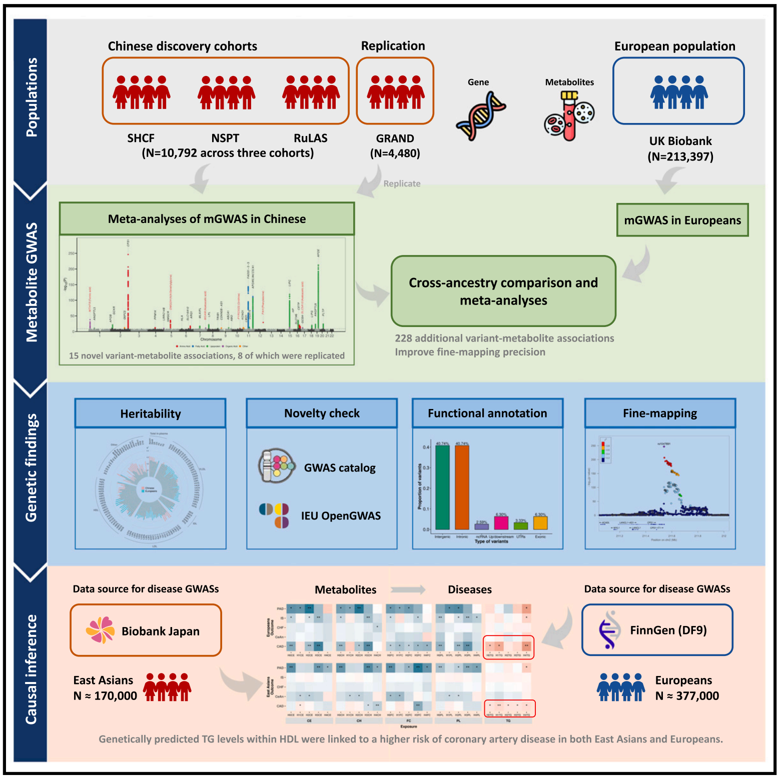
Cell Genomics

Article

Cross-ancestry analyses of Chinese and European populations reveal insights into the genetic architecture and disease implication of metabolites

Graphical abstract


Populations
Chinese discovery cohorts
Replication
European population
Gene
Metabolites
SHCF
NSPT
RuLAS
(N=10,792 across three cohorts)
GRAND
(N=4,480)
UK Biobank
(N=213,397)
Replicate
Metabolite GWAS
Meta-analyses of mGWAS in Chinese
mGWAS in Europeans
Cross-ancestry comparison and meta-analyses
15 novel variant-metabolite associations, 8 of which were replicated
228 additional variant-metabolite associations
Improve fine-mapping precision
Genetic findings
Heritability
Novelty check
GWAS catalog
IEU OpenGWAS
Functional annotation
Fine-mapping
Causal inference
Data source for disease GWASs
Biobank Japan
East Asians
N ≈ 170,000
Metabolites
Diseases
Data source for disease GWASs
FinnGen (DF9)
Europeans
N ≈ 377,000
Genetically predicted TG levels within HDL were linked to a higher risk of coronary artery disease in both East Asians and Europeans.

Authors

Chenhao Lin, Mingfeng Xia,
Yuxiang Dai, ..., Li Jin, Xingjie Hao,
Yan Zheng

Correspondence

lijin@fudan.edu.cn (L.J.),
xingjie@hust.edu.cn (X.H.),
yan_zheng@fudan.edu.cn (Y.Z.)

In brief

Lin et al. explored the genetic basis of blood metabolites in Han Chinese populations and further conducted cross-ancestry comparisons with a European population of UK Biobank data. This study enhances the understanding of genetic architecture underlying metabolites and their roles in complex diseases across diverse ancestral backgrounds.

Highlights

- GWAS identifies and replicates variant-metabolite associations in Chinese individuals
- Cross-ancestry analyses show ethnic differences in metabolite genetics
- Genetic causal inference links metabolites to diseases in different populations

Lin et al., 2025, Cell Genomics *5*, 100810
April 9, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.xgen.2025.100810

Cell Genomics

CellPress
OPEN ACCESS

Article

Cross-ancestry analyses of Chinese and European populations reveal insights into the genetic architecture and disease implication of metabolites

Chenhao Lin,[1,2,11] Mingfeng Xia,[3,11] Yuxiang Dai,[4,11] Qingxia Huang,[1,11] Zhonghan Sun,[1] Guoqing Zhang,[5,6] Ruijin Luo,[7] Qianqian Peng,[5] Jinxi Li,[1] Xiaofeng Wang,[1,8] Huandong Lin,[3] Xin Gao,[3] Huiru Tang,[1] Xia Shen,[1,9] Sijia Wang,[5] Li Jin,[1,*] Xingjie Hao,[10,*] and Yan Zheng[1,4,12,*]

[1]State Key Laboratory of Genetic Engineering, Human Phenome Institute, Center for Evolutionary Biology, and School of Life Sciences, Fudan University, Shanghai 200433, China
[2]College of Food Science, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China
[3]Department of Endocrinology and Metabolism, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[4]Department of Cardiology, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[5]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[6]National Genomics Data Center& Bio-Med Big Data Center, University of Chinese Academy of Sciences, Chinese Academy of Science, Shanghai 200031, China
[7]Shanghai Southgene Technology Co., Ltd., Shanghai 201203, China
[8]Fudan University-the People's Hospital of Rugao Joint Research Institute of Longevity and Aging, Rugao, Jiangsu 226500, China
[9]Center for Intelligent Medicine Research, Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University, Guangzhou, Guangdong 511400, China
[10]Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China
[11]These authors contributed equally
[12]Lead contact
*Correspondence: lijin@fudan.edu.cn (L.J.), xingjie@hust.edu.cn (X.H.), yan_zheng@fudan.edu.cn (Y.Z.)
https://doi.org/10.1016/j.xgen.2025.100810

SUMMARY

Differential susceptibilities to various diseases and corresponding metabolite variations have been documented across diverse ethnic populations, but the genetic determinants of these disparities remain unclear. Here, we performed large-scale genome-wide association studies of 171 directly quantifiable metabolites from a nuclear magnetic resonance-based metabolomics platform in 10,792 Han Chinese individuals. We identified 15 variant-metabolite associations, eight of which were successfully replicated in an independent Chinese population (n = 4,480). By cross-ancestry meta-analysis integrating 213,397 European individuals from the UK Biobank, we identified 228 additional variant-metabolite associations and improved fine-mapping precision. Moreover, two-sample Mendelian randomization analyses revealed evidence that genetically predicted levels of triglycerides in high-density lipoprotein were associated with a higher risk of coronary artery disease and that of glycine with a lower risk of heart failure in both ancestries. These findings enhance our understanding of the shared and specific genetic architecture of metabolites as well as their roles in complex diseases across populations.

INTRODUCTION

Human metabolites play a crucial role in maintaining physiological homeostasis and health. Circulating metabolite levels are strongly influenced by dietary habits, lifestyle, and genetic background.[1] Many metabolites exhibit high heritability and are proximal to the clinical endpoints, making them effective intermediate phenotypes that link genetic susceptibility and diseases.[2] Revealing the genetic underpinnings of metabolite levels helps provide insights into the disease etiology.

In recent years, large-scale cohort consortia and biobanks have facilitated genetic research in metabolites and disease susceptibility. Previous genome-wide association studies (GWASs) have identified hundreds of genetic regions associated with circulating metabolite levels, primarily focusing on European populations.[3–15] However, cross-ancestry variations in genetic architecture, such as allele frequencies and linkage disequilibrium (LD) patterns, necessitate the inclusion of diverse populations to overcome ancestral biases in GWAS results.[12,16] Investigation of the genetic determinants of the human metabolome within non-European populations is crucial for identifying loci and understanding the unique genetic factors that underscore differences in disease susceptibility across ancestries.[15]

Here, we aimed to identify the genome-wide determinants of quantified plasma metabolites in 10,792 Han Chinese individuals and further reexamined these associations in 213,397 European individuals from the UK Biobank. We replicated the associations in an independent population of 4,480 Chinese individuals. In addition, we performed a cross-ancestry meta-analysis by pooling the results from our Chinese participants and the European individuals from the UK Biobank. Subsequently, we assessed the potential causal associations of metabolites with various human diseases in East Asian individuals through a two-sample Mendelian randomization (MR) approach, using summary data from Biobank Japan,[17] and attempted to replicate the findings in European populations as well (Figure 1). Our findings provide unique insights into the genetic basis of human metabolism and its implications for metabolic disorders across different ethnic populations.

RESULTS

Overview of genetic results for circulating metabolites in Chinese individuals

We included 10,792 Han Chinese individuals from three community-based cohorts, predominantly middle-aged with an average age of 62.5 years, and 42.1% of the participants were men (Table S1). Metabolomics profiling was performed using the same nuclear magnetic resonance (NMR)-based metabolomics platform. After quality control, 171 directly quantifiable metabolites, primarily lipoprotein subclasses and amino acids, were included in the subsequent analyses (see STAR Methods; Figure 2A; Table S2).

We first estimated the SNP-based heritability of each metabolite using the GREML algorithm in the GCTA software.[18] The median heritability was 0.23, with values ranging from 0.02 for 2-aminobutyric acid to 0.35 for glycine (Figure 2B; Table S3). We performed GWASs on the levels of these metabolites in each cohort separately and then meta-analyzed the results using the fixed-effect model. In total, we identified 52,439 variant-metabolite associations involving 5,117 variants and 159 metabolites at the study-wide significance threshold ($p < 5 \times 10^{-8}/29 = 1.72 \times 10^{-9}$, Bonferroni correction for 29 principal components that collectively explained 95% of the variance of the metabolomics data; Figures 2C and S1; Table S4). No evidence of excessive test statistic inflation or population stratification was observed (genomic inflation factor λ_{gc} median = 1.01, range = 0.99–1.03; Table S5). The major findings are publicly accessible on a website for their visualization and exploration (see https://www.biosino.org/gwas/).

Next, we defined the significant genomic locus by a stepwise process and determined the independent signals by LD-based clumping (see STAR Methods). We identified 632 metabolite-associated loci, which were then consolidated into 37 genomic loci across metabolites, collectively harboring 1,536 independent signals and involving 259 distinct variants (Figure 2C; Table S6). Furthermore, we identified 48 conditionally independent associations involving 11 additional distinct variants using the GCTA-COJO algorithm and found that six out of the 37 genomic loci harbored at least one additional conditionally independent signal. In summary, we identified a total of 1,584 (1,536 + 48) lead variant-metabolite associations (Table S6), which were the focus of our subsequent analyses. We further prioritized 38 most likely causal genes for the 1,584 lead associations based on biological relevance of the related metabolites or function of the lead SNP or variants in high LD ($r^2 > 0.8$; Table S6).

Genetic architecture of metabolite levels and fine-mapping

We observed substantial evidence of both polygenicity and pleiotropy, consistent with previous reports in European individuals.[9,10,13,15] Of the 159 metabolites with significant genetic associations, 140 were associated with two or more loci (Figure S2A). Meanwhile, each locus was associated with a median of two metabolites, and 13 of the 37 genomic loci were associated with at least 10 metabolites (Figure S2B; Table S6). For example, the locus in the *APOE* gene, which encodes the crucial factor for cholesterol regulation (i.e., apolipoprotein E protein), was associated with 113 metabolites, primarily lipoprotein subclasses.

The 1,584 lead variant-metabolite associations involved 270 lead variants, the majority of which were intronic or intergenic (Figure 3A), consistent with the overall functional distribution of all identified variants (Figure S3). Among these, 14 lead variants were identified as exonic nonsynonymous variants, which generally exhibited a lower minor allele frequency (MAF) and a higher absolute effect size compared to other variants (Figures 3B and 3C). The most significant association was observed between glycine and the missense variant rs1047891 (beta = 0.71 for allele A, $p = 2.90 \times 10^{-247}$) in the *CPS1* gene, which encodes the rate-limiting enzyme of the urea cycle. This variant was also associated with creatine (beta = 0.28 for allele A, $p = 4.27 \times 10^{-51}$), a downstream product of glycine metabolism. Furthermore, a strong correlation was observed between the genetic effect estimate and MAF (Figure 3D). Of the 270 lead variants, 20 were predicted to be deleterious, with a combined annotation-dependent depletion score >12.37.[19,20]

Of the 632 metabolite-associated loci, 416 (65.8%) were fine-mapped to credible sets of 2–10 variants, and 31 (4.9%) were even to one single nominated causal variant, involving 6 missense and 2 intronic variants (see STAR Methods; Figure S4A; Table S7). For example, the missense variant rs7412 in the *APOE* gene was the most likely causal variant for 22 lipoproteins, mainly low-density lipoprotein (LDL) subfractions. The abovementioned missense variant rs1047891 in *CPS1* was significantly associated with glycine and creatine with a posterior probability of >0.99.

Cross-population comparison of metabolite-variant associations in European ancestry

To assess whether our findings are population specific, we conducted GWASs of metabolites in individuals of European ancestry using the UK Biobank dataset (see STAR Methods). Among the 213,397 European participants included in the analyses, the average age was 57.0 years, and 46.5% were men (Table S1). Circulating metabolites were measured using a similar NMR metabolomics platform, resulting in 116 overlapping metabolites between the Chinese and the European populations (Table S2).

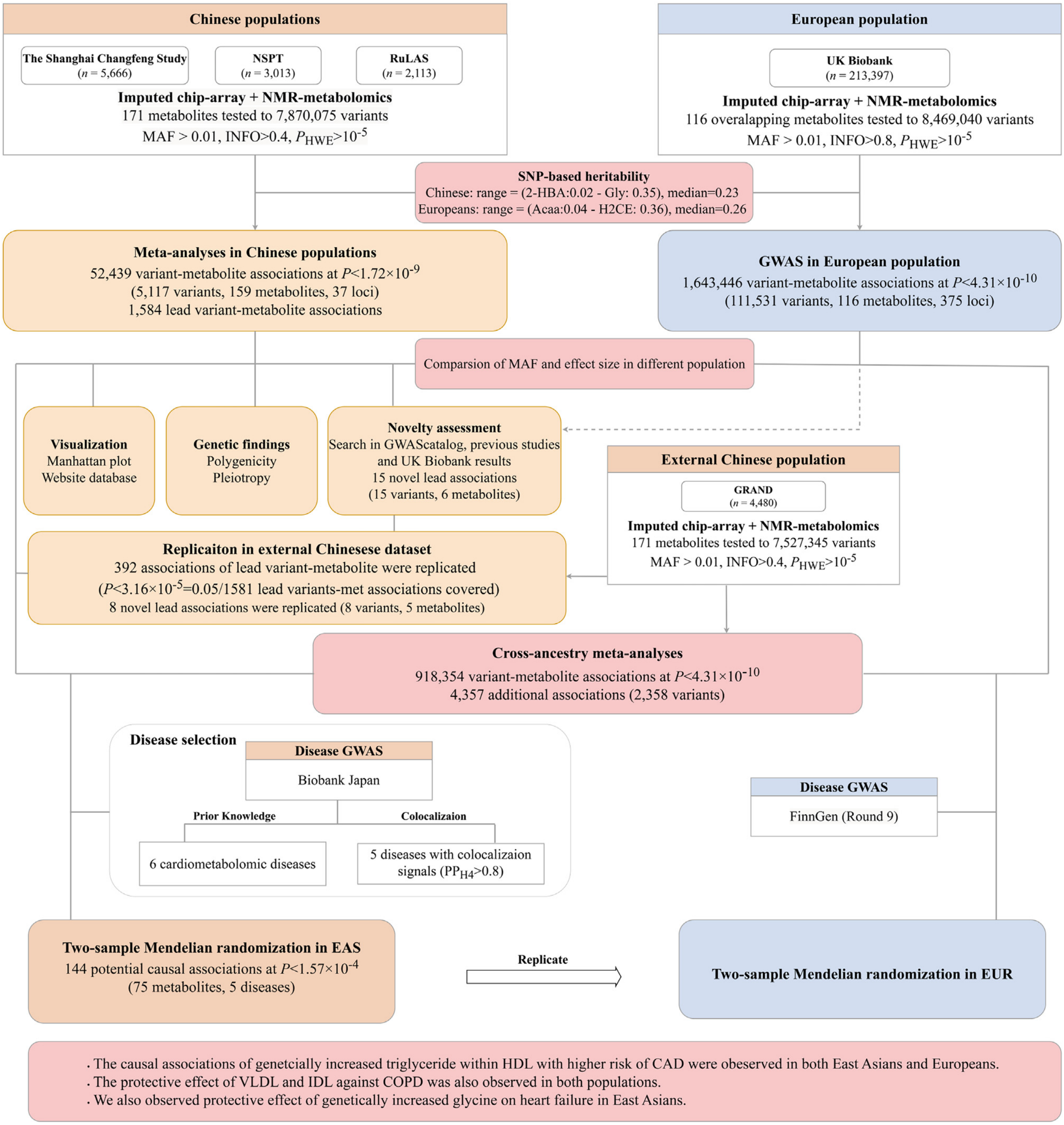
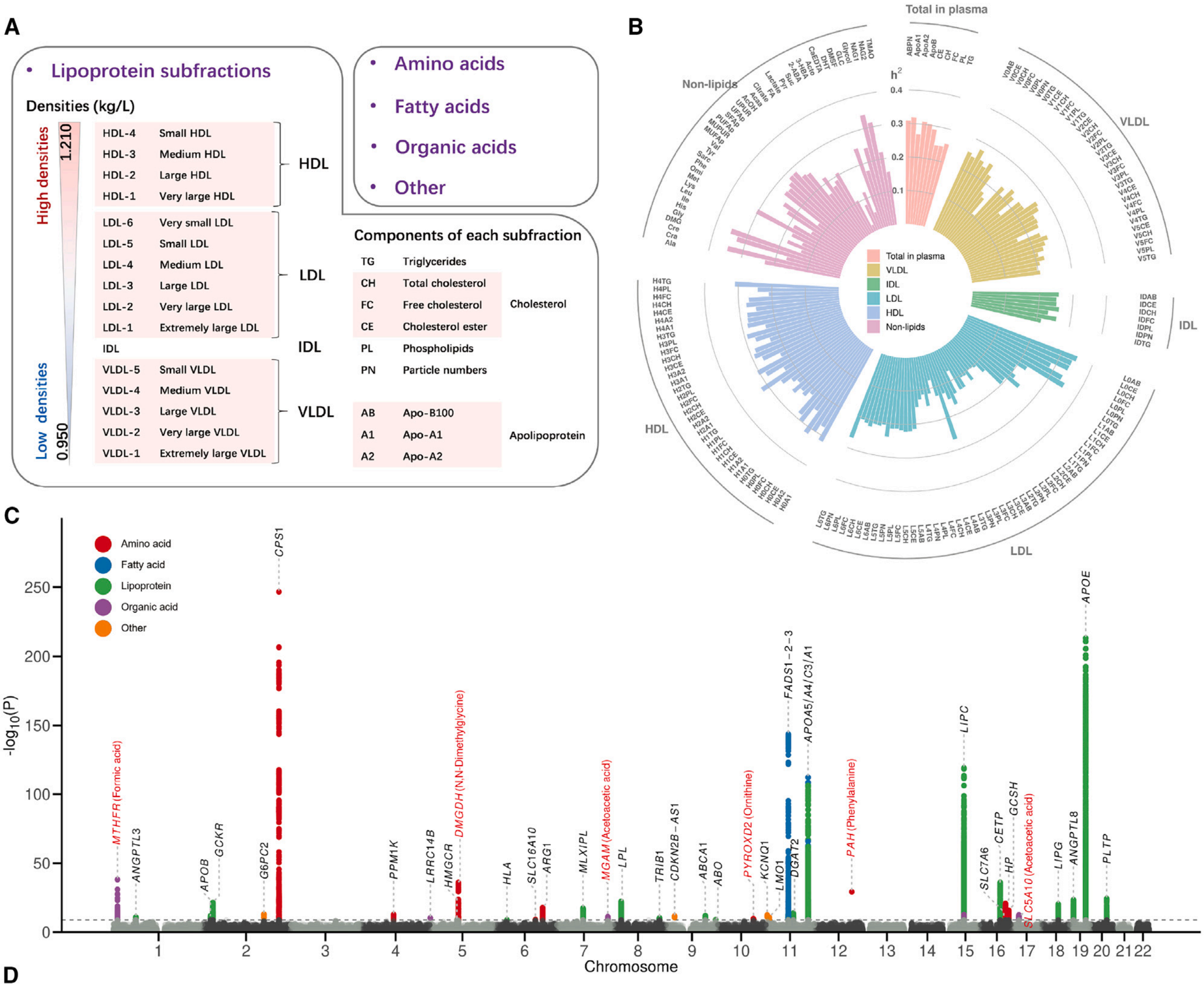

Chinese populations
The Shanghai Changfeng Study (n = 5,666)
NSPT (n = 3,013)
RuLAS (n = 2,113)
Imputed chip-array + NMR-metabolomics
171 metabolites tested to 7,870,075 variants
MAF > 0.01, INFO>0.4, P_HWE>10^-5
European population
UK Biobank (n = 213,397)
Imputed chip-array + NMR-metabolomics
116 overalapping metabolites tested to 8,469,040 variants
MAF > 0.01, INFO>0.8, P_HWE>10^-5
SNP-based heritability
Chinese: range = (2-HBA:0.02 - Gly: 0.35), median=0.23
Europeans: range = (Acaa:0.04 - H2CE: 0.36), median=0.26
Meta-analyses in Chinese populations
52,439 variant-metabolite associations at P<1.72×10^-9
(5,117 variants, 159 metabolites, 37 loci)
1,584 lead variant-metabolite associations
GWAS in European population
1,643,446 variant-metabolite associations at P<4.31×10^-10
(111,531 variants, 116 metabolites, 375 loci)
Comparsion of MAF and effect size in different population
Visualization
Manhattan plot
Website database
Genetic findings
Polygenicity
Pleiotropy
Novelty assessment
Search in GWAScatalog, previous studies and UK Biobank results
15 novel lead associations
(15 variants, 6 metabolites)
External Chinese population
GRAND (n = 4,480)
Imputed chip-array + NMR-metabolomics
171 metabolites tested to 7,527,345 variants
MAF > 0.01, INFO>0.4, P_HWE>10^-5
Replicaiton in external Chinesese dataset
392 associations of lead variant-metabolite were replicated
(P<3.16×10^-5=0.05/1581 lead variants-met associations covered)
8 novel lead associations were replicated (8 variants, 5 metabolites)
Cross-ancestry meta-analyses
918,354 variant-metabolite associations at P<4.31×10^-10
4,357 additional associations (2,358 variants)
Disease selection
Disease GWAS
Biobank Japan
Prior Knowledge
Colocalizaion
6 cardiometabolomic diseases
5 diseases with colocalizaion signals (PP_H4>0.8)
Disease GWAS
FinnGen (Round 9)
Two-sample Mendelian randomization in EAS
144 potential causal associations at P<1.57×10^-4
(75 metabolites, 5 diseases)
Replicate
Two-sample Mendelian randomization in EUR
. The causal associations of genetcially increased triglyceride within HDL with higher risk of CAD were obeserved in both East Asians and Europeans.
. The protective effect of VLDL and IDL against COPD was also observed in both populations.
. We also observed protective effect of genetically increased glycine on heart failure in East Asians.

Figure 1. Study design and workflow

In this study, we conducted meta-analyses of metabolite GWASs in 10,792 Chinese individuals and validated the findings in an independent external cohort of 4,480 Chinese individuals. Metabolite GWASs were also performed in 213,397 individuals of European ancestry from the UK Biobank, followed by cross-population comparisons and cross-ancestry meta-analyses. A two-sample MR approach was applied to infer causal relationships between metabolites and diseases in both East Asian and European populations.

We randomly selected 20,000 individuals of European ancestry from the UK Biobank and calculated their SNP-based heritability of metabolites. The heritability among European participants was comparable to that observed in Chinese participants, with a median value of 0.26 (Figures 4A and S5; Table S3). The GWAS in UK Biobank European individuals identified 1,643,446 variant-metabolite associations at a more stringent threshold ($p < 4.31 \times 10^{-10} = 5 \times 10^{-8}/116$ metabolites),



D

Table of the eight novel lead associations identified in the Chinese discovery set and validated in the Chinese replication set

Metabolite	Chr	rsID	Candidate gene	EA	NEA	Discovery set			Replication set			UK Biobank		AF (gnomAD V3)		Novelty check	Possible scenarios for novelty
						EAF	Beta	P	EAF	Beta	P	EAF	P	EAF_{EAS}	EAF_{EUR}		
Formic acid	1	rs1801133	*MTHFR*	A	G	0.414	-0.21	1.01×10^{-38}	0.430	-0.12	3.97×10^{-9}	Metabolite unmeasured		0.287	0.335	Novel locus	Not measured in other mGWASs
N,N-Dimethylglycine	5	rs10942892	*DMGDH*	A	G	0.101	0.13	1.53×10^{-9}	0.101	0.16	7.43×10^{-7}	Metabolite unmeasured		0.098	0.061	Novel lead variant	Novel mechanism
Acetoacetic acid	7	rs28429916	*MGAM*	G	A	0.287	-0.11	7.36×10^{-12}	0.285	-0.08	6.54×10^{-4}	0.205	0.08	0.281	0.220	Novel locus	Novel mechanism
Ornithine	10	rs2147896	*PYROXD2*	A	G	0.424	-0.09	2.04×10^{-10}	0.430	-0.07	8.00×10^{-4}	Metabolite unmeasured		0.435	0.649	Novel locus	Not measured in other mGWASs
Phenylalanine	12	rs118092776	*PAH*	T	C	0.014	0.66	7.66×10^{-30}	0.016	0.35	2.87×10^{-5}	Variant untyped		0.015	0.001	Novel lead variant	MAF difference across ethnicity
Acetoacetic acid	17	rs28667787	*SLC5A10*	T	C	0.264	0.11	1.52×10^{-12}	0.260	0.08	5.48×10^{-4}	0.445	0.91	0.268	0.460	Novel locus	Novel mechanism
Acetoacetic acid	17	rs1737941	*SLC5A10*	G	T	0.332	0.11	7.50×10^{-13}	0.332	0.08	8.24×10^{-4}	0.449	0.21	0.334	0.473	Novel locus	Novel mechanism
Acetoacetic acid	17	rs115496750	*SLC5A10*	G	A	0.164	0.14	4.60×10^{-13}	0.160	0.09	1.00×10^{-3}	0.021	0.42	0.151	0.021	Novel locus	Novel mechanism

Figure 2. Summary of genetic associations of metabolites in 10,792 Chinese participants

(A) Summary information of metabolites included in the current study. VLDL, very-low-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; HDL, high-density lipoprotein.

(B) SNP-based heritability for each metabolite in 10,792 Chinese participants. The SNP-based heritability was calculated using the GREML method in GCTA tools. The height of the polar bar plot indicates the phenotypic variance of each metabolite explained by the SNP chip variants. The metabolites are marked with different colors based on metabolite classes.

(C) Manhattan plot displaying chromosomal positions (x axis) of significant associations ($p < 1.72 \times 10^{-9}$). Colors indicate different metabolite classes. Putative causal genes with associations are highlighted in red.

(D) Table of the eight lead associations identified in the discovery set ($p < 1.72 \times 10^{-9}$) and validated in the replication set ($p < 0.003$) of Chinese individuals. Chr, chromosome; EA, effect allele; NEA, non-effect allele; EAF, effect allele frequency; AF, allele frequency; MAF, minor allele frequency; mGWAS, metabolite genome-wide association study.

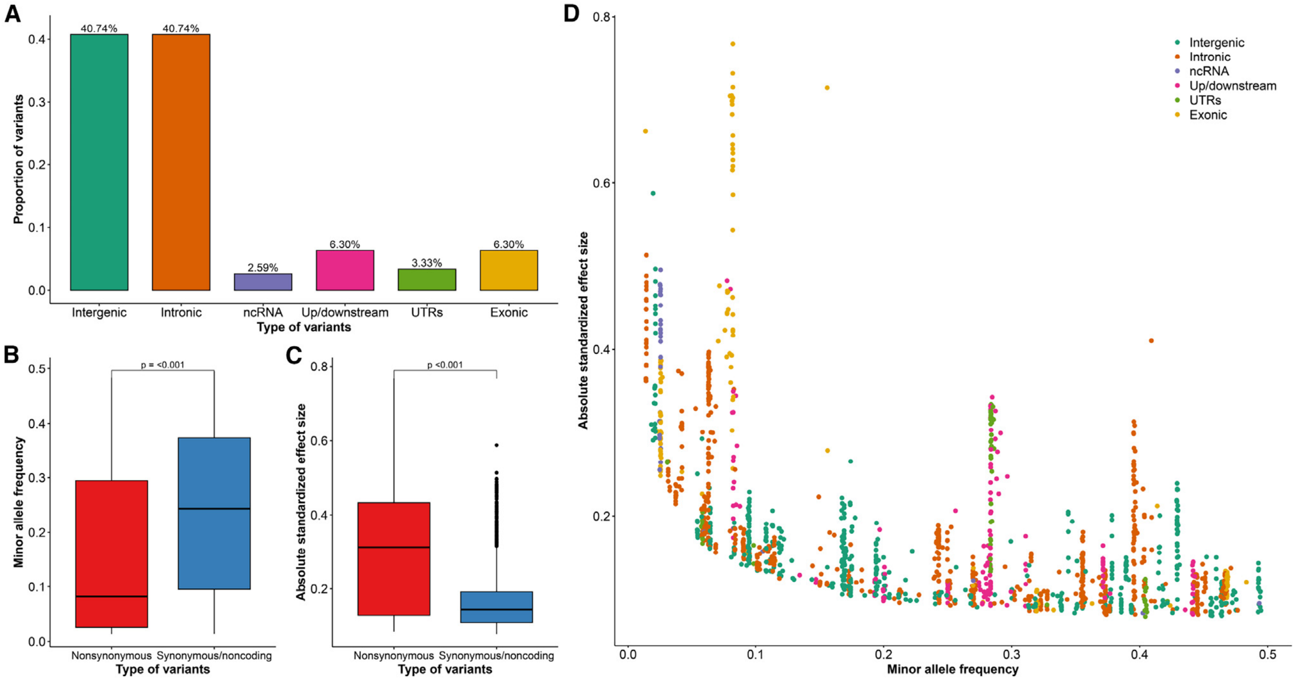


Figure 3. Genetic architecture of metabolite levels

(A) Distribution of functional annotations of lead variants in the current study. Colors indicate different types of functional annotations.

(B and C) Distribution of MAF (B) and effect size (C) based on the type of variants (nonsynonymous [red] versus synonymous or noncoding [blue]) identified in the present study.

(D) Scatterplot of MAF versus effect size for independent associations, with variants colored by functional annotation.

involving 8,726 lead variant-metabolite associations (Figure S6; Table S4). The fine-mapping analyses found 1,204 (13.8%) credible sets that contained one single nominated causal variant (Table S8).

To validate our findings from the Chinese population in the UK Biobank European individuals, we found that 582 of the 1,584 lead variant-metabolite associations identified in Chinese individuals were not observed in European individuals (Figure S7), primarily due to differences in metabolite profiles (376 associations) and low MAF (MAF < 0.01) or the complete absence of variants (206 associations involving 39 variants; Tables S6 and S9). For example, the intergenic variant rs117497152, linked to multiple lipoproteins in the Chinese population, was undetected in European individuals due to a large MAF difference (1,698-fold, $\text{MAF}_{\text{EAS}} = 0.08$ vs. $\text{MAF}_{\text{NFE}} = 4.6 \times 10^{-5}$ in gnomAD; Table S9). Of the 1,002 (=1,584 − 582) associations common to both populations, 833 were significant ($p < 4.99 \times 10^{-5} = 0.05/1{,}002$; Figure S7). The lack of significance for the remaining 169 associations was primarily attributed to small effect sizes and/or low MAF in European individuals despite the European sample being about 20 times larger than the Chinese population. Among the 833 significant associations, for example, the association of rs10421035, an intronic variant in *NECTIN2,* with LDL components (L0PN and L5PN), exhibited a 10-fold larger effect size in Chinese compared to European individuals (Figure 4B; Table S6). These findings underscore the importance of incorporating diverse populations in genetic studies of metabolites, not only including European populations but also expanding to other groups, such as East Asian individuals, to better capture a broader range of genetic influences on metabolic traits.

Identification of variant-metabolite associations and replication in an independent Chinese population

To evaluate the novelty of the 1,584 lead variant-metabolite associations in our Chinese populations, we examined the associations of 270 lead variants and their LD partners ($r^2 > 0.1$ within a 500-kb window; see STAR Methods) in previous GWASs of metabolites and clinical lipids (Table S10) as well as our metabolite GWAS in the UK Biobank. This analysis revealed 15 associations involving 6 non-lipid metabolites (Table S6).

We sought to replicate these associations in an independent Chinese population ($n = 4{,}480$, mean age = 57.7 years, 68.8% male; Table S1). All metabolites from the discovery set were included in the replication cohort because both used the same metabolomics profiling platform. Among the 15 associations identified, 8 associations involving 5 metabolites remained significant after multiple comparison correction ($p < 0.05/15 = 0.003$), with consistent directions of effect in both discovery and replication sets (Figure 2D; Table S11).

Of these, acetoacetic acid was associated with two loci: *MGAM* (rs28429916, $\text{EAF}_{\text{Dis}} = 0.287$, $\text{EAF}_{\text{Rep}} = 0.285$) and *SLC5A10* (rs11652575, $\text{EAF}_{\text{Dis}} = 0.164$, $\text{EAF}_{\text{Rep}} = 0.160$), while the other metabolites were each linked to a single gene. Acetoacetic acid, produced in the liver during the breakdown of fatty acids, is linked to carbohydrate metabolism (*MGAM*) and glucose transport (*SLC5A10*). Impairment in carbohydrate or

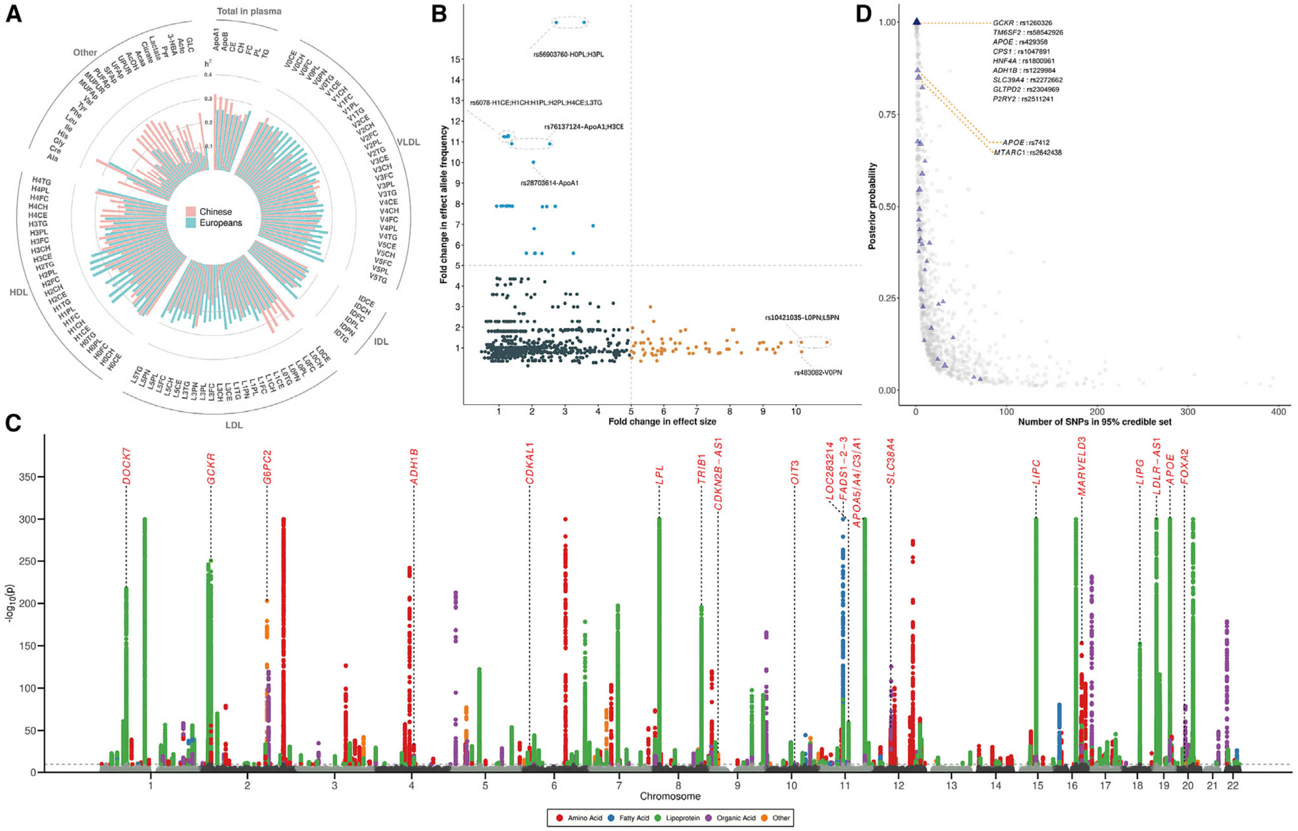


Figure 4. Results of cross-ancestry GWAS meta-analysis of metabolites in Chinese participants (n = 15,272) and European participants (n = 213,397)

(A) SNP-based heritability for each metabolite in Chinese cohorts and European individuals from the UK Biobank.

(B) Comparison of the effect size and effect allele frequency for 833 lead variant-metabolite associations in Chinese and European populations. Associations with a variant of fold change in effect allele frequency larger than 5 are highlighted in blue, and those with a fold change in effect size larger than 5 are highlighted in orange. Associations with a fold change > 10 are explicitly labeled.

(C) Manhattan plot of cross-ancestry meta-analyses. Colors indicate metabolite classes. Loci that harbored 228 additional associations that did not reach genome-wide significance threshold ($p > 5 \times 10^{-8}$) in original GWAS are highlighted in red.

(D) Distribution of 95% credible set size (x axis) against the maximum posterior probability of variants in each locus (y axis). The blue triangles mark missense SNPs, with size proportional to correlated metabolite numbers. Association mapping into small (≤5 SNPs) credible sets with a high posterior probability (≥80%) is explicitly labeled.

oxidative energy production may promote fatty acid breakdown, leading to increased acetoacetic acid under specific metabolic conditions.[21]

We also identified an association between ornithine and variant rs2147896 (EAF_{Dis} = 0.424, EAF_{Rep} = 0.430) on *PYROXD2*, a gene linked previously to asymmetric dimethylarginine,[5] trimethylamine,[22] and N6-methyllysine metabolism.[23] *PYROXD2* encodes a protein that may facilitate the conversion of metabolites containing CH-NH groups into ornithine.[23] In addition, missense SNP rs1801133 (EAF_{Dis} = 0.414, EAF_{Rep} = 0.430) in the *MTHFR* gene was associated with formic acid, a metabolite characterized less frequently in previous metabolomics studies. *MTHFR* is involved in one-carbon metabolism, potentially regulating formic acid levels.

Furthermore, we discovered previously overlooked associations involving loci that had been reported in prior studies. For example, while the *PAH* gene is known to be associated with phenylalanine,[13,15] we identified that the unreported nonsynonymous variant rs118092776 in *PAH* is linked to phenylalanine. The higher MAF of this variant in East Asian individuals (MAF_{EAS} = 0.0135 vs. MAF_{EUR} = 0.0006 from gnomAD) likely accounts for its omission in studies focusing exclusively on European populations. Additionally, we replicated the lead association of rs10942892 in *DMGHD* (which encodes an enzyme catalyzing the oxidative demethylation of dimethylglycine) with N,N-dimethylglycine, a derivative of amino acids produced during choline metabolism.

A more comprehensive understanding of the genetic architecture of circulating metabolites through cross-ancestry GWAS meta-analysis

To gain a more comprehensive, generalizable, and detailed understanding of the genetic influences on circulating metabolites, we conducted a cross-ancestry meta-analysis by pooling data

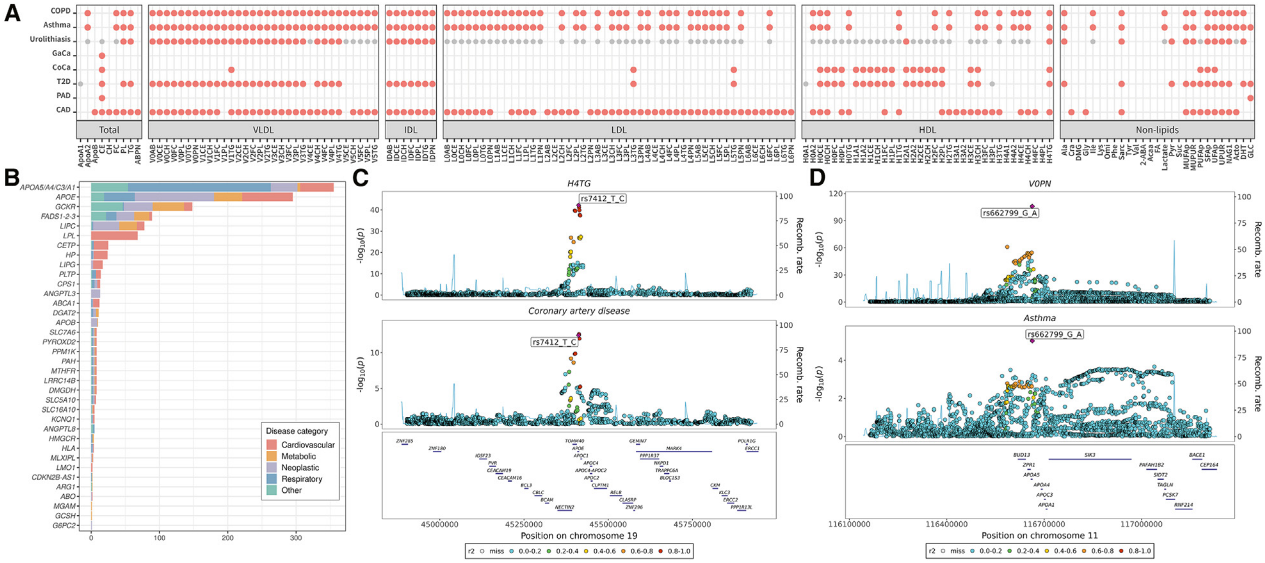


Figure 5. Colocalization of metabolite-related loci with complex disease

(A) Overview of the colocalization analyses between metabolites and different diseases. Red dots represent $PP_{H4} > 0.8$, and gray dots represent PP_{H4} between 0.6 and 0.8.

(B) Distribution of colocalization signals between different metabolite-related loci and diseases. The bars are color coded to represent different disease categories.

(C) Colocalization of H4TG and CAD in the *APOE* locus.

(D) Colocalization of V0PN and asthma in the *APOA5* locus. The color of the dots represents the degree of LD.

from our four Chinese cohorts (n = 15,272), including three discovery cohorts and one replication cohort, with the UK Biobank population of European ancestry (n = 213,397). This approach resulted in a large sample size of 228,669 participants and 116 overlapping metabolites. To account for ancestry-related heterogeneity in variant associations, we performed multi-ancestry meta-regression using MR-MEGA,[24] which incorporates genetic ancestry as a covariate (see STAR Methods).

Among the 5,480,458 shared autosomal variants with MAF > 1% between the Chinese and European populations, this cross-ancestry meta-analysis identified 918,394 significant variant-metabolite associations ($p < 4.31 \times 10^{-10} = 5 \times 10^{-8}/$116 metabolites) involving 5,781 lead associations, 1,138 lead variants, and 280 genomic loci (Figure 4C). Among these significant associations, 228 were newly identified and did not reach the genome-wide significance threshold in the individual GWAS ($p > 5 \times 10^{-8}$ in both the Chinese and European GWASs; Table S12). We further conducted cross-ancestry fine-mapping using an approximate Bayesian factor, identifying 1,095 signals (18.9% of 5,781) with a single putative causal variant (posterior probability > 95%; Table S13). The median size of the 95% credible set was seven variants (Figure S4B). Missense SNPs with posterior probability > 80% were identified in genes such as *GCKR*, *TM6SF2*, *APOE*, *CPS1*, *HNF4A*, *ADH1B*, *SLC39A4*, *GLTPD2*, *P2RY2*, and *MTARC1* (Figure 4D). As expected, the multi-ancestry GWAS produced smaller credible sets sizes and higher posterior probabilities compared with the single-ancestry GWAS (Figure S4B).

Of note, the exonic variant rs2429467 in *SLC38A4*, which exhibited an MAF discrepancy between populations (MAF_{EAS} = 0.26, $P_{Chinese} = 2.14 \times 10^{-5}$; MAF_{EUR} = 0.05, $P_{UK\ Biobank} = 2.50 \times 10^{-5}$; $P_{cross\text{-}ancestry} = 1.02 \times 10^{-10}$; Table S12), was associated with circulating levels of glycine in the cross-ancestry meta-analysis. Although previous research[25] reported associations between other variants in *SLC38A4* (which encodes a transmembrane transporter of small amino acids[26]) and glycine, this missense variant has not been reported before in relation to glycine before. Similarly, four additional exonic variants were identified in the cross-ancestry meta-analysis (Table S12), none of which reached significance in individual GWASs due to the relatively small sample size in the Chinese population and the low MAF in European individuals. This highlights the distinct benefit of cross-ancestry meta-analysis, which bolsters statistical power to detect associations involving alleles with diverse frequencies and effect sizes across different ethnic populations.

Colocalization of metabolite loci with human diseases in East Asian individuals

To investigate whether metabolites share genetic determinants with diseases, we conducted Bayesian colocalization analyses.[27] Among the 21 diseases included in the analyses (see STAR Methods), we examined their colocalization with 632 metabolite-associated genetic loci identified in Chinese individuals during the discovery stage. The posterior probability for colocalization (PP_{H4}) was calculated to determine whether two traits share a causal variant within a region. We identified 531 colocalization signals involving 145 metabolites and 8 diseases ($PP_{H4} > 0.8$; Figure 5A; Table S14), including coronary artery disease (CAD), peripheral artery disease (PAD), type 2 diabetes (T2D), asthma, chronic obstructive pulmonary disease

(COPD), colorectal cancer, gastric cancer, and urolithiasis. Previous studies have reported several colocalization signals between metabolites and diseases, particularly for cardiovascular diseases such as CAD and T2D, across multiple genomic regions.[28–30] Among these, CAD had the most colocalization signals ($n = 170$; Figure 5B; Table S14). Notable lipoprotein-related loci in the *APOA5/A4/C3/A1* gene cluster, *APOE*, and *LPL* genes showed colocalization with CAD (Figure 5C). However, studies examining metabolite-disease colocalization for diseases like asthma and COPD are less common. In our study, we identified colocalization signals between asthma and COPD with multiple lipoproteins in the *APOA5/A4/C3/A1* and *APOE* genes. For example, 104 colocalization signals were associated with asthma, most of which were in the *APOA5/A4/C3/A1* gene cluster. Specifically, V0PN and asthma may share the causal variant rs662799 with a high posterior probability ($PP_{H4} = 0.98$; Figure 5D).

Potential causal associations of metabolites with disease phenotypes revealed by two-sample MR analyses

Although emerging associations of metabolites with health outcomes have been reported, the causality of these associations remains largely unclear. To address this, we performed two-sample MR analyses in East Asian individuals, integrating our data with the diseases GWASs from approximately 200,000 Japanese individuals from Biobank Japan.[17] Based on prior knowledge and colocalization signals, 11 diseases were included in the MR analyses (see STAR Methods, Table S15; Figure S8A).

In the MR analyses among East Asian individuals, we identified 144 potential causal associations between metabolites and diseases after multiple comparison corrections ($p < 1.57 \times 10^{-4} = 0.05/[11 \text{ diseases} \times 29 \text{ PCs of metabolites}]$) involving 75 metabolites and 5 diseases (Figure 6A; Table S16). This included several associations not reported previously. Asthma was associated with the highest number of genetically predicted metabolites ($n = 54$), followed by CAD ($n = 51$), COPD ($n = 25$), PAD ($n = 11$) and T2D ($n = 3$). The MR-Egger intercept test showed no evidence of unbalanced pleiotropy (all $p > 0.05$ for intercepts; Table S16), and leave-one-out analyses showed that the causal associations remained robust even when excluding any single SNP (Figure S9).

For lipids, genetically predicted levels of 28 subclasses of very-low-density lipoproteins (VLDLs; odds ratio [OR] range per standard deviation: 1.14–1.45), 6 subclasses of intermediate-density lipoproteins (IDLs; 1.19–1.25), and 3 subclasses of LDL (1.48–1.73) were consistently associated with a higher risk of CAD (Table S16). Subclasses of high-density lipoproteins (HDLs) showed different patterns. In line with previous observational studies on clinically measured HDL cholesterol, its genetically predicted levels were, in general, inversely associated with CAD risk (Figure 6B). However, we found evidence that genetically predicted triglycerides in HDL were consistently associated with a higher risk of CAD. Specifically, each standard deviation increase in genetically predicted levels of triglycerides in very large HDL (H1TG) was associated with a 9% elevated risk of CAD (OR [95% confidence interval (CI)]: 1.09 [1.04–1.14], $p = 1.35 \times 10^{-4}$; Table S16). For PAD, most genetically predicted levels of cholesterol esters, total cholesterol, free cholesterol, and phospholipids in HDL were associated with a reduced risk (OR range: 0.66–0.80) while triglycerides in small HDL (H4TG) were linked to a higher risk (Figure 6B). These results highlight the value of detailed profiling of lipid components for disease risk reclassification. Additionally, we found that evidence that genetically predicted levels of several lipids were associated with a reduced risk of asthma and COPD, primarily involving subclasses of VLDL and IDL (Figure 6C). These observations were largely reproducible in European populations through analyses integrating metabolite GWASs from European individuals in the UK Biobank and disease GWASs from FinnGen (DF9; Figures 6B and 6C; Table S17).

For non-lipid metabolites, several potential causal associations were also observed. The causal association between glycine and a lower risk of CAD, reported previously in European individuals,[25] was successfully replicated in East Asian individuals (Figure 6D). Furthermore, we report evidence that genetically predicted glycine was suggestively associated with a lower risk of heart failure in both East Asian and European individuals (OR [95% CI] per standard deviation: 0.93 [0.88–0.97], $p = 1.52 \times 10^{-3}$ for East Asian; 0.96 [0.91–0.96], $p = 4.73 \times 10^{-3}$ for European; Figure 6D). Moreover, genetically predicted levels of N-acetyl-glycoprotein 1 (NAG1), a marker of acute phase glycoproteins, were associated with a reduced risk of asthma in East Asian individuals (OR: 0.72 [0.65–0.80], $p=1.68 \times 10^{-10}$; Figure 6E). N-acetylglucosamine, a component of NAG1, promotes vascular calcification by upregulating the osteogenic transcription factor Runx2 and activating Akt[31,32] while also exhibiting anti-allergic effects by reducing histamine release, interleukin-1β production, and nuclear factor κB activation.[33] Additionally, genetically predicted levels of sarcosine were associated with an increased risk of CAD (OR: 1.30 [1.15–1.47], $p=2.2 \times 10^{-5}$; Figure 6F) but a lower risk of asthma (OR: 0.74 [0.65–0.86], $p = 4.67 \times 10^{-5}$; Figure 6G) in East Asian individuals. However, NAG1 and sarcosine were not measured in the UK Biobank metabolomics platform.

DISCUSSION

In this metabolite GWAS involving 10,792 Chinese and 213,397 European individuals using a similar NMR-based metabolomics profiling platform, we identified 15 variant-metabolite associations, 8 of which were replicated in an independent Chinese cohort of 4,480 individuals. In addition, two-sample MR analyses revealed potential causal associations of certain circulating metabolites with complex diseases. Notably, genetically predicted triglycerides within HDL were associated with a higher risk of CAD, while genetically predicted levels of VLDL and IDL showed protective association with COPD in both Chinese and European populations. Furthermore, we identified a potential protective effect of glycine on heart failure. These findings deepen our understanding of human metabolic diversity across populations and highlight the influence of ethnic variation on disease susceptibility.

Our findings of the heritability for lipoproteins in European individuals (ranging from 0.17 to 0.37) echoed previous reports using a similar NMR platform in European populations.[8,9] In

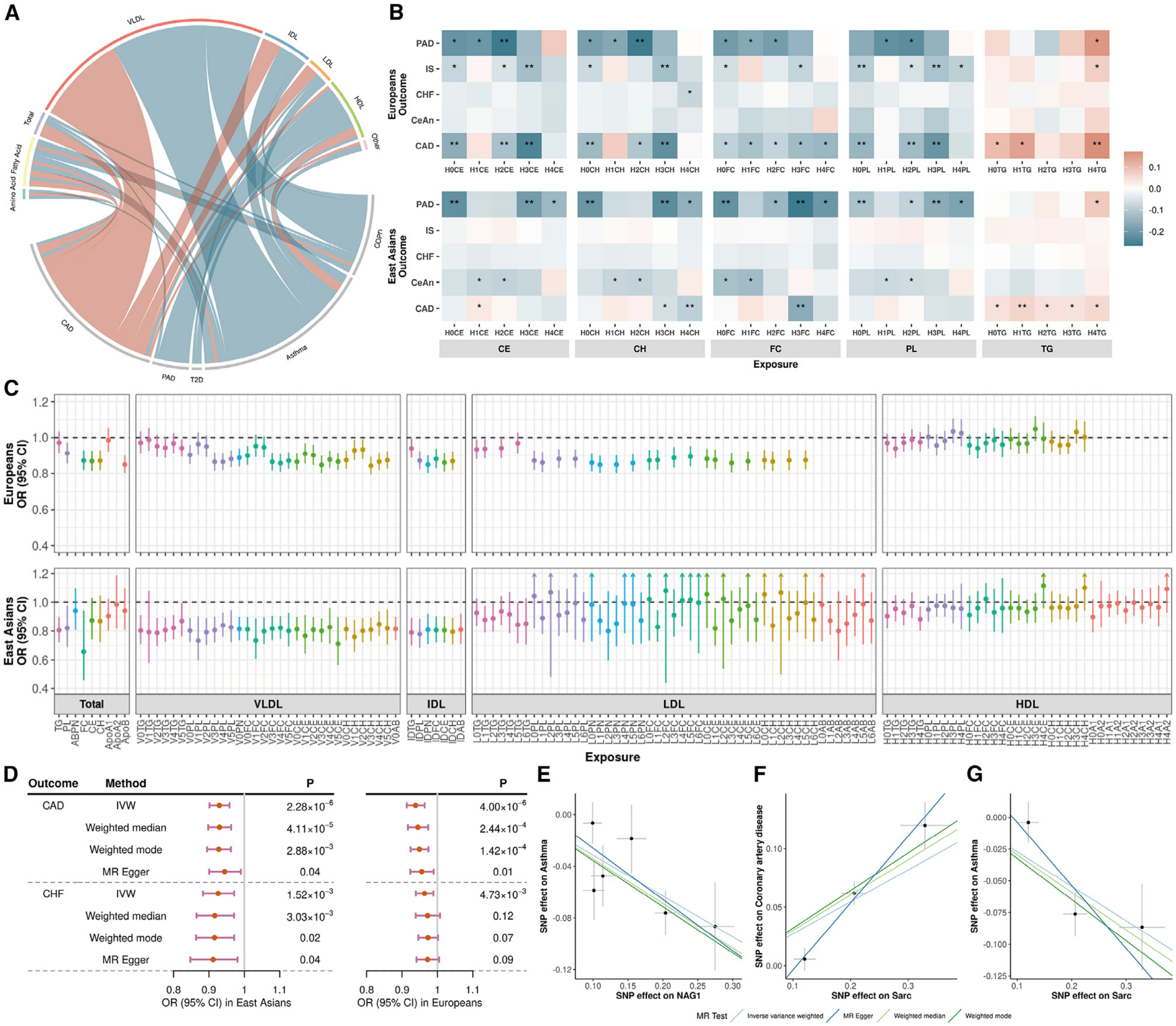


Figure 6. Results of two-sample MR analyses

(A) Overview of 144 potential causal associations between circulating metabolites (top) and diseases (bottom) in East Asian individuals. ORs greater than 1 are shown in red, while those less than 1 are depicted in blue. CAD, coronary artery disease; COPD, chronic obstructive pulmonary disease; PAD, peripheral artery disease; T2D, type 2 diabetes.

(B) The associations of genetically increased triglycerides in HDL with risk of cardiovascular disease in East Asian (bottom) and European (top) individuals. $^{**}p < 1.57 \times 10^{-4}$, $^{*}p < 0.05$). CE, cholesterol ester; CH, cholesterol; FC, free cholesterol; PL, phospholipid; TG, triglyceride; IS, ischemic stroke; CHF, congestive heart failure; CeAn, cerebral aneurysm.

(C) The associations of genetically predicted higher levels of lipids with risk of COPD in East Asian (bottom) and European (top) individuals.

(D) The associations of genetically predicted higher levels of glycine with risk of CAD and CHF in East Asian (left) and European (right) individuals.

(E–G) MR results of (E) NAG1 on risk of asthma, (F) sarcosine on CAD, and (G) sarcosine on asthma.

contrast, the genetic background of metabolites in East Asian populations has been explored less extensively. We report that the heritability for lipoproteins in Chinese individuals ranged from 0.07 to 0.35, which is comparable to that observed in European individuals. However, there are notable differences in metabolite-variant associations between these populations. Among the 1,584 significant lead associations identified in the Chinese population, 206 were absent in the UK Biobank, primarily due to a MAF of <0.01. Some metabolite-related variants that are prevalent in Chinese populations remain underrepresented in European GWAS. Among the 1,002 shared associations, 10.8% were not replicated due to minimal absolute effect sizes (<0.01) in European individuals. When comparing effect sizes across populations, caution is needed due to differences in allele frequencies, especially for low-MAF variants. These differences may lead to less accurate effect estimates, particularly in smaller sample sizes. Additionally, cross-ancestry meta-analyses have yielded deeper genetic insights and substantially enhanced the precision of fine-mapping compared to single-ancestry GWASs. These findings highlight the urgent need for metabolite GWASs in non-European populations.

Through MR analyses, we identified several potential causal associations between metabolites and diseases. While the relationship between LDL cholesterol and CAD risk is well established, the role of HDL cholesterol remains controversial. Observational studies typically suggested a beneficial effect of HDL cholesterol on cardiovascular health,[34] but clinical trials of drugs designed to raise circulating HDL cholesterol levels, such as cholesteryl ester transfer protein inhibitors, showed minimal effects on cardiovascular events.[35] With recent advances in metabolite profiling technologies, several studies have reported positive associations of TG within HDL particles with CAD risk,[36–38] though causality remains unclear. Our study provides causal evidence linking TG within HDL particles to an increased CAD risk in both East-Asian and European populations. These findings highlight the importance of the structure and components of HDL particles in CAD prevention.

Increasing clinical evidence has suggested that dysregulated lipid metabolism may contribute to the onset and progression of COPD.[39–42] A study utilizing NMR-based metabolomics identified lipoproteins, particularly LDL and VLDL, as key biomarkers for distinguishing COPD patients from healthy controls.[43] In our MR analyses, we report that genetically predicted higher circulating levels of VLDL and IDL subclasses were associated with a reduced risk of COPD in East Asian individuals, with similar findings replicated in European populations. However, the mechanisms underlying these associations remain unclear. Larger-scale metabolomics studies and further mechanistic research are needed to elucidate the causal links between altered lipid metabolism and lung dysfunction.

Our research has several notable strengths. It represents the largest metabolite GWAS conducted in a non-European population, helping to address the European-centric bias in genetic research on human metabolites. Using a consistent NMR metabolomics platform, genetic chip array, and data processing methods across both the discovery dataset (10,792 individuals) and replication dataset (4,480 individuals) ensured high data consistency. In addition, our cross-population comparison and meta-analysis, which incorporated extensive European data, provided an unprecedented perspective on genetic diversity and metabolite variation. The application of two-sample MR further enabled us to estimate the causal impact of metabolites on disease, revealing critical cross-ancestry differences and enriching our comprehension of disease mechanisms.

In conclusion, we have identified genetic loci of circulating metabolites in a large Chinese population and compared the genetic influences on metabolites across Chinese and European populations. Our findings also suggest potential causal associations between metabolites and a variety of complex diseases. These results deepen our understanding of the genetic determinants of circulating metabolites across different ancestries and enhance our knowledge of disease etiology.

Limitations of the study

First, although our sample is the largest for East Asian individuals, it remains smaller than those from European populations. Fasting status could also have influenced the genetic associations,[15] as samples from Chinese individuals were collected while fasting, whereas UK Biobank samples were not. Despite adjustments for fasting time, this discrepancy may have hindered cross-population validation. Additionally, the broad age range in the discovery dataset could have introduced age-related effects on metabolites. Furthermore, while we applied rigorous procedures to exclude outliers and select instrumental variables, the presence of pleiotropy in metabolite-related variants means that our MR analysis results may still be impacted by horizontal pleiotropy. Therefore, it is important to interpret the causal inference with caution. We conducted multiple sensitivity analyses to address this, and future studies could benefit from advanced multi-response MR methods to jointly model multiple outcomes.[44] Meanwhile, additional validation through larger observational studies and functional research is needed.

RESOURCE AVAILABILITY

Lead contact

Requests for further information should be directed to the lead contact, Yan Zheng (yan_zheng@fudan.edu.cn).

Materials availability

This study did not generate unique materials.

Data and code availability

The GWAS summary statistics from the meta-analysis results reported in this paper have been deposited in the Genome Variation Map[45] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation[46] under accession number GVP000053 (https://bigd.big.ac.cn/gvm/getProjectDetail?Project=GVP000053). The data were approved by the Ministry of Science and Technology of the People's Republic of China (2025BAT00220). UK Biobank data are available upon application (https://www.ukbiobank.ac.uk/). Other publicly available datasets used in this study are listed in the key resources table. The key codes of statistical analyses are accessible on GitHub (https://github.com/zhenglab2025/metabolite-GWAS), and additional codes and software used in this paper can also be found in the key resources table.

ACKNOWLEDGMENTS

This research was supported by the National Key R&D Program of China (2021YFA1301000, 2022YFC3400700, and 2022YFA0806400), the Shanghai Municipal Science and Technology Major Project (grants 2023SHZDZX02 and 2017SHZDZX01), a grant from the Shanghai Municipal Health Commission (2022JC012), the Program of Shanghai Academic Research Leader (22XD1423300), the Shanghai Oriental Youth Elite Talents Program (QNZH2024009), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. The data analysis server is supported by the Human Phenome Data Center of Fudan University, and we thank the center staff for their support. UK Biobank approved this research under application 54294. We gratefully acknowledge the contributors to the data used in this work.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.Z.; methodology, C.L., M.X., Y.D., Q.H., and X.H.; investigation, C.L.; formal analysis, C.L. and X.H.; writing – original draft, C.L. and Y.Z.; writing – review & editing, C.L., Y.Z., M.X., Y.D., Q.H., X.H., Z.S., Q.P., J.L. X.S., S.W., L.J., and X.H.; design and implementation of the web server, G.Z. and R.L.; funding acquisition, L.J. and Y.Z.; resources, Y.Z., X.H., M.X., Y.D., X.W., H.L., X.G., H.T., and L.J.; supervision, L.J., X.H., and Y.Z. All co-authors contributed to the interpretation of results and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - The Shanghai Changfeng Study
  - The Rugao Longevity and Aging Study
  - The National Survey of Physical Traits cohort
  - Validation in an external independent Chinese cohort
  - UK Biobank
- METHOD DETAILS
  - Genotyping, quality control, and imputation of genetic data
  - Nuclear magnetic resonance-based metabolomics profiling and data processing in Chinese cohorts
  - Metabolites measurement in European individuals from the UK Biobank
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Genome-wide association study and meta-analysis
  - Signal selection and locus definition
  - Stratified analyses in Chinese discovery cohorts with different sample types
  - Metabolite GWAS in the external Chinese validation cohort
  - Metabolite GWAS in European individuals from the UK Biobank
  - SNP-based heritability estimation
  - Conditional analysis
  - Statistical fine-mapping
  - Cross-ancestry meta-analysis and fine-mapping
  - Annotation of variants and genes
  - Identifications of associations
  - Colocalization between metabolites and Biobank Japan disease traits
  - Causal effects between metabolites and diseases estimated by two sample Mendelian randomization
  - Multivariable Mendelian randomization

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2025.100810.

Received: September 25, 2024
Revised: January 22, 2025
Accepted: February 17, 2025
Published: March 20, 2025

REFERENCES

1. Nicholson, G., Rantalainen, M., Maher, A.D., Li, J.V., Malmodin, D., Ahmadi, K.R., Faber, J.H., Hallgrímsdóttir, I.B., Barrett, A., Toft, H., et al. (2011). Human metabolic profiles are stably controlled by genetic and environmental variation. Mol. Syst. Biol. *7*, 525. https://doi.org/10.1038/msb.2011.57.
2. Suhre, K., and Gieger, C. (2012). Genetic variation in metabolic phenotypes: study designs and applications. Nat. Rev. Genet. *13*, 759–769. https://doi.org/10.1038/nrg3314.
3. Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., Deloukas, P., Erdmann, J., et al.; CARDIoGRAM (2011). Human metabolic individuality in biomedical and pharmaceutical research. Nature *477*, 54–60. https://doi.org/10.1038/nature10354.
4. Kettunen, J., Tukiainen, T., Sarin, A.P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.P., Kangas, A.J., Soininen, P., Würtz, P., Silander, K., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat. Genet. *44*, 269–276. https://doi.org/10.1038/ng.1073.
5. Rhee, E.P., Ho, J.E., Chen, M.H., Shen, D., Cheng, S., Larson, M.G., Ghorbani, A., Shi, X., Helenius, I.T., O'Donnell, C.J., et al. (2013). A genome-wide association study of the human metabolome in a community-based cohort. Cell Metab. *18*, 130–143. https://doi.org/10.1016/j.cmet.2013.06.013.
6. Shin, S.Y., Fauman, E.B., Petersen, A.K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.P., et al. (2014). An atlas of genetic influences on human blood metabolites. Nat. Genet. *46*, 543–550. https://doi.org/10.1038/ng.2982.
7. Draisma, H.H.M., Pool, R., Kobl, M., Jansen, R., Petersen, A.K., Vaarhorst, A.A.M., Yet, I., Haller, T., Demirkan, A., Esko, T., et al. (2015). Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. Nat. Commun. *6*, 7208. https://doi.org/10.1038/ncomms8208.
8. Kettunen, J., Demirkan, A., Würtz, P., Draisma, H.H.M., Haller, T., Rawal, R., Vaarhorst, A., Kangas, A.J., Lyytikäinen, L.P., Pirinen, M., et al. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat. Commun. *7*, 11122. https://doi.org/10.1038/ncomms11122.
9. Gallois, A., Mefford, J., Ko, A., Vaysse, A., Julienne, H., Ala-Korpela, M., Laakso, M., Zaitlen, N., Pajukanta, P., and Aschard, H. (2019). A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. Nat. Commun. *10*, 4788. https://doi.org/10.1038/s41467-019-12703-7.
10. Lotta, L.A., Pietzner, M., Stewart, I.D., Wittemans, L.B.L., Li, C., Bonelli, R., Raffler, J., Biggs, E.K., Oliver-Williams, C., Auyeung, V.P.W., et al. (2021). A cross-platform approach identifies genetic regulators of human metabolism and health. Nat. Genet. *53*, 54–64. https://doi.org/10.1038/s41588-020-00751-5.
11. Surendran, P., Stewart, I.D., Au Yeung, V.P.W., Pietzner, M., Raffler, J., Wörheide, M.A., Li, C., Smith, R.F., Wittemans, L.B.L., Bomba, L., et al. (2022). Rare and common genetic determinants of metabolic individuality and their effects on human health. Nat. Med. *28*, 2321–2332. https://doi.org/10.1038/s41591-022-02046-0.
12. Yin, X., Chan, L.S., Bose, D., Jackson, A.U., VandeHaar, P., Locke, A.E., Fuchsberger, C., Stringham, H.M., Welch, R., Yu, K., et al. (2022). Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. Nat. Commun. *13*, 1644. https://doi.org/10.1038/s41467-022-29143-5.
13. Chen, Y., Lu, T., Pettersson-Kymmer, U., Stewart, I.D., Butler-Laporte, G., Nakanishi, T., Cerani, A., Liang, K.Y.H., Yoshiji, S., Willett, J.D.S., et al. (2023). Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. Nat. Genet. *55*, 44–53. https://doi.org/10.1038/s41588-022-01270-1.
14. Feofanova, E.V., Brown, M.R., Alkis, T., Manuel, A.M., Li, X., Tahir, U.A., Li, Z., Mendez, K.M., Kelly, R.S., Qi, Q., et al. (2023). Whole-Genome Sequencing Analysis of Human Metabolome in Multi-Ethnic Populations. Nat. Commun. *14*, 3111. https://doi.org/10.1038/s41467-023-38800-2.
15. Karjalainen, M.K., Karthikeyan, S., Oliver-Williams, C., Sliz, E., Allara, E., Fung, W.T., Surendran, P., Zhang, W., Jousilahti, P., Kristiansson, K., et al. (2024). Genome-wide characterization of circulating metabolic biomarkers. Nature *628*, 130–138. https://doi.org/10.1038/s41586-024-07148-y.
16. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human

Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004.

17. Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., Sakaue, S., Matoba, N., Low, S.K., Okada, Y., et al. (2020). Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Nat. Genet. *52*, 669–679. https://doi.org/10.1038/s41588-020-0640-3.

18. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375. https://doi.org/10.1038/ng.2213.

19. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315. https://doi.org/10.1038/ng.2892.

20. Amendola, L.M., Dorschner, M.O., Robertson, P.D., Salama, J.S., Hart, R., Shirts, B.H., Murray, M.L., Tokita, M.J., Gallego, C.J., Kim, D.S., et al. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. Genome Res. *25*, 305–315. https://doi.org/10.1101/gr.183483.114.

21. Nelson, A.B., Queathem, E.D., Puchalska, P., and Crawford, P.A. (2023). Metabolic Messengers: ketone bodies. Nat. Metab. *5*, 2062–2074. https://doi.org/10.1038/s42255-023-00935-3.

22. Schlosser, P., Li, Y., Sekula, P., Raffler, J., Grundner-Culemann, F., Pietzner, M., Cheng, Y., Wuttke, M., Steinbrenner, I., Schultheiss, U.T., et al. (2020). Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. Nat. Genet. *52*, 167–176. https://doi.org/10.1038/s41588-019-0567-8.

23. Rhee, E.P., Surapaneni, A., Zheng, Z., Zhou, L., Dutta, D., Arking, D.E., Zhang, J., Duong, T., Chatterjee, N., Luo, S., et al. (2022). Trans-ethnic genome-wide association study of blood metabolites in the Chronic Renal Insufficiency Cohort (CRIC) study. Kidney Int. *101*, 814–823. https://doi.org/10.1016/j.kint.2022.01.014.

24. Magi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M.I., Cogent-Kidney Consortium, T.D.G.C., and Morris, A.P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum. Mol. Genet. *26*, 3639–3650. https://doi.org/10.1093/hmg/ddx280.

25. Wittemans, L.B.L., Lotta, L.A., Oliver-Williams, C., Stewart, I.D., Surendran, P., Karthikeyan, S., Day, F.R., Koulman, A., Imamura, F., Zeng, L., et al. (2019). Assessing the causal association of glycine with risk of cardio-metabolic diseases. Nat. Commun. *10*, 1060. https://doi.org/10.1038/s41467-019-08936-1.

26. Bröer, S. (2014). The SLC38 family of sodium-amino acid co-transporters. Pflügers Arch *466*, 155–172. https://doi.org/10.1007/s00424-013-1393-y.

27. Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. *17*, e1009440. https://doi.org/10.1371/journal.pgen.1009440.

28. Yuan, S., Xu, F., Li, X., Chen, J., Zheng, J., Mantzoros, C.S., and Larsson, S.C. (2023). Plasma proteins and onset of type 2 diabetes and diabetic complications: Proteome-wide Mendelian randomization and colocalization analyses. Cell Rep. Med. *4*, 101174. https://doi.org/10.1016/j.xcrm.2023.101174.

29. Klimentidis, Y.C., Arora, A., Newell, M., Zhou, J., Ordovas, J.M., Renquist, B.J., and Wood, A.C. (2020). Phenotypic and Genetic Characterization of Lower LDL Cholesterol and Increased Type 2 Diabetes Risk in the UK Biobank. Diabetes *69*, 2194–2205. https://doi.org/10.2337/db19-1134.

30. Ottensmann, L., Tabassum, R., Ruotsalainen, S.E., Gerl, M.J., Klose, C., Widén, E., FinnGen; Simons, K., Ripatti, S., and Pirinen, M. (2023). Genome-wide association analysis of plasma lipidome identifies 495 genetic associations. Nat. Commun. *14*, 6934. https://doi.org/10.1038/s41467-023-42532-8.

31. Heath, J.M., Sun, Y., Yuan, K., Bradley, W.E., Litovsky, S., Dell'Italia, L.J., Chatham, J.C., Wu, H., and Chen, Y. (2014). Activation of AKT by O-linked N-acetylglucosamine induces vascular calcification in diabetes mellitus. Circ. Res. *114*, 1094–1102. https://doi.org/10.1161/CIRCRESAHA.114.302968.

32. Wright, J.N., Collins, H.E., Wende, A.R., and Chatham, J.C. (2017). O-GlcNAcylation and cardiovascular disease. Biochem. Soc. Trans. *45*, 545–553. https://doi.org/10.1042/BST20160164.

33. Yoon, H.S., Byun, J.W., Shin, J., Kim, Y.H., and Choi, G.S. (2019). Therapeutic Effect of Glucosamine on an Atopic Dermatitis Animal Model. Ann. Dermatol. *31*, 538–544. https://doi.org/10.5021/ad.2019.31.5.538.

34. Goldstein, J.L., and Brown, M.S. (2015). A century of cholesterol and coronaries: from plaques to genes to statins. Cell *161*, 161–172. https://doi.org/10.1016/j.cell.2015.01.036.

35. Kaur, N., Pandey, A., Negi, H., Shafiq, N., Reddy, S., Kaur, H., Chadha, N., and Malhotra, S. (2014). Effect of HDL-raising drugs on cardiovascular outcomes: a systematic review and meta-regression. PLoS One *9*, e94585. https://doi.org/10.1371/journal.pone.0094585.

36. Boren, J., Taskinen, M.R., Bjornson, E., and Packard, C.J. (2022). Metabolism of triglyceride-rich lipoproteins in health and dyslipidaemia. Nat. Rev. Cardiol. *19*, 577–592. https://doi.org/10.1038/s41569-022-00676-y.

37. Tzoulaki, I., Castagné, R., Boulangé, C.L., Karaman, I., Chekmeneva, E., Evangelou, E., Ebbels, T.M.D., Kaluarachchi, M.R., Chadeau-Hyam, M., Mosen, D., et al. (2019). Serum metabolic signatures of coronary and carotid atherosclerosis and subsequent cardiovascular disease. Eur. Heart J. *40*, 2883–2896. https://doi.org/10.1093/eurheartj/ehz235.

38. Nordestgaard, B.G. (2016). Triglyceride-Rich Lipoproteins and Atherosclerotic Cardiovascular Disease: New Insights From Epidemiology, Genetics, and Biology. Circ. Res. *118*, 547–563. https://doi.org/10.1161/CIRCRESAHA.115.306249.

39. Freyberg, J., Landt, E.M., Afzal, S., Nordestgaard, B.G., and Dahl, M. (2023). Low-density lipoprotein cholesterol and risk of COPD: Copenhagen General Population Study. ERJ Open Res. *9*, 00496-2022. https://doi.org/10.1183/23120541.00496-2022.

40. Zafirova-Ivanovska, B., Stojkovikj, J., Dokikj, D., Anastasova, S., Debresliovska, A., Zejnel, S., and Stojkovikj, D. (2016). The Level of Cholesterol in COPD Patients with Severe and Very Severe Stage of the Disease. Open Access Maced. J. Med. Sci. *4*, 277–282. https://doi.org/10.3889/oamjms.2016.063.

41. Li, H., Liu, Y., Wang, L., Shen, T., Du, W., Liu, Z., Chen, R., and Hu, M. (2016). High apolipoprotein M serum levels correlate with chronic obstructive pulmonary disease. Lipids Health Dis. *15*, 59. https://doi.org/10.1186/s12944-016-0228-1.

42. Burkart, K.M., Manichaikul, A., Wilk, J.B., Ahmed, F.S., Burke, G.L., Enright, P., Hansel, N.N., Haynes, D., Heckbert, S.R., Hoffman, E.A., et al. (2014). APOM and high-density lipoprotein cholesterol are associated with lung function and per cent emphysema. Eur. Respir. J. *43*, 1003–1017. https://doi.org/10.1183/09031936.00147612.

43. Zheng, H., Hu, Y., Dong, L., Shu, Q., Zhu, M., Li, Y., Chen, C., Gao, H., and Yang, L. (2021). Predictive diagnosis of chronic obstructive pulmonary disease using serum metabolic biomarkers and least-squares support vector machine. J. Clin. Lab. Anal. *35*, e23641. https://doi.org/10.1002/jcla.23641.

44. Zuber, V., Lewin, A., Levin, M.G., Haglund, A., Ben-Aicha, S., Emanueli, C., Damrauer, S., Burgess, S., Gill, D., and Bottolo, L. (2023). Multi-response Mendelian randomization: Identification of shared and distinct exposures for multimorbidity and multiple related disease outcomes. Am. J. Hum. Genet. *110*, 1177–1199. https://doi.org/10.1016/j.ajhg.2023.06.005.

45. Li, C., Tian, D., Tang, B., Liu, X., Teng, X., Zhao, W., Zhang, Z., and Song, S. (2021). Genome Variation Map: a worldwide collection of genome

variations across multiple species. Nucleic Acids Res. *49*, D1186–D1191. https://doi.org/10.1093/nar/gkaa1005.

46. CNCB-NGDC Members and Partners (2022). Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. Nucleic Acids Res. *50*, D27–D38. https://doi.org/10.1093/nar/gkab951.

47. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature *613*, 508–518. https://doi.org/10.1038/s41586-022-05473-8.

48. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311. https://doi.org/10.1093/nar/29.1.308.

49. Cerezo, M., Sollis, E., Ji, Y., Lewis, E., Abid, A., Bircan, K.O., Hall, P., Hayhurst, J., John, S., Mosaku, A., et al. (2025). The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. Nucleic Acids Res. *53*, D998–D1005. https://doi.org/10.1093/nar/gkae1070.

50. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. Elife *7*, e34408. https://doi.org/10.7554/eLife.34408.

51. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. Nature *625*, 92–100. https://doi.org/10.1038/s41586-023-06045-0.

52. Zaslavsky, L., Cheng, T., Gindulyte, A., He, S., Kim, S., Li, Q., Thiessen, P., Yu, B., and Bolton, E.E. (2021). Discovering and Summarizing Relationships Between Chemicals, Genes, Proteins, and Diseases in PubChem. Front. Res. Metr. Anal. *6*, 689059. https://doi.org/10.3389/frma.2021.689059.

53. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

54. Wishart, D.S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B.L., et al. (2022). HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Res. *50*, D622–D631. https://doi.org/10.1093/nar/gkab1062.

55. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290. https://doi.org/10.1038/ng.3190.

56. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

57. Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6. https://doi.org/10.1038/nmeth.2307.

58. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529. https://doi.org/10.1371/journal.pgen.1000529.

59. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

60. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191. https://doi.org/10.1093/bioinformatics/btq340.

61. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164. https://doi.org/10.1093/nar/gkq603.

62. Lin, S.H., Brown, D.W., and Machiela, M.J. (2020). LDtrait: An Online Tool for Identifying Published Phenotype Associations in Linkage Disequilibrium. Cancer Res. *80*, 3443–3446. https://doi.org/10.1158/0008-5472.CAN-20-0985.

63. Benner, C., Spencer, C.C.A., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics *32*, 1493–1501. https://doi.org/10.1093/bioinformatics/btw018.

64. Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. *16*, e1008720. https://doi.org/10.1371/journal.pgen.1008720.

65. Verbanck, M., Chen, C.Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat. Genet. *50*, 693–698. https://doi.org/10.1038/s41588-018-0099-7.

66. Gao, X., Hofman, A., Hu, Y., Lin, H., Zhu, C., Jeekel, J., Jin, X., Wang, J., Gao, J., Yin, Y., and Zhao, N. (2010). The Shanghai Changfeng Study: a community-based prospective cohort study of chronic diseases among middle-aged and elderly: objectives and design. Eur. J. Epidemiol. *25*, 885–893. https://doi.org/10.1007/s10654-010-9525-6.

67. Wu, Q., Huang, Q.X., Zeng, H.L., Ma, S., Lin, H.D., Xia, M.F., Tang, H.R., and Gao, X. (2021). Prediction of Metabolic Disorders Using NMR-Based Metabolomics: The Shanghai Changfeng Study. Phenomics *1*, 186–198. https://doi.org/10.1007/s43657-021-00021-2.

68. Liu, Z., Wang, Y., Zhang, Y., Chu, X., Wang, Z., Qian, D., Chen, F., Xu, J., Li, S., Jin, L., and Wang, X. (2016). Cohort Profile: The Rugao Longevity and Ageing Study (RuLAS). Int. J. Epidemiol. *45*, 1064–1073. https://doi.org/10.1093/ije/dyv101.

69. Wang, F., Luo, Q., Chen, Y., Liu, Y., Xu, K., Adhikari, K., Cai, X., Liu, J., Li, Y., Liu, X., et al. (2022). A Genome-Wide Scan on Individual Typology Angle Found Variants at SLC24A2 Associated with Skin Color Variation in Chinese Populations. J. Invest. Dermatol. *142*, 1223–1227.e14. https://doi.org/10.1016/j.jid.2021.07.186.

70. Shalaimaiti, S., Dai, Y.-X., Wu, H.-Y., Qian, J.-Y., Zheng, Y., Yao, K., and Ge, J.-B. (2021). Clinical and Genetic Characteristics of Coronary Artery Disease in Chinese Young Adults: Rationale and Design of the ProspectiveGenetic characteristics of coRonaryArtery disease in ChiNese young aDults (GRAND) Study. Cardiol. *6*, 65–73. https://doi.org/10.4103/2470-7511.312594.

71. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.

72. Jimenez, B., Holmes, E., Heude, C., Tolson, R.F., Harvey, N., Lodge, S.L., Chetwynd, A.J., Cannet, C., Fang, F., Pearce, J.T.M., et al. (2018). Quantitative Lipoprotein Subclass and Low Molecular Weight Metabolite Analysis in Human Serum and Plasma by (1)H NMR Spectroscopy in a Multilaboratory Trial. Anal. Chem. *90*, 11962–11971. https://doi.org/10.1021/acs.analchem.8b02412.

73. Xia, M., Zeng, H., Wang, S., Tang, H., and Gao, X. (2021). Insights into contribution of genetic variants towards the susceptibility of MAFLD revealed by the NMR-based lipoprotein profiling. J. Hepatol. *74*, 974–977. https://doi.org/10.1016/j.jhep.2020.10.019.

74. Julkunen, H., Cichońska, A., Tiainen, M., Koskela, H., Nybo, K., Mäkelä, V., Nokso-Koivisto, J., Kristiansson, K., Perola, M., Salomaa, V., et al. (2023). Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. Nat. Commun. *14*, 604. https://doi.org/10.1038/s41467-023-36231-7.

75. Fuller, H., Zhu, Y., Nicholas, J., Chatelaine, H.A., Drzymalla, E.M., Sarvestani, A.K., Julián-Serrano, S., Tahir, U.A., Sinnott-Armstrong, N., Raffield, L.M., et al. (2023). Metabolomic epidemiology offers insights into disease

aetiology. Nat. Metab. *5*, 1656–1672. https://doi.org/10.1038/s42255-023-00903-x.

76. Richardson, T.G., Leyden, G.M., Wang, Q., Bell, J.A., Elsworth, B., Davey Smith, G., and Holmes, M.V. (2022). Characterising metabolomic signatures of lipid-modifying therapies through drug target mendelian randomisation. PLoS Biol. *20*, e3001547. https://doi.org/10.1371/journal.pbio.3001547.

77. Skrivankova, V.W., Richmond, R.C., Woolf, B.A.R., Yarmolinsky, J., Davies, N.M., Swanson, S.A., VanderWeele, T.J., Higgins, J.P.T., Timpson, N.J., Dimou, N., et al. (2021). Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. JAMA *326*, 1614–1621. https://doi.org/10.1001/jama.2021.18236.

78. Richardson, T.G., Sanderson, E., Palmer, T.M., Ala-Korpela, M., Ference, B.A., Davey Smith, G., and Holmes, M.V. (2020). Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. PLoS Med. *17*, e1003062. https://doi.org/10.1371/journal.pmed.1003062.

79. Zhao, Q., Wang, J., Miao, Z., Zhang, N.R., Hennessy, S., Small, D.S., and Rader, D.J. (2021). A Mendelian randomization study of the role of lipoprotein subfractions in coronary artery disease. Elife *10*, e58361. https://doi.org/10.7554/eLife.58361.

80. Sanderson, E., Davey Smith, G., Windmeijer, F., and Bowden, J. (2019). An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int. J. Epidemiol. *48*, 713–727. https://doi.org/10.1093/ije/dyy262.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
Illumina Infinium Chinese Genotyping Array	WeGene Co., Ltd.	N/A
Illumina Infinium Global Screening Array	WeGene Co., Ltd.	N/A
Deposited data		
GWAS summary statistics for metabolites	This paper	National Genomics Data Center: https://bigd.big.ac.cn/gvm/getProjectDetail?Project=GVP000053
Web server for the visualization of GWAS results	This paper	https://www.biosino.org/gwas/
UK Biobank data	UK Biobank	Available from the UK Biobank on application: www.ukbiobank.ac.uk/
GWAS summary statistics for diseases from BioBank Japan Project	BioBank Japan Project[17]	http://jenger.riken.jp/result
GWAS summary statistics for diseases from FinnGen (DF9)	Finngen[47]	https://www.finngen.fi/en/access_results
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/grc
1000 Genomes Project (Phase 3, v5)	The 1000 Genomes Project	https://www.internationalgenome.org/
NCBI dbSNP database	Sherry et al.[48]	https://www.ncbi.nlm.nih.gov/snp/
GWAS Catalog database	GWAS Catalog team[49]	https://www.ebi.ac.uk/gwas
MRC IEU OpenGWAS database	Hemani et al.[50]	https://gwas.mrcieu.ac.uk/
The Genome Aggregation Database (gnomAD)	The gnomAD Team[51]	https://gnomad.broadinstitute.org/
Combined Annotation Dependent Depletion (CADD)	Rentzsch et al.[19]	https://cadd.gs.washington.edu/
PubChem Chemical-Gene Co-occurrences in Literature database	Zaslavsky et al.[52]	https://pubchem.ncbi.nlm.nih.gov
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Kanehisa et al.[53]	https://www.genome.jp/kegg/
Human Metabolome Database (HMDB)	Wishart et al.[54]	https://hmdb.ca
Software and algorithms		
R (4.2.2)	R Foundation	https://www.r-project.org/
BOLT-LMM (2.4.1)	Loh et al.[55]	https://alkesgroup.broadinstitute.org/BOLT-LMM/
PLINK (2.0)	Chang et al.[56]	https://www.cog-genomics.org/plink/2.0/
SHAPEIT2	Delaneau et al.[57]	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
IMPUTE2	Howie et al.[58]	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
KING (v2.2.2)	Manichaikul et al.[59]	https://www.kingrelatedness.com/
METAL	Willer et al.[60]	https://github.com/statgen/METAL
ANNOVAR	Wang et al.[61]	https://annovar.openbioinformatics.org/en/latest/
LDtrait	Lin et al.[62]	https://ldlink.nih.gov/
FINEMAP (1.4.2)	Benner et al.[63]	http://christianbenner.com/
GCTA (1.94.0 beta)	Yang et al.[18]	https://yanglab.westlake.edu.cn/software/gcta
MR-MEGA	Magi et al.[24]	https://genomics.ut.ee/en/tools
R package TwoSampleMR (0.5.6)	Hemani et al.[50]	https://mrcieu.github.io/TwoSampleMR/
R package coloc (5.1.0)	Wallace et al.[64]	https://chr1swallace.github.io/coloc/
R package MR-PRESSO	Verbanck et al.[65]	https://github.com/rondolab/MR-PRESSO

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bruker IVDr Lipoprotein Subclass Analysis	Bruker Biospin	https://www.bruker.com/en/products-and-solutions/mr/nmr-clinical-research-solutions/b-i-lisa.html
Bruker IVDr Quantification in Plasma/Serum	Bruker Biospin	https://www.bruker.com/en/products-and-solutions/mr/nmr-clinical-research-solutions/b-i-quant-ps.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The Shanghai Changfeng Study

The Shanghai Changfeng Study is a prospective community-based cohort study, which enrolled 6,595 individuals that aged 45 years or older from June 2009 to December 2012 in Changfeng community, Shanghai, China. Participants underwent all clinical examination in local health community center and provided demographic data, lifestyle factors and medical history. Biomaterials including fasting blood and urine were collected and stored at −80°C. The study was conducted with the official approval of the ethical committee of Zhongshan Hospital, Fudan University, and each participant has provided written informed consent. Detailed description of the study design has been published elsewhere.[66,67] A total of 5,666 Han Chinese individuals who underwent both metabolic profiling and genome-wide genotyping were included in this analysis.

The Rugao Longevity and Aging Study

The Rugao Longevity and Aging Study (RuLAS) is a population-based, two-arm observational cohort study conducted in Rugao city, Jiangsu Province, China. A detailed description was provided in previous publication.[68] The study was approved by Human Ethics Committee of the School of Life Sciences, Fudan University. In the baseline, 1960 elderly participants aged 70–84 years were recruited between November 2014 and December 2014. Demographic, physiological and clinical data, as well as several biomaterials were collected. Three times follow-up were conducted within 5 years after baseline. In December 2019, the total number of participants reached 2200 after the Wave 4 follow-up. A subsample of 2113 Han Chinese individuals with both metabolomics data and genotyping data available were included in the current study.

The National Survey of Physical Traits cohort

The National Survey of Physical Traits (NSPT) cohort is a population-based prospective cohort that recruited Han Chinese participants in 2015–2018 from three cities: Nanning, Guangxi province, Taizhou, Jiangsu Province and Zhengzhou, Henan province.[69] In total, 3564 samples of Han Chinese participants were collected. The NSPT cohort study was approved by Ethics Committees of Fudan University and the Shanghai Institutes for Biological Sciences. All participants provided written informed consent. Our analysis set comprised 3013 Han Chinese individuals that had both metabolomics data and genotyping data.

Validation in an external independent Chinese cohort

The Genetic characteristics of coRonary Artery disease in ChiNese young aDults (GRAND) study is a multicenter, hospital-based observational clinical study.[70] The GRAND study recruited adult participants from 38 cardiac centers across China between May 2017 and April 2019. In-hospital clinical data were obtained from electronic medical records. At admission, each participant provided a fasting blood sample and completed a basic questionnaire on their lifestyle, dietary habits, and physical activity. The study protocol was approved by the central ethics committee at Zhongshan Hospital (Fudan University, Shanghai, China), as well as by the ethics committees at all participating centers (B2017-051). All participants provided written informed consent.

UK Biobank

UK Biobank is a large-scale prospective cohort study that includes around 500,000 participants from across the United Kingdom, aged 40 to 69 at the time of enrollment. The study provides comprehensive phenotypic data, covering biological measurements, lifestyle factors, and clinical blood biomarkers. Information about the cohort, as well as the data generation and imputation processes, has been described in detail elsewhere.[71] UK Biobank has approval from the North West Multi-centre Research Ethics Committee as a Research Tissue Bank approval. All participants provided written informed consent. This study applied for and received approval from the UK Biobank (Application Number #54294).

METHOD DETAILS

Genotyping, quality control, and imputation of genetic data

The DNA quality control and genotyping were carried out separately for each discovery cohort of Chinese individuals at the WeGene Clinical Laboratory, Shenzhen. The participants in the RuLAS cohort were array genotyped on Illumina Infinium Chinese Genotyping

Array BeadChip (Illumina WeGene V3 Arrays, ∼700k SNPs) and the other two Chinese discovery cohorts were genotyped on Illumina Infinium Global Screening Array BeadChip (Illumina WeGene V2 Arrays, ∼700k SNPs). Genotyping was performed by Illumina iScan System follow the manufacturer's instructions. Genetic data cleaning and quality control were carried out using PLINK.[56] We excluded samples with missing call rates >5% or with a mismatch between genetically inferred sex and self-reported sex. On the variant-level, we filtered out SNPs that had genotype missingness >2% and that deviated from Hardy-Weinberg equilibrium (HWE, $p < 10^{-5}$).

Genotype imputation was also performed for each cohort separately. The chip genotype data were pre-phased using SHAPEIT2[57] and then imputed to the 1000 Genome Project Phase 3 V5 reference panel using IMPUTE2.[58] Quality control was carried out again after imputation. The Bi-allelic SNPs with call rate >98%, HWE-P>10^{-5}, MAF >1% and imputation quality score (INFO) > 0.4 were retained in further analysis. All genomic positions were based on hg19/GRCh37. The total numbers of SNPs before and after imputation for each cohort were shown in Table S1. Joint PCA of Chinese cohorts were shown in Figure S10.

Nuclear magnetic resonance-based metabolomics profiling and data processing in Chinese cohorts

The serum or plasma metabolic traits for each Chinese cohort were measured separately using the same quantitative high-throughput NMR metabolomics platform.[67,72] Briefly, fasting blood metabolomics profiling was performed on a 600 MHz AVANCE III NMR spectrometer equipped with a BBI probe (Bruker Biospin GmbH, Germany), following the method reported previously.[72,73] The analytical samples were stored at −80°C and thawed at 24°C within 30 min, and all subsequent operations were performed upon the ice. Two NMR spectra were acquired at 310K using a standard NOESYGPPR1D and LEDBPGPPR2S1D pulse sequence, respectively, with 98K data points and 32 transients. Using the Bruker IVDr Lipoprotein Subclass Analysis (B.I.LISA) method (Bruker Biospin), 132 lipoprotein parameters were quantified. Lipoprotein subclasses were systematically categorized and sequentially labeled according to their density (Figure 2A), including five categories of very-low-density lipoproteins (VLDL 1–5, from extremely large to small particle size), one category of intermediate-density lipoprotein (IDL), six categories of low-density lipoproteins (LDL 1–6, from extremely large to very small particle size), and four categories of high-density lipoproteins (HDL 1–4, from very large to small particle size). A higher positive number within a category indicates a smaller particle diameter, while the number 0 represents the total sum of that specific component. In addition, lipoprotein particles were classified according to their biochemical composition, such as apolipoprotein A1/A2/B100 (A1/A2/AB), total cholesterol (CH), cholesterol esters (CE), free cholesterol (FC), phospholipids (PL), and triglycerides (TG). Additionally, 41 other metabolites, including amino acids, ketone bodies, glucose, and carboxylic acids, were quantified from the same spectra using the Bruker IVDr Quantification in Plasma/Serum (B.I.Quant-PS) method (Bruker Biospin). Meanwhile, six ratio parameters for fatty acids and two N-acetyl-glycoproteins (NAG1, NAG2) were obtained from the diffusion-edited spectra. A total of 181 metabolites were directly measured by this platform. All metabolites, except for formic acid and dimethyl sulfone, have been previously characterized in earlier GWASs using NMR or other metabolomics platforms. Lipoprotein measurements by the NMR-metabolomics platform were generally concordant with clinical measurements (Figure S11).

Of the 181 metabolites measured, we excluded 10 that had missing measurement in more than 60% of the samples. No participants were removed on the basis of having more than 30% missing metabolomic data, as no individual exceeded this threshold. The missing values of each metabolite were then imputed with half of the lowest detected values after quality control. All metabolites were transformed to normal distribution using inverse rank-based normal transformation and used for association testing. A total of 171 metabolites were included in the following GWAS. We compared the metabolites included in the current study with those from previous studies (NMR or other metabolomic platforms) to determine if they had been measured before (Table S2).

Metabolites measurement in European individuals from the UK Biobank

Metabolite profiling in the UK Biobank was conducted by Nightingale Health, measuring metabolic biomarkers for EDTA plasma samples with NMR-based metabolomics platforms.[74] Briefly, two NMR spectra were recorded for each plasma sample using a 500 MHz NMR spectrometer (Bruker AVANCE IIIHD), including a presaturated proton NMR spectrum (mainly for proteins and lipids within various lipoprotein particles) and a T2-relaxation-filtered spectrum (for detecting low-molecular-weight metabolites). Using the NMR-metabolomics platform similar to that used in the Chinese population, 116 overlapping metabolites were directly measured in the UK Biobank. In general, the NMR-based metabolomic profiles from both the Chinese and UK Biobank populations are largely consistent, with differences observed only for a few specific metabolites and lipoprotein components. For a small subset of low-molecular-weight metabolites, such as organic acid or amino acid, differences may arise from variations in the spectra. The major discrepancies are seen in lipid-related metabolites, particularly because the UK Biobank NMR platform does not differentiate the apolipoproteins within each lipoprotein class (e.g., Apo-A1 and Apo-A2 in HDL, and ApoB in IDL, LDL, and VLDL). Detailed information on metabolites is provided in Table S2.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genome-wide association study and meta-analysis

A GWAS was performed for 171 metabolites in each discovery cohort of Chinese individuals, totaling up to 10,792 individuals with both NMR metabolomics data and genomic data available. To maximize the sample size, the GWAS was carried out using linear mixed model (LMM) under an additive genetic model in BOLT-LMM (2.4.1) software.[55] SNPs with MAF < 1% and low imputation quality (INFO < 0.4) were excluded to avoid false positive association in the analyses. When the BOLT-LMM method failed to estimate the heritability (values close to zero or one), alternative analyses were carried out using PLINK software with linear regression model excluding related individuals. The individuals with second-degree or closer (kinship coefficient > 0.0884) relatives were inferred by KING (v2.2.2) software[59] and were removed when GWASs were run in PLINK. All the analyses were adjusted for age, sex, usage of lipid-lowering medications and top five genetic ancestry principal components. Top five principal components were selected as covariates because they accounted for the main variance of population stratification in the Chinese discovery cohort (Figure S12).

For each metabolite, association results from each participating cohort were then combined using inverse-variance-weighted fixed-effect meta-analysis based on effect sizes and standard errors in METAL software.[60] The meta-analyses were conducted under genomic control correction as implemented in METAL and the results were restricted to variants that presented in more than two cohorts. Heterogeneity among different cohorts was estimated by the I^2 statistics. The study-wide statistical significance threshold was set to $p < 1.72 \times 10^{-9}$ (standard genome-wide significance threshold $5 \times 10^{-8}/29$, correcting for 29 principal components that explained over 95% of variance of the metabolite data in the largest Shanghai Changfeng cohort, Figure S1) in the meta-analyses. The genomic inflation factor for each metabolite was calculated as the ratio of the median of the empirically observed chi-squared test statistics to the expected median (Table S5).

Signal selection and locus definition

For each metabolite, we ranked all associated variants at study-wide significance threshold ($p < 1.72 \times 10^{-9}$) and denoted the variant with lowest p value as the metabolite-sentinel variant. Corresponding metabolite-associated locus was defined as ±500 kilobase (Kb) windows centered around the metabolite-sentinel variant and significantly associated variants were then assigned to the locus. We repeated the procedure for the unassigned variants until no further significant variants remained. Overlapping or physically close (within 250 kb) metabolite-associated loci were merged. Across metabolites, overlapping metabolite-associated loci within chromosomes were finally merged.

Due to the potential existence of multiple independent signals within a single locus, we utilized an LD-based clumping approach to determine independent signals at a given locus. We defined independent signals with $r^2 < 0.1$ in PLINK using 1000 Genomes phase 3 v5 EAS [GRCh37/hg19] as reference panel. After each clumping, variants with the smallest p value were identified as lead SNPs.

Stratified analyses in Chinese discovery cohorts with different sample types

A recent large-scale NMR metabolite GWAS study has suggested that sample type may affect research findings.[15] In our sensitivity analysis, we conducted stratified analyses by blood sample type with comparable population size (serum, Shanghai Changfeng study, $n = 5{,}666$; plasma, meta-analysis of NSPT and RuLAS, $n = 5{,}126$, Table S1). We observed a high correlation in the effect sizes of 1,584 independently significant associations between the two blood sample types ($R^2 = 0.928$, $p < 2.2 \times 10^{-16}$; Figure S13 and Table S18).

Metabolite GWAS in the external Chinese validation cohort

We further attempted to replicate our findings in an external Chinese cohort. The participants were array genotyped on Illumina Infinium Chinese Genotyping Array BeadChip (Illumina WeGene V3 Arrays, ~700k SNPs). The imputation and quality control of genotype data were consistent with Chinese discovery cohorts mentioned above. The metabolites were also profiled using the same quantitative high-throughput NMR metabolomics platform described in the discovery dataset. A total of 4,480 participants with genotype data and metabolomic data were included in the replication set. The association analyses were conducted in BOLT-LMM (2.4.1) software. The analyses were adjusted for age, sex and top five genetic ancestry principal components, usage of lipid-lowering medication, and CAD status.

Metabolite GWAS in European individuals from the UK Biobank

We utilized the UK Biobank baseline metabolomics data from the latest release in 2023, comprising 280,000 individuals approximately. It is noteworthy that the majority were non-fasting samples, with an average fasting time of 3 h (Figure S14). We applied an inverse rank-based normal transformation to normalize the value of metabolites. We conducted GWAS analysis of metabolites in individuals of European ancestry using UK Biobank dataset. The UK Biobank study has been described in detail.[71] Following central quality control procedures of UK Biobank,[71] we excluded (1) individuals who had withdrawn consent, (2) non-Caucasian participants, (3) those without genotyping data or NMR-metabolomic data, (4) those with sex chromosome aneuploidy, (5) those with a mismatch between genetically inferred sex and self-reported sex, and (6) outliers for heterozygosity or missing rate, and eventually included 213,397 participants in the analyses (Table S1 and Figure S15). Analyses were limited to variants that did not deviate from HWE ($p > 10^{-5}$), with an INFO >0.8 and MAF >1%. GWASs were performed in PLINK software, with the adjustment of age, sex, fasting time, genotyping chips, assessment centers, metabolomic profiling batch, usage of lipid-lowering medications, and the top ten principal components (Figure S16). Due to the larger sample size of the UK Biobank, we applied a more conservative Bonferroni-corrected threshold ($p < 4.31 \times 10^{-10} = 5 \times 10^{-8}/116$ metabolites) for significance.

SNP-based heritability estimation

For Chinses populations, a genetic relationship matrix (GRM) was calculated using all autosomal SNPs with INFO>0.4, HWE-P>10^{-5}, and MAF>1% using GCTA-GRM. For UK Biobank, we randomly selected 20,000 unrelated European participants with both NMR-metabolomic and imputed genotype data to build GRM as well. SNP-based heritability of each metabolite was calculated using GCTA-REML with the same covariates as GWASs.

Conditional analysis

To identify additional independent signals, we carried out stepwise conditional analyses (–cojo-slct) using the results of meta-analyses for each metabolite in GCTA software.[18] We used genotype data from the Shanghai Changfeng Study, the largest participating cohort in the meta-analysis of discovery dataset, as the reference for LD calculation. The threshold of significance was set at $p < 1.72 \times 10^{-9}$ and the collinearity cutoff was $r^2 = 0.9$ within a 10Mb window.

Statistical fine-mapping

For each of the metabolite-associated loci, statistical fine-mapping was carried out using the FINEMAP (1.4.2) with the shotgun stochastic search algorithm[63] to identify potential causal variants. According to the FINEMAP's instruction, the pairwise LD matrix of SNPs for the Chinese population was estimated using PLINK based on genotype data from the Shanghai Changfeng Study. For computational efficiency, the LD matrix for European individuals was estimated based on a randomly selected group of 20,000 unrelated European individuals from UK Biobank who passed quality control as mentioned above in SNP-based heritability estimation.

For each locus, the FINEMAP outputs the Bayes factor and the posterior probability of each variant being causal for the association. To construct a 95% credible set of potential causal variants, variants were ranked in descending order of posterior probability, and then were included until the cumulative posterior probability exceeded 0.95.

Cross-ancestry meta-analysis and fine-mapping

We conducted cross-ancestry GWAS meta-analyses using genome-wide summary statistics from all four Chinese cohorts combined with European summary statistics from the UK Biobank. To account for ancestral heterogeneity among these studies, we applied the MR-MEGA algorithm[24] in the cross-ancestry meta-analyses, which allows for heterogeneity between diverse ancestry groups and includes axes of genetic variation among ancestry as covariates in the model.

We further conducted fine-mapping for cross-ancestry GWAS results. In order to identify the most likely causal variants for each genetic locus, we ranked the variants within ±500 kb of the lead variant based on their Bayes factors obtained from MR-MEGA analysis. The posterior probability for each variant was calculated by dividing cumulative Bayes factor of ranked variants by the total cumulative Bayes factor in each locus. We accumulated the posterior probability values from the largest one until the joint probability $\geq$95%, and the list of SNPs formed the 95% credible set. Subsequently, we utilized the Wilcoxon rank-sum test to investigate whether the precision of identifying causal variants is enhanced through cross-ancestry fine-mapping, as compared to an analysis focused solely on East Asian-specific data.

Annotation of variants and genes

We annotated all variants using the ANNOVAR software[61] in hg19 coordinates. ANNOVAR categories identify the SNP's gene hit or nearby, genic position (for example, intron, exon, intergenic) and associated function. Meanwhile, the allele frequency of the variants in East Asian and Non-Finn European individuals were annotated with gnomAD genome database.[51] The CADD (v1.3) phred-scaled score was used as a measure of predicted deleteriousness.[19] A variant was considered deleterious when the CADD score was above the suggested threshold of 12.37.[20]

We attempted to assign the most likely causal gene to each lead SNPs based on the following process: (1) We first retrieved protein-coding genes within a 1 Mb region of lead variants. Subsequently, based on the information from associated metabolites, we queried databases including HMDB,[54] the KEGG pathway database,[53] and the PubChem Chemical-Gene Co-occurrences in Literature database[52] to determine whether these genes were involved in clear biological processes related to the associated metabolites. (2) In case of the absence of a clear biological fit, we further checked if either the variant itself or its LD partner ($R^2 > 0.8$) was a missense/splice/stop-gained variant for a known gene. (3) Finally, for unassigned variants, we defined the nearest gene as the likely causal gene.

Identifications of associations

Based on the insights from metabolite epidemiological studies,[75] we compiled a list of previously published GWASs for association queries. This list included (1) GWASs of NMR-based metabolites, (2) GWASs of clinically measured lipids, and (3) other platform-based metabolite GWAS, from European, East Asian, South Asians, and Middle Eastern populations (Table S10). Given our use of the latest individual-level NMR data from the UK Biobank (n$\sim$200,000) for GWAS, we did not include NMR GWAS studies that solely used the UK Biobank as a single cohort.[76] To assess the novelty of our metabolite associations, we utilized LDtrait[62] (last updated on May 1, 2024) and MRC IEU OpenGWAS database[50] to determine whether each lead SNP along with its LD partners ($r^2 > 0.1$ within 500Kb using genotypes of EAS or EUR ancestry from 1000 Genomes Project phase 3 v5 as the LD reference panel) had been previously linked to the corresponding metabolite. Variant–metabolite pairs were considered novel if neither these lead variants nor their LD partners were previously associated ($p < 5 \times 10^{-8}$) with the same metabolites or lipoproteins in prior studies, including our metabolite GWAS in the UK Biobank.

Colocalization between metabolites and Biobank Japan disease traits

In order to test for shared causal variants between metabolites and Biobank Japan disease traits, we carried out Bayesian colocalization analyses using the R package 'coloc' (v5.1.0).[27] From the Biobank Japan GWAS of 42 diseases,[17] we excluded 20 diseases, including 7 sex-specific diseases which require sex-specific summary statistics for analysis, 3 infectious diseases with very limited genetic inheritance, 3 allergic diseases which were tightly associated with immunity rather than lipoproteins, and 8 diseases with insufficient cases (prevalence <1%), resulting in 21 diseases included in the following analyses (Table S15). We initially employed the Sum of Single Effects (SuSiE) regression-based colocalization (COLOC-SuSiE) method, which can identify multiple potential causal variants within a region. When SuSiE failed to converge after 10,000 iterations, we reverted to using COLOC-single. COLOC-SuSiE requires a matrix of linkage disequilibrium values, which we generated from the EAS samples in the 1000 Genomes phase 3 data with PLINK. For each metabolite, SNPs within the 1Mb range of the lead SNP from the meta-analyses were extracted and included in the colocalization analyses. We used the default priors for the analyses. The posterior probability values were estimated for H_0 (neither trait has a genetic association), H_1/H_2 (only the metabolite or disease has a genetic association); H_3 (two traits are associated but with different causal variants in the region), and H_4 (two traits share one causal variant in the region). Posterior probability values for H_4 greater than 0.8 were considered colocalized.

Causal effects between metabolites and diseases estimated by two sample Mendelian randomization

Two-sample MR analyses were performed to estimate the causal effects between circulating metabolites and diseases in East Asian individuals (Figure S8). We followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)-MR guidelines[77] (Table S19). We utilized the meta-analysis of metabolite GWAS from Chinese populations as the source for exposure. Given that Japanese population shared a similar genetic architecture with Chinese population, we obtained disease GWASs from Biobank Japan as the source for outcomes.[17] In short, the GWASs were performed in approximately 200,000 Japanese individuals using a generalized LMM using SAIGE adjusted for age, sex and the top five principal components (based on GRCh37/hg19).[17] Specifically, there was no overlapping of individuals between the cohorts for exposure and outcome.

We adopted two parallel criteria to select the 11 diseases in the final analysis: (1) Based on the characteristics of the NMR-metabolite platform (focusing on lipids) and prior knowledge on the disease risk influenced by lipids, we selected 6 cardiometabolic diseases. (2) We performed pairwise colocalization analyses between the metabolites and 21 included diseases in Biobank Japan, further including 5 diseases demonstrating colocalization signals with metabolites ($PPH_4 > 0.8$), i.e., asthma, chronic obstructive pulmonary disease, colorectal cancer, gastric cancer and urolithiasis. The detailed information can be found in Figure S8A and Table S14.

For each metabolite, we selected SNPs at a standard genome-wide significant threshold ($p < 5 \times 10^{-8}$) and LD $r^2 < 0.1$ for clumping (1000 Genomes phase 3 v5 EAS [GRCh37/hg19] as reference panel) to build instrumental variables. Summary statistics from our Meta-analyses and BBJ were harmonized to make alignment on effect alleles before every MR analysis. Palindromic SNPs (i.e., SNPs with A/T or G/C) with intermediate allele frequencies (MAF>0.42) were directly excluded. To assess the strength of selected instrumental SNPs and avoid weak instrument bias, we calculate F statistic for each instrumental variable and SNPs with F statistic >10 were considered as strong instruments. Meanwhile, we also applied MR-PRESSO method[65] to detect and filter out any outliers SNP that may contribute to horizontal pleiotropy. SNPs after these stringent selections were then used as instrumental variables in the subsequent two sample MR.

For all metabolomic traits, we used the inverse variance-weighted method to perform univariable MR. Meanwhile, the MR-Egger regression, the weighted-median method, and the weight-mode method were used as routine sensitivity analyses. Considering potential bias due to pleiotropy that documented in metabolite GWAS,[10,13,15] we employed two strategies based on different metabolite categories (Figure S8B). For lipids, we additionally performed multivariable MR, adjusting each lipid for TC, TG, LDL-c and HDL-c, respectively.[78,79] For non-lipid metabolic traits, we constructed a stricter IV for each metabolite following recent practice,[15] excluding pleiotropic variants that had more than two associations across all 171 metabolites identified in our study (the criterion was based on the median number of metabolite associations per lead SNP). A statistically significant causal effect was defined by a Bonferroni-corrected threshold ($p < 1.57 \times 10^{-4}$ = 0.05/[11 diseases×29 PCs of metabolites]) and required consistency in the direction of association across the various above-mentioned MR methods.

Heterogeneity was assessed using Cochran's Q statistics. The presence of potential directional horizontal pleiotropy was assessed by intercept term in MR-Egger method. Leave-one-out analysis were performed by removing each SNP in turn to determine whether the causal estimate was driven by any single SNP. All the TSMR analyses were performed in R software (4.2.2) with the "TwoSampleMR" package (0.5.6).[50]

The significant causal association pairs of metabolite and disease identified in East Asian individuals were then replicated in European populations. The instrumental variables were extracted from the metabolite GWAS in European individuals from the UK Biobank following instrumental variable selection procedure above. The summary statistics for diseases were obtained from FinnGen (DF9)[47] and the detailed information can be found in Table S15. The MR analyses was performed as described above.

Multivariable Mendelian randomization

For lipid metabolic traits, we performed multivariable MR, adjusting each lipid for TC, TG, LDL-c and HDL-c, respectively.[78,79] In each model, we combined SNPs at a standard genome-wide significant threshold ($p < 5 \times 10^{-8}$) for each lipid into a set of instrumental variables. We then clumped these SNPs and re-extracted the final clumped SNPs from the original summary statistics of each lipid. Data harmonization was also performed as above described. The conditional F statistics[80] for each lipoprotein was calculated to assess the instrument strength. The direct effect of each lipid on outcome in the multivariable MR was estimate using the IVW method.