Shared Genetic Architecture Contributes to Risk of Major Cardiovascular Diseases

Siim Pauklin (siim.pauklin@ndorms.ox.ac.uk)
University of Oxford https://orcid.org/0000-0001-8367-3670
Jun Qiao
https://orcid.org/0000-0001-7038-4538
Liuyang Cai
Southern University of Science and Technology
Minjing Chang
Shanxi Medical University https://orcid.org/0000-0001-8875-9254
Can Wang
Shanxi Medical University
Rong Zhao
The Second Hospital of Shanxi Medical University https://orcid.org/0000-0003-2918-0247
Shan Song
The Second Hospital of Shanxi Medical University https://orcid.org/0000-0002-7460-9252
Ning Tan
Guangdong Provincial People's Hospital https://orcid.org/0009-0001-0589-5822
Pengcheng He
Guangdong Provincial People's Hospital https://orcid.org/0000-0001-7706-7105
Lei Jiang
Guangdong Provincial People's Hospital https://orcid.org/0000-0002-9509-370X
Yuliang Feng
Southern University of Science and Technology

Article

Keywords:

Posted Date: September 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3261702/v1

Abstract

The extensive co-occurrence of cardiovascular diseases (CVDs), as evidenced by epidemiological studies, aligns with positive genetic correlations identified in comprehensive genetic investigations. However, the precise nature and mechanisms governing these multifaceted effects remain elusive. By assessing genome-wide and local genetic correlations, polygenic overlaps, and causal connections, we aimed to shed light on common genetic underpinnings among major CVDs. Employing a multi-trait analysis, we pursued diverse strategies to unveil shared genetic elements. These encompassed genomic loci, single-nucleotide polymorphisms (SNPs), genes, biological pathways, functional categories, and protein targets with pleiotropic implications. Our study confirmed elevated genetic resemblance across CVDs and pinpointed 40 genomic loci with pleiotropic influence across multiple CVDs. Notably, 11 of these loci presented consistent evidence from both Metasoft and HyPrColoc's multitrait colocalization analyses, displaying congruent directional effects. Examination of genes linked to these genomic loci unveiled robust associations with circulatory system development processes. Intriguingly, distinct patterns predominantly driven by atrial fibrillation, coronary artery disease, and venous thromboembolism underscored significant disparities between clinically-defined CVD classifications and underlying biological mechanisms. In summary, our findings provide invaluable insights into shared genetic mechanisms spanning CVDs. This knowledge holds potential to guide a biologically-informed restructuring of cardiovascular nosology and innovative therapeutic advancements.

Introduction

Cardiovascular diseases (CVDs), principally coronary artery disease and stroke, are the leading global causes of mortality and disability[1]. These diseases exhibit a complex genetic pattern, implying that children of parents with CVDs are at an increased risk of developing a wide range of CVDs, not just the specific disease their parents had[2,3]. Furthermore, the presence of multiple simultaneous diseases, known as comorbidity, is a common occurrence among CVDs.

Recent large-scale genomic studies have established substantial influence of genetic variation on the risk of various CVDs[4]. These diseases are highly polygenic, with each gene imparting a small genetic effect that cumulatively affects disease risk in concert with environmental risk factors. In combination with environmental risk factors, these genes cumulatively impact the risk of disease. Genome-wide association studies (GWASs) have pinpointed numerous genetic risk loci for major CVDs such as atrial fibrillation (AF), coronary artery disease (CAD), venous thromboembolism (VTE), heart failure (HF), peripheral artery disease (PAD), and stroke[5–10]. Strikingly, these loci often exert effects across multiple phenotypes, revealing a remarkable interconnectedness among CVDs. However, while pleiotropic effects are widely recognized, their systematic presence and downstream consequences within the context of CVDs have yet to be comprehensively assessed[11].

In addition, the common genetic variants identified to date only account for a small fraction of the overall genetic contribution to disease risk. The ongoing quest for new risk loci associated with traits necessitates continually expanding sample sizes. Unfortunately, challenges in recruiting large study populations have notably hindered variant discovery for certain traits, such as PAD, which lags behind more common CVDs like CAD and Stroke[12]. In light of these challenges, the concept of meta-analysis has emerged, combining diverse cohorts with similar or genetically correlated traits to amplify the study sample size[13]. Recent statistical advancements, exemplified by Multitrait analysis of genome-wide association analysis (MTAG)[14], enhance the capacity to detect pleiotropic variants associated with multiple diseases, thereby unraveling the genetic basis for shared risk across diseases.

To date, cross-disease GWAS meta-analyses have predominantly focused on psychiatric and autoimmune disorders[15–20]. Given the pervasive comorbidities observed in epidemiological studies and the existing evidence indicating a shared genetic background across multiple CVDs, the need for a comprehensive large-scale analysis becomes evident.

In the present study, we leveraged MTAG to conduct an integrative analysis of the largest available GWAS datasets encompassing six major CVDs in individuals of European ancestry: AF, CAD, VTE, HF, PAD, and Stroke, which included the identification of novel risk loci and gene prioritization, to characterize the underlying biology of the novel risk loci in the context of CVDs. Furthermore, we explored pleiotropic effects within these extensively expanded datasets generated by MTAG analysis, specifically investigating cross-trait-associated single-nucleotide polymorphisms (SNPs), genes, biological pathways, and protein targets across a participant count averaging over 1.2 million participants. This study addressed four central inquiries regarding the shared genetic basis of these six CVDs (Fig. 1): i) Can we identify shared genetic architectures amidst the diverse landscape of these clinically distinct CVDs? ii) Can we detect additional genomic loci for multiple CVDs (i.e. pleiotropic loci), and whether some of loci exhibit opposite allelic effects across CVDs? iii) Can we identify functional features of the pleiotropic loci that could account for their widespread impact on cardiovascular pathology? iv) Can we identify causal plasma proteins as possible therapeutic targets spanning major CVDs? Collectively, these findings will ultimately reshape our understanding of cardiovascular nosology, spotlight potential cardiovascular physiological mechanisms predisposing to specific clinical presentations, and provide crucial insights into the prevention and treatment of CVDs.

Results

Shared heritability among the six CVDs

Following the harmonization and filtering of SNPs shared between GWAS summary statistics, we used LDSC to calculate SNP-based heritability (h^2_{SNP}) and genome-wide genetic correlations (rgs). The power of GWASs and the strength of their genetic signal varied considerably among the six CVDs, with the median h^2_{SNP} being 1.4% (range: 0.6%-3.2%, Fig. 2a and Supplementary Table 2a). Estimated r_gs were moderate to strong (median = 0.435, range: 0.148–0.677) and always positively significant at a Bonferroni-corrected significance level of $P = 3.33\times10^{-3}$ (Fig. 2b and Supplementary Table 2b). Substantial r_gs were observed both between pairs of CVDs with low h^2_{SNP} (for example, for heart failure [HF] and Stroke, h^2_{SNP} is 0.6% and 0.8%; $r_g = 0.552$) as well as relatively high h^2_{SNP} (for example, for atrial fibrillation [AF] and coronary artery disease [CAD], h^2_{SNP} is 2.5% and 3.2%; $r_g = 0.197$).

Among the six CVDs, HF and CAD exhibited augmented estimated higher estimated polygenicity, indicative of more extensive involvement of genetically correlated genomic regions compared to other CVDs. Conversely, PAD displayed superior discoverability while exhibiting less pronounced polygenicity in relation to other CVDs. Consequently, PAD displayed relatively fewer genomic loci causally associated with varying effect sizes, in contrast to other CVDs (Supplementary Table 3b). Notably, the estimated sample size needed to achieve 90% h^2_{SNP} was over 24 times smaller for PAD compared to HF (Supplementary Fig. 1). Overall, there were substantial polygenic overlaps among the six CVDs with the median Dice coefficient being 0.362 (range: 0.207–0.732, Fig. 2c, Supplementary Fig. 2, and Supplementary Table 3b), both in the presence of moderate positive genetic correlation (CAD and HF) and minimal genetic correlation (AF and PAD). For example, the polygenic overlap between CAD and HF was particularly striking, with 1,397 (SD = 254) shared variants, representing 93.3% variants influencing CAD and 60.9% variants influencing HF, mirroring their robust positive genetic correlations (rg = 0.699, SE = 0.0129). Comparatively fewer shared variants (74, SD = 11) were noted between AF and PAD, representing 14.0% and 18.9% of variants influencing AF and PAD, respectively, with a weak positive genetic correlation (rg = 0.140, SE = 0.0150).

To identify genomic regions harboring variants associated with multiple CVDs, local genetic correlations were calculated within regions displaying univariate signals for CVDs using the LAVA approach. Among 896 defined genomic regions, 495 (55.25%) exhibited univariate genetic signals ($P < 1 \times 10^{-4}$) for multiple CVDs, with 21 regions (2.34%) displaying signals for more than half of the CVDs (Fig. 2d and Supplementary Table 4). Yet, only 24 distinct genomic regions reached significance for local genetic correlations after Bonferroni correction, and merely 2 regions displayed significant correlation for more than two CVDs (P = 0.05 / no. of bivariate tests = 0.05 / 577 = 8.67 × 10 − 5; Fig. 2e and Supplementary Table 5).

The investigation extended to colocalization analysis through GWAS-PW to evaluate whether paired CVD associations emanated from the same signal. GWAS-PW identified 196 unique genomic regions suggestive of variants influencing pairs of CVDs (PPA_3 > 0.9, Supplementary Table 6), signifying shared regions and causal variants. Notably, 50 unique genomic regions demonstrated robust for loci shared by more than two traits. Strikingly, these regions encompassed many genomic loci not deemed significant in the bivariate LAVA analysis of CVD pairs. Meanwhile, only 3 unique genomic regions (chr4: 109978983–111733579, chr9: 21674033–23091193, and chr9: 136042491–136770926) were significantly captured by both bivariate LAVA and GWAS-PW (Supplementary Fig. 3).

Subsequently, the construction of the latent causal variable (LCV) model allowed the estimation of partial genetic causality between significantly correlated CVD pairs. Convincing evidence (|GCP| > 0.6, P < 0.05 / no. of CVD pairs = 0.05 / 15 = 3.33×10^{-3}) for partial genetic causality emerged for two CVD pairs (Supplementary Table 7). By considering the sign of the genetic correlation, it was plausible to infer that CVD could elevate HF risk, while VTE might have an adverse impact on CVD. Notably, the LCV results hinted at a partial causal influence of VTE (GCP = 0.649, $P = 6.31 \times 10^{-3}$) on Stroke, although it did not withstand multiple testing corrections ($P < 3.33 \times 10^{-3}$). Furthermore, a trait pair (AF-HF, GCP = 0.575, $P = 1.07 \times 10^{-3}$) exhibited trends towards partial genetic causation, suggesting AF might partially influence VTE, but fell short of the stringent GCP threshold (|GCP| > 0.6). To validate the reproducibility of partial causal association findings for these four CVD pairs, LHC-MR method was employed, yet all trait pairs demonstrated bias due to reverse causality (Supplementary Fig. 4 and Supplementary Table 8).

In all, these findings reinforced the genetic resemblance among CVDs, implicating the plausible presence of shared SNPs, genes, or biological pathways contributing to general CVD predisposition.

Functional annotation of the MTAG-identified loci

In previous studies, cross-trait meta-analysis has proven valuable in augmenting the statistical power to uncover pleiotropic SNPs, shared genetic underpinnings, and common biological annotations across distinct psychiatric disorders and immunological diseases. In line with this, we undertook cross-trait GWAS meta-analyses employing the MTAG methodology, with the overarching objective of enhancing the identification of novel genetic associations for each trait by pinpointing pleiotropic SNPs that contribute to the genetic susceptibility of cardiovascular diseases. As a result, we observed substantial augmentation in the maximal effective sample sizes, with increments from1,030,836 to 1,076,012 for AF; 1,165,690 to 1,237,236 for CAD; 1,500,861 to 1,631,058 for VTE; 977,323 to 1,219,100 for HF; 511,634 to 843,568 for PAD; and 1,308,460 to 1,735,909 for Stroke (Table 1 and Fig. 3a-f). Importantly, examination of heritability Z-scores for MTAG results indicated greater polygenic signals compared to the original GWAS outcomes. Furthermore, the LDSC intercepts were approximately one, suggesting that the increase in mean χ2 statistics was due to polygenicity and not due to stratification or other confounding biases. The Q-Q plots and λGC of all MTAG and GWAS results are shown in Supplementary Figs. 5 and 6. The λGC values for AF, CAD, VTE, HF, PAD, and Stroke ranged from 1.096 to 1.421 in the GWAS results, and from 1.014 to 1.446 in the MTAG results, suggesting no evidence of inflation was found. Supplementary Note 2 and 3 listed Detailed information about novel genomic loci and mapped genes based on MTAG results of six CVDs (Supplementary Figs. 7 and 8, Supplementary Table 11–18).

Table 1
Statistical summary of GWAS and MTAG results.

	AF		CAD		VTE		HF		PAD		Stroke
	GWAS	MTAG	GWAS	MTAG	GWAS	MTAG	GWAS	MTAG	GWAS	MTAG	GWAS
mean χ2	1.5451	1.4888	1.7648	1.8020	1.5698	1.4830	1.1591	1.2062	1.1025	1.1253	1.2280
h2	0.0250	0.0245	0.0324	0.0345	0.0182	0.0183	0.0080	0.0118	0.0094	0.0157	0.0060
λGC	1.2932	1.2697	1.4210	1.4460	1.3203	1.2498	1.1270	1.1555	1.0957	1.0135	1.1876
Intercept	1.0538	1.0025	1.0218	1.0067	1.0511	0.9501	1.0073	0.9797	1.0157	0.9795	1.0738
Ratio	0.0978	0.0052	0.0302	0.0083	0.0898	< 0	0.0416	< 0	0.1560	< 0	0.3245
Sample size	1,030,836	1,076,012	1,165,690	1,237,236	1,500,861	1,631,058	977,323	1,219,100	511,634	843,568	1,308,460
Total SNPs	6,925,287	5,782,223	6,925,287	5,782,223	6,925,287	5,782,223	6,925,287	5,785,906	6,925,287	5,785,906	6,925,287
nSig SNPs	10,784	9,422	15,789	13,739	9,070	7,155	272	540	92	562	941
nGenomic locus	117	120	203	194	91	80	11	13	4	48	24
nLead SNPs	749	668	1049	992	765	646	21	42	8	78	50
nInd Sig SNPs	224	214	362	339	222	196	12	14	5	55	25
nGenes mapped by position	397	375	662	642	447	399	38	47	21	73	67
nGenes mapped by MAGMA	298	281	551	558	401	334	20	49	5	27	43
nGenes mapped by eqtl	352	329	647	651	345	319	30	39	7	59	56
nGenes mapped by TWAS	155	147	284	295	204	189	11	14	5	15	16

mean χ2, Mean chi-squared value; h2, SNP-based heritability; λGC, LambdaGC of 1 indicates no inflation of the median χ2 association statistic; Intercept, LD intercept of 1 indicates that inflation is not due to population stratification biases; Ratio, (intercept-1) / (mean chi-squared value-1), the ratio indicates what pe the inflation can be ascribed to other sources than polygenicity; nSig SNPs, Number of genome-wide significant SNPs ($P < 5 \times 10^{-8}$); nGenomic locus, Numbe genomic locus; nLead SNPs, Number of lead SNPs in the genomic locus ($P < 5 \times 10^{-8}$ and $r^2 < 0.1$); nInd Sig SNPs, Number of the independent significant SNI genomic locus ($P < 5 \times 10^{-8}$ and $r^2 < 0.6$); nGenes mapped by position, Number of genes mapped by position using FUMA; nGenes mapped by MAGMA, Numl mapped by MAGMA; nGenes mapped by eqtl, Number of genes mapped by eqtl using FUMA; nGenes mapped by TWAS, Number of genes mapped by TWAS.

Characterizing shared biological mechanisms among CVDs

In light of the challenges posed by prior GWAS results in discerning shared biological mechanisms among CVDs, our subsequent efforts involved a comprehensive assessment of genetic overlap within the MTAG results of the six CVDs, spanning genomic loci, SNPs, genes, and pathways.

Among the noteworthy observations, a total of 40 pleiotropic loci showed genetic signals for multiple CVDs, with 9 (22.0%) demonstrating this phenomenon for more than half of the considered CVDs (Supplementary Table 19). However, when testing whether these overlapping loci shared causal variants through multitrait colocalization analysis from HyPrColoc, only 13 (3.32%) regions exhibited strong colocalization evidence (PP > 0.70). Out of these, 11 regions were corroborated by Metasoft analysis (Supplementary Fig. 9).

The most prominent pleiotropic locus we identified is in an intron of the lipoprotein(a) gene (*LPA*) on 6q25.3. This locus, encompassing the shared causal SNP rs10455872, demonstrated colocalization across all CVDs except MTAG_VTE. Notably, this causal variant has been established as linked to elevated lipoprotein (a) (Lp[a]) levels, in addition to an increased risk of CVDs across the general population[21]. Although this region displayed associations solely with MTAG_CAD and MTAG_PAD ($P_{meta} = 1.11 \times 10^{-143}$, Fig. 4a) through MetaSoft analysis, it underscored the significant influence of this locus. Another notable region, 7p21.1 surrounding SNP rs8084351, located at the intergenic region of twist family bHLH transcription factor 1 gene [*TWIST1*] and histone deacetylase 9 [*HDAC9*]), exhibited the second-most pronounced pleiotropic association. This region demonstrated a strong connection with prevalent vascular diseases, including MTAG_CAD, MTAG_PAD, and MTAG_Stroke)($P_{meta} = 2.20 \times 10^{-41}$, Fig. 4b). The signal in our analysis was associated with artery aorta eQTLs for *TWIST1* rather than *HDAC9* (eQTL association FDR for *TWIST1* = 5.54×10^{-4}), supporting *TWIST1* as a plausible candidate gene. Tests have demonstrated

that the rs2107595 locus can regulate the expression of *TWIST1* and provide evidence for a functional role in vascular smooth muscle cells[22]. Further exploration led to the identification of the most pleiotropic novel locus at the11p11.2 region, positioned upstream of hydroxysteroid 17-beta dehydrogenase 12 gene (*HSD17B12*). This locus exhibited association with MTAG_CAD, MTAG_HF, and MTAG_Stroke. Although it fell short of the stringent PP threshold (PP = 0.626, less than 0.7), the shared causal SNP rs7944241 showcased noteworthy significance (P_{meta} = 6.27× 10^{-28}, Fig. 4c). *HSD17B12*, a gene predominantly expressed in platelets, plays a pivotal role in lipid metabolism and fatty acid biosynthesis[23].

A pivotal aspect of our analysis encompassed investigating the presence of opposite directional effects of shared causal variants among the 13 pleiotropic loci across various diseases. We identified only one locus with evidence of antagonistic effects on two or more CVDs. This locus, encompassing rs28929474, an exonic variant located at the Z-allele of serpin family A member 1 gene (*SERPINA1*) within the 14q32.31 region, showed opposing directional effects on MTAG_CAD and MTAG_VTE. However, this result did not surpass the significant threshold in Metasoft analysis (P_{meta} = 0.260 > 0.05; Fig. 4d, Supplementary Fig. 10). Notably, all remaining 12 loci exhibited the same directional effect on diseases, including 10 susceptible loci and 2 protective loci, in line with their strong genome-wide genetic correlation.

Of the 819 Lead SNPs, 1,063 susceptibility genes by MAGMA, and 588 tissue-specific genes by TWAS that achieved genome-wide significance (GWS) for any of the six CVDs, only a fraction of these demonstrated significance across multiple CVDs. Specifically, only 92 (11.23%) Lead SNPs (P< 5 × 10^{-8}; Supplementary Table 20), 160 (15.05%) susceptibility genes (P< 2.84 × 10^{-6}), and 73 (12.41%) tissue-specific genes exhibited GWS for more than one CVD. Of these, only 18 (2.20%) Lead SNPs, 20 (1.88%) susceptibility genes, and 4 (0.68%) tissue-specific genes reached GWS for more than half of CVDs (Supplementary Tables 18 and 20). In conclusion, the observation that more than one CVD showed genetic signals in the given overlapped region with high-supporting evidence of colocalization is further supported by direct comparisons of SNP- and gene-based results among the six CVDs.

While the observation of a relatively limited number of overlapping genes and regions among cardiovascular diseases is apparent, the distinct genes associated with these conditions could potentially converge within shared biological pathways or display enrichment in similar tissues or functional categories. To explore this notion, we undertook a comprehensive analysis, encompassing gene set assessment based on MAGMA-derived susceptibility genes, pathway enrichment analysis using tissue-specific genes from Metascape, and a closer examination of tissues and functional characteristics utilizing SNP-based heritability data from LDSC-SEG. Following meticulous correction for 9,398 tests across the six CVDs (that is, 7,744 and 1,654 gene sets for GO BP and Reactome, respectively; αBON = .05 / 9,398 / 6 = 8.87 × 10^{-7}; Supplementary Table 21), we observed little convergence among sets of CVDs for gene sets, of which only 3 gene sets (involved in circulatory system development, circulatory system process, and heart development) were enriched in more than one CVD and no gene set showed significant enrichment for more than two traits. Subsequently, more lenient enrichment tests of gene function using Metascape confirmed little convergence across CVDs' tissue-specific genes (Supplementary Table 22). Similarly, After correcting separately for tissues or functional categories and six CVDs (that is, 49 tissues and 489 functional categories; αBON for tissues = .05 / 49 / 6 = 1.70 × 10^{-4}; αBON for functional categories = .05 / 489 / 6 = 1.70 × 10^{-5}; Supplementary Figs. 11 and 12 and Supplementary Tables 23 and 24), no test surpassed the significance threshold. Taking a less stringent approach, only correcting separately for tissues or functional categories, there were still no convergence among CVDs for tissues and functional categories. To assess whether certain functional genomic categories (i.e. regulatory or functional features in specific tissues or cell types) show enrichment among multiple CVDs, we performed functional enrichment analysis using GARFIELD. After correction for the 1,005 tested functional categories across all CVDs (αBON = .05 / 1,005 / 6 = 8.29 × 10^{-6}; Supplementary Fig. 13 and Supplementary Table 25), much more significant enrichment of genetic signals was observed for exon region as well as for 217 (21.6%) DNaseI hypersensitive sites, 69 (6.87%) chromatin accessibility peaks, 34 (3.38%) histone-modified regions, 16 (1.59%) transcription-factor footprints, 8 (0.796%) chromatin states, 1 TFBS, and 1 FAIRE of different tissues or cell types in more than one CVD except MTAG_HF, MTAG_PAD, and MTAG_Stroke. This finding suggested that a considerable part of our characterizing shared signals of functional categories was primarily driven by AF, CAD, and VTE, the CVDs with the strongest genetic signals, hindering interpretation of the characterizing shared signals of functional categories as representing 'general liability to CVDs'. Similarly, Of the 92 Lead SNPs, 160 susceptibility genes by MAGMA, and 73 tissue-specific genes by TWAS that were GWS for more than one CVD, 57 (61.96%) Lead SNPs, 58 (36.25%) susceptibility genes, and 53 (72.60%) tissue-specific genes were GWS for more than one CVD except MTAG_HF, MTAG_PAD, and MTAG_Stroke.

Ultimately, these findings corroborate our hypothesis, indicating that the analytic signal characterizing shared biological mechanisms was notably influenced by AF, CAD, and VTE, rather than universally representing the genetic variance inherent in multiple CVDs.

Causal proteins identified by Proteome-wide MR and Colocalization analysis

We conducted an in-depth examination of the MR associations between 1,773 proteins with available index cis-acting variants (cis-pQTL) and the risk of CVD outcomes using the SMR approach. Through this analysis, we identified 1202 unique protein-cardiovascular-disease pairs at the marginal significance level (P < 0.05 for SMR analysis), including 223 protein-MTAG_AF pairs, 323 protein-MTAG_CAD pairs, 265 protein-MTAG_VTE pairs, 148 protein-MTAG_HF pairs, 105 protein-MTAG_PAD pairs, and 138 protein-MTAG_Stroke pairs (Supplementary Table 26). Subsequently, following the removal of associations that did not pass the HEIDI test, as well as employing sensitivity analysis using multi-SNPs-SMR and multiple testing corrections with a 5% false discovery rate (FDR), we identified a total of 173 proteins associated with CVDs. Among these, 20, 70, 47, 1, 1, and 10 proteins were significantly associated with MTAG_AF, MTAG_CAD, MTAG_VTE, MTAG_HF, MTAG_PAD, and MTAG_Stroke, respectively (Fig. 5a and Supplementary Table 26). Further investigation was conducted to ascertain whether the circulating proteins identified in the analysis shared causal variants with CVDs. Strong evidence of colocalization was observed between 43 proteins and CVDs (Supplementary Fig. 14 and Supplementary Table 26). More specifically, MTAG_AF demonstrated high support for colocalization with 6 proteins including LRIG1, GUSB, DUSP13B, SPON1, CFL2, and TNFSF12. For MTAG_CAD and MTAG_VTE, we found 15 and 17 proteins exhibited robust colocalization evidence. Likewise, MTAG_HF, MTAG_PAD, and MTAG_Stroke demonstrated high support for colocalization with 1, 1, and 3 proteins, respectively, including SWAP70 for MTAG_HF, PCSK9 for MTAG_PAD, and HINT1, TMEM106B, and SWAP70 for MTAG_Stroke. A total of 12 proteins were identified in more than one CVD, of these, 4 proteins showed strong evidence of colocalization including CALB2 and F2 for MTAG_CAD and MTAG_VTE, SWAP70 for MTAG_HF and MTAG_Stroke, and TMEM106B for MTAG_CAD and MTAG_Stroke. Only F2 for MTAG_CAD and MTAG_VTE were validated with

high support of evidence (PP > 0.70) using multitrait colocalization analysis from HyPrColoc, where index SNP rs3136516 (an intronic variant in the *F2* gene) was also identified as the shared causal variant (Fig. 5b-d).

Druggability of identified causal proteins

To leverage this genetic information into new potential therapeutic targets for CVDs, we investigated the potential druggability of 39 causal proteins pinpointed in colocalization analysis. Our investigation unearthed a collection of 39 distinct drugs that target 7 unique proteins, as outlined in Supplementary Table 27. Of these drugs, a notable subset consisted of 11 medications designed to target 3 unique proteins specifically tailored for the treatment of the corresponding CVDs. This includes F2, which pertains to for both MTAG_CAD and MTAG_VTE, as well as F11 for MTAG_VTE, and PCSK9 for MTAG_PAD. A prominent example is the drug Dabigatran/Dabigatran etexilate, a direct thrombin inhibitor that targets coagulation factor II (F2). This agent has garnered approval for treating various CVDs such as VTE and CAD[24]. However, it's important to note that despite Dabigatran's favorable risk-benefit profile, it might be associated with major bleeding events, including gastrointestinal bleeding. For the coagulation factor XI inhibitor targeting F11, namely Abelacimab and Milvexian, evidence suggests heightened effectiveness and safety in treating VTE[25]. Additionally, monoclonal antibodies Alirocumab and Evolocumab, both PCSK9 inhibitors, have gained approval for reducing low-density lipoprotein (LDL) cholesterol levels, thereby safeguarding against an array of CVDs, including PAD[26]. Drugs initially designed to address autoimmune conditions, targeting SYK and TNFSF12, also emerge as potential therapeutic avenues for VTE[27,28]. Despite a lack of drug-related data for 32 unique proteins within the OpenTargets database, a compelling 18 (56.25%) of them present as promising candidates for future treatment prospects, exhibiting highly reliable druggability tiers. Whether agents modulating these proteins could be repurposed for the prevention or treatment of CVDs still needs to be examined in clinical trials.

Discussion

Our study represents a significant advancement in the understanding of cardiovascular diseases, leveraging publicly available summary statistics to conduct the largest cross-disease GWAS meta-analysis of CVDs to date. By pooling data from over 1.2 million participants across six distinct CVDs, we meticulously examined the overlap in genomic loci, SNPs, genes, gene sets, tissue types, functional categories, and protein targets among these conditions. This comprehensive approach unveiled locally shared genetic mechanisms that might otherwise have evaded detection, yielding four key insights into the shared genetic underpinnings of CVDs.

First, our findings demonstrated extensive genome-wide and local genetic correlations, along with polygenic overlaps, across different CVDs, supported by various methods such as LDSC, LAVA, GWAS-PW, and MiXeR. Notably, our Mendelian randomization analyses utilizing LCV and LHCMR illuminated that the relationship between most pairs of CVDs is potentially mediated by horizontal pleiotropy-a shared genetic basis-rather than a causal vertical pleiotropy. This points to a substantial genetic correlation between multiple CVDs, suggesting a broader genetic framework underlying the risk of cardiovascular pathology. To overcome limitations in previous cross-disease genetic overlap research, we employed MTAG analysis to enhance power for detecting pleiotropic variants, which allowed us to elucidate the genetic foundation for shared risk across diseases. Remarkably, we identified novel genomic loci for various CVDs, significantly expanding the effective sample sizes for these conditions.

Second, our variant-level analyses lent strong support to the existence of substantial pleiotropy, with approximately 11.23% of Lead SNPs influencing more than one of the examined CVDs. Intriguingly, we identified 13 loci displaying particularly extensive pleiotropic profiles across CVDs, corroborated by multitrait colocalization analysis from HyPrColoc. Among these, the most highly pleiotropic locus was *LPA*, known for encoding apolipoprotein(a), a key component of Lp(a), strongly associated with increased CVD risk[21,29,30]. This exemplifies the complex genetic interplay underlying related diseases. Notably, we also detected a locus with opposing effects on two CVDs, highlighting the intricate genetic relationships that extend beyond overall genetic correlations.

Third, we unearthed extensive evidence linking cross-disease genetics of CVDs to circulatory system development. Alongside *LPA*, pleiotropic genes like *IL6R* and *SMAD7* further highlighted the connection between pleiotropy and genetic effects on circulatory system development[31–34]. Our functional enrichment analyses using GARFIELD underscored that the shared functional categories signal was chiefly driven by AF, CAD, and VTE, confirming the two distinct groups of CVDs: those characterized by more genetic signals and shared biological mechanisms (AF, CAD, and VTE), and those exhibiting the opposite (HF, PAD, and Stroke). This observation challenges the classical categorical classification of CVDs.

Fourth, our exploration delved into the causal roles of circulating proteins in CVDs through proteome-wide MR and colocalization analyses. We identified numerous circulating proteins with strong causal evidence, including F2, which emerged as a potential drug target for VTE with CAD.

However, our study has certain limitations. While our dataset is extensive, larger samples could enhance power for detecting cross-disorder effects. We addressed sample overlap and accounted for comorbidity among diseases. Further datasets are required to expand the analyses of genetic effects across CVDs. Additionally, our study predominantly focused on individuals of European ancestry and more research is needed to validate findings in other populations. Moreover, the applicability of our findings to non-European populations remains uncertain. Finally, while GWAS designs capture common variant aspects of genetic architecture, studies examining copy-number variants and rare mutations are needed for a comprehensive understanding.

In conclusion, our study sheds light on the intricate genetic relationships underlying various cardiovascular diseases, offering insights into shared genetic mechanisms, pleiotropy, and the significance of circulatory system development. This work not only enhances our understanding of CVD genetics but also highlights potential therapeutic avenues and the need for further research to unravel the complex interplay of genetics and environment in cardiovascular health.

Methods

Study Design

Figure 1 presents the workflow for this study.

Data Sources and Quality Control

For this study, we collected GWAS summary statistics for six major CVDs. Detailed information about these diseases and their publication sources is available in Supplementary Table 1 and Supplementary Note 1[5–10]. Only data from European ancestry populations were used. Pre-analysis quality control steps were performed on summary statistics, including (1) aligning to 1000 Genomes Project v3 Europeans reference of hg19 genome build; (2) excluding non-autosomal SNPs; (3) filtering out SNPs without rsID or with duplicated rsID; (4) keeping only biallelic SNPs with minor allele frequency (MAF) > 0.01. To assure fair and interpretable comparisons across the six CVDs, we filtered all summary statistics to include only those SNPs that were available for all six CVDs, i.e., 6,925,287 SNPs. Additional data processing procedures were also carried out according to the corresponding requirements of different methods in subsequent analysis.

Estimating SNP-based heritability and genome-wide genetic correlation using LDSC

We used univariate linkage disequilibrium (LD) score regression (LDSC) to estimate the SNP-based heritability (h^2_{SNP}, i.e. the proportion of the phenotypic variance in a trait can be explained by common genetic variants included in the analysis) of each of the six CVDs[35]. All GWAS summary statistics were reformatted to the pre-computed LD scores of the 1000 Genomes European reference (https://github.com/bulik/ldsc, v1.0.1). SNPs were excluded if they did not intersect with the reference panel, or if they were located in the MHC region (CHR:6 26–35 Mb), had a MAF < 1% or INFO score < 0.3. Bivariate LDSC was used to compute genetic correlations (rgs, i.e. the proportion of genetic variance shared by two traits divided by the square root of the product of their SNP-based heritability estimates) between the six CVDs summary statistics, with Bonferroni-corrected significant threshold set at $P < 3.33 \times 10^{-3}$ (.05/15)[36]. Bivariate LDSC, like single-trait LDSC, only requires summary statistics and produces estimates that are unbiased by sample overlap.

Estimating polygenicity and polygenic overlap using MiXeR

We used univariate Gaussian causal mixture modeling method (MiXeR) analysis to estimate the number of trait-influencing variants (also referred to as “causal” variants) for each trait (i.e., polygenicity) and the average magnitude of additive genetic associations among these variants (i.e., discoverability)[37]. As recommended, we used the 1000 Genomes Project v3 for European samples as a reference panel (https://github.com/precimed/mixer, v1.3) and excluded the MHC region (CHR:6 26–35 Mb). Bivariate analysis models were used to determine additive genetic effects as a mixture of four bivariate Gaussian components: (a) variants not influencing either trait; variants uniquely influencing either the (b) first; or (c) second trait (unique disease-influencing variants); and (d) variants influencing both traits (shared disease-influencing variants)[38]. Results were presented as Venn diagrams displaying the proportion of shared and unique variants that explained 90% of SNP heritability for each GWAS. To evaluate polygenic overlap, MiXeR was implemented to calculate the Dice coefficient (i.e., the ratio of shared variants to the total number of variants). Model fit evaluated via the Akaike information criterion (AIC) was based on the maximum likelihood of GWAS z scores and was visualized by modeled versus observed conditional quantile-quantile (Q-Q) plots.

Estimating local genetic correlation using LAVA and GWAS-PW

We used univariate local analysis of [co]variant association (LAVA) to test for local genetic signals within each of CVD (that is, the local h^2_{SNP}). We used the genomic regions defined as autosomal LD blocks (N = 2,495) by *Werme* et al.[39], which are characterized by having minimum LD across regions, a minimum of 2,500 variants included on each LD block, and with an average LD block size of 1 million bases (Mb). The LD reference panel based on the 1000 Genomes Project v3 for European samples was used following the protocol described in the original article. The detection of valid and interpretable local rg requires the presence of a sufficient local genetic signal. To this end, we used a threshold of $P < 1 \times 10^{-4}$ so as to filter out clearly nonassociated loci (or loci clearly devoid of any univariate heritability) that nonetheless may contribute to the overall h^2_{SNP} to get a more global overview of the shared genetic signal. This resulted in a total of 577 conducted bivariate tests as several tests will be conducted within the same region when multiple CVDs show univariate signals in the same region. *P*-values of the local genetic correlations were corrected for the total number of bivariate tests conducted (threshold: $P = 0.05$ / no. of bivariate tests = $0.05 / 577 = 8.67 \times 10^{-5}$). We also considered sample overlap across GWAS datasets in the analysis, by including the pair-wise genetic covariance estimated by bivariate LDSC and further standardizing it into a correlation matrix.

We utilised a Bayesian pleiotropy association test implemented in the pairwise GWAS (GWAS-PW) to identify shared genomic regions among CVDs[40]. Like LAVA, the 1000 Genomes Project v3 for European samples was used as the LD reference panel, and regions were constructed by partitioning the genome into 2,495 semi-independent blocks of ~ 1 Mb. For each region, GWAS-PW method estimates the posterior probability (PPA) by modelling the probabilities that (i, ii) the locus is associated with either the first (PPA_1) or second trait (PPA_2); (ii) the locus is associated with both traits via same causal variant (PPA_3); and (iii) the locus is associated with both traits but through different causal variants (PPA_4). Pairs of CVDs were considered to have genetic correlations at the local region if PPA_3 was larger than 0.9.

Mendelian randomization analysis using LCV and LHCMR

Genome-wide or local genetic correlation or polygenic overlap may reveal important insights into shared biology between two traits; however, this should not be interpreted as implying a causal relationship in either direction. To evaluate evidence for a causal relationship, we performed the Latent Causal Variable (LCV) model to evaluate partial genetic causality between two traits[41]. The LCV approach leverages the bivariate effect size distribution of SNPs in two GWAS and their LD scores to estimate a posterior mean GCP, such that, evidence of partial genetic causality can be distinguished from genetic correlation. Weak GCP estimates close to zero for genetically correlated traits imply that their relationship is potentially mediated by horizontal pleiotropy, whereby there are shared pathways, but the two traits do not likely exhibit vertical pleiotropy by acting within the same pathway. LCV estimates GCP ranging from − 1 to 1 where a value > 0 implies partial genetic causality of trait one on trait two and vice versa. A strong estimate of the posterior genetic causality proportion (GCP) was defined

as significantly different from zero (one-sided t-test) and an absolute GCP estimate > 0.6, with Bonferroni-corrected significant threshold set at $P < 3.33 \times 10^{-3}$ (.05/15). LCV corrects for heritability and genetic correlation between traits and is not limited by sample overlap. Notably, LCV assumes a single latent variable mediating the effect between two phenotypes, and that the effect can be only unidirectional.

Hence it may be confounded by bidirectional causal effects or by the presence of several latent variables. Given these limitations of LCV method, we also performed an additional assessment for pairs of CVDs that exhibited evidence for partial genetic causality from LCV model using the LHC-MR method. The Latent Heritable Confounder MR (LHC-MR), a recently developed Mendelian randomization method, utilizes all genome-wide variants to assess causal estimates rather than genome-wide significant loci only, to improve statistical power and correct for sample overlap, and control for correlated and uncorrelated horizontal pleiotropy[42]. The LHC-MR method extends the standard two-sample MR by assessing bidirectional causal relationships by dividing the association between an exposure and an outcome trait into four different effects: the causal effect of the exposure on the outcome, the causal effect of the outcome on the exposure, the effect of confounders that affect the outcome through the exposure (i.e. vertical pleiotropy), and the effect of confounders that affect the outcome independently of the exposure (i.e. correlated horizontal pleiotropy). Thus, the unbiased bidirectional causal effect between these two traits is estimated simultaneously along with the confounder effect on each trait, which make LHC-MR more precise at estimating causal effects compared to standard MR methods (i.e., MR egger, weighted median, inverse variance weighted [IVW], simple mode, and weighted mode).

Multi-trait analysis of GWAS using MTAG

Multi-trait analysis of GWAS (MTAG) applies generalized inverse-variance-weighted meta-analysis for multiple correlated traits and aims to detect novel genetic associations for each trait by boosting statistical power by borrowing the correlations among correlated traits[14]. Briefly, MTAG takes summary statistics from single-trait GWAS as inputs and produces trait-specific effects for one common set of SNPs, and the resulting *P*-value could be interpreted and used like those in single-trait GWAS. MTAG incorporated bivariate LDSC to account for possibly unknown sample overlap among the GWASs of multiple correlated traits. MTAG relies on the key homogeneous assumption that all SNPs across traits share the equal variance-covariance matrix of effect sizes, but the estimator of MTAG can be still consistent even if this assumption is violated when some SNPs influence only a subset of the traits. We calculate the upper bound for the false discovery rate ('maxFDR') to evaluate the overall inflation due to violation of the key homogeneous assumption. The 'maxFDR' values were 1.656×10^{-3}, 3.092×10^{-4}, 2.798×10^{-3}, 0.134, 0.354, and 0.569 for AF, CAD, VTE, HF, PAD, and Stroke for the first MTAG analysis. The 'maxFDR' values of HF, PAD, and Stroke were greater than 0.05, suggesting the overall inflation due to no violation of the homogeneous assumption. We then reperformed an MTAG analysis of HF, PAD, and Stroke, with new 'maxFDR' values were 0.0398, 0.0419, and 0.0585.

To investigate if violations of the assumptions of equal SNP heritability for each trait and perfect genetic covariance between traits biased our MTAG results, we performed pairwise cross-trait meta-analysis using cross phenotype association (CPASSOC) as a sensitivity analysis[43]. CPASSOC assumes the presence of heterogeneous effects across traits and estimates the cross-trait statistic SHet and *P*-value through a sample size-weighted, fixed-effect meta-analysis of GWAS summary statistics. We denote the summary statistics from single-trait GWAS as GWAS_AF, GWAS_CAD, GWAS_VTE, GWAS_HF, GWAS_PAD, and GWAS_Stroke, respectively, and the summary statistics from MTAG analysis as MTAG_AF, MTAG_CAD, MTAG_VTE, MTAG_HF, MTAG_PAD, and MTAG_Stroke. Independent SNPs that were genome-wide significant ($P < 5 \times 10^{-8}$) in the cross-trait meta-analyses using both MTAG and CPASSOC were considered to be further verified.

Genomic loci definition and functional annotation using FUMA

FUMA is an online platform that annotates SNPs to their biological functionality and maps implicated genes[44]. We performed FUMA annotation with default settings and used the 1000 Genomes Project v3 of European samples as a reference panel. Before annotation, FUMA first defines independent significant SNPs that have a genome-wide significant *P*-value (5×10^{-8}) and are independent at r2 < 0.6 within 1 Mb. Based on LD information, a subset of these independent significant SNPs is labeled as lead SNPs (independent from each other at r2 < 0.1). Genomic loci were identified by merging the LD blocks of lead SNPs that are closely located to each other (less than 250 kb apart). The top lead SNP was defined as the SNP with the lowest *P*-value in a specific region. Biological functionality for lead SNPs, including potential regulatory functions (RegulomeDB score), deleteriousness score (CADD score), effects on gene functions (using ANNOVAR), and mRNA expression levels (using eQTL data), were also estimated by FUMA. In detail, nearest genes and functional consequence of each SNP on gene functions were annotated based on ANNOVAR. Combined Annotation Dependent Depletion (CADD) score indexes the deleteriousness of variants computed based on 67 annotation resources. SNPs with the CADD score higher than 12 were considered to confer deleterious effects. The RegulomeDB provides a categorical score that describes how likely a SNP is likely to play a regulatory role based on the integration of high-throughput datasets. The RDB score of 1a suggests the strongest evidence, while the score 7 represents the least support for a regulatory potential. eQTL mapping provides significant cis-SNP-gene pairs (up to 1Mb apart) in CVD-related tissue types from GTEx. SNPs were mapped to genes based on their physical position in the genome and expression quantitative trait locus (eQTL) associations (obtained from 10 related tissues in GTExv8). The SNP locations were defined in reference to the human genome Build 37 (GRCh37/hg19), and only protein-coding genes were included in the analysis. In addition, genome-wide significant SNPs from CPASSOC and original GWAS results were also annotated by FUMA for comparison.

Testing overlap of genomic risk loci across CVDs

FUMA-annotated genomic risk loci were overlapped based on the location on the chromosome to find overlapped genomic risk loci across more than one CVD. We further performed Multi-trait colocalization analysis on these overlapped genomic loci by HyPrColoc (Hypothesis Prioritisation in multi-trait Colocalization) method to identify potential shared causal variants, thus indicating the potential biological mechanisms among CVDs[45]. HyPrColoc, as an extension of the colocalization method, allows colocalization analysis for multiple traits, which adopts a deterministic Bayesian divisive clustering algorithm to identify clusters of colocalized traits and candidate causal variants in the same genomic locus and provides the posterior probability (PP) of colocalization for each cluster. Only a genomic risk locus with PP larger than 0.7 and the traits identified by HyPrColoc consistent with the traits by overlapping the genomic risk loci was declared as a colocalized locus. Next, we estimated posterior probabilities (known as the m-value) for each of the pleiotropic loci to quantify the best-fit

model of cross-disease genotype-phenotype relationships using MetaSoft[46] with the fixed effects model (FE). M-value > 0.9 indicated that a particular variant had an effect on a given disease, while m-value < 0.1 predicted that the SNP does not have an effect on the disease.

Gene-level analysis using MAGMA and TWAS

Moving beyond SNP-level studies, we performed gene-based association analysis using Multi-marker analysis of genomic annotation (MAGMA), which yields gene-based *P*-values through the evaluation of the joint association effect of all SNPs within a gene while accounting for LD between SNPs[47]. The SNP locations were defined in reference to the human genome Build 37 (GRCh37/hg19), and only protein-coding genes (17,363) were included in the analysis. We then defined boundaries of gene length within '±10 kb outside the gene', consistent with the default setting of position mapping using FUMA. We applied multiple testing corrections to correct for the number of protein-coding genes (threshold: $P = 0.05$ / no. of protein-coding genes = 0.05 / 17,636 = 2.84×10^{-6}), and additionally also indicate which results survive additional correction for the number of CVDs (threshold: $P = 0.05$ / no. of protein-coding genes / no. of CVDs = 0.05 / 17,636 / 6 = 4.73×10^{-7}).

In addition, we leveraged transcriptome-wide association study (TWAS) analysis using FUSION to further investigate the tissue-specific genes based on the 10 CVD-related tissues[48]. Briefly, we used FUSION method to identify risk genes associated with CVDs by integrating MTAG-based summary statistics with pre-computed gene expression weights reference of trait-relevant tissues from the Genotype-Tissue Expression project version 8 (GTEx v8) while considering LD structures[48]. The cis-genetic components of tissue-specific gene expression were imputed from the MTAG-based summary statistics using five linear models (namely, best linear unbiased prediction [BLUP], Bayesian sparse linear mixed model [BSLMM], least absolute shrinkage and selection operator [LASSO], Elastic Net [ENET], and top SNPs). Then, the calculated gene expression weights calculated by the best-performing prediction model were combined with MTAG results to identify significant associations between gene expression levels and CVDs. For each tissue, we also excluded non-protein coding genes and genes with duplicated names. Finally, 6,954 Adipose Subcutaneous-specific genes, 5,513 Adipose Visceral Omentum-specific genes, 5,666 Artery Aorta-specific genes, 2,626 Artery Coronary-specific genes, 7,042 Artery Tibial-specific genes, 1,888 Cells_EBV-transformed lymphocytes-specific genes, 4,826 Heart Atrial Appendage-specific genes, 4,410 Heart Left Ventricle-specific genes, 2,529 Liver-specific genes, and 6,162 Whole Blood-specific genes were included for further analysis. The significance threshold of TWAS associations was corrected with the Bonferroni correction (such as $P_{\text{Adipose Subcutaneous}} < 0.05 / 6{,}954 = 7.19 \times 10^{-6}$, $P_{\text{Whole Blood}} < 0.05 / 6{,}162 = 8.11 \times 10^{-6}$). These gene-mapping analyses were conducted on the GWAS summary statistics and on the MTAG-based summary statistics of all six CVDs.

Pathway-level analysis using MAGMA, LDSC-SEG, and GARFIELD

Through MAGMA's gene set analysis, we tested 9,398 gene sets (canonical pathways and biological processes sets from the Molecular Signatures Database (MSigDB, v.2023.1; C2: CP REACTOM & C5: GO BP)[49]. We then applied multiple testing corrections to correct for the number of gene sets tested (0.05 / (7,744 + 1,654)), and additionally also indicate which results survive additional correction for the number of CVDs (0.05/(7,744 + 1,654)/6). The statistical tests conducted were all two-sided.

LDSC applied to specifically expressed genes (LDSC-SEG) was used to identify enrichments in tissue-specific gene expression and chromatin modification, with several gene sets including multi-tissue gene expression (including Genotype-Tissue Expression [GTEx] and Franke lab data) and multi-tissue chromatin (including both Roadmap Epigenomics and ENCODE data)[50]. For tissue-specific gene expression, we include annotations constructed based on RNA sequencing data from human tissues from GTEx, with Bonferroni-corrected significant threshold set at $P < 1.73 \times 10^{-4}$ (.05 / 48 / 6). For tissue-specific histone marks, we included annotations constructed based on data from the Roadmap Epigenetics Project for narrowly defined peaks for DNase hypersensitivity, H3K27ac, H3K4me1, H3K4me3, H3K9ac, and H3K36me3 chromatin, with Bonferroni-corrected significant threshold set at $P < 1.70 \times 10^{-5}$ (.05 / 489 / 6).

We used GWAS Analysis of Regulatory and Functional Information Enrichment with LD correction (GARFIELD, a powerful enrichment tool that compares different sets of annotation marks from specific cell types or tissues) to correlate the GWAS summary statistics with various regulatory or functional annotations and find features relevant to a phenotype of interest under different GWAS *P*-value thresholds, adjusting for LD, MAF, and distance to transcription start site (TSS)[51]. The annotations included 1,005 features extracted from ENCODE, GENCODE, and Roadmap Epigenomics projects, including DNase I hypersensitivity hotspots, chromatin accessibility peaks, transcription-factor footprints, formaldehyde-assisted isolation of regulatory elements (FAIRE), histone modifications, chromatin segmentation states, genic annotations, and transcription-factor binding sites (TFBS), among others, in a number of publicly available cell lines or tissues. Enrichment *P*-value is determined empirically through a permutation procedure accounting for associated regions structures on the basis of the number of SNPs and mean LD. We also assessed multiple testing corrections to correct for the number of functional categories tested (0.05 / 1,005), and additionally also indicate which results survive additional correction for the number of CVDs (0.05 / 1,005 / 6).

Proteome-Wide Mendelian Randomization Study using SMR and colocalization analysis

Summary-level statistics of genetic associations with levels of 4907 plasma proteins were extracted from a large-scale cis-protein quantitative trait loci (cis-pQTL) study in 35,559 Icelanders[52]. Further information on the GWAS can be found in the original publication. For each protein, we included cis-acting SNPs (ie, SNPs located within ± 1 Mb) window from the gene body of the target gene. In this study, only proteins with cis-pQTLs available at the genome-wide significance level ($p < 5 \times 10^{-8}$) were included in the MR and colocalization analysis.

Summary data-based Mendelian randomization (SMR) is a mendelian randomization method that uses summary-level data to test if an exposure variable (i.e. gene expression) and outcome (i.e., trait) are associated because of a shared causal variant, using genome-wide significant SNPs as instrumental variables[53]. A significant SMR association could be explained by a causal effect (i.e. the causal variant influences disease risk via changes in gene expression), pleiotropy (i.e. the causal variant has pleiotropic effects on gene expression and disease risk) or linkage (i.e. different causal variants exist for gene expression and disease). Thus, to distinguish causality or pleiotropy from linkage, the heterogeneity in dependent instruments (HEIDI) method was applied to each tested

single nucleotide variant. The association with *P*-value in HEIDI test < 0.01 was considered likely caused by pleiotropy and thus removed from further analyses. SMR test using multiple SNPs (multi-SNPs-SMR) instrumented for each protein was employed as a sensitivity analysis to avoid the potential bias caused by analyzing only a single SNP, which provided significance levels ($P_{smr-multi}$ < 0.05) to strengthen the evidence from the primary analysis[54]. We also used the false discovery rate (FDR) at = 0.05 based on Benjamini–Hochberg (BH) method for multiple testing. The odds ratios (ORs) and corresponding confidence intervals (CIs) of the associations between proteins and the outcomes were calculated using the Wald ratio and the delta method, respectively. SMR and HEIDI analysis were performed using the SMR software tool (v1.3.1).

Colocalization analysis is a crucial complementary analysis while investigating MR that considers the overlapping characteristics between two or more causal variants. We conducted colocalization analysis using the 'coloc' R package to test whether identified associations between causal proteins and each of CVD were driven by the same causal variant or linkage disequilibrium[55]. For each locus, Bayes factors were computed to test the posterior probability (PP) of five mutually exclusive hypotheses: (1) No SNP is associated with either trait 1 or trait 2 (H0); (2) Only trait 1 has a causal SNP, whereas no SNP is associated with trait 2 (H1); (3) Only trait 2 has a causal SNP, whereas no SNP is associated with trait 1 (H2); (4) Both traits have independent and different causal SNPs (H3); (5) Both traits have a shared causal SNP (H4). The posterior probability (PP), represented by PP0, PP1, PP2, PP3, and PP4, quantifies support for each hypothesis. We set prior probabilities of the SNP being associated with trait1 (causal plasma proteins) only (p1) at 1×10^{-4}; the probability of the SNP being associated with trait2 only (p2) at 1×10^{-4}; and the probability of the SNP being associated with both traits (p12) at 1×10^{-5}. Two traits were considered to have strong evidence of colocalization if the posterior probability for shared causal variants (PPH4) was > 0.70.

Druggable proteins identification

To determine the potential druggability of the identified proteins, we searched identified plasma proteins (based on the results of colocalization analysis) in the OpenTargets publicly available data, which covered the target-disease evidence derived from genetic associations, somatic mutations, known drugs, differential expression, animal models, pathways, and systems biology from 22 reputable sources[56]. We selected these drugs for which there was evidence of an association with the protein of interest, which were classified by the clinical trial phase reported on the ClinicalTrials website. In addition, we also cross-referenced our results with the list of druggable genes to ensure consistency and replicability.

Declarations

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Genome-wide summary statistics on AF, HF, and Stroke are available at the GWAS Catalog (GCST90104539, GCST009541, and GCST90104539). GWAS summary statistics on CAD and PAD are publicly available for download at the Cardiovascular Disease Knowledge Portal (CVDKP) website: https://cvd.hugeamp.org/datasets.html. Genome-wide summary statistics on VTE and pQTL are obtained from the deCODE genetics website: https://www.decode.com/summarydata/.

Code availability:

All software (and version, where applicable) used to conduct the analyses in this paper are freely available online:

LDSC (v1.0.1; https://github.com/bulik/ldsc), MiXeR (v1.3; https://github.com/precimed/mixer), LAVA (v0.1.0; https://github.com/josefin-werme/LAVA), GWAS-PW(v0.21; https://github.com/joepickrell/gwas-pw), LCV (https://github.com/lukejoconnor/LCV); LHC-MR (v0.0.0.9000; https://github.com/LizaDarrous/lhcMR), MTAG (v0.9.0 beta; https://github.com/JonJala/mtag), CPASSOC (http://hal.case.edu/~xxz10/zhu-web/), FUMA (v1.5.4; http://fuma.ctglab.nl/), HyPrColoc(v1.0; https://github.com/jrs95/hyprcoloc), MetaSoft (http://genetics.cs.ucla.edu/meta/), MAGMA (v.1.08; https://ctg.cncr.nl/software/magma), TWAS (http://gusevlab.org/projects/fusion/), GARFIELD (v2; https://www.ebi.ac.uk/birney-srv/GARFIELD/), SMR (v1.31; https://yanglab.westlake.edu.cn/software/smr/), COLOC (v5.2.1; https://github.com/chr1swallace/coloc), and R (v.4.1.3; https://www.r-project.org/).

Acknowledgements

This study was supported by Shenzhen Pengcheng Peacock Plan (To Y.F), National Natural Science Foundation (Grant no. 82170339 and 82270241), NSFC Incubation Project of Guangdong Provincial People's Hospital (Grant no. KY0120220021), Natural Science Foundation of Guangdong Province (Grant no. 2023B1515020082) (To L.J.), Cancer Research UK Career Development Fellowship (Grant no. C59392/A25064) (To S.P.), and Center for Computational Science and Engineering at Southern University of Science and Technology. The funder had no role in the design, implementation, analysis, interpretation of the data, approval of the manuscript, and decision to submit the manuscript for publication.

Author contributions

Y.F., L.J., and S.P. conceptualized, supervised this project, and wrote the manuscript. J.Q., M.C., and, C.W. performed the main analyses and wrote the manuscript. L.C., R.Z., and S.S. performed statistical analysis and assisted with interpretation of results. N.T. and P.H. provided expertise in cardiovascular biology. All authors discussed the results and commented on the paper.

Competing interests

All authors declare no competing interests.

References

1. Roth, G.A. et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol* 76, 2982-3021 (2020).
2. Hajar, R. Genetics in Cardiovascular Disease. *Heart Views* 21, 55-56 (2020).
3. Lloyd-Jones, D.M. et al. Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. *JAMA* 291, 2204-11 (2004).
4. Arking, D.E. & Chakravarti, A. Understanding cardiovascular disease through the lens of genome-wide association studies. *Trends Genet* 25, 387-94 (2009).
5. Nielsen, J.B. et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet* 50, 1234-1239 (2018).
6. Aragam, K.G. et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat Genet* 54, 1803-1815 (2022).
7. Ghouse, J. et al. Genome-wide meta-analysis identifies 93 risk loci and enables risk prediction equivalent to monogenic forms of venous thromboembolism. *Nat Genet* 55, 399-409 (2023).
8. Shah, S. et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun* 11, 163 (2020).
9. van Zuydam, N.R. et al. Genome-Wide Association Study of Peripheral Artery Disease. *Circ Genom Precis Med* 14, e002862 (2021).
10. Mishra, A. et al. Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature* 611, 115-123 (2022).
11. Benn, M. & Nordestgaard, B.G. From genome-wide association studies to Mendelian randomization: novel opportunities for understanding cardiovascular disease causality, pathogenesis, prevention, and treatment. *Cardiovasc Res* 114, 1192-1208 (2018).
12. Palotie, A., Widen, E. & Ripatti, S. From genetic discovery to future personalized health research. *N Biotechnol* 30, 291-5 (2013).
13. Evangelou, E. & Ioannidis, J.P. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* 14, 379-89 (2013).
14. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* 50, 229-237 (2018).
15. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* 179, 1469-1482.e11 (2019).
16. Romero, C. et al. Exploring the genetic overlap between twelve psychiatric disorders. *Nat Genet*54, 1795-1802 (2022).
17. Grotzinger, A.D. et al. Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic and molecular genetic levels of analysis. *Nat Genet* 54, 548-559 (2022).
18. Ellinghaus, D. et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet* 48, 510-8 (2016).
19. Li, Y.R. et al. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat Med* **21**, 1018-27 (2015).
20. Demela, P., Pirastu, N. & Soskic, B. Cross-disorder genetic analysis of immune diseases reveals distinct gene associations that converge on common pathways. *Nat Commun* 14, 2743 (2023).
21. Nordestgaard, B.G. et al. Lipoprotein(a) as a cardiovascular risk factor: current status. *Eur Heart J* 31, 2844-53 (2010).
22. Nurnberg, S.T. et al. Genomic profiling of human vascular cells identifies TWIST1 as a causal gene for common vascular diseases. *PLoS Genet* 16, e1008538 (2020).
23. Nukala, S.B. et al. Differentially Expressed Proteins in Primary Endothelial Cells Derived From Patients With Acute Myocardial Infarction. *Hypertension* 74, 947-956 (2019).
24. Schulman, S. et al. Treatment of acute venous thromboembolism with dabigatran or warfarin and pooled analysis. *Circulation* 129, 764-72 (2014).
25. Fredenburgh, J.C. & Weitz, J.I. News at XI: moving beyond factor Xa inhibitors. *J Thromb Haemost* 21, 1692-1702 (2023).
26. Natarajan, P. & Kathiresan, S. PCSK9 Inhibitors. *Cell* 165, 1037 (2016).
27. Zarrin, A.A., Bao, K., Lupardus, P. & Vucic, D. Kinase inhibition in autoimmunity and inflammation. *Nat Rev Drug Discov* 20, 39-63 (2021).
28. Croft, M. & Siegel, R.M. Beyond TNF: TNF superfamily cytokines as targets for the treatment of rheumatic diseases. *Nat Rev Rheumatol* 13, 217-233 (2017).
29. Trinder, M., Uddin, M.M., Finneran, P., Aragam, K.G. & Natarajan, P. Clinical Utility of Lipoprotein(a) and LPA Genetic Risk Score in Risk Prediction of Incident Atherosclerotic Cardiovascular Disease. *JAMA Cardiol* 6, 1-9 (2020).
30. Rider, D.A. et al. Pre-clinical assessment of SLN360, a novel siRNA targeting LPA, developed to address elevated lipoprotein (a) in cardiovascular disease. *Atherosclerosis* 349, 240-247 (2022).
31. Bick, A.G. et al. Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. *Circulation* 141, 124-131 (2020).
32. Georgakis, M.K. et al. Associations of genetically predicted IL-6 signaling with cardiovascular disease risk across population subgroups. *BMC Med* 20, 245 (2022).
33. Wei, L.H. et al. Deficiency of Smad7 enhances cardiac remodeling induced by angiotensin II infusion in a mouse model of hypertension. *PLoS One* 8, e70195 (2013).
34. Wei, L.H. et al. Smad7 inhibits angiotensin II-induced hypertensive cardiac remodelling. *Cardiovasc Res* 99, 665-73 (2013).

35. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47, 1236-41 (2015).
36. Bulik-Sullivan, B.K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47, 291-5 (2015).
37. Holland, D. et al. Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet* 16, e1008612 (2020).
38. Frei, O. et al. Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nat Commun* 10, 2417 (2019).
39. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C.A. An integrated framework for local genetic correlation analysis. *Nat Genet* 54, 274-282 (2022).
40. Pickrell, J.K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* 48, 709-17 (2016).
41. O'Connor, L.J. & Price, A.L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat Genet* 50, 1728-1734 (2018).
42. Darrous, L., Mounier, N. & Kutalik, Z. Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. *Nat Commun* 12, 7274 (2021).
43. Zhu, X. et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet* 96, 21-36 (2015).
44. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 8, 1826 (2017).
45. Foley, C.N. et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat Commun* 12, 764 (2021).
46. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* 88, 586-98 (2011).
47. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 11, e1004219 (2015).
48. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 48, 245-52 (2016).
49. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739-40 (2011).
50. Finucane, H.K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* 50, 621-629 (2018).
51. Iotchkova, V. et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat Genet* 51, 343-353 (2019).
52. Ferkingstad, E. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* 53, 1712-1721 (2021).
53. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 48, 481-7 (2016).
54. Wu, Y. et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* 9, 918 (2018).
55. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 10, e1004383 (2014).
56. Finan, C. et al. The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* 9(2017).

Figures

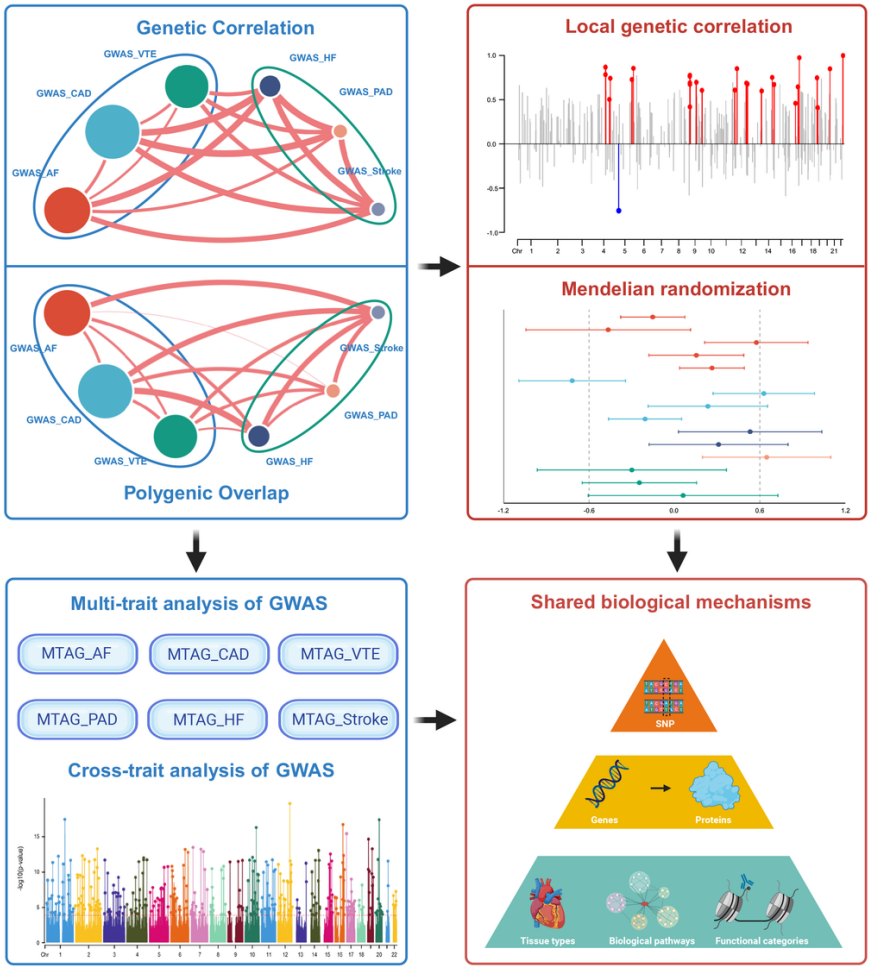

Figure 1

Schematic representation of analyses performed for all 6 major cardiovascular diseases in the current study.

This figure illustrates the comprehensive analytical approach undertaken for all six major cardiovascular diseases within this study. Initial GWAS findings were sourced from various repositories. Preceding the integration of cardiovascular diseases in a multi-trait analysis, genetic characteristics such as SNP-based heritability, genome-wide or local genetic correlations, polygenic overlap, and causal associations were estimated individually. The outcomes of SNP and genomic loci analyses from the multi-trait assessment, along with cross-trait analysis for replication, were juxtaposed against the results from individual cardiovascular diseases, unearthing novel pleiotropic loci. To comprehensively characterize the genetic overlaps for each of the six cardiovascular diseases, various methods were employed. Firstly, we explored the convergence of SNPs, genomic loci, and mapped genes. Subsequent stages encompassed biological pathway and functional category analyses, evaluating the enrichment of genetic signals across 9,398 distinct gene sets and 49 tissue types for each disease. The enrichment assessments were further expanded using LDSC-SEG and GARFIELD, encompassing 489 and 1,005 functional genomic categories respectively. Ultimately, the overlap within circulating proteins was assessed via proteome-wide Mendelian randomization and colocalization analyses, which offer preclinical indicators for drug development strategies. The diagram was generated using BioRender (www.biorender.com) and has been included with permission for publication.

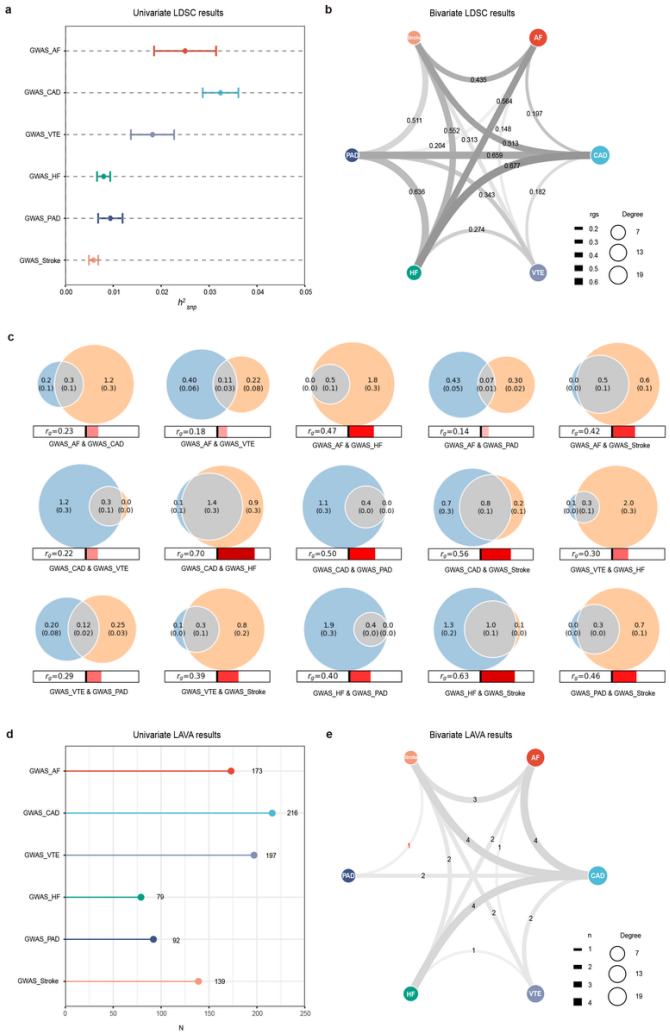

Figure 2

SNP-based heritability, genetic correlations, and polygenic overlap of the six cardiovascular diseases.

(a) Error-bar plot of the SNP-based heritability (h^2_{SNP}) point estimates for the six cardiovascular diseases, computed using univariate LDSC. **(b)** Network visualization of the Bonferroni-corrected significant global rgs among the cardiovascular diseases, computed using bivariate LDSC. Connections represent significant r_gs, with correlation value along connections, thicker lines denoting stronger correlations, and dark grey denoting more significant correlations. The size of the nodes is weighed by the sample size and h^2_{SNP} of the given cardiovascular disease (size = h^2_{SNP} × sqrt(N)). **(c)** Venn diagrams depicting the unique (blue and orange) and shared (grey) causal variants associated with pairs of cardiovascular diseases. Polygenic overlap is represented in gray. The numbers indicate the estimated quantity of causal variants (in thousands) per component, explaining 90% of SNP heritability in each of cardiovascular disease, with standard error in parentheses. The size of the circles reflects the degree of polygenicity for each trait. The estimated genetic correlation for pairs of cardiovascular diseases is also shown below the corresponding Venn diagram, with an accompanying directional scale (blue shades for negative values and red shades for positive values). **(d)** Lollipop plot of the number of sufficient local genetic signals for the six cardiovascular diseases, computed using univariate LAVA. **(e)** Network visualization of the Bonferroni-corrected significant local r_gs between pairs of cardiovascular diseases, computed using bivariate LAVA. Connections represent how many significant local r_gs were identified between pairs of cardiovascular diseases, with thicker connections denoting higher numbers and dark grey denoting more significant correlations. Numbers in black font indicate positive local r_gs, whereas numbers in red font indicate negative local r_gs. The orientation and the size of the nodes were set to mirror that of the network visualization of global r_gs.

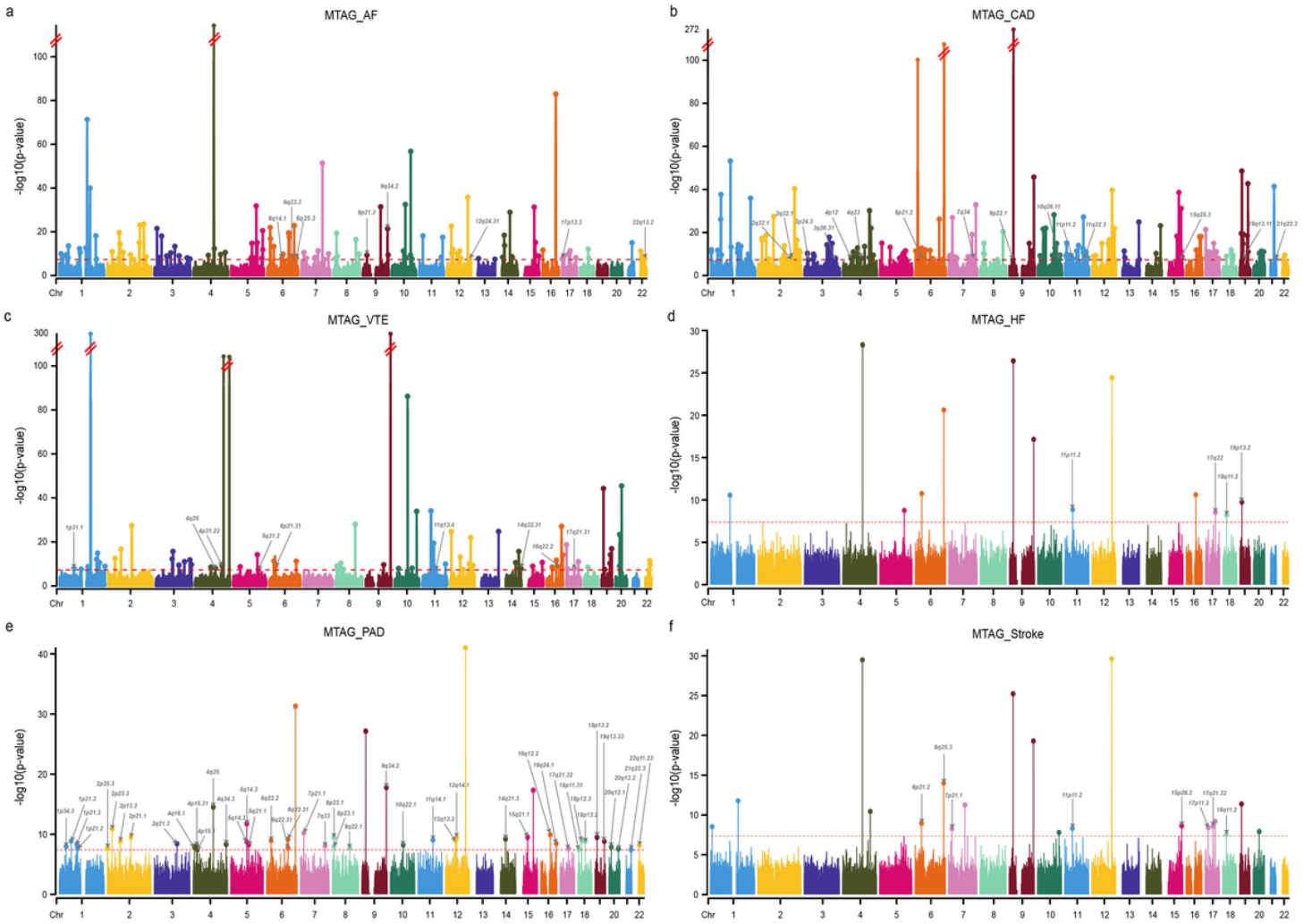

Figure 3

Manhattan plots for MTAG results of each of the six cardiovascular diseases.

The X-axis is the chromosomal position, and the Y-axis is negative log10 transformed *P*-values for each SNP. The cytoband annotations for the newly identified genomic loci are in gray. The dotted lines in red indicate the genome-wide significant *P*-value of −log10 (5×10^{-8}). Colorful dots indicate an independent genome-wide significant association with the smallest *P*-value (Top lead SNP). Only SNPs shared across all summary statistics were included.

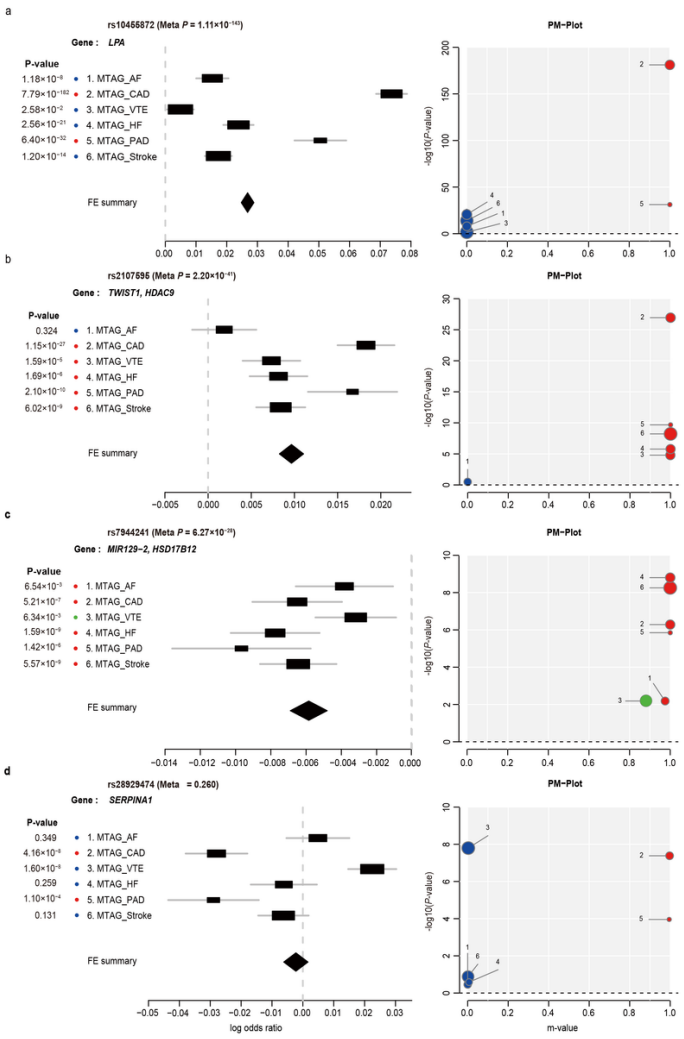

Figure 4

Profile of cardiovascular disease associations for illustrative pleiotropic loci: (a) rs10455872 on 6q25.3; (b) rs2107595 on 7p21.1; (c) rs7944241 on 11p11.2; and (d) rs28929474 on 1q24.2.

For each locus, cardiovascular disease-specific effects of the casual SNP are shown using ForestPMPlot. The first panel is the forest plot, displaying disease-specific association *P*-value, log odds ratios (ORs), and standard errors of the SNP. The meta-analysis *P*-value and the corresponding summary statistic are displayed on the top and the bottom of the forest plot, respectively. The second panel is the PM-plot in which X-axis represents the m-value, the posterior probability that the effect exists in each disease, and the Y-axis represents the disease-specific association *P*-value as -log10(*P*-value). Diseases are depicted as a dot whose size represents the sample size of individual GWAS. Diseases with estimated m-values of at least 0.9 are colored in red suggesting that the SNP does have an effect on the disease, while those with m-values less than 0.1 are marked in blue suggesting that the SNP does not have an effect on the disease. And diseases with estimated m-values between 0.1 and 0.9 are marked in green suggesting that the SNP does have an uncertain effect on the disease.

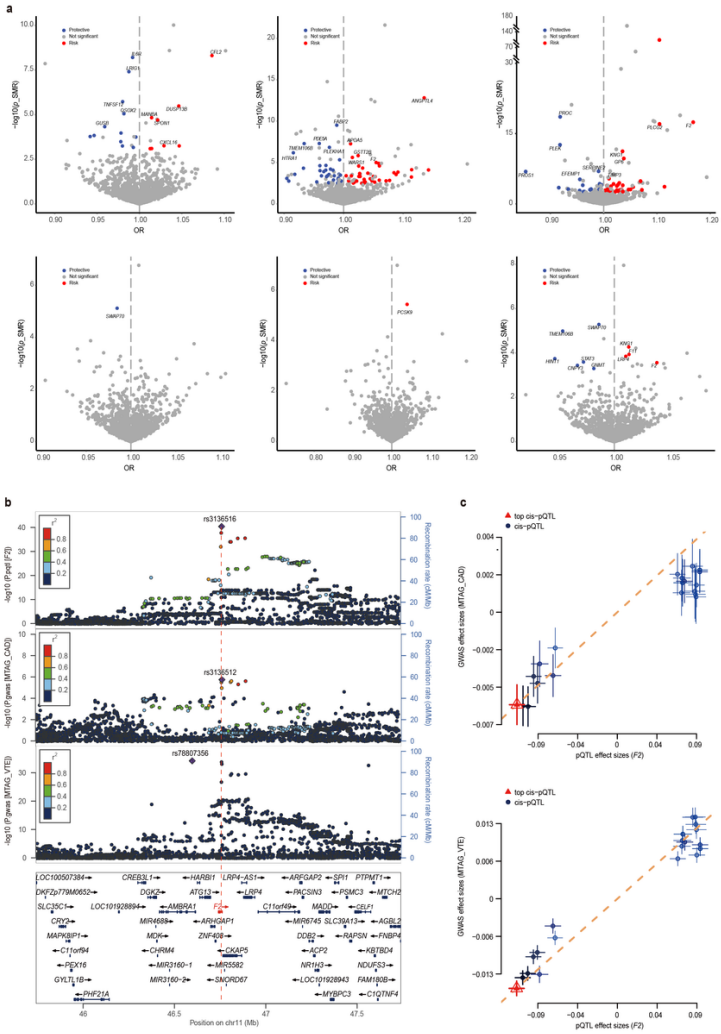

Figure 5

Proteome-wide MR and Colocalization analysis of the six cardiovascular diseases.

(a) Volcano plot of associations between 1,773 proteins and each of cardiovascular diseases. Association strength between proteins and cardiovascular diseases according to their effect. Annotated proteins passed the threshold of both Proteome-wide MR and Colocalization analysis. The red and blue colors represent positive and negative effects, respectively. **(b)** Regional association plots at F2, showing MTAG_CAD and MTAG_VTE association and cis-pQTLs for F2 (top). Variants are color-coded based on the LD r2 with the sentinel variants (purple dot in a diamond shape). **(c)** Scatter plots showing that higher F2 protein levels in the blood tend to have an increased risk of CAD and VTE. Each point represents a SNP, the x value of a SNP is its β effect size on a protein and the horizontal error bar represents the SE around the Beta. The y value of the SNP is its β effect size on disease risk and the vertical error bar represents the SE around its β. The dashed line represents the SMR estimate (a line with the intercept of 0 and the slope of β from the SMR test).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable.xlsx
- ReportingSummary.pdf
- SupplementaryInformation.pdf