Blood plasma proteome-wide association study implicates novel proteins in the pathogenesis of multiple cardiovascular diseases

Jia-Hao Wang[1#], Shan-Shan Dong[1#], Hao-An Wang[1], Shao-Shan Liu[1], Xiaoyi Ma[2], Ren-Jie Zhu[1], Wei Shi[1], Hao Wu[1], Ke Yu[1], Tian-Pei Zhang[1], Cong-Ru Wang[1], Yan Guo[1*], Tie-Lin Yang[1*]

[1]Key Laboratory of Biomedical Information Engineering of Ministry of Education, Key Laboratory of Biology Multiomics and Diseases in Shaanxi Province Higher Education Institutions, Biomedical Informatics & Genomics Center, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, P. R. China, 710049

[2]Department of Biostatistics, School of Public Health and Health Professions, The State University of New York at Buffalo

#These authors contributed equally to this study.

*Corresponding authors

Address Correspondence to:

Dr. Tie-Lin Yang and Dr. Yan Guo

E-mail: yangtielin@xjtu.edu.cn, guoyan253@xjtu.edu.cn

Abstract

Cardiovascular diseases (CVD) are the leading cause of global mortality, but current treatments are only effective in a subset of individuals. To identify new potential treatment targets, we present here the first PWAS for 26 CVDs using plasma proteomics data of the largest cohort to date (53,022 individuals from the UK Biobank Pharma Proteomics Project (UKB-PPP) project).

The GWAS summary data for 26 CVDs spanning 3 categories (16 cardiac diseases, 5 venous diseases, 5 cerebrovascular diseases, up to 1,308,460 individuals). We also conducted replication analyses leveraging two other independent human plasma proteomics datasets, encompassing 7,213 participants from the Atherosclerosis Risk in Communities (ARIC) study and 3,301 individuals from the INTERVAL study.

We identified 94 genes that are consistent with being causal in CVD, acting via their cis-regulated plasma protein abundance. 34 of 45 genes were replicated in at least one of the replication datasets. 41 of the 94 genes are novel genes not implicated in original GWAS. 91.48% (86/94) proteins are category-specific, only two proteins (ABO, PROCR) were associated with diseases in all three CVD categories. Longitudinal analysis revealed that 37 proteins exhibit stable expression in plasma. In addition, PBMC scRNA-seq data analysis showed that 23 of the 94 genes were stably expressed in CD14+ monocytes, implicating their potential utility as biomarkers for CVD disease status. Drug repurposing analyses showed that 39 drugs targeting 23

genes for treating diseases from other systems might be considered in further research.

In conclusion, our findings provide new insights into the pathogenic mechanisms of CVD and offering promising targets for further mechanistic and therapeutic studies.

Keywords: CVD; Human blood plasma proteomes; Causal proteins; PWAS

Introduction

Cardiovascular diseases (CVD) are a group of disorders of the heart and blood vessels. As the leading global cause of mortality [1, 2], CVD took an estimated 17.9 million lives in 2019, accounting for 32% of all deaths worldwide[3]. In clinical practice, there are some commonly used treatments for CVD, such as the use of statins to reduce cardiovascular morbidity and mortality. However, current treatments are not suitable for all patients and have certain risks of side effects[4]. Therefore, there is an urgent and critical need for new therapeutic targets for CVD[5].

Proteins, as the final products of gene expression, are the main functional components of biological processes[6]. In addition, most therapeutic agents target proteins[7]. Therefore, understanding their relationship with diseases is crucial for effective treatments[8-11]. However, direct observational studies linking protein abundance to phenotypes can be confounded or represent reverse causation[12]. Proteome-wide association study (PWAS) is a powerful strategy to solve this problem. It uses single-nucleotide polymorphisms (SNPs) to genetically impute proteins and relate them to genome wide association study (GWAS) summary statistics of a trait to provide evidence of causality[13, 14]. The genetic models are restricted to the cis-region of the protein, reducing the risk of confounding by horizontal pleiotropy (independent of the protein). Further summary data-based Mendelian randomization (SMR)[15] or colocalization analyses[16] can be used to identify genes contribute to disease pathogenesis through modulating protein abundance. This integrative analytical

approach has been employed to identify novel potential therapeutic targets for neurological disorders[13, 14, 17-19] using brain proteomics data. However, PWAS for CVD is still limited.

Using plasma proteomics data of the largest cohort to date (53,022 individuals from the UK Biobank Pharma Proteomics Project (UKB-PPP) project[20]), we present here the first PWAS for multiple CVDs. The study design is shown in Figure 1. We applied the above analytic approaches to the discovery dataset consisting of human plasma proteomic and genetic data from UKB-PPP[20] and the GWAS of 26 CVDs spanning 3 categories (16 cardiac diseases, N = 234,829~1,030,836; 5 venous diseases, N = 388,830~484,598; 5 cerebrovascular diseases, N = 484,598~1,308,460)[21-26]. Additionally, we conducted replication analyses leveraging two independent human plasma proteomics datasets[27, 28], encompassing 7,213 participants from the Atherosclerosis Risk in Communities (ARIC) study and 3,301 individuals from the INTERVAL study, to ensure robustness and reproducibility. For functional interpretation of the identified proteins, enrichment analyses were performed to detect the pathways associated with CVD. Longitudinal stability analysis at plasma and cell-type level was used to assess the expression stability of the proteins. We also pharmacologically annotate the proteins of interest with approved drugs to assess their feasibility as treatment targets.

Results

Discovery PWAS of multiple CVDs

We generated the human blood plasma proteome model based on 53,022 UKB-PPP[20] participants. The initial proteomes data include total 2,923 proteins. After quality control, 1,715 proteins with significant SNP-based heritability ($P < 0.05$, $h^2 > 0$), were used for PWAS. The correlation R^2 between the model's predictive power and heritability for each gene was 0.88 (Supplementary Figure S1), supporting the accuracy of our protein estimation model.

The plasma proteome model results were integrated with the 26 CVD GWAS data using the FUSION pipeline[29]. Detail information of the 26 GWAS datasets is shown in Supplementary Table S1. We performed genetic correlation analysis and the results showed that diseases belong to the same category are usually with higher correlation (Supplementary Figure S2).

As shown in Supplementary Table S2, we identified 341 significant protein-CVD pairs of associations after multiple testing corrections ($P < 2.92\times10^{-5}$, 0.05/1,715 proteins) (Fig 2A). Among these associations, 87 genes are located within 1 Mb of each other. With the goal of identifying multiple independent associations, we performed conditional analyses using a regression with summary statistics approach [27]. 27 pairs of associations no longer significant associations were removed (Supplementary Table S3). Finally, we obtained 314 independent and significant PWAS association signals, including 155 unique proteins associated with CVD. The

number of associated genes for each phenotype is shown in Figure 2B. Taking heart failure as an example, PWAS identified 48 proteins associated with it (Fig 2B). 18 genes from 7 loci were subjected to conditional analysis. Six genes (*SORT1, SHISA5, PDE5A, PGF, FURIN, DDT*) were no longer significant after conditional analysis and were removed from subsequent analyses (Fig 2C). Finally, we obtained 42 genes associated with heart failure.

Replication PWAS using the proteomics data from ARIC and INTERVAL

For the 155 proteins identified from the discovery dataset, 75 were detected in the at least one of the replication proteomic datasets. The number of detected proteins in the ARIC and INTERVAL project was 64 and 50, respectively. For these proteins, we incorporated their previously built human plasma protein models from the FUSION website[29] and the CVD GWAS datasets to perform replication PWAS.

In the ARIC dataset, the results showed that 46 proteins were associated with CVD ($P < 7.81 \times 10^{-4}$, 0.05/64). As for the INTERVAL dataset, the number of successfully replicated proteins was 37 ($P < 1.00 \times 10^{-3}$, 0.05/50). As shown in supplementary Table S4, the significant association of 55 proteins (73.33%, 55/75) were replicated in at least one dataset with the same effect direction as the discovery PWAS, and 23 proteins were replicated in both datasets.

Causal-analysis of the proteins identified by PWAS

We employed two independent but supplementary approaches (SMR and colocalization) to further evaluate the causality of the 155 proteins[15, 16] from the UKB-PPP dataset. The SMR and its accompanying heterogeneity in dependent instruments (HEIDI) test was used to test whether PWAS-significant genes were associated with CVD via their cis-regulated protein abundance. The SMR results showed that the cis-regulated protein abundance mediates the association between genetic variants and CVD for 125 unique proteins. However, HEIDI results argued against a causal role for 55 genes due to linkage disequilibrium (Supplementary Table S5). Therefore, 70 unique proteins have evidence consistent with a causal role in CVD by SMR/HEIDI.

The colocalization test was used to examine the posterior probability for a shared causal variant between a pQTL and CVD GWAS for the PWAS-significant genes. The colocalization analysis identified 74 proteins with shared causal variant between pQTL and CVD GWAS (posterior probability PPH4 ≥ 0.7). We kept proteins with evidence from either SMR or colocalization analysis, and finally a total of 94 proteins were remained for subsequent analysis (Fig 3A, Supplementary Table S5).

Combining evidence for replication and results of causality tests, 45 of 94 causal proteins were detected in the at least one of replication proteomic datasets. 75.56% (34/45) proteins were found to be with evidence of both replication and causality (Table 1). There were 28 proteins replicated in the ARIC dataset and 21 proteins replicated in the INTERVAL dataset. In particular, 15 proteins were replicated in both

datasets. For example, ABO, F10, IL6R and PROC.

Common proteins associated with diseases in three CVD categories

91.49% (86/94) proteins were identified in only one disease category (Fig 3B). Only 2 proteins (ABO, PROCR) were associated with diseases in all three CVD categories. As shown in Figure 3A, ABO showed significant positive associations with multiple diseases from the three categories. PROCR was negatively associated with 2 cardiac diseases and 1 cerebrovascular disease, and positively associated with 2 venous diseases. The inconsistency in association direction might because PROCR is linked to anti-inflammatory and anticoagulant functions[30].

Novelty of the CVD causal genes

To assess the novelty of the 94 potentially causal genes, we checked the lowest p-values for the SNPs within 2 Mb window of these genes using the summary statistics from the CVD GWAS. We found that 41 genes were not located within 2 Mb of a significant GWAS signal ($P < 5 \times 10^{-8}$), suggesting that these 41 genes are novel genes not implicated in the original GWAS (Fig 3C). 25 of the novel genes were have not been detected in other CVD GWAS either. For example, PROC was found to be associated with venous thromboembolism (Fig 3D top) and COMT was found to be associated with three cardiac diseases (hypertension, statin medication and angina pectoris; Supplementary Figure S3). All 26 CVD GWAS data didn't detect their association with CVD diseases. The rest 16 novel genes were not implicated in the

original GWAS but have been detected in the GWAS of other CVD. For example, our PWAS results showed that FN1 was associated with coronary heart disease and coronary atherosclerosis (Fig3D bottom). The GWAS of these two diseases didn't detect the association signal of this gene, but it was found to be associated with heart failure, myocardial infarction, coronary revascularization, and coronary artery bypass grafting in their GWAS data. Our results further expand the important role of FN1 in multiple cardiac diseases.

Gene ontology enrichment analysis

To further elucidate the molecular mechanisms underlying the 94 identified proteins, we carried out a non-redundant gene ontology (GO) biological processes enrichment analysis using WebGestalt 2024[31, 32]. The results (Fig 4A) showed that genes associated with cardiac disease enriched in 12 pathways. 58.33% (7/12) of these pathways belong to three categories (immunity/inflammation, lipid-related process, and vessel/blood-related process). Genes associated with venous diseases were found to be significantly enriched in 5 biological pathways, and three of them belong to the vessel/blood-related process, particularly the coagulation process. No significant pathway was detected for the genes associated with cerebrovascular disease due to the limited number of genes.

Protein-protein interactions (PPI) network analysis

To investigate the connectivity for the 94 proteins, we performed network-based

analysis using STRING[33]. The minimum required interaction score was 0.4. We constructed a PPI network with 30 nodes and 37 edges, primarily comprising 3 protein communities. (Fig 4B). The proteins associated with cardiac disease are mainly within the network with FN1and APOE as the core proteins. Consistent with pathway enrichment analysis results, these proteins are mostly in the community of immunity/inflammation and lipid-related process. The network of proteins associated with venous disease is mainly driven by 6 proteins (F2, F10, F11, PROC, PROCR and PROS1) involved in blood/vessel-related process, especially in the coagulation processes. The network for venous disease proteins is distinct from that of cardiac disease proteins, and the two networks are connected by F2 and PROS1. Interactions among the proteins associated with cerebrovascular diseases are relatively sparse. Complete information about communities is presented in Supplementary Table S7.

Mouse phenotypic annotation of potential causal genes.

We further evaluated whether 94 proteins were associated with CVD-related phenotype in mouse using the Mouse Genome Informatics (MGI) database[34]. We performed phenotype enrichment analysis using the Fisher's exact test. Consistent with the pathway enrichment and network analysis results, mutations in genes associated with cardiac diseases are enriched in phenotypes related to immunity/inflammation, lipid-related process, and vessel/blood-related process (Fig 4C). Mutations in genes associated with venous disease are enriched in phenotypes related to vessel/blood-related process (Fig 4C). These results further support the

involvement of the identified proteins in CVD pathogenesis.

Evaluate longitudinal stability at protein and single-cell level

To evaluate the expression stability of the 94 proteins, we performed longitudinal analysis using data using plasma proteomics data and peripheral blood mononuclear cells (PBMC) single cell RNA-seq (scRNA-seq) data from GEO dataset GSE190992.

The plasma proteomics data were collected from 6 healthy, non-smoking Caucasian donors over a 10-week period. 44 of the 94 proteins were detected in this dataset. Among them, 84.09% (37/44) proteins exhibit stable expression in plasma (median coefficient of variation < 10%, Fig4 D left). Fluctuations in the plasma levels of these proteins might serve as potential markers of disease status.

The PBMC scRNA-seq data were collected weekly from four donors over the course of six weeks. We found 24 genes exhibited stable expression in at least one cell type (median coefficient of variation < 10% in at least one cell type across all donors. Fig4 D right). Notably, 23 of the 24 genes stably expressed in CD14+ monocytes. As per previous studies, monocytes play a crucial role in both local ischemia and inflammatory responses, which are closely linked to the development of cardiovascular diseases[35-37].

Cell-type specific expression of the CVD causal genes

To investigate whether these genes show distinct enrich across different cell types, we utilized PBMC RNA-seq data obtained from the plasma of another 11 healthy donors to examine the specific expression patterns of these genes. Among 94 CVD causal genes, 39 were enriched in one or more cell types (FDR adjust $P < 0.05$ and logFC > 1.5, Supplementary Table 9), include CD4 T cells, CD14+ monocytes, Platelet, Natural killer cell and other monocyte. A total of 21 genes were highly expressed in CD14+ monocytes, and half of these genes (11 out of 21) were also found to be stably expressed in CD14+ monocytes through stability analysis.

Drug repurposing analyses identified potential therapeutic targets for CVD

To investigate the potential drug target genes, we construct a gene-drug-disease network (Fig4 D). The results showed that 25 of the 94 proteins are the targets of 53 drugs with completed or currently undergoing clinical trials (Supplementary Table 10). 14 drugs have already been used for treating circulatory system disorders. For example, Drotrecogin alfa targeting *F2*, *PROCR* and *PROS1* are currently one of the efficacious treatments for managing cerebrovascular ischemic events[38]. The rest 39 drugs for treating diseases from other systems might be considered in further drug repurposing research. For example, Menadione targeting PROC are currently used for treating vitamin K deficiency and prostate cancer.

Discussion

In this study, we integrated data from 26 CVD GWAS along with three large-scale human plasma protein datasets to conduct a comprehensive PWAS analysis. Collectively, we identified 186 significant independent protein-disease association pairs, involving 94 unique proteins associated with CVD. Among these proteins, 41 proteins are novel proteins not implicated in original GWAS. We also elucidated potential biological mechanisms underlying CVD and provided potential new targets for CVD drug development.

The PWAS analysis identified 96 genes that are consistent with being causal in CVD, including 41 novel genes not implicated in original GWAS. For example, PROC was newly found to be associated with venous thromboembolism. PROC is a vitamin K-dependent enzyme that plays a crucial role in regulating human thrombosis and hemostasis[39]. Consist with our results, previous studies have demonstrated that reduced PROC levels in plasma can be used as a marker of increased risk of venous thrombosis[40, 41]. In the PPI network, we also demonstrated that PROC, together with coagulation factors such as F2 and F10, forms a venous-related network. Longitudinal stability analysis showed that this protein is stably expressed in blood plasma. Currently, two drugs (Menadione and Cupric Chloride) targeting PROC have passed clinical trials for the treatment of conditions such as fungal infections, prostate cancer, and vitamin K deficiency[42]. Further studies are needed to explore the potential of these drugs for treating venous thromboembolism.

91.49% (86/94) of the identified proteins are category-specific, suggesting that the underlying pathogenesis mechanisms of the three disease categories are different. Only two proteins (ABO and PROCR) were associated with diseases in all three CVD categories. ABO was found to be positively associated with multiple cardiovascular diseases. Consistently, epidemiological studies have reported that ABO is associated with a wide range of diseases, including cardiovascular ailments, malignancies, and infectious conditions[43,44]. PROCR is positively associated with venous disorders but negatively associated with stroke and coronary artery disease. PROCR is a receptor for activated protein C, which is a serine protease activated by and involved in the blood coagulation pathway. Consistent with our results, GWAS studies have shown that the minor G allele of rs867186 at this gene is correlated with a higher risk of venous thromboembolism[45, 46] but a lower risk of CAD[47, 48]. A previous study[30] has shown that PROCR linked to CAD through anti-inflammatory mechanisms and to VTE through pro-thrombotic mechanisms.

The longitudinal stability analysis showed that 37 of the 44 detected proteins (84.09%) exhibit stable expression in plasma, suggesting that they might serve as potential markers of disease status. In addition, PBMC scRNA data analysis identified 24 genes exhibited stable expression in at least one cell type and 23 of the 24 genes stably expressed in CD14+ monocytes, highlighting the important role of CD14+ monocytes in CVD development. Consistently, previous studies have associated increased frequency of the CD14+ monocytes clinical CVD events and plaque vulnerability[49, 50].

Monocyte density of CD14 was found to be higher in patients with moderate severe heart failure in comparison with normal or mild LV impairment [51, 52]. These results suggested that CD14+ monocytes might be used as markers for CVD.

Our study has several limitations. First, since the current available proteomics and GWAS are mainly derived from European populations, our results are mainly applicable to the European population. Second, we focused on cis-regulatory elements when constructing models to assess protein influences. This is a common choice for current researchers, because the current sample size of proteomics may not be sufficient to detect the trans effect. With larger scale data available in future, models considering both cis and trans effects can be constructed.

In summary, using the largest available proteomics data from UKB-PPP projects (a total of 1,715 inheritable proteins from 53,202 individuals), we performed a PWAS study for 26 CVDs. We identified 94 genes that contribute to CVD pathogenesis through modulating their plasma protein abundance. These genes may serve as potential targets for future mechanistic and therapeutic studies aimed at finding effective treatments for CVD.

Methods

Human plasma proteomic and genetic data in UKB.

We generated the human blood plasma proteome models from 53,022 participants of

European ancestry of the UKB-PPP. The sample was selected in two batches from Consortium members and UK Biobank cohort and the proteomic profiling was performed using standard Olink proteomics pipeline using Proximity Extension Assay [20]. antibodies matched to unique complementary oligonucleotides, which were bound to their respective target proteins, underwent quantification through next-generation sequencing. Following rigorous quality control measures, the normalized protein expression (NPX) values were computed using the Inter-Plate Control method. This NPX score effectively served as a quantitative measure of protein abundance in our samples.

Genotype data matching the protein dataset underwent genotyping, imputation, and quality control steps as detailed in previous work [53]. This included sex discrepancy, sex chromosome aneuploidy, and heterozygosity checks, with imputed variants filtered for INFO scores >0.7. All chromosomal positions were updated to the hg38 assembly using LiftOver [54]. Genotyping quality control was executed using PLINK2.0 software[55]. Participants exhibiting over 5% missing genotypic data were removed from consideration. Moreover, variants displaying deviations from the Hardy-Weinberg equilibrium (with p-values less than $1x10^{-8}$), a genotype missing rate exceeding 5%, a minor allele frequency below 1%, or those not classified as SNPs, were also excluded from the analysis.

Following the preprocessing of both genotype and protein datasets, we adopted the FUSION software to train the proteome model and we only consider the subset comprising 1,190,321 SNPs from the HapMap3 project [56]. SNPs situated up to 500 kb

away from either end of genes were defined as cis-SNPs. The model further incorporated adjustments for protein expression based on gender and age as covariates to refine the association analysis and account for potential confounding variables.

CVD GWAS summary association statistics

Our GWAS data mainly comes from the GWAS catalog [22 25 21, 23, 24, 26] and FinnGen [57] database. In accordance with the ICD-10 standard of circulatory disorders, we selected GWAS studies involving a minimum of 5,000 cases. When multiple studies of the same condition were identified, we opted for those with the largest sample sizes. This stringent selection procedure resulted in a final cohort of 26 unique GWAS for our investigation. Based on the distinct pathophysiological mechanisms, we categorized the diseases into cardiac diseases, venous disease, and cerebrovascular disease.

Statistical approach

Proteome-wide association studies (PWAS). We used the standard processes in the FUSION software[29] to construct protein models and incorporate GWAS data for our PWAS analysis. After applying the previously outlined quality control measures to screen the sample and genotype data, we utilized GCTA software[58] to estimate the SNP-based heritability for individual proteins. To expedite calculations, a random subset of approximately 10,000 individuals was selected from the full cohort for each protein's heritability estimation. From the analysis of 2,923 proteins, 1,715 displayed

statistically significant heritability ($h^2 > 0$, $p < 0.05$). We then employed the FUSION software to estimate the impact of SNPs on protein abundance using multiple predictive models (top1, lasso, enet) [29] and select the most predictive model as the final predictor. Finally, we obtained a total of 1,715 distinct protein models encoded by different genes and we applied the Bonferroni correction for multiple testing. Consequently, proteins with a P-value threshold of 2.92×10^{-5} (0.05/1,715) were deemed statistically significant in our discovery PWAS analysis.

We then performed the replication PWAS analysis in two other publicly available data sets. The modeling methodologies for these datasets have been documented in prior research, the Atherosclerosis Risk in Communities (ARIC) study dataset included 4,483 protein measurements from 7,213 European participants [27], while the INTERVAL study retained information on 3,170 proteins for 3,301 individuals [28]. Subsequent to the heritability filtering phase, an ensemble of 2,379 proteins (1,348 in ARIC and 1,031 in INTERVAL) was selected for incorporation into our replication verification.

Causal analysis. We adopted two independent frameworks to rigorously ascertain the causal inference of the proteins implicated in our PWAS findings. For the Bayesian colocalization analysis [16], we employed the COLOC module embedded within the FUSION software suite. The COLOC tool operates by estimating the posterior probability indicating that the same causal variant underlies both GWAS and pQTL.

Within the colocalization analysis framework, a comprehensive set of five hypotheses (H0 through H4) are scrutinized. Notably, hypothesis H4 posits the existence of a SNP that acts as a shared causal driver for both pQTL and GWAS. In our study, we defined causality for proteins identified through the COLOC analysis as those exhibiting a posterior probability for Hypothesis H4 exceeding 0.7. We subsequently employed the SMR [15] approach to further validate the causal relationships inferred from the PWAS and GWAS. For this SMR analysis, we leveraged recently published pQTL data [20], which were derived from UKB-PPP study, complemented by independently obtained GWAS data [21-24, 26, 57] on cardiovascular disease, which were also considered in our PWAS. Our determination of significant causal relationships relied on an adjusted P <0.05 for the SMR analysis and the unadjusted P>0.05 from the HEIDI test.

PPI and GO enrichment

For the investigation of causal genes implicated in three diseases, we employed the STRING [33] database to perform an extensive network analysis. The Markov cluster (MCL) algorithm was used with the following parameters: inflation parameter—1.5. Subsequently, the derived network was refined and visually optimized utilizing Cytoscape [59] software. In this visualization, node size corresponds to the degree of connectivity for each gene, indicative of its interaction frequency within the network, while distinct colors denote different gene categories, facilitating categorical distinction and interpretation. Additionally, we conducted functional enrichment analysis for causal genes pertinent in three categories diseases using the WebGestalt [32]

online platform, focusing on the GO BP pathways. We select the pathways with $P < 0.05$ (with FDR adjusted) and number of overlap genes > 3 as the significant result.

Longitudinal Data Stability Analysis

We conducted a longitudinal analysis utilizing the data set GSE190992 from the Gene Expression Omnibus (GEO) database [60]. Specific details regarding the data collection methodology and information have been reported in detail in a previous publication. The data encompass proteomics measurements over a 10-week period for 6 healthy donors, as well as single-cell data collected over a six-week period for 4 of these donors. For each donor, we calculated the coefficient of variation (CV) for each gene at both the proteome and single-cell levels as a measure of stability (CV = standard deviation / mean × 100%). We selected thresholds of 10% as criteria for stable gene expression in proteome data and single-cell data, respectively. These genes that exhibit stable expression in plasma, along with the associated cell types, can be considered as more reliable biomarkers for early screening and prediction of CVD.

Cell-type specific expression of the CVD causal genes

We utilized scRNA-seq data from 11 healthy donors sourced from the GEO database (GSE244515). During the preprocessing phase, we filtered out cells that expressed less than 200 genes or had a mitochondrial gene content exceeding 15%. For cell type annotation, we normalized the count matrices using the LogNormalize method with a scaling factor of 10,000, which also helped in identifying variable features. To align

datasets from different samples and mitigate batch effects, we applied the Harmony integration method. These procedures were carried out using Seurat package version 4.4.0 within the R environment. After quality control and normalization, the data comprised a total of 27,484 genes across 371,086 cells. For the 94 CVD causal genes, we used the Wilcoxon rank sum test to compare the expression levels between the cells of interest and other cells. We applied FDR correction to the P values derived from the multiple tests, with the total number of tests set to 27,484 genes. Finally, we retained the significant results of FDR *P* value < 0.05 and logFC > 1.5, and thought that the expression was specific expression in cells.

Mouse genome informatics and Drug analysis

MGI database [34] serves as a global repository for murine research, offering a comprehensive integration of genetic, genomic, and biological information. This platform fosters investigations into human health and disease by facilitating insights garnered from mouse models and we demonstrated many of the gene deletion mouse models exhibit phenotypes associated with circulatory system disease. We enriched the mouse phenotype using the Fisher's exact test method, and retained phenotypes with more than three overlap genes, P < 0.05 (with FDR adjusted) and OR > 1. Furthermore, we constructed a gene-drug-disease interaction network by integrating gene-drug associations from the DrugBank [61] database and drug-disease relationships from the Therapeutic Target Database (TTD) [62]. Our network focused exclusively on drugs with approved clinical efficacy and excluding those with discontinued

development at any stage.

Code availability

All software and datasets in our study are publicly available online.

Acknowledgments

This research has been conducted using the UK biobank resource under application number 46387.

Funding

This work was supported by the National Natural Science Foundation of China (32370653, and 82372458); Innovation Capability Support Program of Shaanxi Province (2022TD-44); Key Research and Development Project of Shaanxi Province (2022GXLH-01-22), and the Fundamental Research Funds for the Central Universities. This study was also supported by the High-Performance Computing Platform and Instrument Analysis Center of Xi'an Jiaotong University.

Data and Resource Availability

The GWAS summary data of CVD we used in this article were available from GWAS catalog (https://www.ebi.ac.uk/gwas/downloads/summary-statistics) and FinnGen study (https://finngen.gitbook.io/documentation/data-download). Download links for all datasets are provided in Table S1.

Author Contributions

J.-H.W. and S.-S.D. designed this project. J.-H.W., A.-H.W., and C.-R.W. conducted the computational work. J.-H.W. wrote the manuscript. S.-S.D. and A-H.W. revised the manuscript. J.-H.W., H-A.W. and S.-S.D. summarized the tables and figures. R.-J.Z., W.S., H.W., K.Y, T.-P.Z., X.Y. M. and S-S.L. collected the public data. Y.G. and T.-L.Y. supported and supervised this project.

Ethical approval and consent to participate

All datasets were publicly available, and ethical approval and informed consent were acquired for all original studies.

Competing interests

The authors declare that they have no conflict of interest.

References

1. Townsend N, Kazakiewicz D, Lucy Wright F, et al. Epidemiology of cardiovascular disease in Europe. *Nat Rev Cardiol*. Feb 2022;19(2):133-143. doi:10.1038/s41569-021-00607-3

2. Collaborators GBDCoD. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. Nov 10 2018;392(10159):1736-1788. doi:10.1016/S0140-6736(18)32203-7

3. Holmes MV, Richardson TG, Ference BA, Davies NM, Davey Smith G. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat Rev Cardiol*. Jun 2021;18(6):435-453. doi:10.1038/s41569-020-00493-1

4. Ward NC, Watts GF, Eckel RH. Statin Toxicity. *Circ Res*. Jan 18 2019;124(2):328-350. doi:10.1161/CIRCRESAHA.118.312782

5. Li Y, Li Z, Ren Y, et al. Mitochondrial-derived peptides in cardiovascular disease: Novel insights and therapeutic opportunities. *J Adv Res*. Nov 24 2023;doi:10.1016/j.jare.2023.11.018

6. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*. 2012;13(4):227-232. doi:10.1038/nrg3185

7. Yao P, Iona A, Pozarickij A, et al. Proteomic Analyses in Diverse Populations Improved Risk Prediction and Identified New Drug Targets for Type 2 Diabetes. *Diabetes Care*. Jun 1 2024;47(6):1012-1019. doi:10.2337/dc23-2145

8. Lindsey ML, Mayr M, Gomes AV, et al. Transformative Impact of Proteomics on Cardiovascular Health and Disease. *Circulation*. 2015;132(9):852-872. doi:10.1161/cir.0000000000000226

9. Wik L, Nordberg N, Broberg J, et al. Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol Cell Proteomics*. 2021;20:100168. doi:10.1016/j.mcpro.2021.100168

10. Haslam DE, Li J, Dillon ST, et al. Stability and reproducibility of proteomic profiles in epidemiological studies: comparing the Olink and SOMAscan platforms. *Proteomics*. Jul 2022;22(13-14):e2100170. doi:10.1002/pmic.202100170

11. Wang R-S, Maron BA, Loscalzo J. Multiomics Network Medicine Approaches to Precision Medicine and Therapeutics in Cardiovascular Diseases. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2023;43(4):493-503. doi:10.1161/atvbaha.122.318731

12. Brandes N, Linial N, Linial M. PWAS: proteome-wide association study-linking genes and phenotypes by functional variation in proteins. *Genome Biol*. Jul 14 2020;21(1):173. doi:10.1186/s13059-020-02089-x

13. Wingo AP, Liu Y, Gerasimov ES, et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nat Genet*. Feb 2021;53(2):143-146. doi:10.1038/s41588-020-00773-z

14. Wingo TS, Liu Y, Gerasimov ES, et al. Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat Neurosci*. Jun 2021;24(6):810-817. doi:10.1038/s41593-021-00832-6

15. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies

predicts complex trait gene targets. *Nat Genet*. May 2016;48(5):481-7. doi:10.1038/ng.3538

16. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. May 2014;10(5):e1004383. doi:10.1371/journal.pgen.1004383

17. Li SJ, Shi JJ, Mao CY, et al. Identifying causal genes for migraine by integrating the proteome and transcriptome. *J Headache Pain*. Aug 17 2023;24(1):111. doi:10.1186/s10194-023-01649-3

18. Phillips B, Western D, Wang L, et al. Proteome wide association studies of LRRK2 variants identify novel causal and druggable proteins for Parkinson's disease. *NPJ Parkinsons Dis*. Jul 8 2023;9(1):107. doi:10.1038/s41531-023-00555-4

19. Wu BS, Chen SF, Huang SY, et al. Identifying causal genes for stroke via integrating the proteome and transcriptome from brain and blood. *J Transl Med*. Apr 21 2022;20(1):181. doi:10.1186/s12967-022-03377-9

20. Sun BB, Chiou J, Traylor M, et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*. Oct 2023;622(7982):329-338. doi:10.1038/s41586-023-06592-6

21. Nielsen JB, Thorolfsdottir RB, Fritsche LG, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet*. Sep 2018;50(9):1234-1239. doi:10.1038/s41588-018-0171-3

22. Shah S, Henry A, Roselli C, et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun*. Jan 9 2020;11(1):163. doi:10.1038/s41467-019-13690-5

23. Donertas HM, Fabian DK, Valenzuela MF, Partridge L, Thornton JM. Common genetic associations between age-related diseases. *Nat Aging*. Apr 2021;1(4):400-412. doi:10.1038/s43587-021-00051-5

24. Hartiala JA, Han Y, Jia Q, et al. Genome-wide analysis identifies novel susceptibility loci for myocardial infarction. *Eur Heart J*. Mar 1 2021;42(9):919-933. doi:10.1093/eurheartj/ehaa1040

25. Mishra A, Malik R, Hachiya T, et al. Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature*. Nov 2022;611(7934):115-123. doi:10.1038/s41586-022-05165-3

26. Sollis E, Mosaku A, Abid A, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res*. Jan 6 2023;51(D1):D977-D985. doi:10.1093/nar/gkac1010

27. Zhang J, Dutta D, Kottgen A, et al. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat Genet*. May 2022;54(5):593-602. doi:10.1038/s41588-022-01051-w

28. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature*. Jun 2018;558(7708):73-79. doi:10.1038/s41586-018-0175-2

29. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. Mar 2016;48(3):245-52. doi:10.1038/ng.3506

30. Stacey D, Chen L, Stanczyk PJ, et al. Elucidating mechanisms of genetic cross-disease associations at the PROCR vascular disease locus. *Nat Commun*. Mar 9 2022;13(1):1222. doi:10.1038/s41467-022-28729-3

31. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science*. 2021;31(1):8-22. doi:10.1002/pro.4218

32. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. Jul 2 2019;47(W1):W199-W205. doi:10.1093/nar/gkz401

33. Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*. Jan 6 2023;51(D1):D638-D646. doi:10.1093/nar/gkac1000

34. Blake JA, Baldarelli R, Kadin JA, et al. Mouse Genome Database (MGD): Knowledgebase for mouse–human comparative biology. *Nucleic Acids Research*. 2021;49(D1):D981-D987. doi:10.1093/nar/gkaa1083

35. Jaipersad AS, Lip GY, Silverman S, Shantsila E. The role of monocytes in angiogenesis and atherosclerosis. *J Am Coll Cardiol*. Jan 7-14 2014;63(1):1-11. doi:10.1016/j.jacc.2013.09.019

36. Tahir S, Steffens S. Nonclassical monocytes in cardiovascular physiology and disease. *Am J Physiol Cell Physiol*. May 1 2021;320(5):C761-C770. doi:10.1152/ajpcell.00326.2020

37. Ruder AV, Wetzels SMW, Temmerman L, Biessen EAL, Goossens P. Monocyte heterogeneity in cardiovascular disease. *Cardiovasc Res*. Sep 5 2023;119(11):2033-2045. doi:10.1093/cvr/cvad069

38. Southan C, Sharman JL, Benson HE, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein

targets and 6000 ligands. *Nucleic Acids Res*. Jan 4 2016;44(D1):D1054-68. doi:10.1093/nar/gkv1037

39. Dinarvand P, Moser KA. Protein C Deficiency. *Arch Pathol Lab Med*. Oct 2019;143(10):1281-1285. doi:10.5858/arpa.2017-0403-RS

40. Tang W, Stimson MR, Basu S, et al. Burden of rare exome sequence variants in PROC gene is associated with venous thromboembolism: a population-based study. *J Thromb Haemost*. Feb 2020;18(2):445-453. doi:10.1111/jth.14676

41. Manderstedt E, Lind-Hallden C, Hallden C, et al. Classic Thrombophilias and Thrombotic Risk Among Middle-Aged and Older Adults: A Population-Based Cohort Study. *J Am Heart Assoc*. Feb 15 2022;11(4):e023018. doi:10.1161/JAHA.121.023018

42. Kovács KB, Pataki I, Bárdos H, et al. Molecular characterization of p. Asp77Gly and the novel p. Ala163Val and p. Ala163Glu mutations causing protein C deficiency. *Thrombosis research*. 2015;135(4):718-726.

43. Wu O, Bayoumi N, Vickers MA, Clark P. ABO(H) blood groups and vascular disease: a systematic review and meta-analysis. *J Thromb Haemost*. Jan 2008;6(1):62-9. doi:10.1111/j.1538-7836.2007.02818.x

44. Li S, Schooling CM. A phenome-wide association study of ABO blood groups. *BMC Med*. Nov 17 2020;18(1):334. doi:10.1186/s12916-020-01795-4

45. Dennis J, Johnson CY, Adediran AS, et al. The endothelial protein C receptor (PROCR) Ser219Gly variant and risk of common thrombotic disorders: a HuGE review and meta-analysis of evidence from observational studies. *Blood*. Mar 8 2012;119(10):2392-400. doi:10.1182/blood-2011-10-383448

46. Medina P, Navarro S, Bonet E, et al. Functional analysis of two haplotypes of the human endothelial protein C receptor gene. *Arterioscler Thromb Vasc Biol*. Mar 2014;34(3):684-90. doi:10.1161/ATVBAHA.113.302518

47. Howson JMM, Zhao W, Barnes DR, et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat Genet*. Jul 2017;49(7):1113-1119. doi:10.1038/ng.3874

48. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res*. Feb 2 2018;122(3):433-443. doi:10.1161/CIRCRESAHA.117.312086

49. Kashiwagi M, Imanishi T, Tsujioka H, et al. Association of monocyte subsets with vulnerability characteristics of coronary plaques as assessed by 64-slice multidetector computed tomography in patients with stable angina pectoris. *Atherosclerosis*. Sep 2010;212(1):171-6. doi:10.1016/j.atherosclerosis.2010.05.004

50. Tapp LD, Shantsila E, Wrigley BJ, Pamukcu B, Lip GY. The CD14++CD16+ monocyte subset and monocyte-platelet interactions in patients with ST-elevation myocardial infarction. *J Thromb Haemost*. Jul 2012;10(7):1231-41. doi:10.1111/j.1538-7836.2011.04603.x

51. Anker SD, Egerer Kr Fau - Volk HD, Volk Hd Fau - Kox WJ, Kox Wj Fau - Poole-Wilson PA, Poole-Wilson Pa Fau - Coats AJ, Coats AJ. Elevated soluble CD14 receptors and altered cytokines in chronic heart failure. (0002-9149 (Print))

52. Niebauer J, Volk HD, Kemp M, et al. Endotoxin and immune activation in chronic heart failure: a prospective cohort study. *Lancet*. May 29 1999;353(9167):1838-42. doi:10.1016/S0140-6736(98)09286-1

53. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. Oct 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

54. Hinrichs AS, Karolchik D, Baertsch R, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. Jan 1 2006;34(Database issue):D590-8. doi:10.1093/nar/gkj144

55. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7. doi:10.1186/s13742-015-0047-8

56. International HapMap C, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. Sep 2 2010;467(7311):52-8. doi:10.1038/nature09298

57. Kurki MI, Karjalainen J, Palta P, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. Jan 2023;613(7944):508-518. doi:10.1038/s41586-022-05473-8

58. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. Jan 7 2011;88(1):76-82. doi:10.1016/j.ajhg.2010.11.011

59. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. Nov 2003;13(11):2498-504. doi:10.1101/gr.1239303

60. Vasaikar SV, Savage AK, Gong Q, et al. A comprehensive platform for analyzing longitudinal multi-omics data. *Nat Commun*. Mar 27 2023;14(1):1684. doi:10.1038/s41467-

023-37432-w

61. Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res*. Jan 5 2024;52(D1):D1265-D1275. doi:10.1093/nar/gkad976

62. Zhou Y, Zhang Y, Zhao D, et al. TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res*. Jan 5 2024;52(D1):D1465-D1477. doi:10.1093/nar/gkad751

Figure Legends

Figure 1. Workflow of the current study. We collected proteomics data from three different sources: UKB, ARIC and INTERVAL. GWAS summary data for 26 CVDs spanning 3 categories (16 cardiac diseases, 5 venous diseases, 5 cerebrovascular diseases, up to 1,308,460 individuals) were included. We performed PWAS with proteomics data from the three projects followed by Mendelian randomization and colocalization analysis. Functional annotation of the genes identified by PWAS was finally performed.

Figure 2. Result of the PWAS

A. Manhattan plot for the PWAS of CVD. Each dot represents the correlation between a disease and a gene, with the x-axis indicating genomic location and the y-axis showing $-\log_{10}(P)$. The gray horizontal line represents the Bonferroni-corrected significant threshold, $P < 2.92\times10^{-5}$. The significant results of the three categories diseases are shown in red, green, blue, respectively. The labeled genes are the most significant results on each chromosome

B. The number of significant genes in PWAS for 26 CVD diseases. Different colors represent different disease categories. 27 jointly significant genes dropped by conditional analysis (gray).

C. Regional association of PWAS hits in conditional analysis for heart failure. Conditionally significant proteins are CELSR2, GSTM1, SPINK8, DAG1, HYAL1, FABP2, ACYP1, FES, GSTT2B and SUSD2. Top panel in each plot highlights the

marginally associated PWAS genes (blue) and the jointly significant genes (green). Bottom panel shows a regional Manhattan plot of the data before (grey) and after (blue) conditioning on the predicted expression of the green genes. Chr: chromosome.

Figure 3. Results for the causal genes.

A. The heatmap presents whole PWAS results for 94 genes passing causality tests and color depth reflects the association direction and magnitude. Genes identified in replication PWAS are represented by circles in the heatmap. Causal genes are labeled "*" and the novel gene with no significant variant ($P < 5 \times 10^{-8}$) within ±2M window of the gene range in GWAS results are labeled in red.

B. The Venn diagram illustrates the overlap of causal genes across three disease categories.

C. The number of novel genes in different diseases.

D. The top Manhattan plot represents the pQTL and the GWAS results within the *PROC* genomic region for venous thromboembolism. The bottom Manhattan plot represents the pQTL and the GWAS results within the *FN1* genomic region for coronary heart disease and coronary atherosclerosis.

Figure 4. Function annotation of the identified genes

A. The significant enriched Gene Ontology (GO) biological process (BP) terms of the causal genes in different categories. The color of the bar represents the biological function category to which the pathway belongs. Immunity/inflammation (light coral),

lipid-related process (faint yellow), vessel/blood-related process (purple) and other (gray).

B. The network constructed with identified causal genes. Lines represent a physical interaction, and line thickness is proportional to the interaction score. Genes associated with cardiac disease, venous disease and cerebrovascular disease are shown in red, green, and blue, respectively. Genes with more connections are shown with larger size. Community 1 include 16 proteins associated with immunity/inflammation. Community 2 include 6 proteins associated with lipid-related process. Community 3 include 6 proteins associated with vessel/blood-related process, especially the formation of fibrin clot.

C. The significant enrichment results of mouse phenotypes of the causal genes in different categories. The color of the bar represents the biological function category to which the pathway belongs. Immunity/inflammation (light coral), lipid-related process (faint yellow), vessel/blood-related process, (purple) and other (gray).

D. Results of the longitudinal stability analysis at the protein level (left) and single-cell level (right). At the protein level, genes are classified as stable (blue) or variable (red) based on a coefficient of variation (CV) threshold of 10%. Among the 94 causal genes, 37 genes were identified as stable. The color blocks on the left indicate the relevant grouping of the genes in the PWAS results. At the single-cell level, the threshold is set at 10%, with gray representing samples with low average expression (average expression < 0.01 after normalization). 24 genes exhibit stable expression across 19 cell types. Different donors are indicated by different colors.

E. The constructed gene-drug-disease network of causal genes. The colors of the lines in the network signify the category of genes. ICD: International Classification of Diseases.

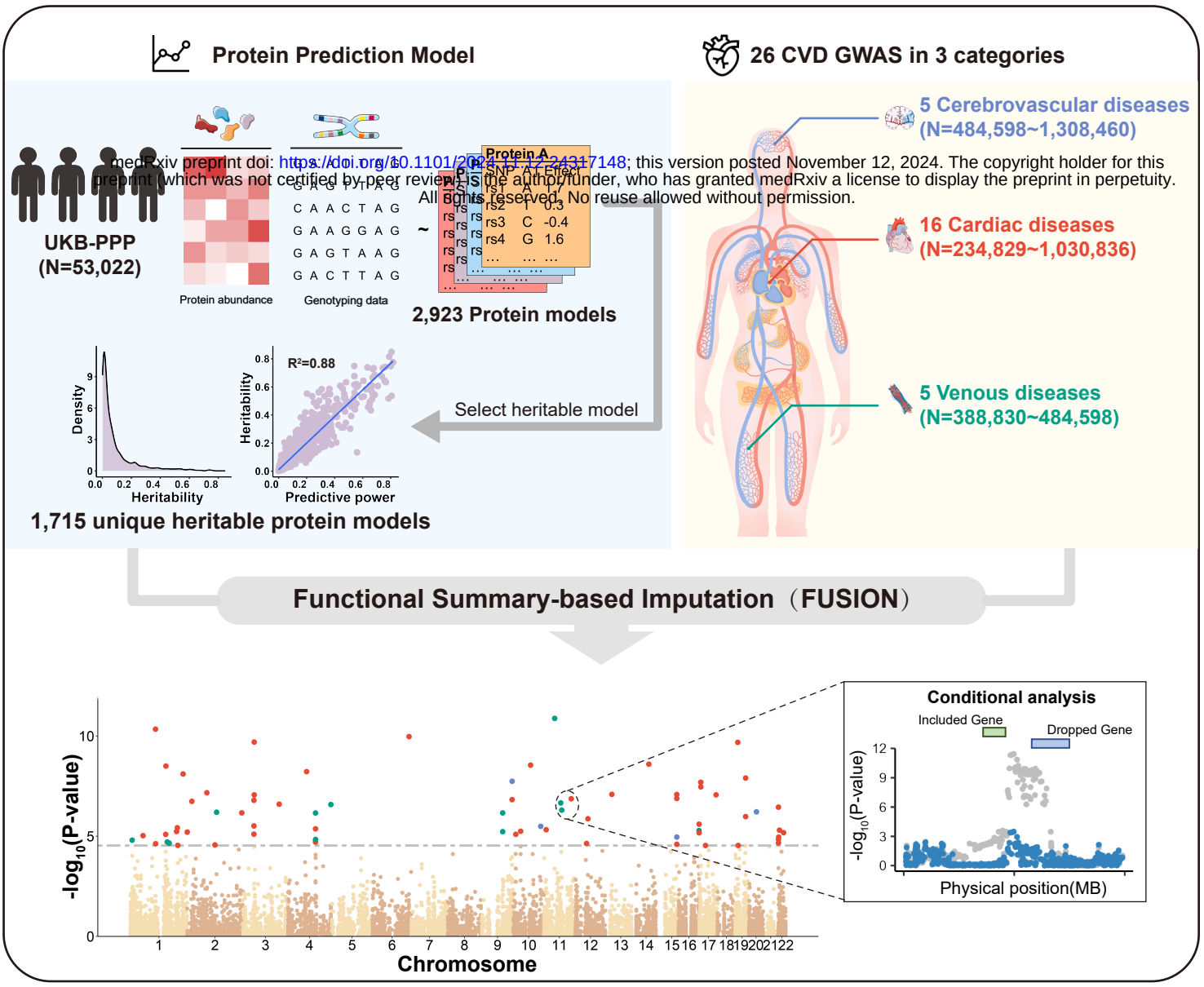
762 Table1. Summary of the 34 replicable CVD causal genes.

Gene	replicated	causal	PHE
ABO	Both	Both	Pulmonary embolism
	Both	COLOC	Deep Venous Thrombosis, Venous Thromboembolism
	INTERVAL	Both	Large artery stroke
	INTERVAL	COLOC	Myocardial infarction, Any stroke, Cardioembolic stroke
ANGPTL3	ARIC	COLOC	Statin medication
ASPN	Both	Both	Venous Thromboembolism
CCDC134	Both	Both	Atrial fibrillation
CD4	INTERVAL	SMR	Coronary atherosclerosis
COL15A1	Both	SMR	Atrial fibrillation
COL6A3	Both	Both	Heart failure
CTSB	INTERVAL	COLOC	Calcific aortic valvular stenosis
DLK1	INTERVAL	COLOC	Diseases of veins
DUSP13	INTERVAL	Both	Atrial fibrillation
ECM1	ARIC	COLOC	Hypertension
EPHA2	Both	Both	Statin medication
F10	Both	Both	Pulmonary embolism
	Both	COLOC	Venous Thromboembolism
F11	ARIC	Both	Deep Venous Thrombosis, Diseases of veins, Any stroke, Cardioembolic stroke
	ARIC	COLOC	Venous Thromboembolism, Pulmonary embolism
F2	Both	Both	Venous Thromboembolism, Pulmonary embolism
	Both	SMR	Deep Venous Thrombosis, Diseases of veins
FABP2	ARIC	SMR	Hypertension, Heart failure, Coronary revascularization
FN1	ARIC	Both	Heart failure, Coronary Heart Disease, Angina pectoris
	ARIC	COLOC	Myocardial infarction, Coronary atherosclerosis, Coronary revascularization, Coronary artery bypass grafting
GAS6	ARIC	COLOC	Statin medication
GSTT2B	INTERVAL	SMR	Heart failure
IL1RN	Both	SMR	Myocardial infarction
IL6R	Both	Both	Aortic aneurysm, Coronary Heart Disease, Coronary atherosclerosis, Angina pectoris, Coronary angioplasty
	Both	COLOC	Coronary revascularization, Coronary artery bypass grafting
	Both	SMR	Atrial fibrillation, Heart failure
INHBB	ARIC	Both	Coronary artery bypass grafting
	Both	COLOC	Statin medication
INHBC	ARIC	Both	Heart failure
ITIH3	ARIC	SMR	Heart failure
KLB	ARIC	SMR	Statin medication
LRIG1	ARIC	Both	Atrial fibrillation
MMP12	Both	Both	Any stroke
	Both	COLOC	Large artery stroke
NRP1	Both	SMR	Coronary Heart Disease
NUDT5	ARIC	SMR	Statin medication, Heart failure
PCSK9	ARIC	Both	Heart failure
	ARIC	COLOC	Statin medication, Myocardial infarction, Coronary Heart Disease, Coronary atherosclerosis, Angina pectoris, Coronary revascularization, Coronary angioplasty, Coronary artery bypass grafting, Valvular heart disease, Diseases of arteries and

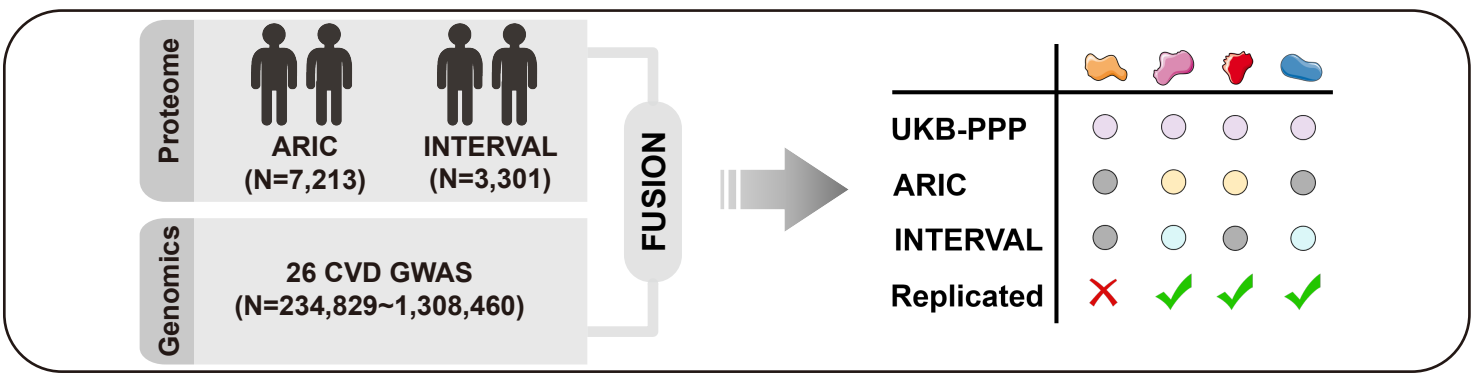
			capillaries
	ARIC	SMR	Calcific aortic valvular stenosis
PROC	Both	Both	Venous Thromboembolism
THBS2	Both	COLOC	Diseases of veins, Varicose veins
	Both	SMR	Aortic aneurysm
TNFSF12	INTERVAL	Both	Atrial fibrillation
UROD	ARIC	Both	Diseases of veins

763

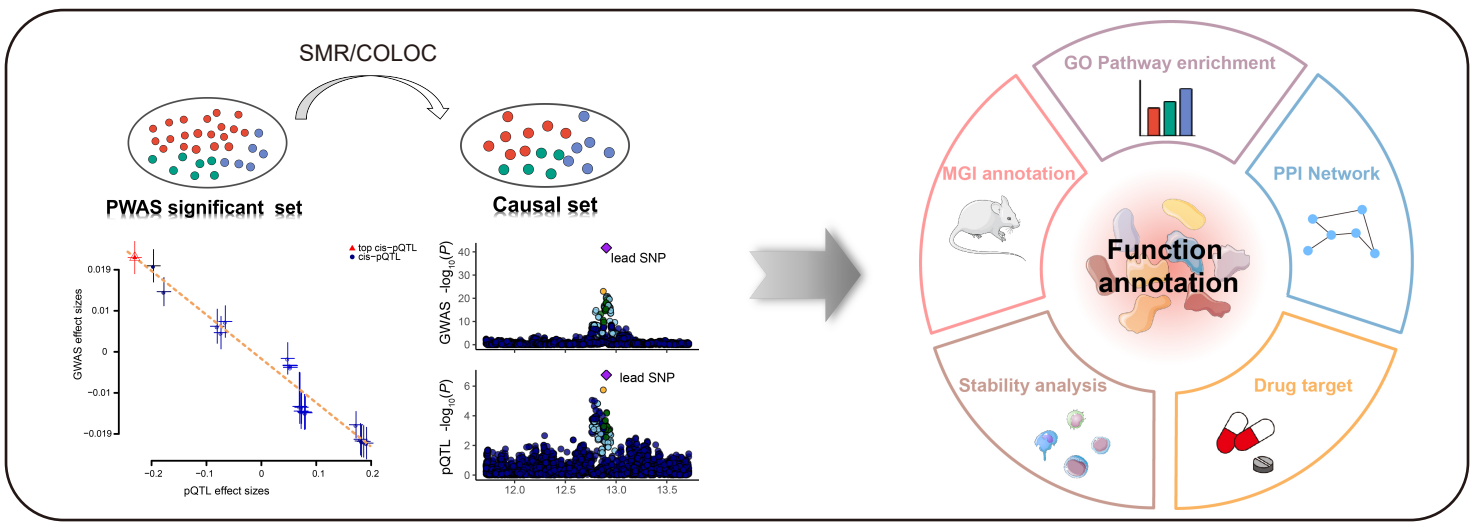
1. Discovery plasma proteome-wide association study (PWAS) analysis



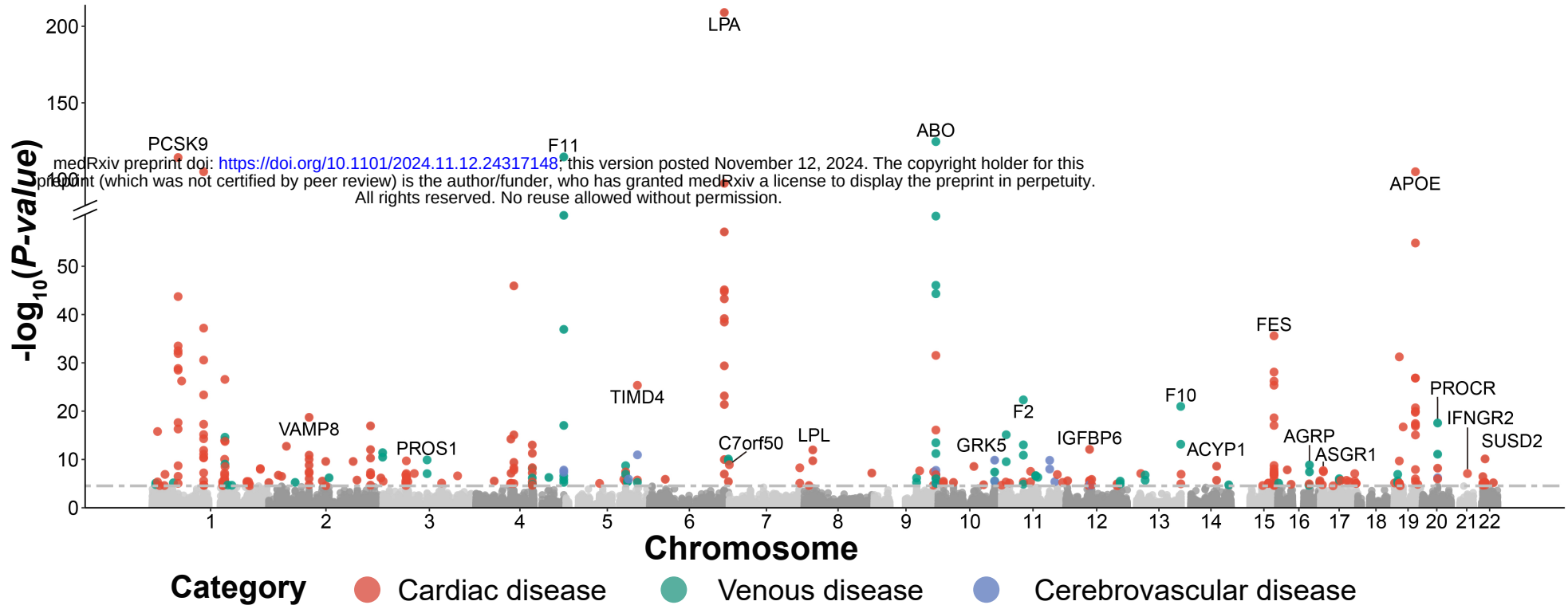
2. Replication PWAS of CVD



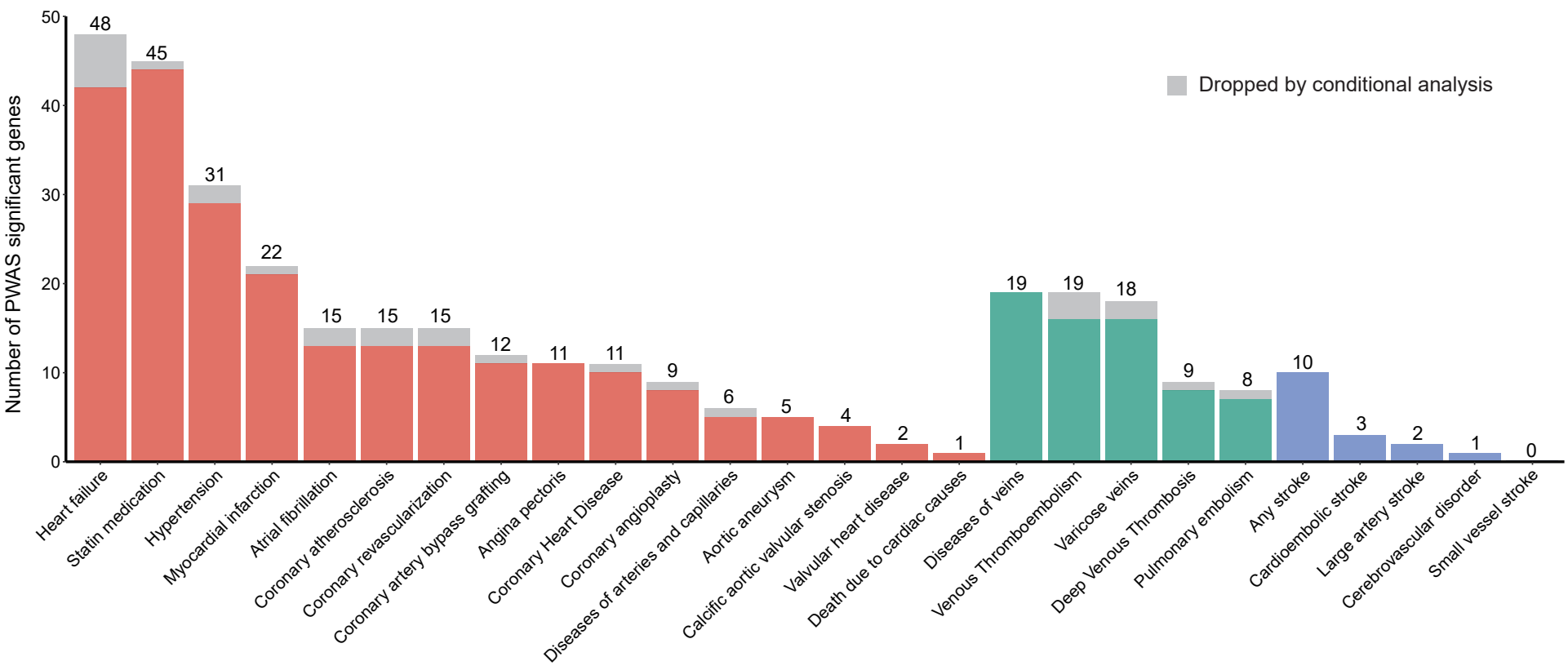
3. Function verification of candidate causal genes



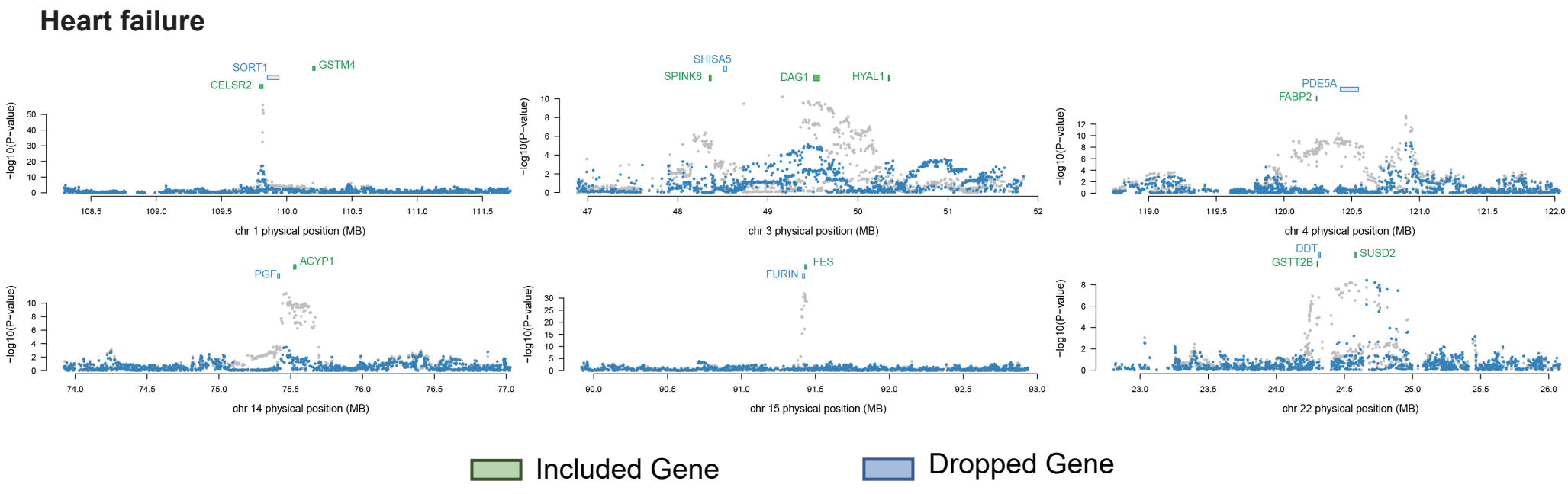
A

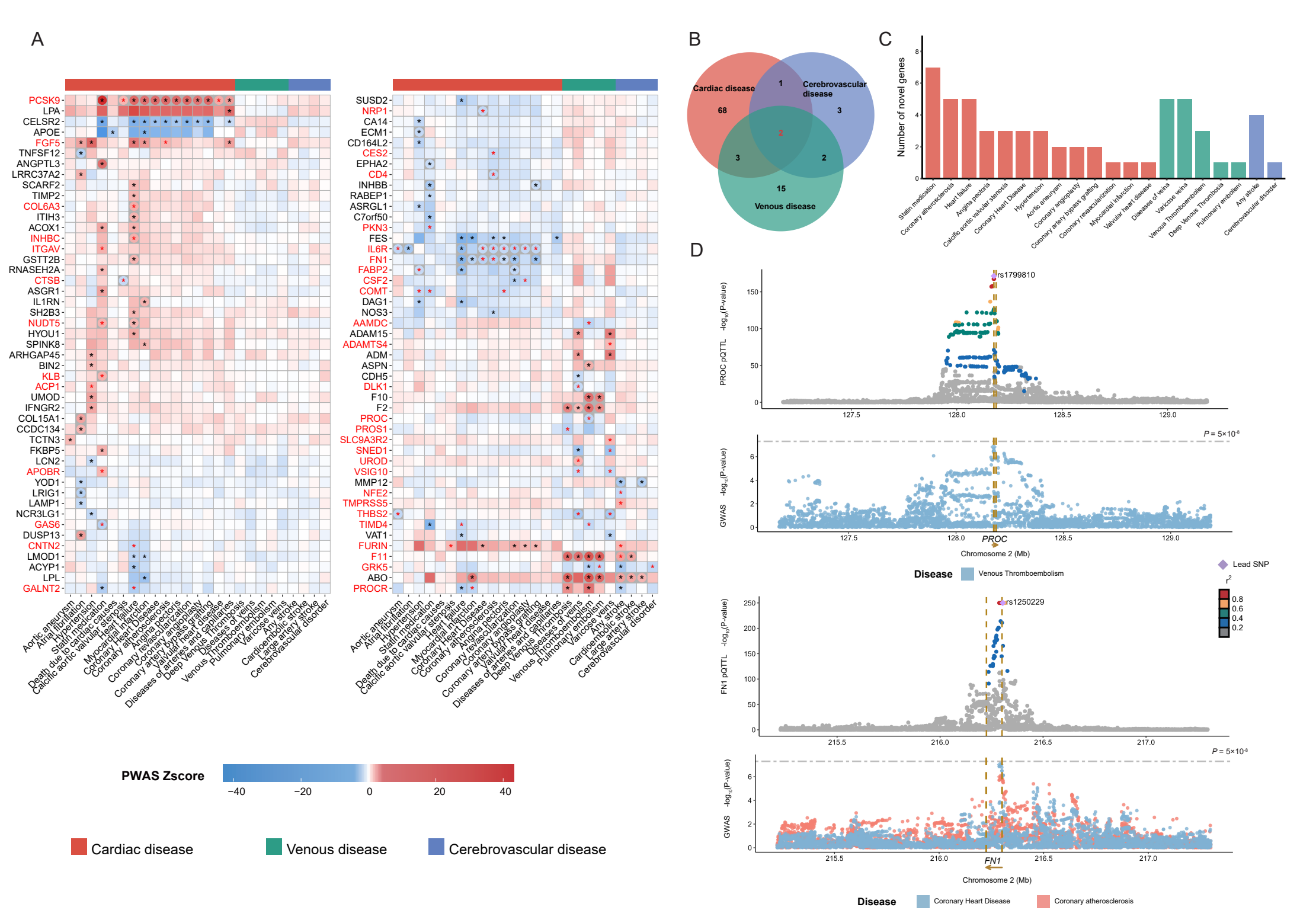


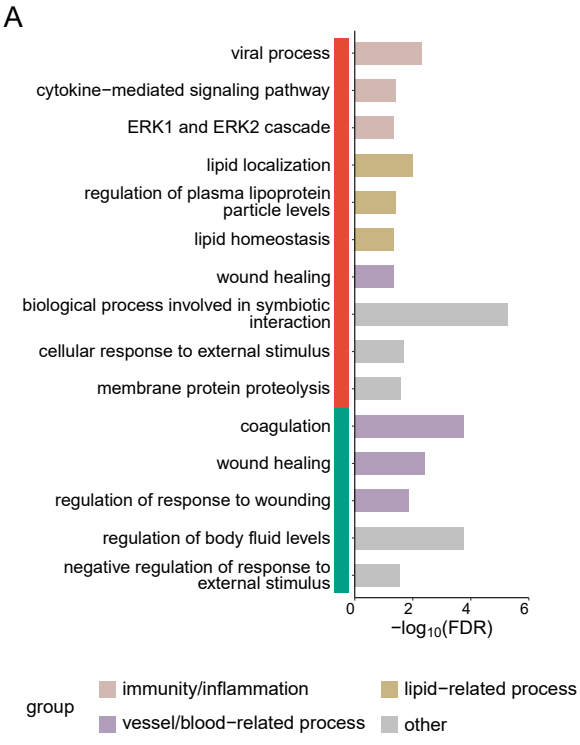
B

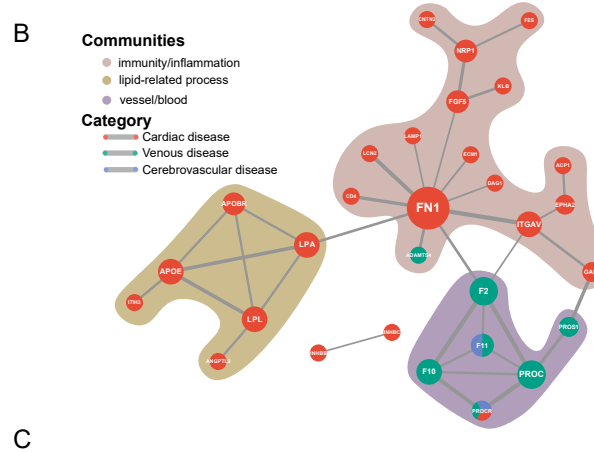


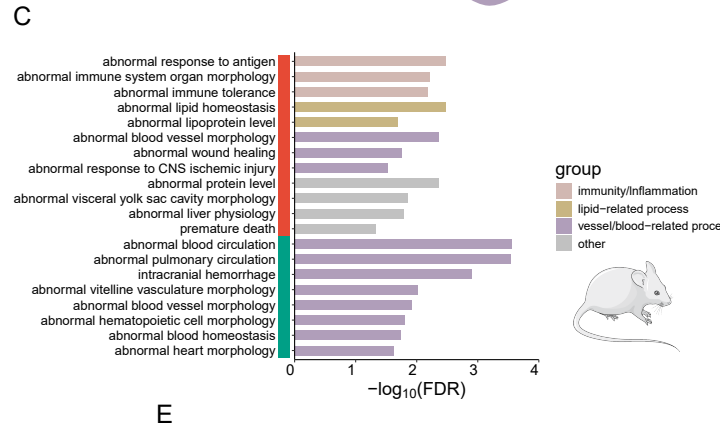
C

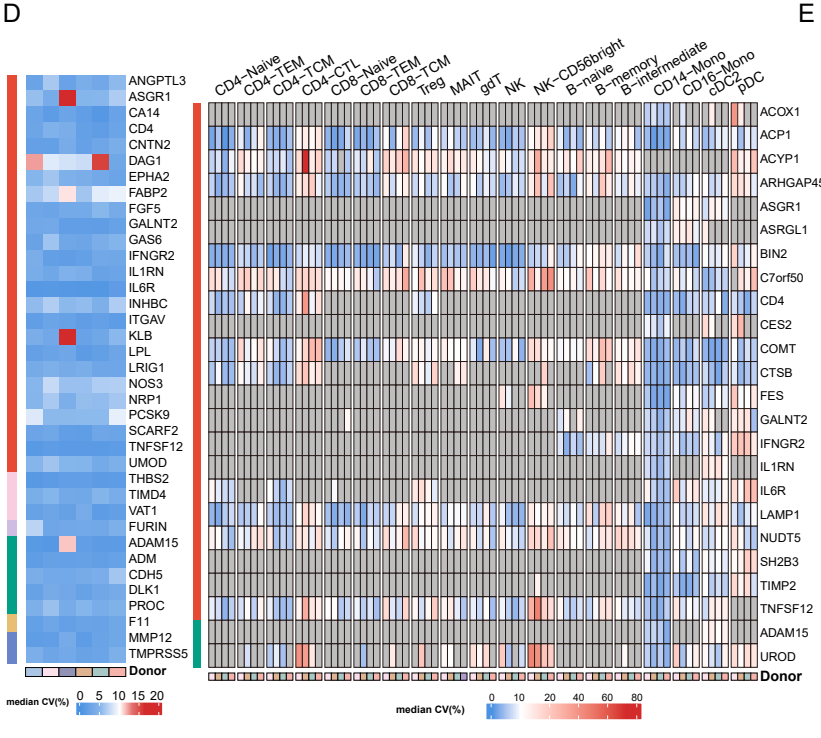


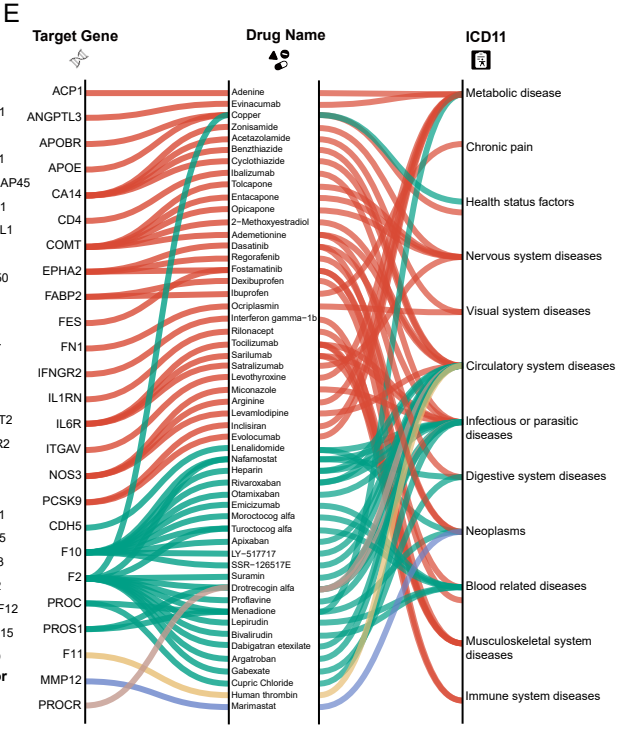

A
PCSK9
LPA
CELSR2
APOE
FGF5
TNFSF12
ANGPTL3
LRRC37A2
SCARF2
TIMP2
COL6A3
ITIH3
ACOX1
INHBC
ITGAV
GSTT2B
RNASEH2A
CTSB
ASGR1
IL1RN
SH2B3
NUDT5
HYOU1
SPINK8
ARHGAP45
BIN2
KLB
ACP1
UMOD
IFNGR2
COL15A1
CCDC134
TCTN3
FKBP5
LCN2
APOBR
YOD1
LRIG1
LAMP1
NCR3LG1
GAS6
DUSP13
CNTN2
LMOD1
ACYP1
LPL
GALNT2
SUSD2
NRP1
CA14
ECM1
CD164L2
CES2
EPHA2
CD4
INHBB
RABEP1
ASRGL1
C7orf50
PKN3
FES
IL6R
FN1
FABP2
CSF2
COMT
DAG1
NOS3
AAMDC
ADAM15
ADAMTS4
ADM
ASPN
CDH5
DLK1
F10
F2
PROC
PROS1
SLC9A3R2
SNED1
UROD
VSIG10
MMP12
NFE2
TMPRSS5
THBS2
TIMD4
VAT1
FURIN
F11
GRK5
ABO
PROCR
Aortic aneurysm
Atrial fibrillation
Hypertension
Statin medication
Death due to cardiac causes
Calcific aortic valvular stenosis
Myocardial infarction
Coronary Heart Disease
Coronary atherosclerosis
Angina pectoris
Coronary revascularization
Coronary artery bypass grafting
Valvular heart disease
Coronary angioplasty
Diseases of arteries and capillaries
Deep Venous Thrombosis
Diseases of veins
Venous Thromboembolism
Pulmonary embolism
Varicose veins
Any stroke
Cardioembolic stroke
Large artery stroke
Cerebrovascular disorder
PWAS Zscore
−40
−20
0
20
40
Cardiac disease
Venous disease
Cerebrovascular disease
B
Cardiac disease
Cerebrovascular disease
Venous disease
68
1
3
2
3
2
15
C
Number of novel genes
Statin medication
Coronary atherosclerosis
Heart failure
Angina pectoris
Calcific aortic valvular stenosis
Coronary Heart Disease
Hypertension
Aortic aneurysm
Coronary angioplasty
Coronary artery bypass grafting
Coronary revascularization
Myocardial infarction
Valvular heart disease
Diseases of veins
Varicose veins
Venous Thromboembolism
Deep Venous Thrombosis
Pulmonary embolism
Any stroke
Cerebrovascular disorder
D
rs1799810
PROC pQTTL -log10(P-value)
GWAS -log10(P-value)
P = 5×10-8
PROC
Chromosome 2 (Mb)
Disease
Venous Thromboembolism
Lead SNP
r2
0.8
0.6
0.4
0.2
rs1250229
FN1 pQTTL -log10(P-value)
FN1
Coronary Heart Disease
Coronary atherosclerosis


A
viral process
cytokine−mediated signaling pathway
ERK1 and ERK2 cascade
lipid localization
regulation of plasma lipoprotein particle levels
lipid homeostasis
wound healing
biological process involved in symbiotic interaction
cellular response to external stimulus
membrane protein proteolysis
coagulation
wound healing
regulation of response to wounding
regulation of body fluid levels
negative regulation of response to external stimulus
0
2
4
6
−log10(FDR)
group
immunity/inflammation
lipid−related process
vessel/blood−related process
other


B
Communities
immunity/inflammation
lipid-related process
vessel/blood
Category
Cardiac disease
Venous disease
Cerebrovascular disease
APOBR
APOE
LPA
LPL
ITIH3
ANGPTL3
FN1
NRP1
FGF5
KLB
LAMP1
LCN2
ECM1
CD4
DAG1
ACP1
EPHA2
ITGAV
GAS6
F2
PROS1
F11
F10
PROC
PROCR
INHBC
INHBA


C
abnormal response to antigen
abnormal immune system organ morphology
abnormal immune tolerance
abnormal lipid homeostasis
abnormal lipoprotein level
abnormal blood vessel morphology
abnormal wound healing
abnormal response to CNS ischemic injury
abnormal protein level
abnormal visceral yolk sac cavity morphology
abnormal liver physiology
premature death
abnormal blood circulation
abnormal pulmonary circulation
intracranial hemorrhage
abnormal vitelline vasculature morphology
abnormal blood vessel morphology
abnormal hematopoietic cell morphology
abnormal blood homeostasis
abnormal heart morphology
0
1
2
3
4
−log10(FDR)
group
immunity/Inflammation
lipid−related process
vessel/blood−related process
other


D
CD4−Naive
CD4−TEM
CD4−TCM
CD4−CTL
CD8−Naive
CD8−TEM
CD8−TCM
Treg
MAIT
gdT
NK
NK−CD56bright
B−naive
B−memory
B−intermediate
CD14−Mono
CD16−Mono
cDC2
pDC
ANGPTL3
ASGR1
CA14
CD4
CNTN2
DAG1
EPHA2
FABP2
FGF5
GALNT2
GAS6
IFNGR2
IL1RN
IL6R
INHBC
ITGAV
KLB
LPL
LRIG1
NOS3
NRP1
PCSK9
SCARF2
TNFSF12
UMOD
THBS2
TIMD4
VAT1
FURIN
ADAM15
ADM
CDH5
DLK1
PROC
F11
MMP12
TMPRSS5
Donor
median CV(%)
0
5
10
15
20
ACOX1
ACP1
ACYP1
ARHGAP45
ASGR1
ASRGL1
BIN2
C7orf50
CD4
CES2
COMT
CTSB
FES
GALNT2
IFNGR2
IL1RN
IL6R
LAMP1
NUDT5
SH2B3
TIMP2
TNFSF12
ADAM15
UROD
Donor
median CV(%)
0
10
20
40
60
80


E
Target Gene
Drug Name
ICD11
ACP1
ANGPTL3
APOBR
APOE
CA14
CD4
COMT
EPHA2
FABP2
FES
FN1
IFNGR2
IL1RN
IL6R
ITGAV
NOS3
PCSK9
CDH5
F10
F2
PROC
PROS1
F11
MMP12
PROCR
Adenine
Evinacumab
Copper
Zonisamide
Acetazolamide
Benzthiazide
Cyclothiazide
Ibalizumab
Tolcapone
Entacapone
Opicapone
2−Methoxyestradiol
Ademetionine
Dasatinib
Regorafenib
Fostamatinib
Dexibuprofen
Ibuprofen
Ocriplasmin
Interferon gamma−1b
Rilonacept
Tocilizumab
Sarilumab
Satralizumab
Levothyroxine
Miconazole
Arginine
Levamlodipine
Inclisiran
Evolocumab
Lenalidomide
Nafamostat
Heparin
Rivaroxaban
Otamixaban
Emicizumab
Moroctocog alfa
Turoctocog alfa
Apixaban
LY−517717
SSR−126517E
Suramin
Drotrecogin alfa
Proflavine
Menadione
Lepirudin
Bivalirudin
Dabigatran etexilate
Argatroban
Gabexate
Cupric Chloride
Human thrombin
Marimastat
Metabolic disease
Chronic pain
Health status factors
Nervous system diseases
Visual system diseases
Circulatory system diseases
Infectious or parasitic diseases
Digestive system diseases
Neoplasms
Blood related diseases
Musculoskeletal system diseases
Immune system diseases
Cardiac disease
Cardiac diseases and Venous disease
Cardiac disease, Venous disease and Cerebrovascular disease
Venous disease
Cardiac disease and Cerebrovascular disease
Cerebrovascular disease
Venous disease and Cerebrovascular disease