Genotype-based "virtual" metabolomics in a clinical biobank identifies novel metabolite-disease associations

Minoo Bagheri
  Vanderbilt University Medical Center

Andrei Bombin
  Vanderbilt University Medical Center

Mingjian Shi
  VUMC   https://orcid.org/0000-0003-1569-9822

Venkatesh L. Murthy
  University of Michigan   https://orcid.org/0000-0002-7901-1321

Ravi Shah
  Vanderbilt University Medical Center

Jonathan D. Mosley
  Vanderbilt University Medical Center   https://orcid.org/0000-0001-6421-2887

Jane F. Ferguson (✉ jane.f.ferguson@vumc.org)
  Vanderbilt University Medical Center   https://orcid.org/0000-0001-6896-1025

Article

Abstract

Circulating metabolites act as biomarkers of dysregulated metabolism, and may inform disease pathophysiology. A portion of the inter-individual variability in circulating metabolites is influenced by common genetic variation. We evaluated whether a genetics-based "virtual" metabolomics approach can identify novel metabolite-disease associations. We examined the association between polygenic scores for 726 metabolites (derived from OMICSPRED) with 1,247 clinical phenotypes in 57,735 European ancestry and 15,754 African ancestry participants from the BioVU DNA Biobank. We probed significant relationships through Mendelian randomization (MR) using genetic instruments constructed from the METSIM Study, and validated significant MR associations using independent GWAS of candidate phenotypes. We found significant associations between 336 metabolites and 168 phenotypes in European ancestry and 107 metabolites and 56 phenotypes among African ancestry. Of these metabolite-disease pairs, MR analyses confirmed associations between 73 metabolites and 53 phenotypes in European ancestry. Of 22 metabolite-phenotype pairs evaluated for replication in independent GWAS, 16 were significant (false discovery rate p<0.05). Validated findings included the metabolites bilirubin and X−21796 with cholelithiasis, phosphatidylcholine(16:0/22:5n3,18:1/20:4) and arachidonate(20:4n6) with inflammatory bowel disease and Crohn's disease, and campesterol with coronary artery disease and myocardial infarction. These associations may represent biomarkers or potentially targetable mediators of disease risk.

INTRODUCTION

Dysregulated metabolism underlies many of the leading causes of morbidity and mortality, causing considerable human suffering, and high healthcare costs.[1−3] The adverse clinical consequences of extreme disruptions of metabolite homeostasis caused by inborn errors of metabolism are well recognized.[4] However, modest, long-term perturbations of metabolites attributable to common genetic variation also may contribute to disease risk. The clinical consequences of these perturbations remains incompletely defined, but may underlie the residual risk that exists for many complex diseases that is not explained by our current knowledge of disease biology and mechanisms[5]. Identifying associations between circulating metabolites and diseases not only has the potential to identify biomarkers that can be used to risk-stratify individuals, but also to provide insight into disease mechanisms and enable targeted therapies.

Genome wide association studies (GWAS) for circulating metabolites measured by broad metabolomic profiling have identified numerous common single nucleotide polymorphisms (SNPs) associated with circulating metabolite levels.[6−10] These data can be repurposed to develop genetic instruments of individual metabolite levels which can be used to test for associations between metabolites and disease[11−13]. High throughput methodologies, such as Phenome-Wide Association Studies (PheWAS), test associations between genetic instruments and large number of clinical phenotypes using Electronic Health Record (EHR)-linked DNA biobanks.[14,15] These approaches can have significant advantages over

traditional epidemiological approaches, allowing for highly-powered analyses which would otherwise be unfeasible due to cost or logistics. In this context, a 'virtual' metabolomics approach provides a powerful tool to identify candidate pathways that could be targeted to modulate diseases, and to advance risk prediction beyond standard genetic models.

To define the broader phenome associated with circulating metabolites, we applied a virtual metabolomics approach that leveraged a large collection of clinical phenotypes derived from Vanderbilt's BioVU EHR-linked biobank. We constructed virtual metabolomes based on metabolite polygenic scores (PGS), to identify clinical diagnoses that shared genetic modulators with metabolites. Mendelian randomization approaches were then used to better define the relationship between candidate metabolite-phenotype pairs. Significant associations were further validated using external data sets. Our data shed light on multiple metabolite-disease relationships and highlight novel pathways for potential therapeutic intervention.

RESULTS

Predicted circulating levels of metabolites associate with a broad range of clinical phenotypes

We tested for associations among PGS for 726 metabolites and up to 1,247 clinical phenotypes in BioVU. There were 336 metabolites significantly associated with 168 phenotypes in European ancestry (**Supplementary Table 1**) and 107 metabolites that were significantly (false discovery rate [FDR] $p < 0.05$) associated with 56 phenotypes in the African ancestry individuals (**Supplementary Table 2**). 78 metabolites, 11 phenotypes and 104 associations overlapped between European and African ancestry individuals. Clinical phenotypes with the highest number of significant metabolite associations included regional enteritis (n = 63), inflammatory bowel disease (n = 59), disorders of lipid metabolism (n = 56), gout (n = 34), and chronic ischemic heart disease (n = 22) in the European ancestry population [Figure 1A]. Within African ancestry, there were multiple associations between metabolites and methicillin resistant Staphylococcus aureus (n = 32; one amino acid, one unknown metabolite and 30 lipids), adult failure to thrive (n = 29), and urinary tract infection (n = 28) [Figure 1B].

Metabolites with the highest number of significant associations with phenotypes in European ancestry included galactonate (n = 36), N-palmitoyl-sphingosine (d18:1/16:0) (n = 25), 1-palmitoyl-2-stearoyl-GPC (16:0/18:0) (n = 17), and cholesterol (n = 16) [Figure 1A]. In African ancestry, phosphatidylcholine (18:0/20:5 16:0/22:5n6) (n = 8), 1-stearoyl-2-meadoyl-GPC (18:0/20:3n9) (n = 8), 1-palmitoyl-2-eicosapentaenoyl-GPC (16:0/20:5) (n = 8), 1-arachidonoyl-GPC (20:4n6) (n = 8), and 1-palmitoyl-2-arachidonoyl-GPC (16:0/20:4) (n = 7) associated with multiple phenotypes [Figure 1B].

Mendelian randomization highlights relationships between circulating lipids and multiple disease phenotypes

For significant metabolite and phenotype pairs from PheWAS of metabolite PGS, we further characterized the associations under an Mendelian randomization (MR) framework. In European ancestry, of the 336 significant metabolites, GWAS summary statistics were available for 280 matched metabolites in the METSIM study. Of the study metabolites with no corresponding match in METSIM, 45 of 56 were unknown/unidentified metabolites. We identified 159 significant associations (FDR < 0.05) among 73 metabolites and 53 phenotypes by the inverse-variance weighted average (IVWA) method (Fig. 1A, Supplementary Table 3). Among these associations were several distinct phenotype groups with a high number of significant associations with metabolites including those related to dyslipidemia (hyperlipidemia [n = 13]; disorders of lipid metabolism [n = 11]; hyperglyceridemia [n = 8]; hypercholesterolemia [n = 8]), gastrointestinal disorders (inflammatory bowel disease [n = 8]; regional enteritis [n = 7]), metabolic disorders (disorders of bilirubin excretion [n = 8]; cholelithiasis and cholecystitis [n = 6], gout and other crystal arthropathies [n = 5]), decreased white blood cell count (n = 5), and nasal polyps (n = 2). The corresponding metabolites were predominately lipids, including 1-palmitoyl-2-palmitoleoyl-GPC (16:0/16:1) (n = 9), palmitoyl-linoleoyl-glycerol (16:0/18:2) (n = 8), palmitoyl sphingomyelin (d18:1/16:0) (n = 8), campesterol (n = 7), cholesterol (n = 6), 2-hydroxybutyrate/2-hydroxyisobutyrate (n = 5) and 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2) (n = 5).

Many of these associations were driven by instruments composed of only one or two SNPs, increasing the likelihood of associations due to SNPs with pleiotropic effects. We thus selected only metabolites with genetic instruments composed of 3 or more SNPs for further validation. Similarly, to avoid spurious associations driven by pleiotropy, we excluded associations with significant heterogeneity (p < 0.05). After applying these exclusion criteria, 47 significant associations (FDR < 0.05) among 32 metabolites and 34 phenotypes remained. A summary of the retained metabolite pairs is presented in Fig. 2 **and Supplementary Table 3**. These metabolites map to four super-pathways, with the majority mapping to lipid pathways. Distinct phenotypes with a high number of significant associations with metabolites included cholecystitis [n = 5], hypercholesterolemia [n = 3] and Inflammatory Bowel Disease (IBD) [n = 2]. Metabolites with a high number of significant associations with phenotypes included campesterol [n = 7], phosphatidylcholine (16:0/22:5n3, 18:1/20:4) [n = 3], bilirubin (E,E) [n = 2], methylsuccinate [n = 2] and X − 21796 [n = 2].

In the African ancestry population, of 107 metabolites with significant associations in the PGS analysis, 85 had available summary statistics in the METSIM study and among unmatched metabolites, 14 were unknown. The IVWA method identified 22 significant (FDR < 0.05) associations comprising of 15 metabolites and 13 phenotypes (Fig. 1B, **Supplementary Table 4**). These included several associations between lipids and infectious or acute inflammatory diseases, including urinary tract infections, sepsis, and fever.

A summary of the associations between the individual SNPs used in the genetic instrument for each metabolite and the clinical phenotypes is presented for European (**Supplementary Table 5**) and African (**Supplementary Table 6**) ancestry individuals.

Validation of the significant association

To validate the significant findings from MR, we tested associations between the metabolite genetic instruments and phenotypes with available external GWAS summary statistics. After excluding associations with significant heterogeneity, < 3 SNPs and non-specific phenotypes (e.g. "Other mental disorder"), there were 15 phenotypes (with 12 associated metabolites) taken forward for further validation from European ancestry (Fig. 3A). There were no suitable external GWAS datasets available to evaluate the significant associations in African ancestry.

Of 22 metabolite-phenotype pairs evaluated, 16 were significant (FDR $p < 0.05$), with the same direction of effect (Fig. 3B **and Supplementary Table 7**). Among the disease associations were bilirubin (E,E) and X − 21796 associated with cholelithiasis, phosphatidylcholine (16:0/22:5n3, 18:1/20:4) and arachidonate (20:4n6) inversely associated with inflammatory bowel disease and Crohn's disease, and campesterol with coronary artery disease (CAD) and myocardial infarction (MI). Phosphatidylcholine (16:0/22:5n3, 18:1/20:4) was associated with low neutrophil count (neutropenia). The significant associations of phosphatidylcholine (16:0/22:5n3, 18:1/20:4) with low neutrophil count (neutropenia) and lipid diagnosis related to hypercholesteremia (total cholesterol) with 1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1) were not consistent among the MR methods, suggesting that they may represent pleiotropy or are spurious.

DISCUSSION

Metabolites are highly relevant integrative markers of health and disease, that can inform disease prediction and pathophysiology. However, measuring metabolites in the large datasets required to robustly interrogate metabolite-phenotype associations is costly, logistically challenging, and often unfeasible. In this "virtual" metabolomics study, we leveraged a state-of-the-art genetic methods in conjunction with large, phenotypically diverse clinical and genetic data sets to interrogate an extended set of metabolites against a broad clinical phenome. Among 726 metabolites analyzed, there were 336 and 107 metabolites that showed significant associations among BioVU participants of European and African ancestries, respectively. Of these, 159 and 22, respectively, were associated under a MR framework using genetic instruments for metabolites constructed in an independent population, suggesting they may be mediators of disease risk. Among associations identified in the European ancestry population, we validated associations for 16 of 22 metabolite-phenotype pairs using phenotypes derived from independent GWAS studies. Among the validated phenotypes were IBD, cholelithiasis, CAD, MI, neutropenia and lipid phenotypes. These analyses highlight the value of applying the "virtual" metabolomic approach in diverse, phenotype-rich biobanks to identify novel associations.

We found consistent associations between gastrointestinal disease phenotypes and bioactive lipids, highlighting both inflammation and resolution of inflammation as important disease mediators. We found inverse associations between phosphatidylcholine (PC) (16:0/22:5n3, 18:1/20:4) and arachidonate (20:4n6) with IBD and Crohn's disease, both inflammatory diseases of the gut mucosa. Circulating phosphatidylcholines have been reported to be reduced in inflammatory bowel disease, suggesting that they may have a protective role in the gut mucosa.[16,17] The protective effects of PCs may be attributed to anti-inflammatory action and prevention of mucosal damage[16], with potential therapeutic application for

IBD.[18] It is important to identify the specific PC involved in protecting the gut mucus against disease. One of the abundant main species of phosphatidylcholines in gut mucus is PC 16:0/18:1.[16] This is consistent with our data indicating that lower genetically-predicted phosphatidylcholine (16:0/22:5n3, 18:1/20:4) associates with IBD and Crohn's disease. Arachidonate (20:4n6) was also associated with IBD. Arachidonic acid is a precursor of eicosanoids, with potential anti-inflammatory activity[19], and has previously been shown to be inversely associated with IBD including UC and Crohn's disease.[20–22]

We observed several other plausible disease specific associations. There were positive associations between bilirubin (E,E) and X−21796 and cholelithiasis (gallstone disease). A causal association has been reported between extreme levels of bilirubin and increased risk of gallstone disease.[23] This could be due to increased efflux of this metabolite into bile and/or the variation in the expression of genes controlling both bilirubin levels and the disease. [23] Bilirubin (E,E) is one of the water soluble isomers of bilirubin that is converted from unconjugated bilirubin (Z,Z) upon exposure to light.[24] As X−21796 has an unknown identity, the associated pathway is unknown. However, SNPs associated with X−21796 map to several members of the *UGT1A* family of genes, which have been associated with bilirubin levels and risk of gallstones[23], and *SLCO1B*, which is involved in bilirubin transport into the liver.[25]. This also highlights the utility of our approach to define the underlying mechanistic basis of associations with unknown metabolites using the underlying genetic data, which is generally not feasible using other standard epidemiological approaches.

Interestingly, the "virtual" metabolomics approach provided us with a considerable opportunity for novel discovery in relation to cardiovascular disease (CVD). Previously, a meta-analysis showed that there is no association between serum concentrations of two common plant sterols (sitosterol and campesterol) and risk of CVD.[26] However, through this large well-powered study, we found a positive association between campesterol and risk of CAD and MI. Campesterol was also strongly associated with most of the phenotypes categorized in the lipid-related disorders group. Several factors have been proposed as the potential mechanisms linking elevated concentration of campesterol and increased risk of these two diseases, including common pathways influencing the absorption of cholesterol and plant sterols in the intestines,[27] shared genetics linking lipoproteins and phytosterols to MI and atherosclerosis, [28][29] poor nutritional status,[30] and poor metabolic health.[31] However, we anticipate that future analyses may validate and explore the mechanistic bases and the underlying pathophysiology of this novel finding.

This unbiased discovery approach allowed us to create and validate a resource of associations which identified metabolites that are biomarkers and potential mediators of several other clinical phenotypes. For instance, we successfully validated an inverse association between the plasmalogen 1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1) and hypercholesterolemia. This metabolite was reported as inversely related to visceral adipose tissue volume and the percentage of fat in the liver and pancreas.[32] We also found associations between 1-palmitoyl-2-stearoyl-GPC (16:0/18:0) and low-density lipoprotein (LDL) and total cholesterol; this metabolite has been found to be positively associated with dyslipidemia.[33] Our data demonstrated that hypertriglyceridemia was positively associated with oleoyl-

linoleoyl-glycerol (18:1/18:2), potentially a novel association. We also found and validated a significant association between phosphatidylcholine (16:0/22:5n3, 18:1/20:4) and low blood cell count (neutropenia). There were other interesting associations we were unable to validate using external data sets due to lack of available data. For instance, we observed positive significant associations between stearidonate (18:4n3) and 1-stearoyl-2-meadoyl-GPC (18:0/20:3n9) and nasal polyps. Dysregulated lipid metabolism has been reported in Nasal polyps.[34] These metabolites potentially represent new biomarkers of this disorder. An inverse association between methylsuccinate and Alzheimer's disease (AD) was not validated, however given published data linking methylsuccinate supplementation to improvement in neuron dysfunction in AD, this may merit further study.

A significant strength of this study was the use of large datasets which have proven robust for discovery of SNPs associated with both metabolites and disease. A further strength is that we utilized genetic approaches that are well-validated for the applications we propose.[35,36] We analyzed data from multiple sources, including independent cohorts using independent metabolite measurement platforms, and analysis in both European and African American populations where possible. This allowed us to maximize discovery through increased sample sizes and a more diverse population sample, to ensure generalizability, reproducibility and rigor of the association.[37] Moreover, validating the observed associations using available external GWAS additionally strengthened our findings.

Our study also has some limitations. An important limitation of a genetics-based association approach is that the association may not be consistent when using directly measured levels of the metabolite. This can be due to pleiotropic associations, such as when a SNP in the predictor tags a genetic locus that is associated with an outcome through a mechanism unrelated to the metabolite, or due to weak instrument bias.[38,39] Further, some metabolites are heavily modulated by environment and homeostatic physiology, which may mask an association. A second limitation is that we could not find GWAS data for all the phenotype showing a significant association with metabolites. This limited the number of total novel findings we could evaluate in external data sets.

In summary, we identified novel metabolite-phenotype associations, and confirmed known relationships between metabolites and disease. Further studies are needed to replicate and clinically validate these findings. This study highlights the utility of a genetics-based "virtual" metabolomics approach in conjunction with DNA biobanks to link metabolites to clinical diseases and clinical diagnoses. As genetic biobanks continue to grow, the potential to discover genetic underpinnings of the metabolome will also expand. This approach can be used to identify additional metabolite-disease associations, uncover novel disease biology and move towards application in clinical populations.

METHODS

Vanderbilt BioVU Study Population

Genetic and phenotypic data were obtained from BioVU, Vanderbilt University Medical Center's (VUMC) DNA Biobank linked to a de-identified electronic health record.[40] The study population comprised individuals of genetic white European (n = 57,735) and African American (n = 15,754) ancestries, 18 years and older who had existing SNP genotyping. Genetic ancestry of individuals was determined using principal component analysis in conjunction with HAPMAP reference sets.[40,41] This study was reviewed by the VUMC Institutional Review Board (IRB) in accordance with the informed consent guidelines and was determined to be non-human subjects research.

Genetic Data and Quality Control

BioVU participants were genotyped on the Illumina Infinium Multi-Ethnic Genotyping Array (MEGA[EX]) platform. Quality control procedures for this population have been described previously.[42] Individuals with a biological sex discrepancy or who were related (one participant from each related pair [pi-hat > 0.2] was randomly excluded) were excluded. Analyses used PLINK v1.9.[43] Genotype imputation was performed using IMPUTE4[44] version 2.3.0 (University of Oxford), using the 10/2014 release of the 1000 Genomes cosmopolitan reference haplotypes. Genetic variants with imputation quality scores less than 0.3 were excluded. Principal components (PCs) to adjust for residual population stratification were generated using SmartPCA.[45]

Phenotype Data

For the BioVU population, the primary analyses examined clinical diagnoses based on PheCodes (v1.2), which are derived from International Classification of Disease (ICD) billing codes (ICD-9-CM and ICD-10 diagnosis codes).[46,47] For each phenotype, cases were defined as participants with at least two PheCode instances in their medical record. Individuals without any closely related PheWAS codes and who fell within the observed age of the cases were used as controls. We analyzed associations for 1,247 and 600 PheCodes with ≥ 100 cases in the European and African ancestry population, respectively.

Specification of a Virtual Metabolome via Human Genetics

OMICSPRED: Validated PGSs for 726 metabolites were obtained from the OMICSPRED resource (www.omicspred.org).[48] These PGS were developed using SNPs that significantly associated with concentrations of human blood metabolites in the INTERVAL cohort (n = 8,153 healthy individuals in England).[49] Briefly, metabolites were measured in plasma by an untargeted mass spectrometry metabolomics platform (Metabolon HD4), and participants were genotyped using the Affymetrix Biobank Axiom array.[50] Bayesian ridge regression was used to develop genetic scores for each metabolite, and scores were validated (Spearman correlation) using an independent validation INTERVAL subset (n = 8,114 non-overlapping participants, 527 validated metabolites) and an external validation cohort (ORCADES, n = 1,007 European participants, 455 validated metabolites).

METSIM: SNP instruments used for validation of the OMICSPRED associations by MR analyses were derived from the independent METSIM Finnish population study using publicly available GWAS summary

statistics for metabolites.[51] This study included 1,391 metabolites quantified in 6,136 non-diabetic male participants. Summary statistics were obtained from the METSIM Metabolomics PheWeb server (https://pheweb.org/metsim-metab).

Polygenic Score Analysis

SNPs associated with each of the 726 OMICSPRED metabolites were used to calculate PGSs as a weighted sum of trait-associated alleles for BioVU subjects described above, with PLINK v2.00a3LM.[52] The association between metabolite PGS and each PheCode phenotype was tested using a multivariable logistic regression model, adjusting for sex and age. All analyses were stratified by genetic ancestry. Within each phenotype, association p-values were adjusted for multiple testing using a Benjamini−Hochberg false discovery rate (FDR) correction, (rstatix v0.7.0 R package).

Mendelian Randomization Analysis to Validate PGS associations

Phenotype and metabolite pairs that were significantly associated (FDR p < 0.05) with PGS through PheWAS in BioVU, were selected for MR analysis. MR tests for associations under three assumptions: (1) the SNPs are associated with the exposure; (2) the SNPs are not associated with confounders; and (3) the SNPs affect the outcome only through the exposure.[53] For these analyses, SNPs associated with the metabolite were used as the exposure instrumental variables. Genetic instruments for each metabolite were selected using a clumping algorithm that selected an LD-reduced (r-square < 0.05) set of SNPs associated with metabolites ($p < 5x10^{-6}$) in the METSIM Study. The association between metabolite-associated SNPs and the BioVU clinical phenotype of interest was computed using an additive logistic regression genetic model that adjusted for age, sex and 10 principal components (PLINK v2.00a3LM software). The IVWA, MR-Egger and weighted median methods, as implemented in the MendelianRandomization R package[54] were used to perform the analyses. Horizontal pleiotropy was determined by a low heterogeneity p-value (p < 0.05) based on the Cochran's Q statistic. P-values were adjusted for multiple testing using a Benjamini−Hochberg FDR correction, per tested phenotype. For non-pleiotropic associations (heterogeneity p > 0.05), we selected significant (FDR p < 0.05) metabolite-phenotype pairs based on the IVWA model, that showed consistent findings across the other MR methods. For associations with evidence of pleiotropy, we used MR-PRESSO to identify and evaluate the contributions of pleiotropic SNPs.

MR Validation in Independent disease-specific GWAS Datasets

We validated significant MR associations using summary statistics from published GWAS datasets, where available. GWAS summary statistics for IBD and Crohn's disease were obtained from a meta-analysis of 59,957 individuals of European ancestry.[55] Summary statistics for cholelithiasis were obtained from FinnGen (19,023 cases, 195,144 controls; FinnGen Consortium Release 5) and UK Biobank (11,632 cases, 289,159 controls)(https://ctg.cncr.nl/software/summary_statistics)[56] For Atopic

dermatitis, GWAS summary statistics were obtained from a multi-ancestry GWAS of 21,399 cases and 95,464 controls from populations of European, African, Japanese and Latino ancestries.[57] Summary statistics for AD were obtained from a meta-analysis of 1,126,563 individuals of European ancestry.[58]

GWAS summary statistics for CAD and MI were downloaded from www.CARDIOGRAMPLUSC4D.ORG[59] which included a GWAS meta-analysis of ~ 185,000 CAD cases and controls with a subgroup analysis in cases with a reported history of myocardial infarction (around 70% of the total number of cases). Summary statistics for neutrophil counts were obtained from a trans-ethnic GWAS meta-analyses of 746,667 participants, including 184,535 non-European individuals.[60] High-density lipoprotein (HDL), LDL, total cholesterol and triglycerides (TG) were obtained from the Global lipids consortium phenotypes (http://lipidgenetics.org/)[61] including 188,577 European, East Asian, South Asian and African ancestry individuals.

All statistical tests were two-sided and analyses used R v.4.0.2. The circlize package was used to create the circular plots.

Declarations

References

1. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. Accessed September 29, 2017. https://data.globalchange.gov/report/national-diabetes-statistics-report-estimates-diabetes-its-burden

2. Mozaffarian D, Benjamin EJ, Go AS, et al. Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. *Circulation*. 2016;133(4):e38-e360.

doi:10.1161/CIR.0000000000000350

3. American Diabetes Association. Economic costs of diabetes in the U.S. in 2012. *Diabetes Care*. 2013;36(4):1033-1046. doi:10.2337/dc12-2625

4. Mootha VK, Hirschhorn JN. Inborn variation in metabolism. *Nat Genet*. 2010;42(2):97-98. doi:10.1038/ng0210-97

5. Lieb W, Enserro DM, Larson MG, Vasan RS. Residual cardiovascular risk in individuals on lipid-lowering treatment: quantifying absolute and relative risk in the community. *Open Heart*. 2018;5(1):e000722. doi:10.1136/openhrt-2017-000722

6. Shin SY, Fauman EB, Petersen AK, et al. An atlas of genetic influences on human blood metabolites. *Nature genetics*. 2014;46:543-550. doi:10.1038/ng.2982

7. Rhee EP, Ho JE, Chen MH, et al. A genome-wide association study of the human metabolome in a community-based cohort. *Cell metabolism*. 2013;18:130-143. doi:10.1016/j.cmet.2013.06.013

8. Rhee EP, Yang Q, Yu B, et al. An exome array study of the plasma metabolome. *Nat Commun*. 2016;7:12360. doi:10.1038/ncomms12360

9. Kettunen J, Demirkan A, Wurtz P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun*. 2016;7:11122. doi:10.1038/ncomms11122

10. Demirkan A, Henneman P, Verhoeven A, et al. Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet*. 2015;11(1):e1004835. doi:10.1371/journal.pgen.1004835

11. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*. 2017;18(2):117-127. doi:10.1038/nrg.2016.142

12. Maher BS. Polygenic Scores in Epidemiology: Risk Prediction, Etiology, and Clinical Utility. *Current Epidemiology Reports*. 2015;2(4):239-244. doi:10.1007/s40471-015-0055-3

13. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology*. 2003;32(1):1-22. doi:10.1093/ije/dyg070

14. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics (Oxford, England)*. 2010;26:1205-1210. doi:10.1093/bioinformatics/btq126

15. Karnes JH, Bastarache L, Shaffer CM, et al. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci Transl Med*. 2017;9. doi:10.1126/scitranslmed.aai8708

16. Treede I, Braun A, Sparla R, et al. Anti-inflammatory Effects of Phosphatidylcholine*. *Journal of Biological Chemistry*. 2007;282(37):27155-27164. doi:10.1074/jbc.M704408200

17. Stremmel W, Vural H, Evliyaoglu O, Weiskirchen R. Delayed-Release Phosphatidylcholine Is Effective for Treatment of Ulcerative Colitis: A Meta-Analysis. *Digestive Diseases*. 2021;39(5):508-515.

doi:10.1159/000514355

18. Ai R, Xu J, Ji G, Cui B. Exploring the Phosphatidylcholine in Inflammatory Bowel Disease: Potential Mechanisms and Therapeutic Interventions. *Curr Pharm Des*. 2022;28(43):3486-3491. doi:10.2174/1381612829666221124112803

19. Marton LT, Goulart R de A, Carvalho ACA de, Barbalho SM. Omega Fatty Acids and Inflammatory Bowel Diseases: An Overview. *Int J Mol Sci*. 2019;20(19). doi:10.3390/ijms20194851

20. Levy E, Rizwan Y, Thibault L, et al. Altered lipid profile, lipoprotein composition, and oxidant and antioxidant status in pediatric Crohn disease. *Am J Clin Nutr*. 2000;71(3):807-815. doi:10.1093/ajcn/71.3.807

21. ROMANATO G, SCARPA M, ANGRIMAN I, et al. Plasma lipids and inflammation in active inflammatory bowel diseases. *Alimentary Pharmacology & Therapeutics*. 2009;29(3):298-307. doi:10.1111/j.1365-2036.2008.03886.x

22. Bugajska J, Berska J, Zwolińska-Wcisło M, Sztefko K. The risk of essential fatty acid insufficiency in patients with inflammatory bowel diseases: fatty acid profile of phospholipids in serum and in colon biopsy specimen. *Arch Med Sci*. 2022;18(4):1103-1107. doi:10.5114/aoms/150041

23. Stender S, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A. Extreme Bilirubin Levels as a Causal Risk Factor for Symptomatic Gallstone Disease. *JAMA Internal Medicine*. 2013;173(13):1222-1228. doi:10.1001/jamainternmed.2013.6465

24. Wang J, Guo G, Li A, Cai WQ, Wang X. Challenges of phototherapy for neonatal hyperbilirubinemia (Review). *Exp Ther Med*. 2021;21(3):231. doi:10.3892/etm.2021.9662

25. Keppler D. The roles of MRP2, MRP3, OATP1B1, and OATP1B3 in conjugated hyperbilirubinemia. *Drug Metab Dispos*. 2014;42(4):561-565. doi:10.1124/dmd.113.055772

26. Genser B, Silbernagel G, De Backer G, et al. Plant sterols and cardiovascular disease: a systematic review and meta-analysis. *Eur Heart J*. 2012;33(4):444-451. doi:10.1093/eurheartj/ehr441

27. Silbernagel G, Fauler G, Hoffmann MM, et al. The associations of cholesterol metabolism and plasma plant sterols with all-cause and cardiovascular mortality. *J Lipid Res*. 2010;51(8):2384-2393. doi:10.1194/jlr.P002899

28. Wang Y, Wang L, Liu X, et al. Genetic variants associated with myocardial infarction and the risk factors in Chinese population. *PLoS One*. 2014;9(1):e86332. doi:10.1371/journal.pone.0086332

29. Scholz M, Horn K, Pott J, et al. Genome-wide meta-analysis of phytosterols reveals five novel loci and a detrimental effect on coronary atherosclerosis. *Nature Communications*. 2022;13(1):143. doi:10.1038/s41467-021-27706-6

30. Strandberg TE, Pitkälä KH. Frailty in elderly people. *Lancet*. 2007;369(9570):1328-1329. doi:10.1016/S0140-6736(07)60613-8

31. Simonen P, Gylling H, Howard AN, Miettinen TA. Introducing a new component of the metabolic syndrome: low cholesterol absorption. *Am J Clin Nutr*. 2000;72(1):82-88. doi:10.1093/ajcn/72.1.82

32. Lind L, Salihovic S, Risérus U, et al. The Plasma Metabolomic Profile is Differently Associated with Liver Fat, Visceral Adipose Tissue, and Pancreatic Fat. *The Journal of Clinical Endocrinology & Metabolism*. 2021;106(1):e118-e129. doi:10.1210/clinem/dgaa693

33. Yousri NA, Suhre K, Yassin E, et al. Metabolic and Metabo-Clinical Signatures of Type 2 Diabetes, Obesity, Retinopathy, and Dyslipidemia. *Diabetes*. 2022;71(2):184-205. doi:10.2337/db21-0490

34. Miyata J, Fukunaga K, Kawashima Y, et al. Dysregulated fatty acid metabolism in nasal polyp-derived eosinophils from patients with chronic rhinosinusitis. *Allergy*. 2019;74(6):1113-1124. doi:10.1111/all.13726

35. Voight BF, Peloso GM, Orho-Melander M, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet*. 2012;380(9841):572-580. doi:10.1016/S0140-6736(12)60312-2

36. Larsson SC, Burgess S, Michaëlsson K. Association of Genetic Variants Related to Serum Calcium Levels With Coronary Artery Disease and Myocardial Infarction. *JAMA*. 2017;318(4):371-380. doi:10.1001/jama.2017.8981

37. Vsevolozhskaya OA, Kuo CL, Ruiz G, Diatchenko L, Zaykin DV. The more you test, the more you find: The smallest P-values become increasingly enriched with real findings as more tests are conducted. *Genet Epidemiol*. Published online September 14, 2017. doi:10.1002/gepi.22064

38. Gianola D, de los Campos G, Toro MA, Naya H, Schön CC, Sorensen D. Do Molecular Markers Inform About Pleiotropy? *Genetics*. 2015;201(1):23-29. doi:10.1534/genetics.115.179978

39. Davies NM, von Hinke Kessler Scholder S, Farbmacher H, Burgess S, Windmeijer F, Smith GD. The many weak instruments problem and Mendelian randomization. *Stat Med*. 2015;34(3):454-468. doi:10.1002/sim.6358

40. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther*. 2008;84(3):362-369. doi:10.1038/clpt.2008.89

41. Gibbs RA, Belmont JW, Hardenbol P, et al. The International HapMap Project. *Nature*. 2003;426(6968):789-796. doi:10.1038/nature02168

42. Ruderfer DM, Walsh CG, Aguirre MW, et al. Significant shared heritability underlies suicide attempt and clinically predicted probability of attempting suicide. *Molecular Psychiatry*. Published online January 4, 2019. doi:10.1038/s41380-018-0326-8

43. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. 2007;81(3):559-575. doi:10.1086/519795

44. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. Schork NJ, ed. *PLoS Genet*. 2009;5(6):e1000529. doi:10.1371/journal.pgen.1000529

45. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904-909. doi:10.1038/ng1847

46. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*. 2013;31(12):1102-1111. doi:10.1038/nbt.2749

47. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*. 2010;26(9):1205-1210. doi:10.1093/bioinformatics/btq126

48. Xu Y, Ritchie SC, Liang Y, et al. *An Atlas of Genetic Scores to Predict Multi-Omic Traits*. Genomics; 2022. doi:10.1101/2022.04.17.488593

49. Xu Y, Ritchie SC, Liang Y, et al. An atlas of genetic scores to predict multi-omic traits. *Nature*. 2023;616(7955):123-131. doi:10.1038/s41586-023-05844-9

50. Shin SY, Fauman EB, Petersen AK, et al. An atlas of genetic influences on human blood metabolites. *Nature Genetics*. 2014;46(6):543-550. doi:10.1038/ng.2982

51. Yin X, Chan LS, Bose D, et al. Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. *Nature Communications*. 2022;13(1):1644. doi:10.1038/s41467-022-29143-5

52. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795

53. Emdin CA, Khera A V., Kathiresan S. Mendelian randomization. *JAMA - Journal of the American Medical Association*. 2017;318(19):1925-1926. doi:10.1001/jama.2017.17219

54. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*. 2018;50(11):1505-1513. doi:10.1038/s41588-018-0241-6

55. de Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet*. 2017;49(2):256-261. doi:10.1038/ng.3760

56. Mi J, Jiang L, Liu Z, et al. Identification of blood metabolites linked to the risk of cholelithiasis: a comprehensive Mendelian randomization study. *Hepatology International*. 2022;16(6):1484-1493. doi:10.1007/s12072-022-10360-5

57. Paternoster L, Standl M, Waage J, et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics*. 2015;47(12):1449-1456. doi:10.1038/ng.3424

58. Wightman DP, Jansen IE, Savage JE, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature Genetics*. 2021;53(9):1276-1282. doi:10.1038/s41588-021-00921-z

59. Nikpay M, Goel A, Won HH, et al. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*. 2015;47(10):1121-1130. doi:10.1038/ng.3396

60. Chen MH, Raffield LM, Mousas A, et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*. 2020;182(5):1198-1213.e14. doi:10.1016/j.cell.2020.06.045

61. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*. 2013;45(11):1274-1283. doi:10.1038/ng.2797
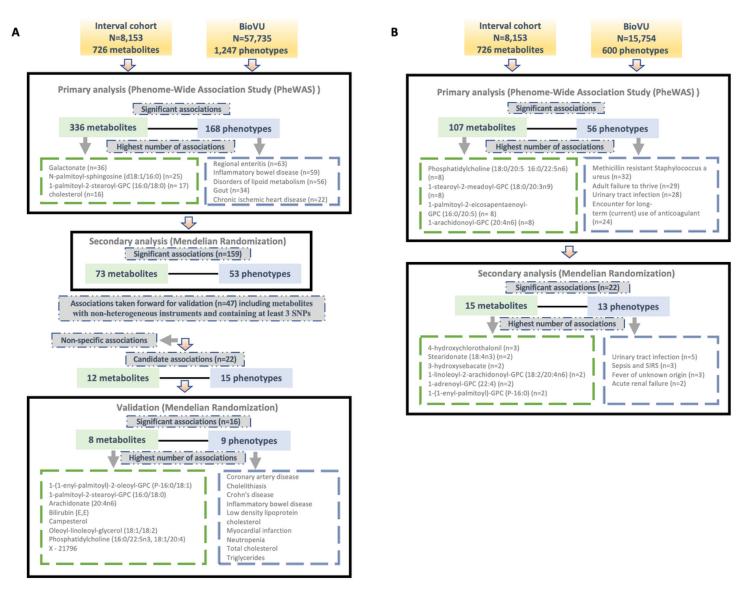
Figures

Figure 1

Overview of the study design and findings in (A) European and (B) African ancestry BioVU participants.
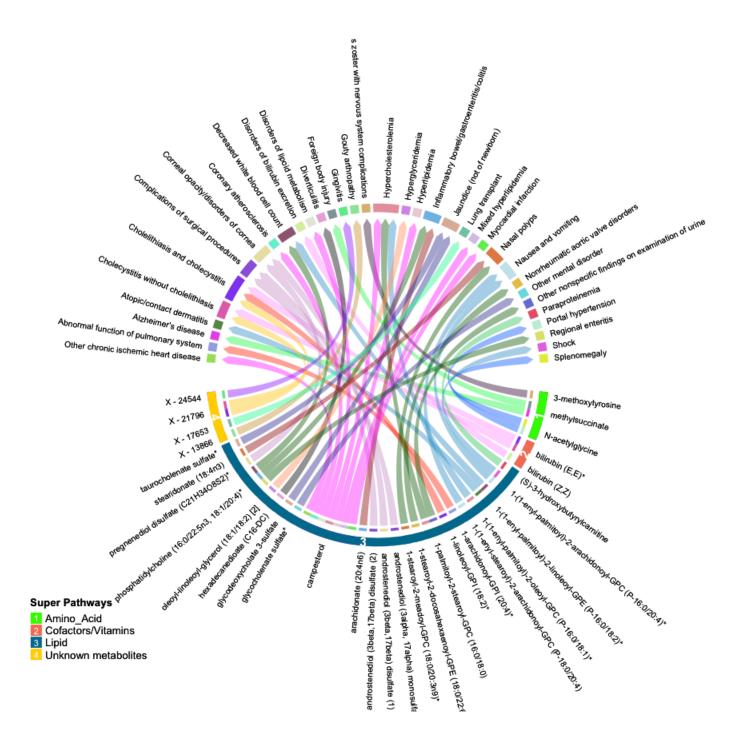
Figure 2

Circular plot summarizing significant associations between circulating metabolites and phenotypes identified by inverse-variance weighted association (IVWA) Mendelian randomization (FDR p<0.05). Metabolites are shown in bottom half of the figure with super-pathways depicted on the outer track (with colors and numbers) and sub-pathways shown as the color of each line (i.e. lines with the same color belong to the same sub pathway). Each color of the outer top track and the inner bottom track corresponds to a specific phenotype.

A

Metabolite Phenotype Associations	Number of SNPs	Beta (95% CI)
Alzheimer's disease		
Methylsuccinate	20	-0.389 (-0.566 to -0.213)
Coronary atherosclerosis		
Campesterol	8	0.193 (0.118 to 0.268)
Myocardial infarction		
Campesterol	8	0.229 (0.129 to 0.330)
Cholecystitis without cholelithiasis		
Bilirubin (E,E)	23	0.274 (0.130 to 0.419)
X - 21796	11	0.497 (0.235 to 0.760)
Cholelithiasis and cholecystitis		
1-linoleoyl-GPI (18:2)	7	0.314 (0.150 to 0.478)
bilirubin (E,E)	23	0.185 (0.103 to 0.268)
X - 21796	11	0.359 (0.199 to 0.520)
Inflammatory bowel disease and other gastroenteritis and colitis		
Phosphatidylcholine (16:0/22:5n3, 18:1/20:4)	8	-0.305 (-0.453 to -0.158)
Arachidonate (20:4n6)	8	-0.306 (-0.461 to -0.151)
Regional enteritis		
Phosphatidylcholine (16:0/22:5n3, 18:1/20:4)	8	-0.368 (-0.538 to -0.197)
Diverticulitis		
Hexadecanedioate (C16-DC)	20	0.275 (0.139 to 0.410)
Atopic/contact dermatitis		
1-(1-enyl-palmitoyl)-2-linoleoyl-GPE (P-16:0/18:2)	4	0.474 (0.275 to 0.672)
Decreased white blood cell count		
Phosphatidylcholine (16:0/22:5n3, 18:1/20:4)	8	-0.277 (-0.423 to -0.130)
1-(1-enyl-palmitoyl)-2-linoleoyl-GPE (P-16:0/18:2)	4	0.642 (0.403 to 0.882)
Disorders of lipoid metabolism		
Campesterol	8	0.189 (0.130 to 0.248)
Hypercholesterolemia		
Campesterol	8	0.214 (0.113 to 0.316)
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)	3	-0.489 (-0.685 to -0.294)
1-palmitoyl-2-stearoyl-GPC (16:0/18:0)	3	0.327 (0.162 to 0.491)
Hypertriglyceridemia		
Oleoyl-linoleoyl-glycerol (18:1/18:2) [2]	6	1.710 (1.191 to 2.229)
Hyperlipidemia		
Campesterol	8	0.188 (0.129 to 0.246)
Mixed hyperlipidemia		
Campesterol	8	0.184 (0.112 to 0.255)

-0.5 0 0.5 1 1.5 2

B

Metabolite Phenotype Associations	Number of SNPs	Beta (95% CI)	FDR P-value
Alzheimer's disease			
Methylsuccinate	19	-0.004 (-0.009 to 0.001)	0.116
Coronary artery disease			
Campesterol	8	0.180 (0.088 to 0.272)	2e-04
Myocardial infarction			
Campesterol	8	0.177 (0.044 to 0.310)	0.014
Cholelithiasis			
1-linoleoyl-GPI (18:2)	5	0.061 (-0.016 to 0.138)	0.147
X - 21796	9	0.129 (0.034 to 0.224)	0.014
Bilirubin (E,E)	18	0.088 (0.051 to 0.124)	6.38e-06
Crohn's disease			
Phosphatidylcholine (16:0/22:5n3, 18:1/20:4)	8	-0.202 (-0.299 to -0.106)	8.72e-05
Arachidonate (20:4n6)	8	-0.211 (-0.322 to -0.100)	3e-04
Inflammatory bowel disease			
Phosphatidylcholine (16:0/22:5n3, 18:1/20:4)	8	-0.127 (-0.182 to -0.072)	1.55e-05
Arachidonate (20:4n6)	8	-0.149 (-0.234 to -0.063)	0.001
Hexadecanedioate (C16-DC)	18	0.003 (-0.039 to 0.044)	0.9
Dermatitis			
1-(1-enyl-palmitoyl)-2-linoleoyl-GPE (P-16:0/18:2)	3	0.134 (-0.006 to 0.275)	0.081
Neutropenia			
1-(1-enyl-palmitoyl)-2-linoleoyl-GPE (P-16:0/18:2)	4	-0.037 (-0.098 to 0.025)	0.26
Phosphatidylcholine (16:0/22:5n3, 18:1/20:4)	8	0.046 (0.022 to 0.069)	2e-04
Low density lipoprotein cholesterol			
Campesterol	5	0.244 (0.147 to 0.341)	3e-06
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)	3	-0.139 (-0.190 to -0.088)	5.71e-07
1-palmitoyl-2-stearoyl-GPC (16:0/18:0)	2	0.410 (0.354 to 0.465)	1.16e-46
Total cholesterol			
Campesterol	5	0.229 (0.135 to 0.323)	5.13e-06
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)	3	0.282 (0.233 to 0.331)	2.07e-28
1-palmitoyl-2-stearoyl-GPC (16:0/18:0)	2	0.609 (0.556 to 0.661)	3.78e-111
Triglycerides			
Campesterol	5	0.028 (-0.016 to 0.071)	4.8e-06
Oleoyl-linoleoyl-glycerol (18:1/18:2) [2]	4	0.679 (0.401 to 0.956)	0.227

-0.2 0 0.2 0.4 0.6 0.8

Figure 3

Summary of association from MR analyses between genetic instruments for metabolites in in METSIM and genetic predisposition of phenotypes derived from (A) BioVU (all significant at false discovery rate (FDR) P-value < 0.05) and (B) validation phenotypes (The effect size and 95% confidence interval (CI) are based on raw p-value. However, the significant results are considered at FDR P-value <0.05).

Supplementary Files

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

- SupplementaryTables.xlsx