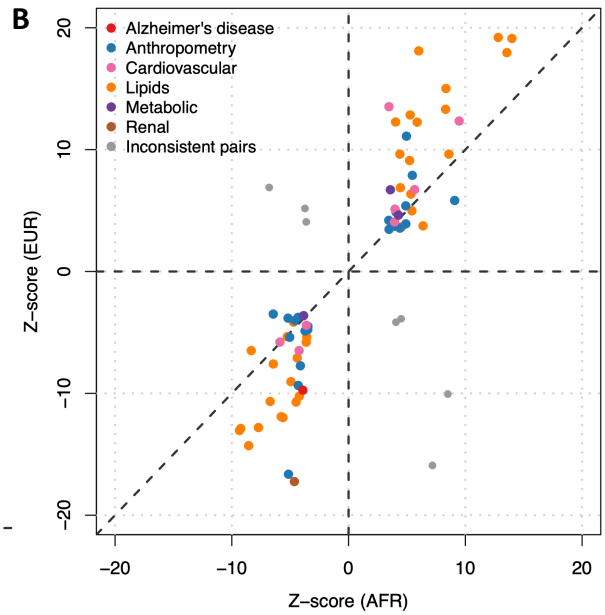

A
Heritability = 0.01
Heritability = 0.05
Heritability = 0.1
Predictived R−squared
Sample size
1000
7000
45000
Method
BLISS PWAS model
Standard PWAS model
B
Individual−level predictive R−squared
Summary−level predictive R−squared
C
BLISS PWAS model
Statistical power
Phenotypic heritability
pQTL sample size
1000
7000
45000
Expression heritability
0.01
0.05
0.1
D
Sample size = 7000
Method
BLISS PWAS model
Standard PWAS model
Expression heritability
0.01
0.05
0.1
E
Phenotypic heritability = 0.005
Expression heritability
Method
BLISS PWAS model
Standard PWAS model
pQTL sample size
1000
45000
7000

Figure 2



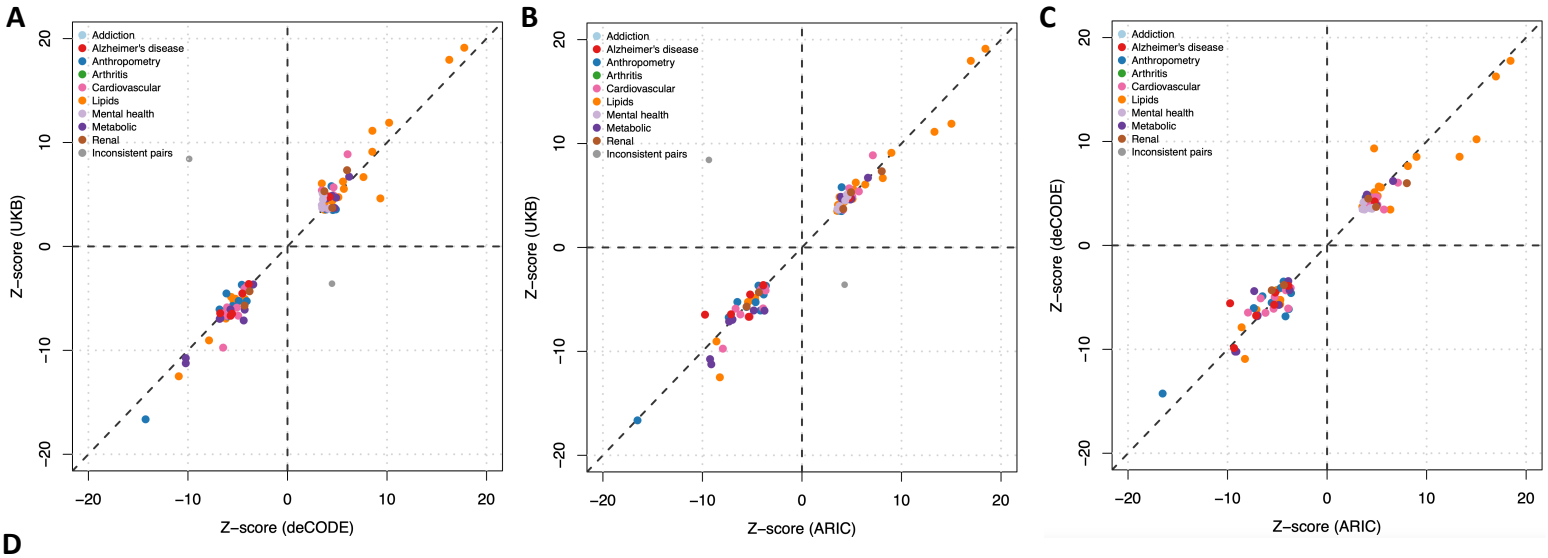


C

Trait(s)	Gene(s)	Beta	*P* value	Beta	*P* value	GWAS (GWAS Catalog)	Colocalization (PPH4 > 0.8)	pQTL MR (FDR < 0.05)
Alzheimer's disease	*TREM2*	$\beta < -0.013$	$P < 2.57\times10^{-8}$	$\beta < -0.017$	$P < 3.67\times10^{-4}$	*TREM2*	*TREM2*	*TREM2*
Venous thromboembolism (VTE)	*ABO, F11, PROC, PROS1, THBD*	$\lvert\beta\rvert > 0.012$	$P < 1.01\times10^{-5}$	$\lvert\beta\rvert > 0.020$	$P < 5.45\times10^{-4}$	*ABO, F11, THBD*	*ABO, F11*	*ABO, F11*
Varicose veins	*ABO*	$\beta > 0.002$	$P < 1.23\times10^{-4}$	$\beta = 0.011$	$P = 1.59\times10^{-8}$	*ABO*	*ABO*	*ABO*
Cardiovascular ideal health score (IHS)	*ERBB4, HP*	$\beta > 0.005$	$P < 4.87\times10^{-5}$	$\beta > 0.020$	$P < 8.40\times10^{-5}$		*HP*	*ERBB4, HP*
Non-alcoholic fatty liver disease (NAFLD)	*IL1RN*	$\beta > 0.025$	$P < 5.47\times10^{-10}$	$\beta = 0.076$	$P = 3.46\times10^{-4}$			*IL1RN*
Type 2 diabetes (T2D)	*MSR1, TREML2*	$\lvert\beta\rvert > 0.008$	$P < 2.92\times10^{-4}$	$\lvert\beta\rvert > 0.011$	$P < 6.08\times10^{-4}$	*TREML2*		*MSR1, TREML2*
Estimated glomerular filtration rate (eGFR)	*UMOD*	$\beta = -0.069$	$P = 1.36\times10^{-66}$	$\beta = -0.102$	$P = 3.42\times10^{-6}$	*UMOD*		*UMOD*
Body mass index (BMI)	*ARL3, PSMD9*	$\lvert\beta\rvert > 0.019$	$P < 2.69\times10^{-5}$	$\lvert\beta\rvert > 0.018$	$P < 5.78\times10^{-4}$	*ARL3*		*ARL3*
Lipids	*ANGPTL3, APOA5, CD300LG, GALNT2, LILRB1, LPL, PLA2G15, PLTP, BCAM, BCAT2, GAS6, GLTPD2, HP, PCSK9, SORT1, TIMD4, PCSK7, UGT1A6, APOA4, LRP4, MSR1*	$\lvert\beta\rvert > 0.005$	$P < 2.46\times10^{-4}$	$\lvert\beta\rvert > 0.005$	$P < 3.40\times10^{-4}$	*ANGPTL3, APOA5, CD300LG, GALNT2, LPL, PLA2G15, PLTP, BCAM, GAS6, GLTPD2, HP, PCSK9, SORT1, TIMD4, PCSK7, UGT1A6, APOA4, LRP4, MSR1*	*ANGPTL3,APOA5, CD300LG, GALNT2, PLA2G15, GAS6, GLTPD2, HP, PCSK9, TIMD4, UGT1A6*	*ANGPTL3,APOA5, CD300LG, GALNT2, PLA2G15, PLTP, GAS6, GLTPD2, HP, PCSK9, TIMD4, UGT1A6, APOA4, MSR1*
Height	*ACE, ADAMTS5, AMBP, C1QL1, DPEP1, EFEMP1, FAM20A, GHR, IGF1R, MANF, MFAP2, MRC2, NAGLU, NOTUM, NPW, PDCD6, SERPINH1, SPATA20, STAT3*	$\lvert\beta\rvert > 0.004$	$P < 5.40\times10^{-4}$	$\lvert\beta\rvert > 0.007$	$P < 5.26\times10^{-4}$	*ACE, ADAMTS5, AMBP, DPEP1, EFEMP1, FAM20A, GHR, IGF1R, MFAP2, MRC2, PDCD6, SERPINH1, SPATA20, STAT3*	*EFEMP1, PDCD6*	*ADAMTS5, AMBP, DPEP1, EFEMP1, FAM20A, GHR, MFAP2, MRC2, PDCD6, SPATA20, STAT3*

European African

Figure 3



D

Trait(s)	Gene(s)	Beta	*P* value	Beta	*P* value	Beta	*P* value	GWAS (GWAS Catalog)	Colocalization (PPH4 > 0.8)	pQTL MR (FDR < 0.05)
Alzheimer's disease	*BCAM, CD55, EPHB4, GRN, LILRB1, SIRPA, TREM2*	$\lvert\beta\rvert > 0.005$	$P < 2.95\times10^{-4}$	$\lvert\beta\rvert > 0.005$	$P < 1.03\times10^{-4}$	$\lvert\beta\rvert > 0.002$	$P < 1.00\times10^{-4}$	*BCAM, CD55, EPHB4, GRN, TREM2*	*GRN, TREM2, SIRPA*	*BCAM, CD55, EPHB4, GRN, LILRB1, SIRPA, TREM2*
Post-traumatic stress disorder (PTSD)	*CTSF, CTSV, CD14*	$\beta > 0.036$	$P < 2.83\times10^{-4}$	$\beta > 0.022$	$P < 3.64\times10^{-4}$	$\beta > 0.014$	$P < 5.59\times10^{-4}$		*CTSF*	*CTSF, CTSV, CD14*
Coronary artery disease (and angiographic burden)	*IL6R, PCSK9, SPARCL1*	$\lvert\beta\rvert > 0.016$	$P < 1.09\times10^{-4}$	$\lvert\beta\rvert > 0.023$	$P < 2.56\times10^{-4}$	$\lvert\beta\rvert > 0.008$	$P < 4.15\times10^{-5}$	*IL6R, PCSK9*	*IL6R, PCSK9*	*IL6R, PCSK9*
Peripheral artery disease	*C2, MMP12*	$\lvert\beta\rvert > 0.006$	$P < 7.05\times10^{-8}$	$\lvert\beta\rvert > 0.012$	$P < 1.09\times10^{-8}$	$\lvert\beta\rvert > 0.007$	$P < 5.57\times10^{-4}$		*MMP12*	*C2, MMP12*
Venous thromboembolism (VTE)	*GP6, NPPB, OBP2B*	$\lvert\beta\rvert > 0.018$	$P < 1.01\times10^{-6}$	$\lvert\beta\rvert > 0.012$	$P < 9.28\times10^{-5}$	$\lvert\beta\rvert > 0.019$	$P < 9.85\times10^{-5}$		*GP6, NPPB*	*GP6, NPPB*
Varicose veins	*FABP2, RSPO3, TNFSF12*	$\lvert\beta\rvert > 0.026$	$P < 2.82\times10^{-4}$	$\lvert\beta\rvert > 0.008$	$P < 3.63\times10^{-5}$	$\lvert\beta\rvert > 0.002$	$P < 5.85\times10^{-5}$	*RSPO3*	*RSPO3*	*FABP2, RSPO3, TNFSF12*
Cardiovascular ideal health score (IHS)	*PCSK9, ENTPD6*	$\beta < -0.025$	$P < 5.16\times10^{-6}$	$\beta < -0.041$	$P < 3.42\times10^{-5}$	$\beta < -0.024$	$P < 1.55\times10^{-5}$		*PCSK9, ENTPD6*	*PCSK9, ENTPD6*
Non-alcoholic fatty liver disease (NAFLD)	*APOH, FCRLB, IL1RN, RSPO3, SPON1**	$\lvert\beta\rvert > 0.009$	$P < 2.68\times10^{-6}$	$\lvert\beta\rvert > 0.017$	$P < 9.48\times10^{-5}$	$\lvert\beta\rvert > 0.009$	$P < 1.08\times10^{-5}$	*RSPO3*	*APOH*	*APOH, FCRLB, IL1RN, RSPO3*
Type 2 diabetes (T2D)	*MLN, NCAN, NELL1, PAM, AGER, BST1*	$\lvert\beta\rvert > 0.005$	$P < 2.53\times10^{-4}$	$\lvert\beta\rvert > 0.004$	$P < 1.65\times10^{-4}$	$\lvert\beta\rvert > 0.002$	$P < 5.87\times10^{-4}$	*NELL1, PAM*	*NELL1, PAM*	*MLN, NCAN, NELL1, PAM, AGER, BST1*
Estimated glomerular filtration rate (eGFR)	*ANGPTL3, DPEP1, IDI2, NCAN, TIMP4*	$\lvert\beta\rvert > 0.012$	$P < 2.08\times10^{-4}$	$\lvert\beta\rvert > 0.028$	$P < 3.44\times10^{-5}$	$\lvert\beta\rvert > 0.003$	$P < 2.11\times10^{-4}$	*DPEP1, IDI2*	*ANGPTL3, IDI2*	*ANGPTL3, DPEP1, IDI2, NCAN, TIMP4*
Body mass index (BMI)	*SCG3, SIGLEC9*	$\beta < -0.006$	$P < 2.38\times10^{-4}$	$\beta < -0.006$	$P < 2.90\times10^{-4}$	$\beta < -0.003$	$P < 5.04\times10^{-4}$	*SCG3, SIGLEC9*		*SCG3, SIGLEC9*
Lipids	*CCL22, COL18A1, DPEP2, LILRA5, NUDT2, PNLIPRP2, RSPO3, IL1RN, CTSB, TNFSF12, PLTP#, ANGPTL3#, GAS6#, PCSK9#*	$\lvert\beta\rvert > 0.011$	$P < 4.31\times10^{-4}$	$\lvert\beta\rvert > 0.010$	$P < 3.96\times10^{-4}$	$\lvert\beta\rvert > 0.006$	$P < 5.62\times10^{-4}$	*CCL22, COL18A1, DPEP2, LILRA5, PLTP, PNLIPRP2, RSPO3, ANGPTL3, GAS6, PCSK9, IL1RN, CTSB, TNFSF12*	*COL18A1, PLTP, RSPO3, ANGPTL3, GAS6, PCSK9, IL1RN*	*CCL22, COL18A1, DPEP2, LILRA5, PLTP, PNLIPRP2, RSPO3, ANGPTL3, GAS6, PCSK9, IL1RN, CTSB, TNFSF12*
Height	*AMBP, BOC, CLPS, DPEP2, EFEMP1, EPHB4, FKBP1B, HYAL1, ISLR2, ITIH3, NCAN, QPCT, TNC, TNFRSF11B, DPEP1**, *ENG**	$\lvert\beta\rvert > 0.074$	$P < 4.62\times10^{-4}$	$\lvert\beta\rvert > 0.035$	$P < 1.70\times10^{-4}$	$\lvert\beta\rvert > 0.069$	$P < 1.95\times10^{-4}$	*AMBP, BOC, EFEMP1, ITIH3, NCAN, QPCT, TNC, TNFRSF11B, DPEP1, ENG*	*BOC, EFEMP1*	*AMBP, BOC, EFEMP1, ITIH3, NCAN, QPCT, TNFRSF11B*

UKB **ARIC** **deCODE** ***: identified by African models only** **#: identified by both European and African models**

Figure 4

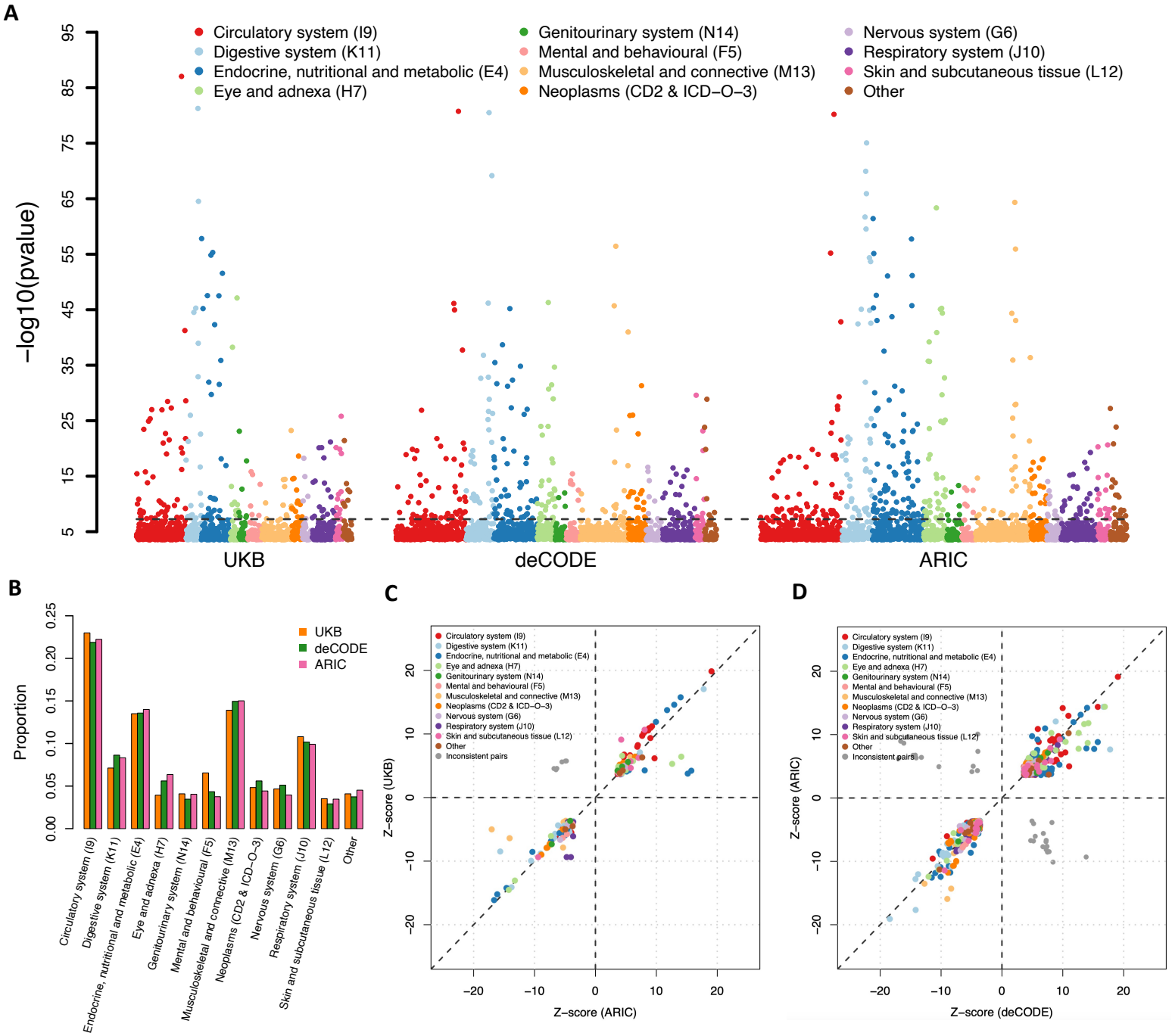

A
Circulatory system (I9)
Digestive system (K11)
Endocrine, nutritional and metabolic (E4)
Eye and adnexa (H7)
Genitourinary system (N14)
Mental and behavioural (F5)
Musculoskeletal and connective (M13)
Neoplasms (CD2 & ICD−O−3)
Nervous system (G6)
Respiratory system (J10)
Skin and subcutaneous tissue (L12)
Other
−log10(pvalue)
UKB
deCODE
ARIC
B
Proportion
UKB
deCODE
ARIC
C
Z−score (UKB)
Z−score (ARIC)
Inconsistent pairs
D
Z−score (ARIC)
Z−score (deCODE)
Inconsistent pairs

Figure 5

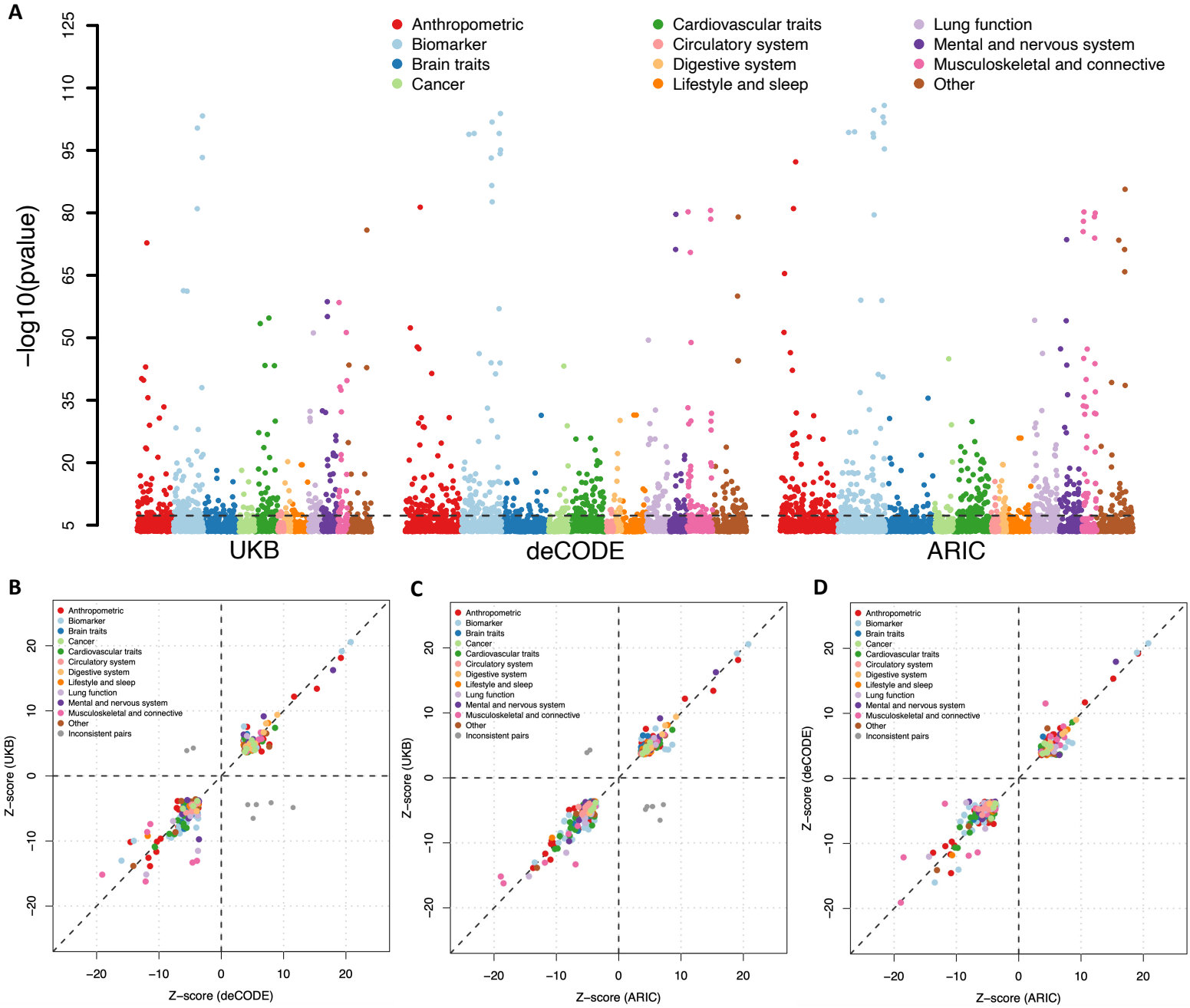

A
Anthropometric
Biomarker
Brain traits
Cancer
Cardiovascular traits
Circulatory system
Digestive system
Lifestyle and sleep
Lung function
Mental and nervous system
Musculoskeletal and connective
Other
−log10(pvalue)
UKB
deCODE
ARIC
B
C
D
Inconsistent pairs
Z−score (UKB)
Z−score (deCODE)
Z−score (ARIC)

Figure 6