

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

3

4 Jia-Hao Wang^{1#}, Shan-Shan Dong^{1#}, Hao-An Wang¹, Shao-Shan Liu¹, Xiaoyi Ma²,
5 Ren-Jie Zhu¹, Wei Shi¹, Hao Wu¹, Ke Yu¹, Tian-Pei Zhang¹, Cong-Ru Wang¹, Yan
6 Guo^{1*}, Tie-Lin Yang^{1*}

7

8 ¹Key Laboratory of Biomedical Information Engineering of Ministry of Education,
9 Key Laboratory of Biology Multiomics and Diseases in Shaanxi Province Higher
10 Education Institutions, Biomedical Informatics & Genomics Center, School of Life
11 Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, P. R. China,
12 710049

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

15

16 #These authors contributed equally to this study.

17 *Corresponding authors

18

19 **Address Correspondence to:**

20 Dr. Tie-Lin Yang and Dr. Yan Guo

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

22

23 **Abstract**

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

29

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

35

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

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

46

47 In conclusion, our findings provide new insights into the pathogenic mechanisms of

48 CVD and offering promising targets for further mechanistic and therapeutic studies.

49

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

51

52

53 **Introduction**

54 Cardiovascular diseases (CVD) are a group of disorders of the heart and blood vessels.
55 As the leading global cause of mortality^{1,2}, CVD took an estimated 17.9 million lives
56 in 2019, accounting for 32% of all deaths worldwide³. In clinical practice, there are
57 some commonly used treatments for CVD, such as the use of statins to reduce
58 cardiovascular morbidity and mortality. However, current treatments are not suitable
59 for all patients and have certain risks of side effects⁴. Therefore, there is an urgent and
60 critical need for new therapeutic targets for CVD⁵.

61

62 Proteins, as the final products of gene expression, are the main functional components
63 of biological processes⁶. In addition, most therapeutic agents target proteins⁷.
64 Therefore, understanding their relationship with diseases is crucial for effective
65 treatments⁸⁻¹¹. However, direct observational studies linking protein abundance to
66 phenotypes can be confounded or represent reverse causation¹². Proteome-wide
67 association study (PWAS) is a powerful strategy to solve this problem. It uses single-
68 nucleotide polymorphisms (SNPs) to genetically impute proteins and relate them to
69 genome wide association study (GWAS) summary statistics of a trait to provide
70 evidence of causality^{13, 14}. The genetic models are restricted to the cis-region of the
71 protein, reducing the risk of confounding by horizontal pleiotropy (independent of the
72 protein). Further summary data-based Mendelian randomization (SMR)¹⁵ or
73 colocalization analyses¹⁶ can be used to identify genes contribute to disease
74 pathogenesis through modulating protein abundance. This integrative analytical

75 approach has been employed to identify novel potential therapeutic targets for
76 neurological disorders^{13, 14, 17-19} using brain proteomics data. However, PWAS for
77 CVD is still limited.

78

79 Using plasma proteomics data of the largest cohort to date (53,022 individuals from
80 the UK Biobank Pharma Proteomics Project (UKB-PPP) project²⁰), we present here
81 the first PWAS for multiple CVDs. The study design is shown in Figure 1. We applied
82 the above analytic approaches to the discovery dataset consisting of human plasma
83 proteomic and genetic data from UKB-PPP²⁰ and the GWAS of 26 CVDs spanning 3
84 categories (16 cardiac diseases, N = 234,829~1,030,836; 5 venous diseases, N =
85 388,830~484,598; 5 cerebrovascular diseases, N = 484,598~1,308,460)²¹⁻²⁶.

86 Additionally, we conducted replication analyses leveraging two independent human
87 plasma proteomics datasets^{27, 28}, encompassing 7,213 participants from the
88 Atherosclerosis Risk in Communities (ARIC) study and 3,301 individuals from the
89 INTERVAL study, to ensure robustness and reproducibility. For functional
90 interpretation of the identified proteins, enrichment analyses were performed to detect
91 the pathways associated with CVD. Longitudinal stability analysis at plasma and cell-
92 type level was used to assess the expression stability of the proteins. We also
93 pharmacologically annotate the proteins of interest with approved drugs to assess their
94 feasibility as treatment targets.

95

96 **Results**

97 ***Discovery PWAS of multiple CVDs***

98 We generated the human blood plasma proteome model based on 53,022 UKB-PPP²⁰
99 participants. The initial proteomes data include total 2,923 proteins. After quality
100 control, 1,715 proteins with significant SNP-based heritability ($P < 0.05$, $h^2 > 0$), were
101 used for PWAS. The correlation R^2 between the model's predictive power and
102 heritability for each gene was 0.88 (Supplementary Figure S1), supporting the
103 accuracy of our protein estimation model.

104

105 The plasma proteome model results were integrated with the 26 CVD GWAS data
106 using the FUSION pipeline²⁹. Detail information of the 26 GWAS datasets is shown
107 in Supplementary Table S1. We performed genetic correlation analysis and the results
108 showed that diseases belong to the same category are usually with higher correlation
109 (Supplementary Figure S2).

110

111 As shown in Supplementary Table S2, we identified 341 significant protein-CVD
112 pairs of associations after multiple testing corrections ($P < 2.92 \times 10^{-5}$, $0.05/1,715$
113 proteins) (Fig 2A). Among these associations, 87 genes are located within 1 Mb of
114 each other. With the goal of identifying multiple independent associations, we
115 performed conditional analyses using a regression with summary statistics approach²⁷.
116 27 pairs of associations no longer significant associations were removed
117 (Supplementary Table S3). Finally, we obtained 314 independent and significant
118 PWAS association signals, including 155 unique proteins associated with CVD. The

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

125

126 *Replication PWAS using the proteomics data from ARIC and INTERVAL*

127 For the 155 proteins identified from the discovery dataset, 75 were detected in the at
128 least one of the replication proteomic datasets. The number of detected proteins in the
129 ARIC and INTERVAL project was 64 and 50, respectively. For these proteins, we
130 incorporated their previously built human plasma protein models from the FUSION
131 website²⁹ and the CVD GWAS datasets to perform replication PWAS.

132

133 In the ARIC dataset, the results showed that 46 proteins were associated with CVD (P
134 $< 7.81 \times 10^{-4}$, 0.05/64). As for the INTERVAL dataset, the number of successfully
135 replicated proteins was 37 ($P < 1.00 \times 10^{-3}$, 0.05/50). As shown in supplementary
136 Table S4, the significant association of 55 proteins (73.33%, 55/75) were replicated in
137 at least one dataset with the same effect direction as the discovery PWAS, and 23
138 proteins were replicated in both datasets.

139

140 *Causal-analysis of the proteins identified by PWAS*

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

150

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

157

158 Combining evidence for replication and results of causality tests, 45 of 94 causal
159 proteins were detected in the at least one of replication proteomic datasets. 75.56%
160 (34/45) proteins were found to be with evidence of both replication and causality
161 (Table 1). There were 28 proteins replicated in the ARIC dataset and 21 proteins
162 replicated in the INTERVAL dataset. In particular, 15 proteins were replicated in both

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

164

165 ***Common proteins associated with diseases in three CVD categories***

166 91.49% (86/94) proteins were identified in only one disease category (Fig 3B). Only 2

167 proteins (ABO, PROCR) were associated with diseases in all three CVD categories.

168 As shown in Figure 3A, ABO showed significant positive associations with multiple

169 diseases from the three categories. PROCR was negatively associated with 2 cardiac

170 diseases and 1 cerebrovascular disease, and positively associated with 2 venous

171 diseases. The inconsistency in association direction might because PROCR is linked

172 to anti-inflammatory and anticoagulant functions³⁰.

173

174 ***Novelty of the CVD causal genes***

175 To assess the novelty of the 94 potentially causal genes, we checked the lowest p-

176 values for the SNPs within 2 Mb window of these genes using the summary statistics

177 from the CVD GWAS. We found that 41 genes were not located within 2 Mb of a

178 significant GWAS signal ($P < 5 \times 10^{-8}$), suggesting that these 41 genes are novel

179 genes not implicated in the original GWAS (Fig 3C). 25 of the novel genes were have

180 not been detected in other CVD GWAS either. For example, PROC was found to be

181 associated with venous thromboembolism (Fig 3D top) and COMT was found to be

182 associated with three cardiac diseases (hypertension, statin medication and angina

183 pectoris; Supplementary Figure S3). All 26 CVD GWAS data didn't detect their

184 association with CVD diseases. The rest 16 novel genes were not implicated in the

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

192

193 ***Gene ontology enrichment analysis***

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

204

205 ***Protein-protein interactions (PPI) network analysis***

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

207 analysis using STRING³³. The minimum required interaction score was 0.4. We
208 constructed a PPI network with 30 nodes and 37 edges, primarily comprising 3
209 protein communities. (Fig 4B). The proteins associated with cardiac disease are
210 mainly within the network with FN1 and APOE as the core proteins. Consistent with
211 pathway enrichment analysis results, these proteins are mostly in the community of
212 immunity/inflammation and lipid-related process. The network of proteins associated
213 with venous disease is mainly driven by 6 proteins (F2, F10, F11, PROC, PROCR and
214 PROS1) involved in blood/vessel-related process, especially in the coagulation
215 processes. The network for venous disease proteins is distinct from that of cardiac
216 disease proteins, and the two networks are connected by F2 and PROS1. Interactions
217 among the proteins associated with cerebrovascular diseases are relatively sparse.
218 Complete information about communities is presented in Supplementary Table S7.

219

220 ***Mouse phenotypic annotation of potential causal genes.***

221 We further evaluated whether 94 proteins were associated with CVD-related
222 phenotype in mouse using the Mouse Genome Informatics (MGI) database³⁴. We
223 performed phenotype enrichment analysis using the Fisher's exact test. Consistent
224 with the pathway enrichment and network analysis results, mutations in genes
225 associated with cardiac diseases are enriched in phenotypes related to
226 immunity/inflammation, lipid-related process, and vessel/blood-related process (Fig
227 4C). Mutations in genes associated with venous disease are enriched in phenotypes
228 related to vessel/blood-related process (Fig 4C). These results further support the

229 involvement of the identified proteins in CVD pathogenesis.

230

231 ***Evaluate longitudinal stability at protein and single-cell level***

232 To evaluate the expression stability of the 94 proteins, we performed longitudinal

233 analysis using data using plasma proteomics data and peripheral blood mononuclear

234 cells (PBMC) single cell RNA-seq (scRNA-seq) data from GEO dataset GSE190992.

235

236 The plasma proteomics data were collected from 6 healthy, non-smoking Caucasian

237 donors over a 10-week period. 44 of the 94 proteins were detected in this dataset.

238 Among them, 84.09% (37/44) proteins exhibit stable expression in plasma (median

239 coefficient of variation < 10%, Fig4 D left). Fluctuations in the plasma levels of these

240 proteins might serve as potential markers of disease status.

241

242 The PBMC scRNA-seq data were collected weekly from four donors over the course

243 of six weeks. We found 24 genes exhibited stable expression in at least one cell type

244 (median coefficient of variation < 10% in at least one cell type across all donors. Fig4

245 D right). Notably, 23 of the 24 genes stably expressed in CD14+ monocytes. As per

246 previous studies, monocytes play a crucial role in both local ischemia and

247 inflammatory responses, which are closely linked to the development of

248 cardiovascular diseases³⁵⁻³⁷.

249

250 ***Cell-type specific expression of the CVD causal genes***

251 To investigate whether these genes show distinct enrich across different cell types, we
252 utilized PBMC RNA-seq data obtained from the plasma of another 11 healthy donors
253 to examine the specific expression patterns of these genes. Among 94 CVD causal
254 genes, 39 were enriched in one or more cell types (FDR adjust $P < 0.05$ and $\log_{2}FC >$
255 1.5, Supplementary Table 9), include CD4 T cells, CD14+ monocytes, Platelet,
256 Natural killer cell and other monocyte. A total of 21 genes were highly expressed in
257 CD14+ monocytes, and half of these genes (11 out of 21) were also found to be stably
258 expressed in CD14+ monocytes through stability analysis.

259

260 *Drug repurposing analyses identified potential therapeutic targets for CVD*

261 To investigate the potential drug target genes, we construct a gene-drug-disease
262 network (Fig4 D). The results showed that 25 of the 94 proteins are the targets of 53
263 drugs with completed or currently undergoing clinical trials (Supplementary Table 10).
264 14 drugs have already been used for treating circulatory system disorders. For
265 example, Drotrecogin alfa targeting *F2*, *PROCR* and *PROSI* are currently one of the
266 efficacious treatments for managing cerebrovascular ischemic events³⁸. The rest 39
267 drugs for treating diseases from other systems might be considered in further drug
268 repurposing research. For example, Menadione targeting PROC are currently used for
269 treating vitamin K deficiency and prostate cancer.

270

271 **Discussion**

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

279

280 The PWAS analysis identified 96 genes that are consistent with being causal in CVD,
281 including 41 novel genes not implicated in original GWAS. For example, PROC was
282 newly found to be associated with venous thromboembolism. PROC is a vitamin K-
283 dependent enzyme that plays a crucial role in regulating human thrombosis and
284 hemostasis³⁹. Consistent with our results, previous studies have demonstrated that
285 reduced PROC levels in plasma can be used as a marker of increased risk of venous
286 thrombosis^{40, 41}. In the PPI network, we also demonstrated that PROC, together with
287 coagulation factors such as F2 and F10, forms a venous-related network. Longitudinal
288 stability analysis showed that this protein is stably expressed in blood plasma.
289 Currently, two drugs (Menadione and Cupric Chloride) targeting PROC have passed
290 clinical trials for the treatment of conditions such as fungal infections, prostate cancer,
291 and vitamin K deficiency⁴². Further studies are needed to explore the potential of
292 these drugs for treating venous thromboembolism.

293

294 91.49% (86/94) of the identified proteins are category-specific, suggesting that the
295 underlying pathogenesis mechanisms of the three disease categories are different.
296 Only two proteins (ABO and PROCR) were associated with diseases in all three CVD
297 categories. ABO was found to be positively associated with multiple cardiovascular
298 diseases. Consistently, epidemiological studies have reported that ABO is associated
299 with a wide range of diseases, including cardiovascular ailments, malignancies, and
300 infectious conditions^{43,44}. PROCR is positively associated with venous disorders but
301 negatively associated with stroke and coronary artery disease. PROCR is a receptor
302 for activated protein C, which is a serine protease activated by and involved in the
303 blood coagulation pathway. Consistent with our results, GWAS studies have shown
304 that the minor G allele of rs867186 at this gene is correlated with a higher risk of
305 venous thromboembolism^{45, 46} but a lower risk of CAD^{47, 48}. A previous study³⁰ has
306 shown that PROCR linked to CAD through anti-inflammatory mechanisms and to
307 VTE through pro-thrombotic mechanisms.

308

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

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

319

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

327

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

334

335 **Methods**

336 **Human plasma proteomic and genetic data in UKB.**

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

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

347 Genotype data matching the protein dataset underwent genotyping, imputation, and
348 quality control steps as detailed in previous work ⁵³. This included sex discrepancy,
349 sex chromosome aneuploidy, and heterozygosity checks, with imputed variants
350 filtered for INFO scores >0.7. All chromosomal positions were updated to the hg38
351 assembly using LiftOver ⁵⁴. Genotyping quality control was executed using PLINK2.0
352 software ⁵⁵. Participants exhibiting over 5% missing genotypic data were removed
353 from consideration. Moreover, variants displaying deviations from the Hardy-
354 Weinberg equilibrium (with p-values less than 1×10^{-8}), a genotype missing rate
355 exceeding 5%, a minor allele frequency below 1%, or those not classified as SNPs,
356 were also excluded from the analysis.

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

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

363

364 **CVD GWAS summary association statistics**

365 Our GWAS data mainly comes from the GWAS catalog^{22 25 21, 23, 24, 26} and FinnGen⁵⁷
366 database. In accordance with the ICD-10 standard of circulatory disorders, we
367 selected GWAS studies involving a minimum of 5,000 cases. When multiple studies
368 of the same condition were identified, we opted for those with the largest sample sizes.
369 This stringent selection procedure resulted in a final cohort of 26 unique GWAS for
370 our investigation. Based on the distinct pathophysiological mechanisms, we
371 categorized the diseases into cardiac diseases, venous disease, and cerebrovascular
372 disease.

373

374 **Statistical approach**

375 *Proteome-wide association studies (PWAS)*. We used the standard processes in the
376 FUSION software²⁹ to construct protein models and incorporate GWAS data for our
377 PWAS analysis. After applying the previously outlined quality control measures to
378 screen the sample and genotype data, we utilized GCTA software⁵⁸ to estimate the
379 SNP-based heritability for individual proteins. To expedite calculations, a random
380 subset of approximately 10,000 individuals was selected from the full cohort for each
381 protein's heritability estimation. From the analysis of 2,923 proteins, 1,715 displayed

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

389

390 We then performed the replication PWAS analysis in two other publicly available data
391 sets. The modeling methodologies for these datasets have been documented in prior
392 research, the Atherosclerosis Risk in Communities (ARIC) study dataset included
393 4,483 protein measurements from 7,213 European participants²⁷, while the
394 INTERVAL study retained information on 3,170 proteins for 3,301 individuals²⁸.
395 Subsequent to the heritability filtering phase, an ensemble of 2,379 proteins (1,348 in
396 ARIC and 1,031 in INTERVAL) was selected for incorporation into our replication
397 verification.

398

399 *Causal analysis.* We adopted two independent frameworks to rigorously ascertain the
400 causal inference of the proteins implicated in our PWAS findings. For the Bayesian
401 colocalization analysis¹⁶, we employed the COLOC module embedded within the
402 FUSION software suite. The COLOC tool operates by estimating the posterior
403 probability indicating that the same causal variant underlies both GWAS and pQTL.

404 Within the colocalization analysis framework, a comprehensive set of five hypotheses
405 (H0 through H4) are scrutinized. Notably, hypothesis H4 posits the existence of a SNP
406 that acts as a shared causal driver for both pQTL and GWAS. In our study, we defined
407 causality for proteins identified through the COLOC analysis as those exhibiting a
408 posterior probability for Hypothesis H4 exceeding 0.7. We subsequently employed the
409 SMR¹⁵ approach to further validate the causal relationships inferred from the PWAS
410 and GWAS. For this SMR analysis, we leveraged recently published pQTL data²⁰,
411 which were derived from UKB-PPP study, complemented by independently obtained
412 GWAS data^{21-24, 26, 57} on cardiovascular disease, which were also considered in our
413 PWAS. Our determination of significant causal relationships relied on an adjusted P
414 <0.05 for the SMR analysis and the unadjusted P>0.05 from the HEIDI test.

415

416 **PPI and GO enrichment**

417 For the investigation of causal genes implicated in three diseases, we employed the
418 STRING³³ database to perform an extensive network analysis. The Markov cluster
419 (MCL) algorithm was used with the following parameters: inflation parameter—1.5.
420 Subsequently, the derived network was refined and visually optimized utilizing
421 Cytoscape⁵⁹ software. In this visualization, node size corresponds to the degree of
422 connectivity for each gene, indicative of its interaction frequency within the network,
423 while distinct colors denote different gene categories, facilitating categorical
424 distinction and interpretation. Additionally, we conducted functional enrichment
425 analysis for causal genes pertinent in three categories diseases using the WebGestalt³²

426 online platform, focusing on the GO BP pathways. We select the pathways with $P <$
427 0.05 (with FDR adjusted) and number of overlap genes > 3 as the significant result.

428

429 *Longitudinal Data Stability Analysis*

430 We conducted a longitudinal analysis utilizing the data set GSE190992 from the Gene
431 Expression Omnibus (GEO) database⁶⁰. Specific details regarding the data collection
432 methodology and information have been reported in detail in a previous publication.
433 The data encompass proteomics measurements over a 10-week period for 6 healthy
434 donors, as well as single-cell data collected over a six-week period for 4 of these
435 donors. For each donor, we calculated the coefficient of variation (CV) for each gene
436 at both the proteome and single-cell levels as a measure of stability ($CV = \text{standard}$
437 $\text{deviation} / \text{mean} \times 100\%$). We selected thresholds of 10% as criteria for stable gene
438 expression in proteome data and single-cell data, respectively. These genes that
439 exhibit stable expression in plasma, along with the associated cell types, can be
440 considered as more reliable biomarkers for early screening and prediction of CVD.

441

442 *Cell-type specific expression of the CVD causal genes*

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

448 datasets from different samples and mitigate batch effects, we applied the Harmony
449 integration method. These procedures were carried out using Seurat package version
450 4.4.0 within the R environment. After quality control and normalization, the data
451 comprised a total of 27,484 genes across 371,086 cells. For the 94 CVD causal genes,
452 we used the Wilcoxon rank sum test to compare the expression levels between the
453 cells of interest and other cells. We applied FDR correction to the P values derived
454 from the multiple tests, with the total number of tests set to 27,484 genes. Finally, we
455 retained the significant results of FDR P value < 0.05 and $\log_{2}FC > 1.5$, and thought
456 that the expression was specific expression in cells.

457

458 **Mouse genome informatics and Drug analysis**

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

470 development at any stage.

471

472 **Code availability**

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

474

475 **Acknowledgments**

476 This research has been conducted using the UK biobank resource under application

477 number 46387.

478

479 **Funding**

480 This work was supported by the National Natural Science Foundation of China

481 (32370653, and 82372458); Innovation Capability Support Program of Shaanxi

482 Province (2022TD-44); Key Research and Development Project of Shaanxi Province

483 (2022GXLH-01-22), and the Fundamental Research Funds for the Central

484 Universities. This study was also supported by the High-Performance Computing

485 Platform and Instrument Analysis Center of Xi'an Jiaotong University.

486

487 **Data and Resource Availability**

488 The GWAS summary data of CVD we used in this article were available from GWAS

489 catalog (<https://www.ebi.ac.uk/gwas/downloads/summary-statistics>) and FinnGen

490 study (<https://finngen.gitbook.io/documentation/data-download>). Download links for

491 all datasets are provided in Table S1.

492

493 **Author Contributions**

494 J.-H.W. and S.-S.D. designed this project. J.-H.W., A.-H.W., and C.-R.W. conducted
495 the computational work. J.-H.W. wrote the manuscript. S.-S.D. and A.-H.W. revised
496 the manuscript. J.-H.W., H.-A.W. and S.-S.D. summarized the tables and figures. R.-
497 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
498 T.-L.Y. supported and supervised this project.

499

500 **Ethical approval and consent to participate**

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

503

504 **Competing interests**

505 The authors declare that they have no conflict of interest.

506

507 **References**

- 508 1. Townsend N, Kazakiewicz D, Lucy Wright F, et al. Epidemiology of cardiovascular
509 disease in Europe. *Nat Rev Cardiol.* Feb 2022;19(2):133-143. doi:10.1038/s41569-021-
510 00607-3
- 511 2. Collaborators GBDCoD. Global, regional, and national age-sex-specific mortality for 282
512 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the
513 Global Burden of Disease Study 2017. *Lancet.* Nov 10 2018;392(10159):1736-1788.
514 doi:10.1016/S0140-6736(18)32203-7
- 515 3. Holmes MV, Richardson TG, Ference BA, Davies NM, Davey Smith G. Integrating
516 genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug
517 development. *Nat Rev Cardiol.* Jun 2021;18(6):435-453. doi:10.1038/s41569-020-00493-1
- 518 4. Ward NC, Watts GF, Eckel RH. Statin Toxicity. *Circ Res.* Jan 18 2019;124(2):328-350.
519 doi:10.1161/CIRCRESAHA.118.312782
- 520 5. Li Y, Li Z, Ren Y, et al. Mitochondrial-derived peptides in cardiovascular disease: Novel
521 insights and therapeutic opportunities. *J Adv Res.* Nov 24
522 2023;doi:10.1016/j.jare.2023.11.018
- 523 6. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic
524 and transcriptomic analyses. *Nature Reviews Genetics.* 2012;13(4):227-232.
525 doi:10.1038/nrg3185
- 526 7. Yao P, Iona A, Pozarickij A, et al. Proteomic Analyses in Diverse Populations Improved
527 Risk Prediction and Identified New Drug Targets for Type 2 Diabetes. *Diabetes Care.* Jun 1
528 2024;47(6):1012-1019. doi:10.2337/dc23-2145

- 529 8. Lindsey ML, Mayr M, Gomes AV, et al. Transformative Impact of Proteomics on
530 Cardiovascular Health and Disease. *Circulation*. 2015;132(9):852-872.
531 doi:10.1161/cir.0000000000000226
- 532 9. Wik L, Nordberg N, Broberg J, et al. Proximity Extension Assay in Combination with
533 Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol Cell*
534 *Proteomics*. 2021;20:100168. doi:10.1016/j.mcpro.2021.100168
- 535 10. Haslam DE, Li J, Dillon ST, et al. Stability and reproducibility of proteomic profiles in
536 epidemiological studies: comparing the Olink and SOMAscan platforms. *Proteomics*. Jul
537 2022;22(13-14):e2100170. doi:10.1002/pmic.202100170
- 538 11. Wang R-S, Maron BA, Loscalzo J. Multiomics Network Medicine Approaches to
539 Precision Medicine and Therapeutics in Cardiovascular Diseases. *Arteriosclerosis,*
540 *Thrombosis, and Vascular Biology*. 2023;43(4):493-503. doi:10.1161/atvbaha.122.318731
- 541 12. Brandes N, Linial N, Linial M. PWAS: proteome-wide association study-linking genes and
542 phenotypes by functional variation in proteins. *Genome Biol*. Jul 14 2020;21(1):173.
543 doi:10.1186/s13059-020-02089-x
- 544 13. Wingo AP, Liu Y, Gerasimov ES, et al. Integrating human brain proteomes with genome-
545 wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nat*
546 *Genet*. Feb 2021;53(2):143-146. doi:10.1038/s41588-020-00773-z
- 547 14. Wingo TS, Liu Y, Gerasimov ES, et al. Brain proteome-wide association study implicates
548 novel proteins in depression pathogenesis. *Nat Neurosci*. Jun 2021;24(6):810-817.
549 doi:10.1038/s41593-021-00832-6
- 550 15. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies

- 551 predicts complex trait gene targets. *Nat Genet.* May 2016;48(5):481-7. doi:10.1038/ng.3538
- 552 16. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation
553 between pairs of genetic association studies using summary statistics. *PLoS Genet.* May
554 2014;10(5):e1004383. doi:10.1371/journal.pgen.1004383
- 555 17. Li SJ, Shi JJ, Mao CY, et al. Identifying causal genes for migraine by integrating the
556 proteome and transcriptome. *J Headache Pain.* Aug 17 2023;24(1):111. doi:10.1186/s10194-
557 023-01649-3
- 558 18. Phillips B, Western D, Wang L, et al. Proteome wide association studies of LRRK2
559 variants identify novel causal and druggable proteins for Parkinson's disease. *NPJ*
560 *Parkinsons Dis.* Jul 8 2023;9(1):107. doi:10.1038/s41531-023-00555-4
- 561 19. Wu BS, Chen SF, Huang SY, et al. Identifying causal genes for stroke via integrating the
562 proteome and transcriptome from brain and blood. *J Transl Med.* Apr 21 2022;20(1):181.
563 doi:10.1186/s12967-022-03377-9
- 564 20. Sun BB, Chiou J, Traylor M, et al. Plasma proteomic associations with genetics and
565 health in the UK Biobank. *Nature.* Oct 2023;622(7982):329-338. doi:10.1038/s41586-023-
566 06592-6
- 567 21. Nielsen JB, Thorolfsdottir RB, Fritsche LG, et al. Biobank-driven genomic discovery
568 yields new insight into atrial fibrillation biology. *Nat Genet.* Sep 2018;50(9):1234-1239.
569 doi:10.1038/s41588-018-0171-3
- 570 22. Shah S, Henry A, Roselli C, et al. Genome-wide association and Mendelian
571 randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun.*
572 Jan 9 2020;11(1):163. doi:10.1038/s41467-019-13690-5

- 573 23. Donertas HM, Fabian DK, Valenzuela MF, Partridge L, Thornton JM. Common genetic
574 associations between age-related diseases. *Nat Aging*. Apr 2021;1(4):400-412.
575 doi:10.1038/s43587-021-00051-5
- 576 24. Hartiala JA, Han Y, Jia Q, et al. Genome-wide analysis identifies novel susceptibility loci
577 for myocardial infarction. *Eur Heart J*. Mar 1 2021;42(9):919-933.
578 doi:10.1093/eurheartj/ehaa1040
- 579 25. Mishra A, Malik R, Hachiya T, et al. Stroke genetics informs drug discovery and risk
580 prediction across ancestries. *Nature*. Nov 2022;611(7934):115-123. doi:10.1038/s41586-022-
581 05165-3
- 582 26. Sollis E, Mosaku A, Abid A, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and
583 deposition resource. *Nucleic Acids Res*. Jan 6 2023;51(D1):D977-D985.
584 doi:10.1093/nar/gkac1010
- 585 27. Zhang J, Dutta D, Kottgen A, et al. Plasma proteome analyses in individuals of European
586 and African ancestry identify cis-pQTLs and models for proteome-wide association studies.
587 *Nat Genet*. May 2022;54(5):593-602. doi:10.1038/s41588-022-01051-w
- 588 28. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome.
589 *Nature*. Jun 2018;558(7708):73-79. doi:10.1038/s41586-018-0175-2
- 590 29. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide
591 association studies. *Nat Genet*. Mar 2016;48(3):245-52. doi:10.1038/ng.3506
- 592 30. Stacey D, Chen L, Stanczyk PJ, et al. Elucidating mechanisms of genetic cross-disease
593 associations at the PROCRA vascular disease locus. *Nat Commun*. Mar 9 2022;13(1):1222.
594 doi:10.1038/s41467-022-28729-3

- 595 31. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER:
596 Making genome-scale phylogenetics accessible to all. *Protein Science*. 2021;31(1):8-22.
597 doi:10.1002/pro.4218
- 598 32. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit
599 with revamped UIs and APIs. *Nucleic Acids Res*. Jul 2 2019;47(W1):W199-W205.
600 doi:10.1093/nar/gkz401
- 601 33. Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein-
602 protein association networks and functional enrichment analyses for any sequenced genome
603 of interest. *Nucleic Acids Res*. Jan 6 2023;51(D1):D638-D646. doi:10.1093/nar/gkac1000
- 604 34. Blake JA, Baldarelli R, Kadin JA, et al. Mouse Genome Database (MGD):
605 Knowledgebase for mouse-human comparative biology. *Nucleic Acids Research*.
606 2021;49(D1):D981-D987. doi:10.1093/nar/gkaa1083
- 607 35. Jaipersad AS, Lip GY, Silverman S, Shantsila E. The role of monocytes in angiogenesis
608 and atherosclerosis. *J Am Coll Cardiol*. Jan 7-14 2014;63(1):1-11.
609 doi:10.1016/j.jacc.2013.09.019
- 610 36. Tahir S, Steffens S. Nonclassical monocytes in cardiovascular physiology and disease.
611 *Am J Physiol Cell Physiol*. May 1 2021;320(5):C761-C770. doi:10.1152/ajpcell.00326.2020
- 612 37. Ruder AV, Wetzels SMW, Temmerman L, Biessen EAL, Goossens P. Monocyte
613 heterogeneity in cardiovascular disease. *Cardiovasc Res*. Sep 5 2023;119(11):2033-2045.
614 doi:10.1093/cvr/cvad069
- 615 38. Southan C, Sharman JL, Benson HE, et al. The IUPHAR/BPS Guide to
616 PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein

- 617 targets and 6000 ligands. *Nucleic Acids Res.* Jan 4 2016;44(D1):D1054-68.
618 doi:10.1093/nar/gkv1037
- 619 39. Dinarvand P, Moser KA. Protein C Deficiency. *Arch Pathol Lab Med.* Oct
620 2019;143(10):1281-1285. doi:10.5858/arpa.2017-0403-RS
- 621 40. Tang W, Stimson MR, Basu S, et al. Burden of rare exome sequence variants in PROC
622 gene is associated with venous thromboembolism: a population-based study. *J Thromb*
623 *Haemost.* Feb 2020;18(2):445-453. doi:10.1111/jth.14676
- 624 41. Manderstedt E, Lind-Hallden C, Hallden C, et al. Classic Thrombophilias and Thrombotic
625 Risk Among Middle-Aged and Older Adults: A Population-Based Cohort Study. *J Am Heart*
626 *Assoc.* Feb 15 2022;11(4):e023018. doi:10.1161/JAHA.121.023018
- 627 42. Kovács KB, Pataki I, Bárdos H, et al. Molecular characterization of p. Asp77Gly and the
628 novel p. Ala163Val and p. Ala163Glu mutations causing protein C deficiency. *Thrombosis*
629 *research.* 2015;135(4):718-726.
- 630 43. Wu O, Bayoumi N, Vickers MA, Clark P. ABO(H) blood groups and vascular disease: a
631 systematic review and meta-analysis. *J Thromb Haemost.* Jan 2008;6(1):62-9.
632 doi:10.1111/j.1538-7836.2007.02818.x
- 633 44. Li S, Schooling CM. A phenome-wide association study of ABO blood groups. *BMC Med.*
634 Nov 17 2020;18(1):334. doi:10.1186/s12916-020-01795-4
- 635 45. Dennis J, Johnson CY, Adediran AS, et al. The endothelial protein C receptor (PROCR)
636 Ser219Gly variant and risk of common thrombotic disorders: a HuGE review and meta-
637 analysis of evidence from observational studies. *Blood.* Mar 8 2012;119(10):2392-400.
638 doi:10.1182/blood-2011-10-383448

- 639 46. Medina P, Navarro S, Bonet E, et al. Functional analysis of two haplotypes of the human
640 endothelial protein C receptor gene. *Arterioscler Thromb Vasc Biol.* Mar 2014;34(3):684-90.
641 doi:10.1161/ATVBAHA.113.302518
- 642 47. Howson JMM, Zhao W, Barnes DR, et al. Fifteen new risk loci for coronary artery
643 disease highlight arterial-wall-specific mechanisms. *Nat Genet.* Jul 2017;49(7):1113-1119.
644 doi:10.1038/ng.3874
- 645 48. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an
646 Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res.* Feb 2
647 2018;122(3):433-443. doi:10.1161/CIRCRESAHA.117.312086
- 648 49. Kashiwagi M, Imanishi T, Tsujioka H, et al. Association of monocyte subsets with
649 vulnerability characteristics of coronary plaques as assessed by 64-slice multidetector
650 computed tomography in patients with stable angina pectoris. *Atherosclerosis.* Sep
651 2010;212(1):171-6. doi:10.1016/j.atherosclerosis.2010.05.004
- 652 50. Tapp LD, Shantsila E, Wrigley BJ, Pamukcu B, Lip GY. The CD14++CD16+ monocyte
653 subset and monocyte-platelet interactions in patients with ST-elevation myocardial infarction.
654 *J Thromb Haemost.* Jul 2012;10(7):1231-41. doi:10.1111/j.1538-7836.2011.04603.x
- 655 51. Anker SD, Egerer Kr Fau - Volk HD, Volk Hd Fau - Kox WJ, Kox Wj Fau - Poole-Wilson
656 PA, Poole-Wilson Pa Fau - Coats AJ, Coats AJ. Elevated soluble CD14 receptors and altered
657 cytokines in chronic heart failure. (0002-9149 (Print))
- 658 52. Niebauer J, Volk HD, Kemp M, et al. Endotoxin and immune activation in chronic heart
659 failure: a prospective cohort study. *Lancet.* May 29 1999;353(9167):1838-42.
660 doi:10.1016/S0140-6736(98)09286-1

- 661 53. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep
662 phenotyping and genomic data. *Nature*. Oct 2018;562(7726):203-209. doi:10.1038/s41586-
663 018-0579-z
- 664 54. Hinrichs AS, Karolchik D, Baertsch R, et al. The UCSC Genome Browser Database:
665 update 2006. *Nucleic Acids Res*. Jan 1 2006;34(Database issue):D590-8.
666 doi:10.1093/nar/gkj144
- 667 55. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation
668 PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.
669 doi:10.1186/s13742-015-0047-8
- 670 56. International HapMap C, Altshuler DM, Gibbs RA, et al. Integrating common and rare
671 genetic variation in diverse human populations. *Nature*. Sep 2 2010;467(7311):52-8.
672 doi:10.1038/nature09298
- 673 57. Kurki MI, Karjalainen J, Palta P, et al. FinnGen provides genetic insights from a well-
674 phenotyped isolated population. *Nature*. Jan 2023;613(7944):508-518. doi:10.1038/s41586-
675 022-05473-8
- 676 58. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait
677 analysis. *Am J Hum Genet*. Jan 7 2011;88(1):76-82. doi:10.1016/j.ajhg.2010.11.011
- 678 59. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated
679 models of biomolecular interaction networks. *Genome Res*. Nov 2003;13(11):2498-504.
680 doi:10.1101/gr.1239303
- 681 60. Vasaikar SV, Savage AK, Gong Q, et al. A comprehensive platform for analyzing
682 longitudinal multi-omics data. *Nat Commun*. Mar 27 2023;14(1):1684. doi:10.1038/s41467-

683 023-37432-w

684 61. Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank Knowledgebase for

685 2024. *Nucleic Acids Res.* Jan 5 2024;52(D1):D1265-D1275. doi:10.1093/nar/gkad976

686 62. Zhou Y, Zhang Y, Zhao D, et al. TTD: Therapeutic Target Database describing target

687 druggability information. *Nucleic Acids Res.* Jan 5 2024;52(D1):D1465-D1477.

688 doi:10.1093/nar/gkad751

689

690

691 **Figure Legends**

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

699

700 **Figure 2. Result of the PWAS**

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

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

710 **C.** Regional association of PWAS hits in conditional analysis for heart failure.

711 Conditionally significant proteins are CELSR2, GSTM1, SPINK8, DAG1, HYAL1,

712 FABP2, ACYP1, FES, GSTT2B and SUSD2. Top panel in each plot highlights the

713 marginally associated PWAS genes (blue) and the jointly significant genes (green).

714 Bottom panel shows a regional Manhattan plot of the data before (grey) and after

715 (blue) conditioning on the predicted expression of the green genes. Chr: chromosome.

716

717 **Figure 3. Results for the causal genes.**

718 **A.** The heatmap presents whole PWAS results for 94 genes passing causality tests and

719 color depth reflects the association direction and magnitude. Genes identified in

720 replication PWAS are represented by circles in the heatmap. Causal genes are labeled

721 “*” and the novel gene with no significant variant ($P < 5 \times 10^{-8}$) within $\pm 2M$ window

722 of the gene range in GWAS results are labeled in red.

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

724 categories.

725 **C.** The number of novel genes in different diseases.

726 **D.** The top Manhattan plot represents the pQTL and the GWAS results within the

727 *PROC* genomic region for venous thromboembolism. The bottom Manhattan plot

728 represents the pQTL and the GWAS results within the *FNI* genomic region for

729 coronary heart disease and coronary atherosclerosis.

730

731 **Figure 4. Function annotation of the identified genes**

732 **A.** The significant enriched Gene Ontology (GO) biological process (BP) terms of the

733 causal genes in different categories. The color of the bar represents the biological

734 function category to which the pathway belongs. Immunity/inflammation (light coral),

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

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

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

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

757 **E.** The constructed gene-drug-disease network of causal genes. The colors of the lines

758 in the network signify the category of genes. ICD: International Classification of

759 Diseases.

760

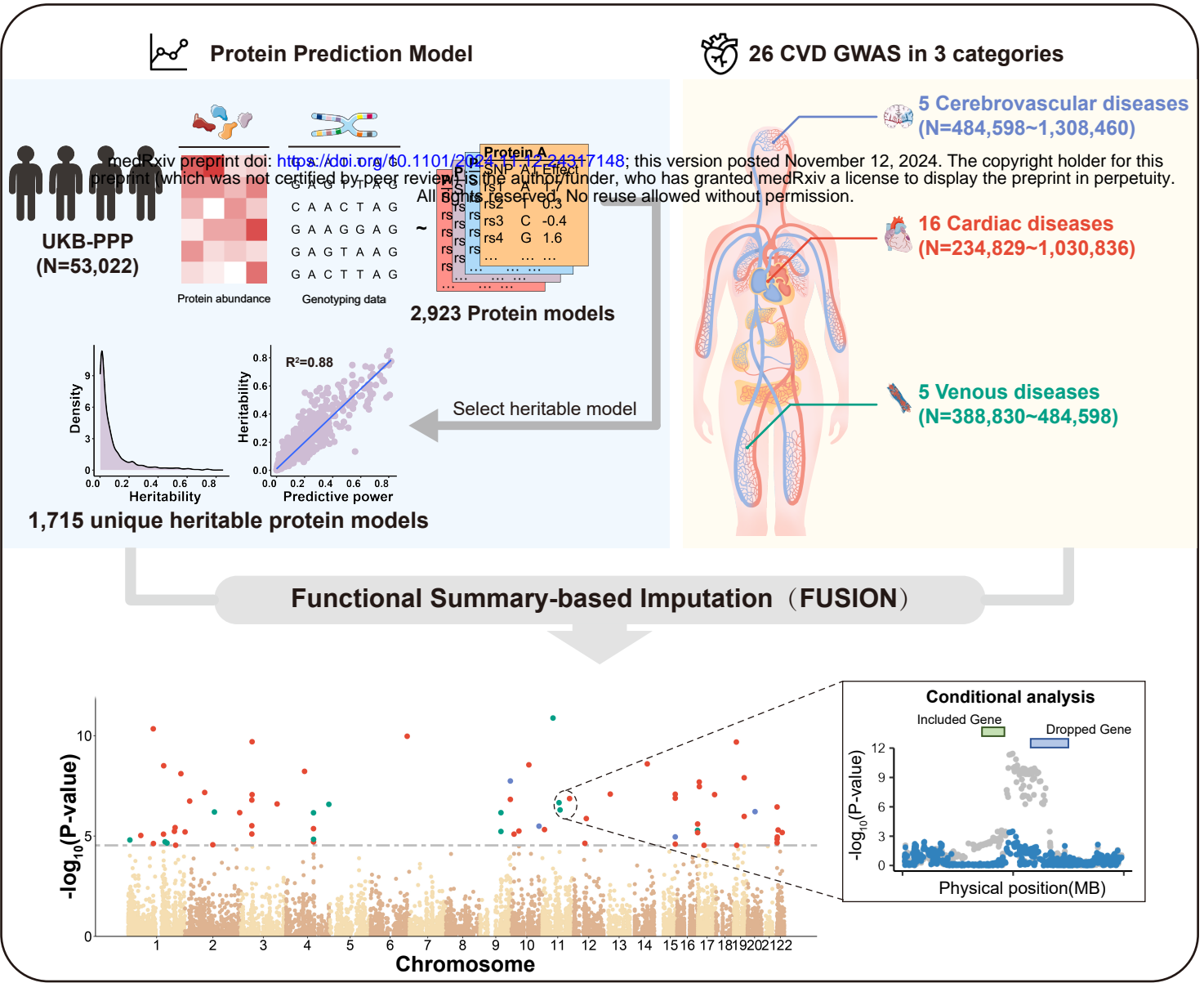
761

762 Table 1. Summary of the 34 replicable CVD causal genes.

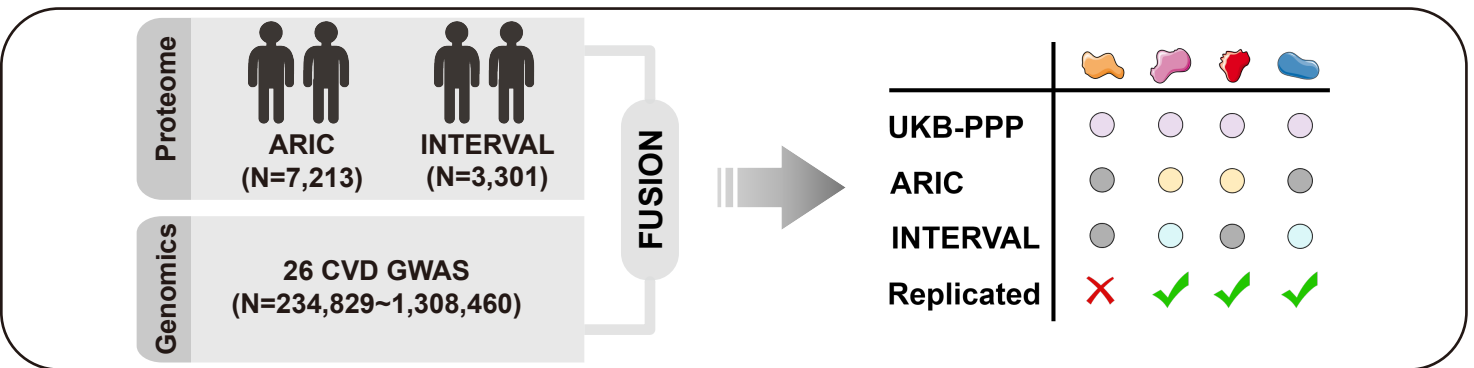
Gene	replicated	causal	PHE
ABO	Both	Both	Pulmonary embolism
	Both	COLOC	Deep Venous Thrombosis, Venous Thromboembolism
	INTERVAL	Both	Large artery stroke
	INTERVAL	COLOC	Myocardial infarction, Any stroke, Cardioembolic stroke
ANGPTL3	ARIC	COLOC	Statin medication
ASPN	Both	Both	Venous Thromboembolism
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

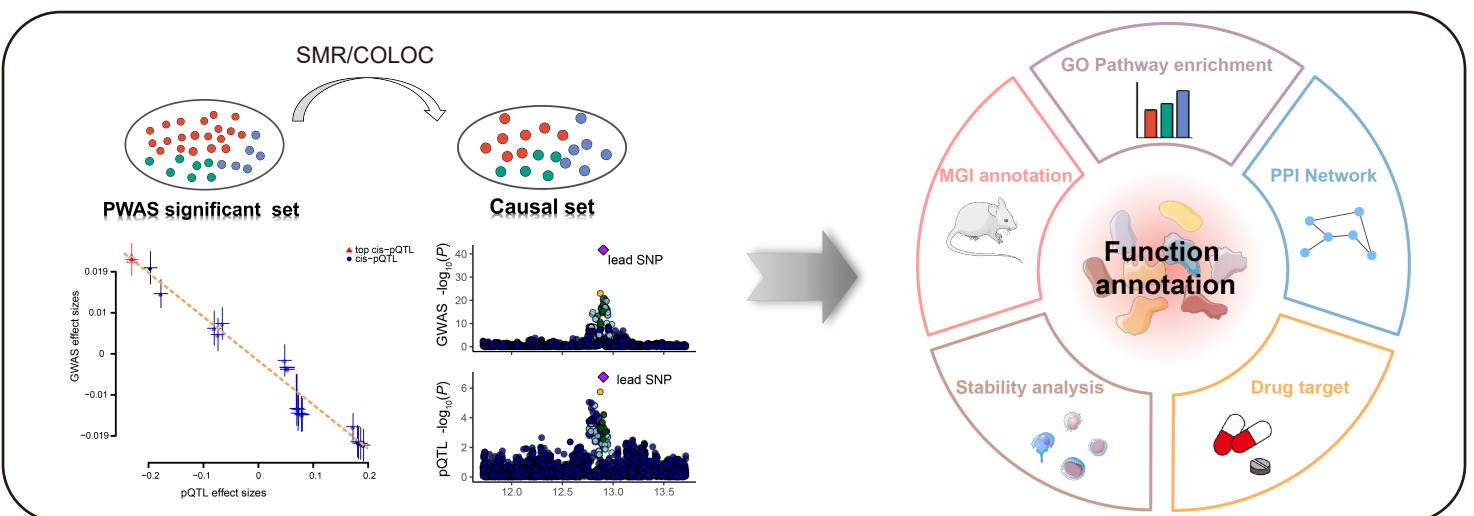
1. Discovery plasma proteome-wide association study (PWAS) analysis



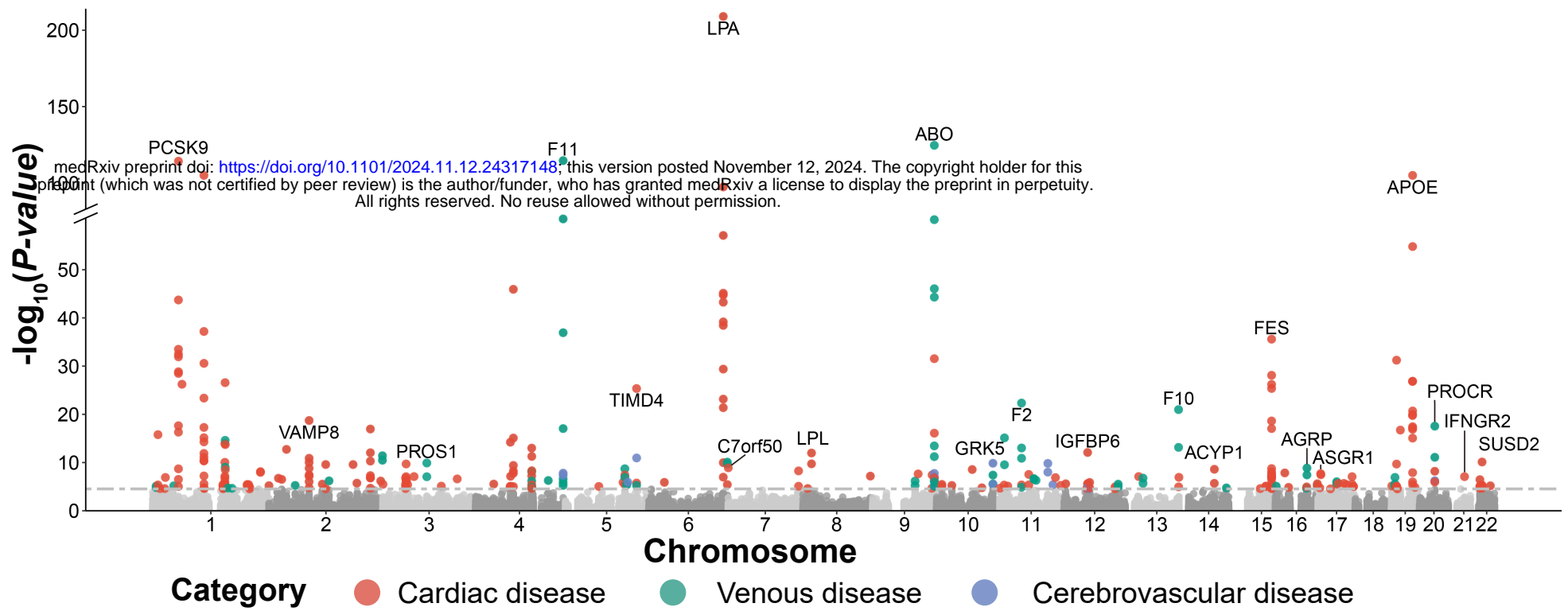
2. Replication PWAS of CVD



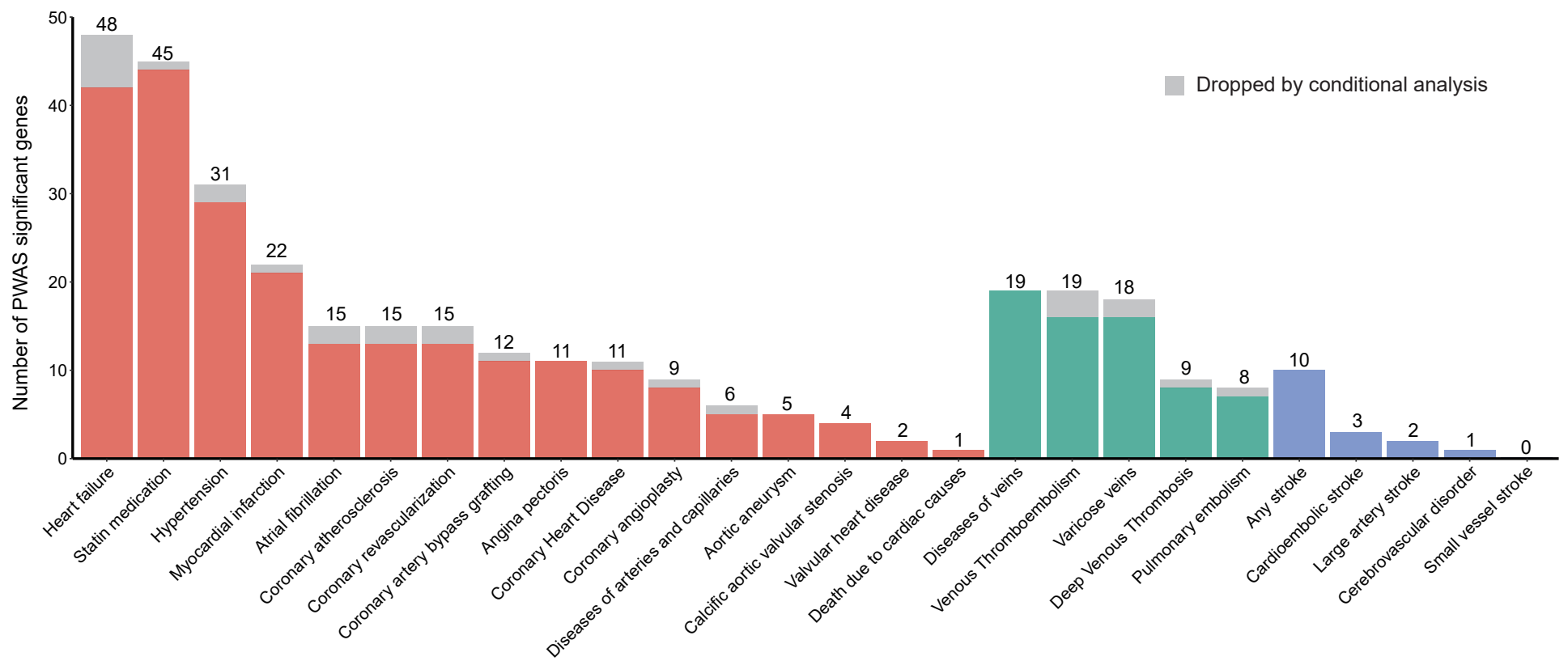
3. Function verification of candidate causal genes



A

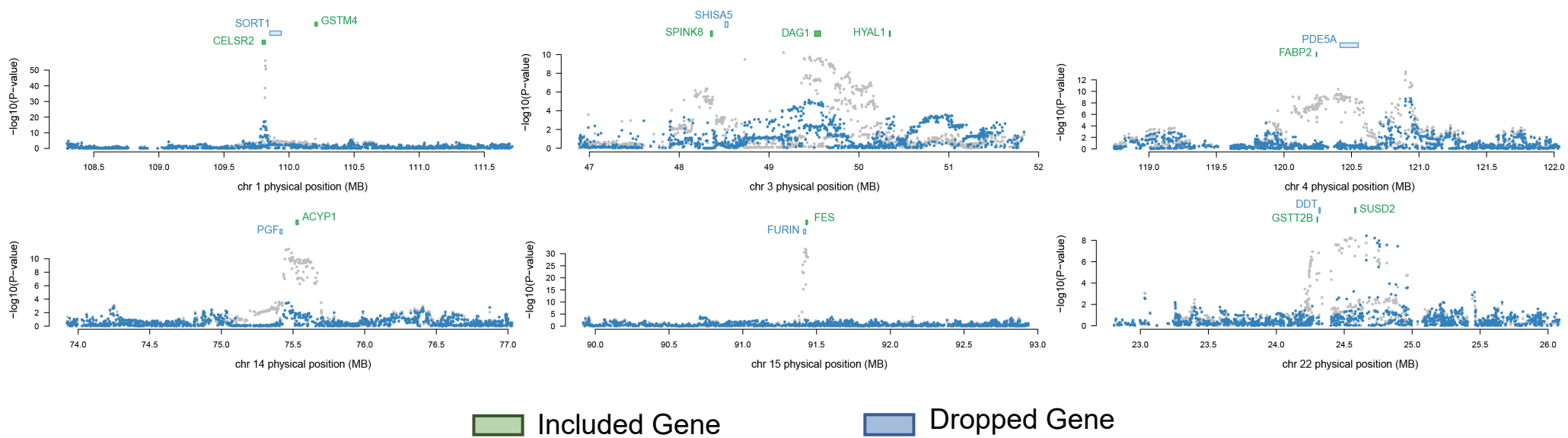


B

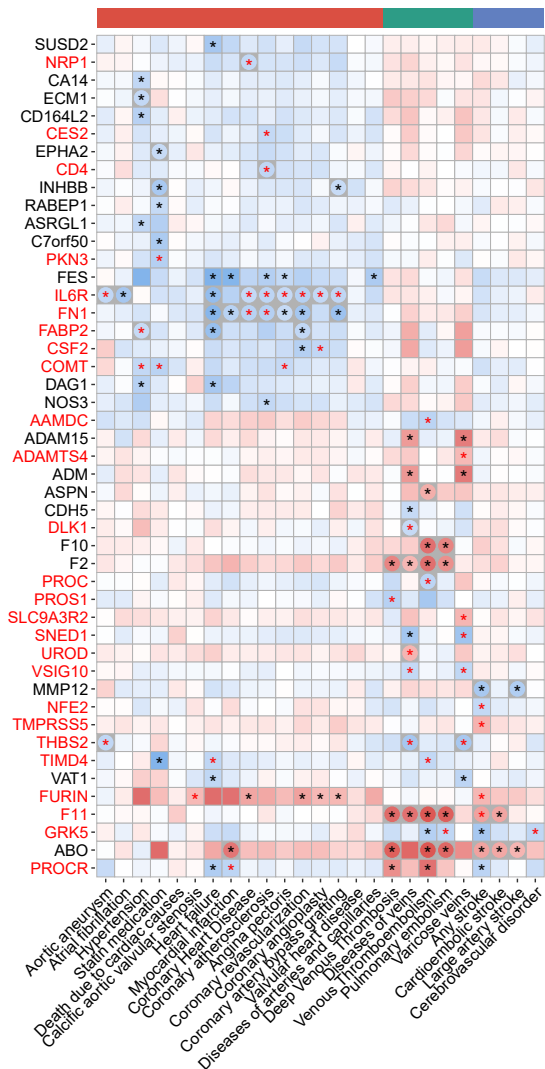
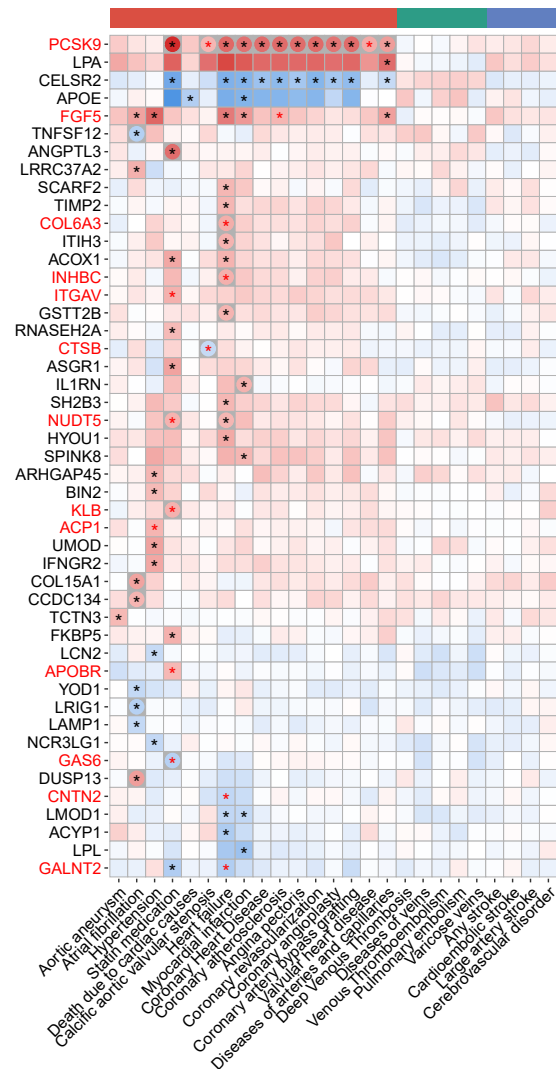


C

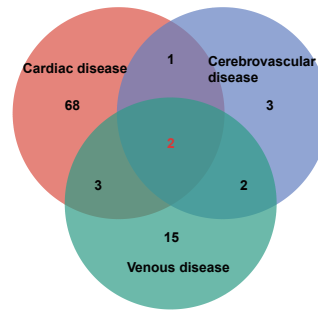
Heart failure



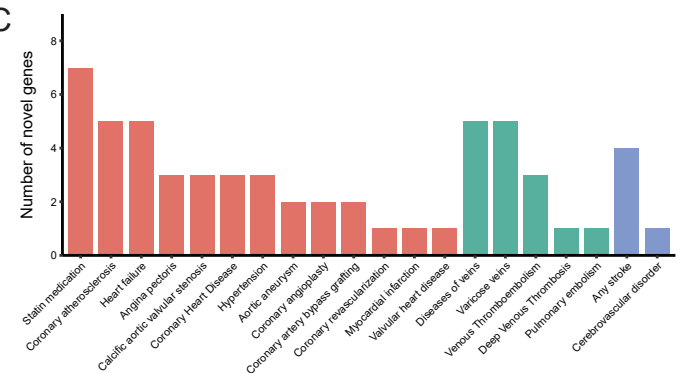
A



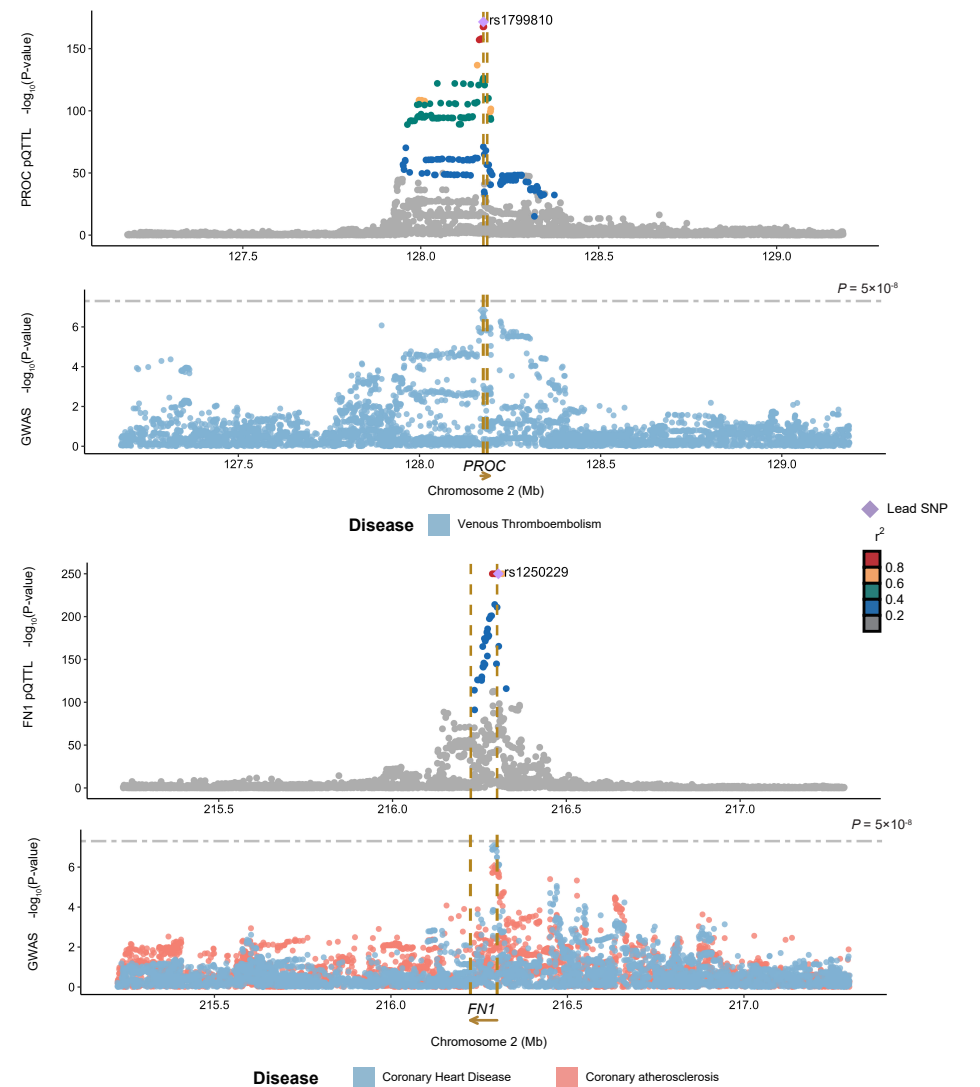
B



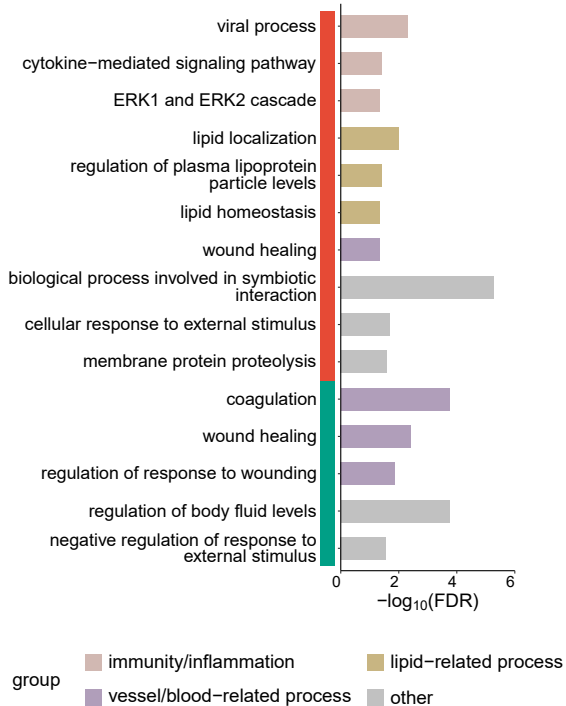
C



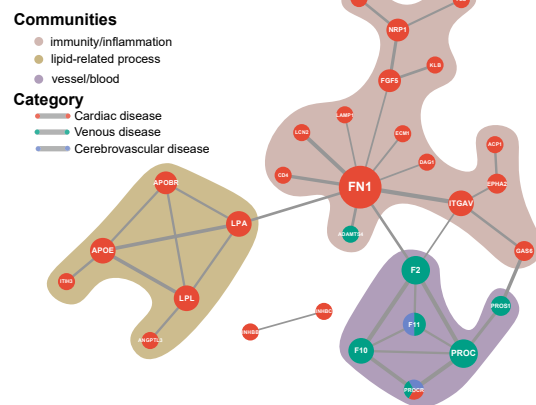
D



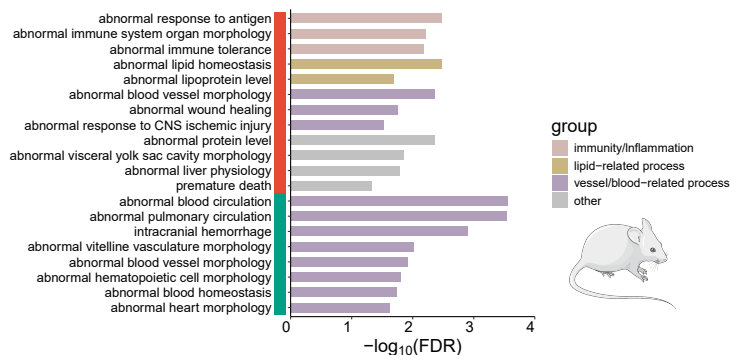
A



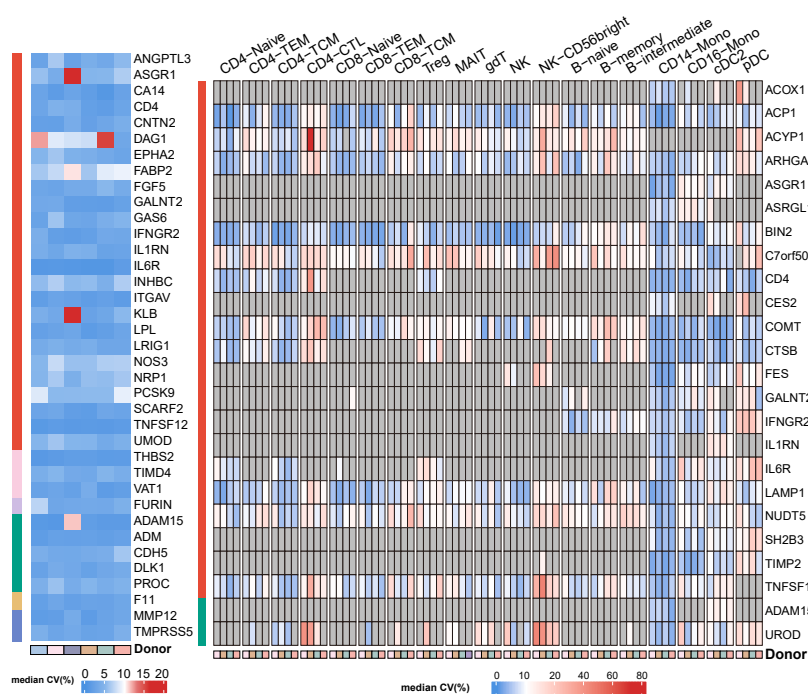
B



C



D



E

