nature medicine

Article

https://doi.org/10.1038/s41591-025-04105-8

Circulating metabolites, genetics and lifestyle factors in relation to future risk of type 2 diabetes

Received: 27 September 2024

Accepted: 6 November 2025

Published online: 14 January 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

The human metabolome reflects complex metabolic states affected by genetic and environmental factors. However, metabolites associated with type 2 diabetes (T2D) risk and their determinants remain insufficiently characterized. Here we integrated blood metabolomic, genomic and lifestyle data from up to 23,634 initially T2D-free participants from ten cohorts. Of 469 metabolites examined, 235 were associated with incident T2D during up to 26 years of follow-up, including 67 associations not previously reported across bile acid, lipid, carnitine, urea cycle and arginine/proline, glycine and histidine pathways. Further genetic analyses linked these metabolites to signaling pathways and clinical traits central to T2D pathophysiology, including insulin resistance, glucose/insulin response, ectopic fat deposition, energy/lipid regulation and liver function. Lifestyle factors—particularly physical activity, obesity and diet—explained greater variations in T2D-associated versus non-associated metabolites, with specific metabolites revealed as potential mediators. Finally, a 44-metabolite signature improved T2D risk prediction beyond conventional factors. These findings provide a foundation for understanding T2D mechanisms and may inform precision prevention targeting specific metabolic pathways.

Diabetes affects 589 million adults globally, and the number is estimated to increase to more than 853 million by 2050[1]. T2D accounts for more than 90% of all diabetes cases, and its pathogenesis involves both polygenic susceptibility and environmental risk factors (for example, diet and lifestyle)[2]. Manifested by insulin resistance, β-cell dysfunction and consequent hyperglycemia, the progression of T2D is characterized by comprehensive yet integrative metabolic changes orchestrated at several organ systems[3]. Identifying the molecular profile characterizing the dysregulated metabolism contributing to T2D, as well as the genetic and environmental determinants of such a metabolic profile, is crucial for understanding T2D etiology, and may inform the design of more effective preventive strategies targeting specific metabolic pathways.

The circulating metabolome is the quantitative collection of small molecules in the blood and provides a comprehensive functional readout of the metabolic homeostasis in each person. In past decades, prospective studies examining circulating metabolites from preselected pathways or, more recently, a broader spectrum of the metabolome, have identified more than 100 metabolites associated with T2D risk[4–6]. These studies collectively highlight important roles of several metabolites in T2D pathophysiology, such as branched-chain amino acids (BCAAs), tryptophan and lysine, specific phospholipids (PLs) and ceramides[4–6]. Recent research further indicated that the blood metabolome can be influenced by genetics[7,8], health conditions, diet and lifestyle[6,9,10], among other factors[10,11]. As such, there is a strong need to integrate multimodal data to better understand how various risk factors are related to disease-associated metabolites to advance precision prevention. However, systematic evaluations of the circulating metabolome associated with T2D risk are lacking, and the genetic and nongenetic contributors to the T2D metabolome have not been investigated in a comprehensive manner.

e-mail: juli@bwh.harvard.edu; qibin.qi@einsteinmed.edu

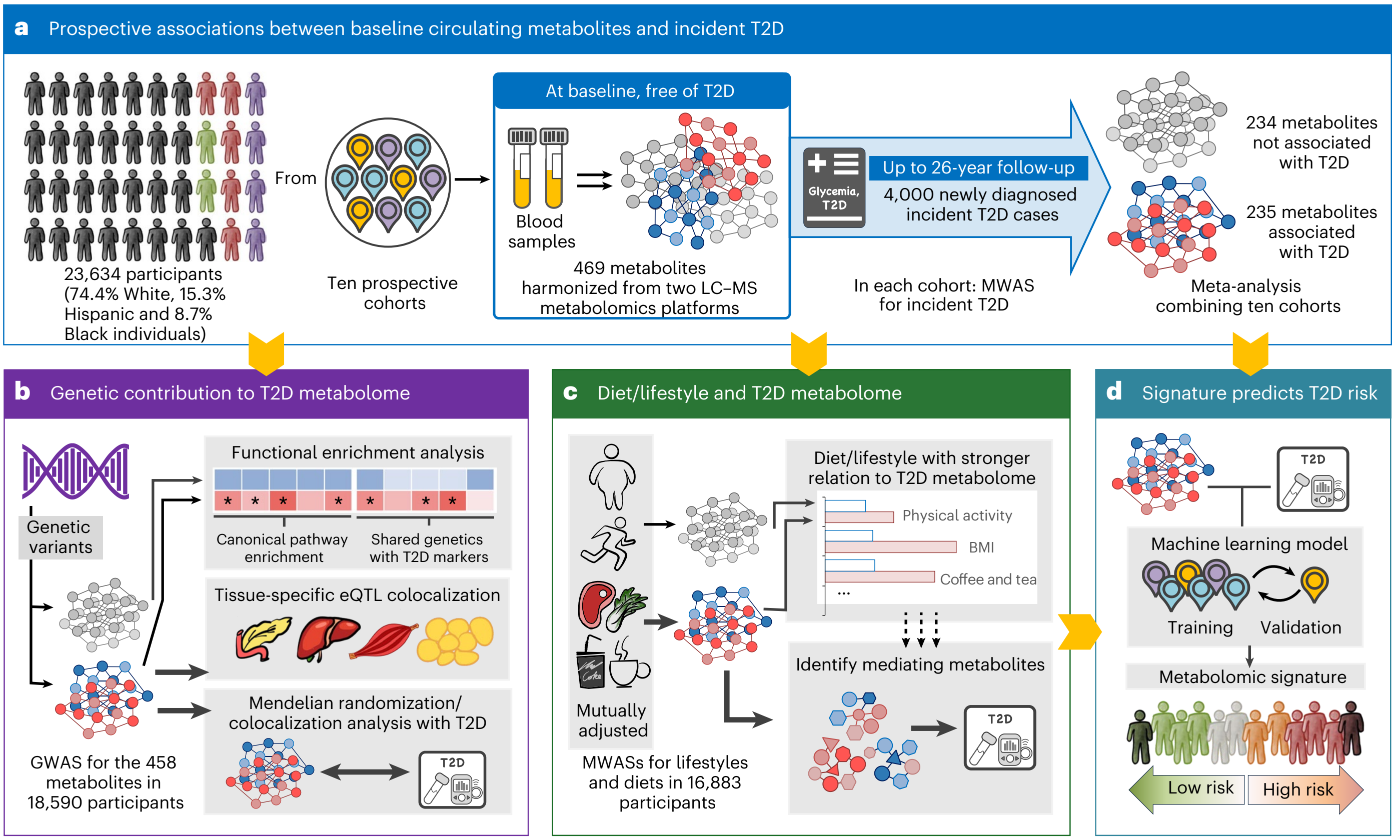

a Prospective associations between baseline circulating metabolites and incident T2D
23,634 participants (74.4% White, 15.3% Hispanic and 8.7% Black individuals)
From
Ten prospective cohorts
At baseline, free of T2D
Blood samples
469 metabolites harmonized from two LC–MS metabolomics platforms
Glycemia, T2D
Up to 26-year follow-up
4,000 newly diagnosed incident T2D cases
In each cohort: MWAS for incident T2D
234 metabolites not associated with T2D
235 metabolites associated with T2D
Meta-analysis combining ten cohorts
b Genetic contribution to T2D metabolome
Genetic variants
GWAS for the 458 metabolites in 18,590 participants
Functional enrichment analysis
Canonical pathway enrichment
Shared genetics with T2D markers
Tissue-specific eQTL colocalization
Mendelian randomization/ colocalization analysis with T2D
T2D
c Diet/lifestyle and T2D metabolome
Mutually adjusted
MWASs for lifestyles and diets in 16,883 participants
Diet/lifestyle with stronger relation to T2D metabolome
Physical activity
BMI
Coffee and tea
...
Identify mediating metabolites
T2D
d Signature predicts T2D risk
T2D
Machine learning model
Training
Validation
Metabolomic signature
Low risk
High risk

Fig. 1 | Study overview. a, To identify blood metabolites associated with incident T2D, we analyzed 469 harmonized metabolites in up to 23,634 participants from ten prospective cohort studies. At baseline, participants were free of T2D and other chronic diseases; and blood metabolome was profiled using the metabolomic platforms at Broad Institute or Metabolon Inc. A metabolome-wide association study (MWAS) for incident T2D was conducted in each cohort; and results from the ten cohorts were combined using meta-analysis, identifying 235 metabolites associated with T2D risk. **b**, We curated meta-analyzed genome-wide association studies (GWASs) for each metabolite using data of up to 18,590 people from eight cohorts, followed by functional analyses, colocalization analyses and Mendelian randomization analyses. **c**, We conducted MWASs for major modifiable risk factors in up to 16,883 participants from five cohorts, identifying metabolites that potentially mediated the associations between risk factors and T2D risk. **d**, We used machine learning analyses to develop a metabolomic signature reflecting the complex metabolic states predictive of long-term T2D risk, which may facilitate the identification of high-risk individuals and precision prevention.

To fill these knowledge gaps, we examined 469 circulating metabolites in a pooled study of 23,634 initially T2D-free and racially/ethnically diverse people from ten prospective cohorts, to identify metabolites associated with incident T2D over up to 26 years of follow-up. We further conducted integrative analyses combining genomic data and diet/lifestyle factors, to systematically elucidate genetic determinants, functional enrichments and potential tissue origin for T2D-associated metabolites; and to illustrate the relationships among diet/lifestyle factors, circulating metabolites and incident T2D. Finally, we derived and validated a multi-metabolite signature that reflected the complex metabolic states predictive of future T2D risk, with the potential to facilitate risk stratification and precision prevention (Fig. 1).

Results

Metabolome-wide association analysis of incident T2D

Our primary analysis included 23,634 participants from ten prospective cohorts free of T2D at study baseline. During up to 26 years of follow-up, 4,000 incident T2D cases were identified (Fig. 1 and Extended Data Table 1). Metabolomic profiling was conducted at either the Broad Institute or Metabolon Inc., and 469 metabolites were harmonized across cohorts for analyses (Supplementary Fig. 1). We conducted metabolome-wide association analysis in each cohort stratified by major racial/ethnic groups (Supplementary Table 1; Methods). In meta-analysis of all subsets, we identified 235 metabolites associated with incident T2D (false discovery rate (FDR) < 0.05), after adjusting for demographic, socioeconomic and clinical factors, including body mass index (BMI) and waist–hip ratio (WHR) (Fig. 2, Extended Data Fig. 1a and Supplementary Table 2). These include 168 previously reported associations and 67 additional significant associations identified in this study (Supplementary Table 3). Aside from glucose, multivariable-adjusted risk ratio (RR) for incident T2D per s.d. increment in circulating levels of metabolites ranged from 0.67 (95% confidence interval (CI), 0.61–0.72) for C22:4 cholesterol ester (CE), to 1.71 (95% CI, 1.60–1.83) for C32:0 diacylglycerol (DAG).

The large number of metabolites associated with T2D risk is expected, given the correlations among metabolites—particularly among lipids (Supplementary Fig. 2) and our large sample size. Results were consistent in multiple sensitivity analyses, including a basic model adjusting for only demographic and socioeconomic factors, and models further adjusting for diet quality and physical activity (PA), blood pressure, blood lipids or estimated glomerular filtration rate (eGFR) (Supplementary Figs. 3–4 and Supplementary Table 2). In stratified meta-analysis by major racial/ethnic groups, associations between most metabolites and T2D risk were comparable between non-Hispanic white individuals and those of other racial and ethnic groups, although a few discrepancies deserve further investigations (for example, acisoga was associated with T2D risk only in Black participants, and C36:1 phosphatidylcholine (PC) only in Hispanic/Latino

https://doi.org/10.1038/s41591-025-04105-8

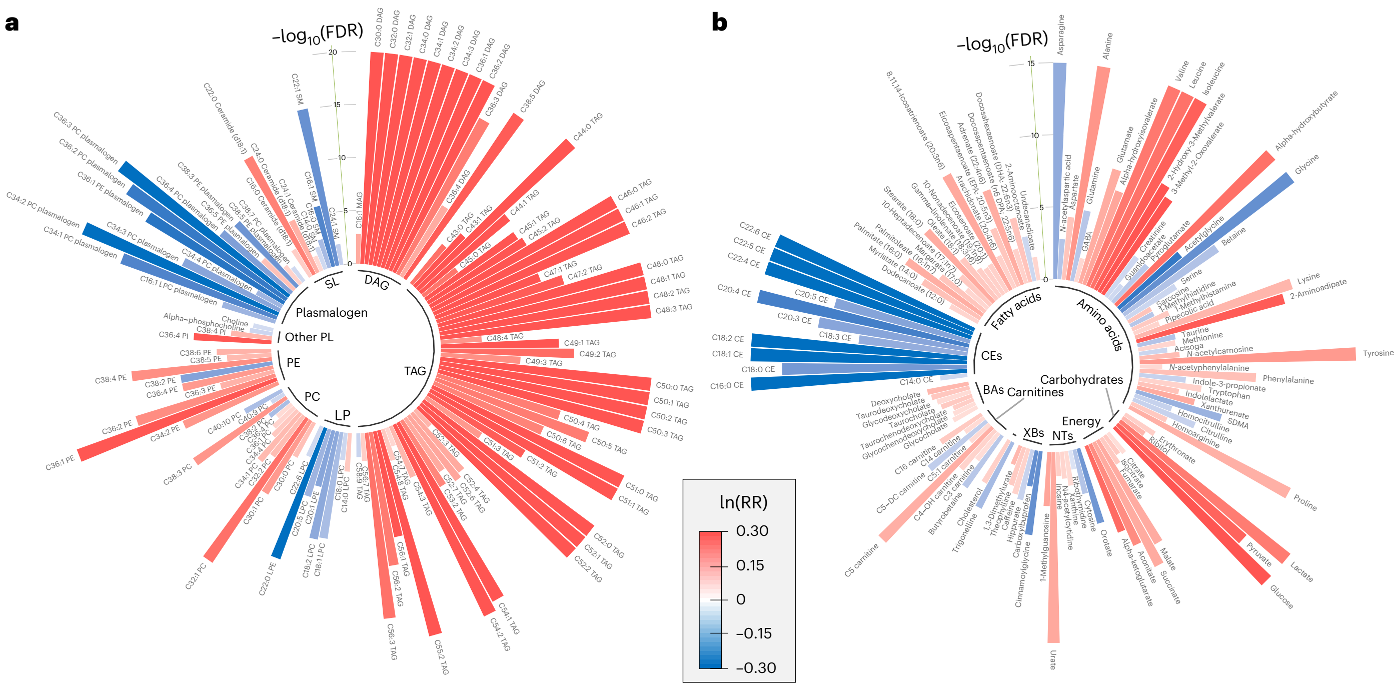


Fig. 2 | Associations between 235 metabolites and incident T2D in meta-analysis of ten prospective cohorts. Circular plots illustrate metabolites associated with incident T2D at FDR < 0.05, by biochemical category. **a**, Results for complex lipids including monoacylglycerols (MAG) and DAG, TAG, LP, PC, PE, other PLs, PL plasmalogens and sphingolipids (SG). **b**, Results for other metabolites, including amino acids, carbohydrates, bioenergetic metabolites, nucleotides (NTs), xenobiotics (XBs), as well as other lipid metabolites including carnitines, BAs, CEs and nonesterified fatty acids. Each bar represents results for one metabolite; red and blue indicate positive and inverse associations, respectively; color depth indicates association magnitude, that is, ln(RR) per s.d. increment in the metabolite, capped at −0.3 to 0.3; and bar height indicates association significance, capped at 10^{-20} in **a** and 10^{-15} in **b**. Analyses were conducted in each cohort by racial/ethnic groups adjusting for age, sex, smoking, alcohol consumption, fasting status, hypertension, dyslipidemia, lipid-lowering medication use, anti-hypertensive medication use, BMI, WHR, family history of T2D and cohort-specific variables, and results were combined using meta-analysis.

adults) (Extended Data Fig. 1b–d and Supplementary Table 4). Further comparison between cohorts applying Broad Institute versus Metabolon platforms also yielded consistent findings for overlapping metabolites (Extended Data Fig. 2 and Supplementary Table 5).

The 235 metabolites associated with incident T2D

We examined 205 complex lipids, identifying 130 significantly associated with incident T2D (including 18 associations not previously reported) and providing a comprehensive characterization of the lipidomic association patterns (Fig. 2a and Supplementary Tables 2 and 3). Positive associations with T2D risk were observed for triacylglycerols (TAG; the main form of energy storage in adipose tissue[12]), DAG (activate signaling cascades triggering hepatic insulin resistance[13]), ceramides (involved in β cell apoptosis and impaired insulin signaling[14]), as well as PC, phosphatidylethanolamine (PE) and phosphatidylinositol (related to insulin resistance[13,15,16]). Inverse associations with T2D risk were found for CEs, lysophospholipids (LPs), unsaturated PL plasmalogens (function as endogenous antioxidants[17]), some sphingomyelins (may counteract effects of ceramides[14]) and alpha-glycerophosphate (involved in glycolysis[18]). In addition, we noted that more double bonds were correlated with lower T2D risk among DAG, TAG and sphingomyelins, but with higher T2D risk among plasmalogens (Extended Data Fig. 3 and Supplementary Table 6), possibly due to functions of constituting fatty acids and/or the insulin-mediated regulation of FA desaturases[19].

Of other lipid signaling pathways, we identified 34 metabolites associated with T2D risk, including 26 significant associations not reported previously (Fig. 2b and Supplementary Tables 2 and 3). Notably, detrimental associations were identified for four primary and three secondary bile acids (BAs), including taurocholate, glycochenodeoxycholate, taurochenodeoxycholate, deoxycholate and taurodeoxycholate, that were not linked previously to T2D risk, adding further evidence to the role of BAs in T2D pathogenesis[20,21]. Significant associations with T2D were also noted for eight carnitine metabolites, including six significant associations not reported before (that is, C3, C5:1, C14, C16, C5–DC, butyrobetaine), supporting its role in energy metabolism[22]. Of the 19 free fatty acids associated with T2D risk, 18 showed positive associations, whereas some of their CE counterparts were inversely associated with T2D risk.

Across other pathways, we identified 43, five, seven and eight metabolites involved in amino acid, carbohydrate, energy and nucleotide metabolism, respectively, and eight other metabolites associated with T2D risk, comprising 23 significant associations not reported previously (Fig. 2b and Supplementary Tables 2 and 3). Key amino acid pathways underlying T2D risk indicated by these associations encompassing alanine, aspartate and glutamate, glycine, serine and threonine, urea cycle, arginine and proline, histidine, BCAA, tryptophan, lysine, phenylalanine and creatine metabolism. Notably, although coffee consumption has been consistently related to lower T2D risk[23], metabolites derived from phytochemicals in coffee showed both positive (for example, caffeine, theophylline and 1,3-dimethylurate) and inverse (for example, trigonelline and hippuric acid) associations with T2D risk.

Shared genetic architecture between metabolites and T2D

To offer new biological insights, we examined the shared genetic architecture between T2D-associated metabolites and T2D risk. Through meta-analyses in up to 18,590 people from eight cohorts, we curated genome-wide association study (GWAS) summary statistics for 458 harmonized (including 233 T2D-associated) metabolites (Fig. 1 and Supplementary Table 7; Methods). We identified one or more genetic loci for 165 T2D-associated metabolites at $P < 1.09 \times 10^{-10}$ (5×10^{-8}

https://doi.org/10.1038/s41591-025-04105-8

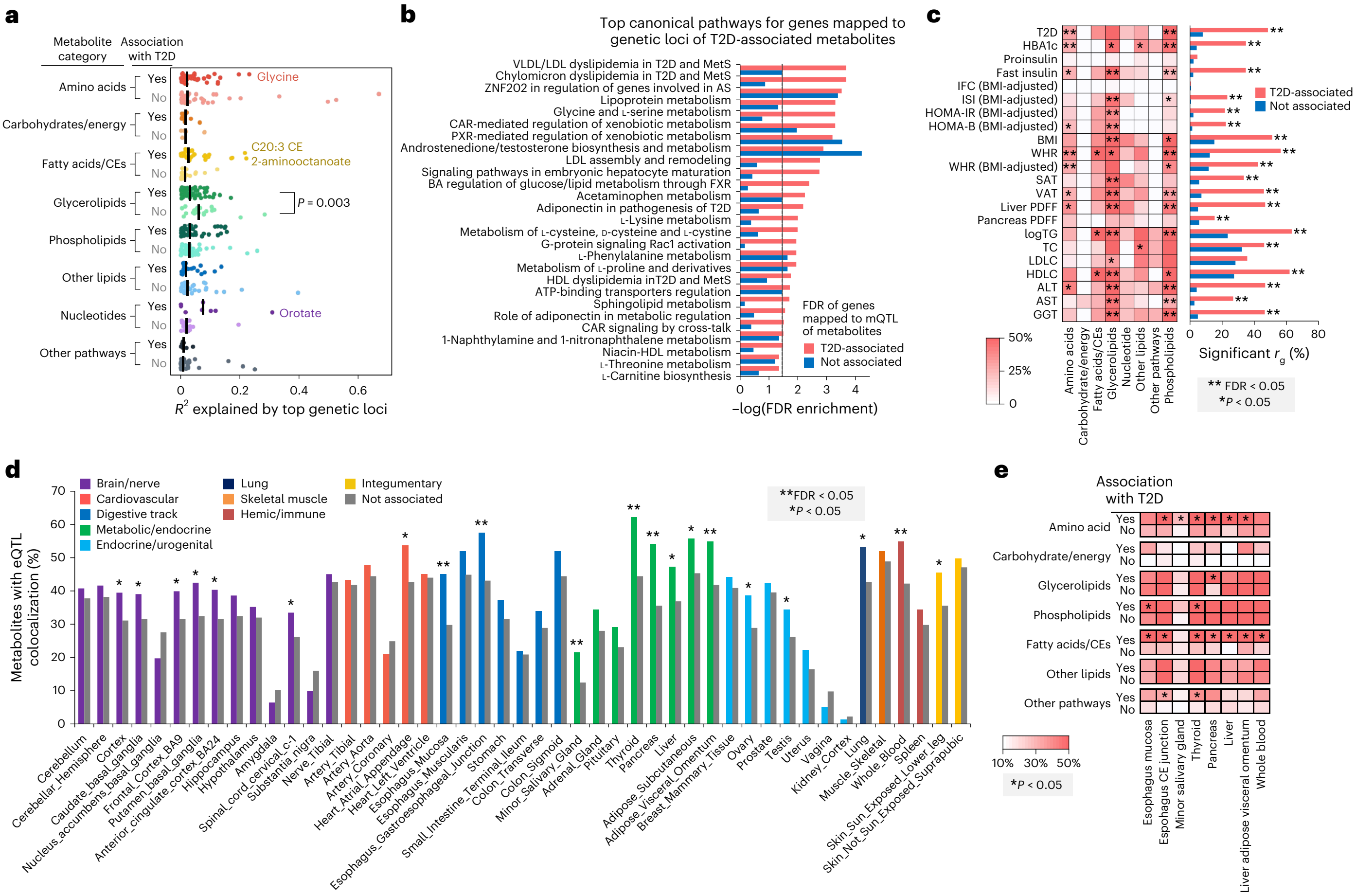


Fig. 3 | Genetic determinants of T2D-associated metabolites. We examined genetic architectures of T2D-associated metabolites based on genome-wide meta-analysis summary statistics. **a**, R^2 explained by genetics comparing the T2D-associated metabolites versus other metabolites by biochemical category (Wilcoxon test, statistical significance defined as two-sided $P < 0.00625$, correcting for eight categories). **b**, Top enriched canonical pathways for genes mapped to mQTLs of T2D-associated metabolites, most of which were not enriched for genes mapped to mQTLs of non-associated metabolites. AS, atherosclerosis; CAR, constitutive androstane receptor; FXR, farnesoid X receptor; MetS, metabolic syndrome; PXR, pregnane X receptor; VLDL, very low-density lipoprotein. **c**, Percentages of metabolites showing nominally significant ($P < 0.05$) genetic correlations (r_g) with traits reflecting T2D pathophysiology, comparing T2D-associated versus non-associated metabolites (two-sided chi-squared test). Barplot shows results for all metabolites (**FDR < 0.05, correcting for 22 traits); and heatmap shows percentage among T2D-associated metabolites by biochemical category (**FDR < 0.05 correcting for 121 comparisons; *$P < 0.05$). ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT, gamma-glutamyltransferase; HDLC, HDL cholesterol; LDLC, LDL cholesterol; TC, total cholesterol; TG, triglycerides. **d**, Proportions of metabolites colocalized (PPH4 > 0.8) with tissue-specific gene expression across 47 human tissues. We tested whether the proportions were higher among T2D-associated metabolites (colors: organ systems) versus non-associated metabolites (gray) using univariant logistic regression (**one-sided FDR < 0.05 correcting for 47 tissue types; *$P < 0.05$). **e**, For tissue types showing enriched genetic colocalizations with T2D-associated metabolites (seven tissue types with FDR < 0.05, plus the main metabolic organ liver with $P < 0.05$), we detailed the enrichment by biochemical category (color depth: proportions among T2D-associated versus non-associated metabolites; *one-sided $P < 0.05$).

correcting for 458 metabolites), with 45% of the identified metabolite–locus pairs not reported by previous major metabolite quantitative trait locus (mQTL) studies[8,24–26] (Supplementary Table 8; Methods).

Most mQTLs were annotated to genes involved in cellular metabolism, synthesis, and/or transportation of the corresponding metabolites. Approximately 23% of the identified loci were associated with multiple T2D-associated metabolites, especially those from the same or closely related pathways (for example, *GCKR*, *FADS1-3* and *ZNF259*) (Extended Data Fig. 4). A brief comparison between racial/ethnic groups suggested that some mQTLs may be specific to certain groups (for example, *OPLAH* for pyroglutamate in Hispanic/Latino adults) (Supplementary Figs. 5 and 6), warranting confirmation by larger *trans*-ancestry studies. Several mQTLs for T2D-associated metabolites overlapped with known T2D risk loci[27], with significant genetic colocalizations observed at several loci (posterior probability for H4 (PPH4) > 0.8). For example, 53 lipids, ten amino acids and two carbohydrates colocalized with T2D at the *GCKR* locus (probably driven by rs1260326). Similar colocalizations with T2D were observed at *LDL* and *APOE* for several complex lipids and at *L17REL* for leucine and valine (Extended Data Fig. 4 and Supplementary Tables 8 and 9).

The number of independent variants and the proportion of variance explained by genetics per metabolite, were generally similar between T2D-associated and non-associated metabolites (mean $r^2 = 4.4\%$ versus 5.1%; $P_{\text{difference}} = 0.47$) (Fig. 3a, Supplementary Figs. 7 and 8 and Supplementary Table 10). The top enriched canonical pathways for mQTLs of T2D-associated metabolites, however, were notably different from those of non-associated metabolites. Genes annotated to mQTLs of T2D-associated metabolites were enriched in pathways closely relevant to T2D pathogenesis—such as lipoprotein metabolism[28,29], adiponectin functions[30], BAs in glucose/lipid regulation[31], insulin response and Rac1 activation[32], as well as glycine and L-serine[33], L-lysine[34], L-cysteine[35] and L-phenylalanine metabolism, whereas most of these pathways were not enriched for mQTLs of non-associated metabolites (Fig. 3b, Extended Data Fig. 5 and Supplementary Table 11).

We then examined genetic correlations (r_g) between metabolites and 21 traits indicative of various T2D pathophysiologic mechanisms, leveraging summary statistics from large-scale GWAS for these traits (Methods). T2D-associated versus non-associated metabolites demonstrated substantially more significant r_g with fasting insulin (17-fold), BMI-adjusted insulin secretion and sensitivity indices (10- to 32-fold), liver enzymes (10- to 11-fold), intrahepatic and visceral fat (9- and 6-fold, respectively), obesity and blood lipids. Such an enrichment seemed to be driven by complex lipids, lipid signaling metabolites and amino acids (Fig. 3c). Significant r_g demonstrated a potential link between metabolites and physiological functions; for example, BCAAs were genetically correlated with traits reflecting insulin resistance, ectopic fat and impaired liver function (Supplementary Table 12).

Genetic colocalization between circulating metabolites and tissue-specific transcriptome

We hypothesized that levels of circulating metabolites may partially reflect biological homeostasis and gene regulations of related metabolic pathways across different organ systems. We therefore conducted a colocalization analysis between circulating mQTLs and tissue-specific *cis*-gene expression quantitative trait loci (eQTLs) of 47 human tissue types (leveraging Genotype-Tissue Expression v.8 data[36]) (Methods). Genetic colocalizations were observed across all tissue types (PPH4 > 0.8), supporting our hypothesis. T2D-associated metabolites, compared to non-associated metabolites, had significantly higher (FDR < 0.05) percentage of colocalizations in seven digestive and metabolic/endocrine tissues, including thyroid (62%), esophagus mucosa (45%), esophagus–gastroesophageal junction (58%), visceral fat (55%), whole blood (55%), pancreas (54%) and salivary gland (21%), and nominally higher ($P < 0.05$) percentage of colocalizations in another 13 tissue types including liver (Fig. 3d and Supplementary Table 13). Such an enrichment of colocalizations seemed to be driven primarily by T2D-associated amino acids, fatty acids and complex lipids (Fig. 3e). Further, each T2D-associated metabolite seemed to be colocalized with gene expressions within several, instead of one specific, metabolic/endocrine tissue types (Extended Data Fig. 6), consistent with the cumulative evidence that T2D development involves integrative biological changes across liver, fat, pancreas and digestive organ systems[3].

We observed several instances where tissue-specific gene expression, circulating metabolites and T2D colocalized at the same potential causal variants, highlighting potential genes and tissue types underlying the observed metabolite–T2D associations. For example, of the 65 metabolites colocalized with T2D at the *GCKR*/*PPM1G*/*IFT172* locus, 61 also colocalized with *PPM1G* expression in pancreas, *IFT172* in thyroid and/or *NRBP1* in esophagus–gastroesophageal junction (likely causal variant rs1260326). Similarly, 34:4 PC colocalized with T2D, as well as *FADS1* expression in liver, visceral fat and esophagus–gastroesophageal junction, and *TMEM258* expression in thyroid, by rs174545 (Supplementary Fig. 9a,b).

Bidirectional Mendelian randomization analysis

We conducted two-sample Mendelian randomization (MR) analyses to infer the potential causal relationships between 233 T2D-associated metabolites and T2D risk, leveraging a published consortium GWAS for T2D[27] (Methods). Genetically predicted circulating levels of 42 lipids and five amino acids were associated with T2D risk (FDR < 0.05), supported by several MR methods (Supplementary Table 14a). Sensitivity analyses selecting genetic instruments using a more stringent *P* threshold did not change the results, but removing variants on the three most recurrent loci (that is, *GCKR*, *ZNF259* and *FADS1-3*) attenuated results especially for lipids, which was expected given the roles of these genes in lipogenesis and lipid metabolism (Supplementary Fig. 10). Of note, genetically predicted T2D was not associated with any metabolite except for glucose—a known diagnostic criterion, rather than an etiological biomarker of T2D (Supplementary Table 14b), supporting that our prospective analysis findings are less likely to be due to reverse causation.

Modifiable risk factors and T2D-associated metabolites

Lifestyle and dietary factors play a pivotal role in metabolism and T2D development[37–39]. We next examined relationships between modifiable risk factors (that is, BMI, smoking, PA and intakes of 15 main food groups, mutually adjusted for one another) with circulating metabolites in up to 16,883 participants (Fig. 1; Methods). BMI accounted for more between-person variation in T2D-associated versus non-associated metabolites ($r^2 = 1.52\%$ versus 0.55%, $P_{\mathrm{difference}} = 1 \times 10^{-13}$), which seemed to be driven by glycerolipids (GLs), PLs and several amino acids (Fig. 4), consistent with their strong genetic correlation with BMI (Fig. 3c). Behavioral factors (especially PA, and red meat, vegetable and coffee/tea consumption) in total explained more variations in T2D-associated versus non-associated metabolites ($r^2 = 7.73\%$ versus 6.57%, $P_{\mathrm{difference}} = 0.029$), especially for GLs, fatty acids, amino acids and bioenergetic metabolites (Fig. 4 and Supplementary Table 15).

Metabolites mediating associations between modifiable risk factors and incident T2D

T2D-associated metabolites (versus non-associated metabolites) seemed to show stronger associations with several baseline risk factors, in a direction that is consistent with the epidemiological associations between risk factors and T2D risk (Fig. 5a–c). For example, among the 235 T2D-associated metabolites, there was a strong, positive correlation ($r = 0.86$) between their association coefficients with baseline BMI and their prospective association coefficients with incident T2D (Fig. 5a). Likewise, positive correlations of association coefficients were observed for risk-increasing behavioral factors such as smoking, and higher consumption of red meat and sugary drinks. In contrast, metabolites associated with higher levels of PA, and higher consumption of coffee/tea and vegetables, tended to be associated with lower T2D risk ($r = -0.65$, -0.46 and -0.34, respectively) (Fig. 5b–c, Extended Data Fig. 7 and Supplementary Table 16).

Four risk factors (BMI, PA, coffee/tea consumption and red meat intake) demonstrated expected prospective associations with T2D risk consistently across our study cohorts (Supplementary Table 17a). We therefore employed a mediation analysis to identify which metabolites, and to what degree, mediated the associations between these risk factors and incident T2D. For BMI and PA, we identified 148 and 50 metabolites, respectively, potentially mediating their associations with T2D risk (Fig. 5d–e and Supplementary Table 17a). Notably, many of these metabolites have been linked, in our genetic analyses, to T2D-related traits such as intrahepatic and visceral fat, lipids and liver enzymes, and to tissue types such as visceral fat, pancreas and thyroid, among others (Fig. 5g and Supplementary Tables 12, 13 and 17a). We found eight metabolites (including C22:0 ceramide, C32:0 DAG and C36:2 PC Plasmalogen) as potentially causal mediators between BMI and T2D risk, based on mediation analysis and two-step MR analysis (Supplementary Fig. 11 and Supplementary Table 17b). These findings suggest that obesity and PA may affect T2D risk through metabolic modulations related to visceral and intrahepatic fat deposition, liver and endocrine dysfunction, and lipid dysregulation.

We identified 74 metabolites as potential mediators between coffee/tea consumption and lower T2D risk, comprising several complex lipids, hippuric acid, isoleucine and glycine (Fig. 5f and Supplementary Table 17a). Hippuric acid is formed through hepatic glycine conjugation of benzoic acid, which is generated by the gut microbiota from polyphenols such as chlorogenic acids and epicatechins (abundant in coffee and tea)[40,41], highlighting a potential host–microbe interplay in polyphenol metabolism and metabolic health. We also identified six lipids as potential mediators between red meat intake and T2D risk, including lipids linked to ectopic fat and lipid dysregulation in our genetic analyses (Supplementary Tables 12, 13 and 17a).

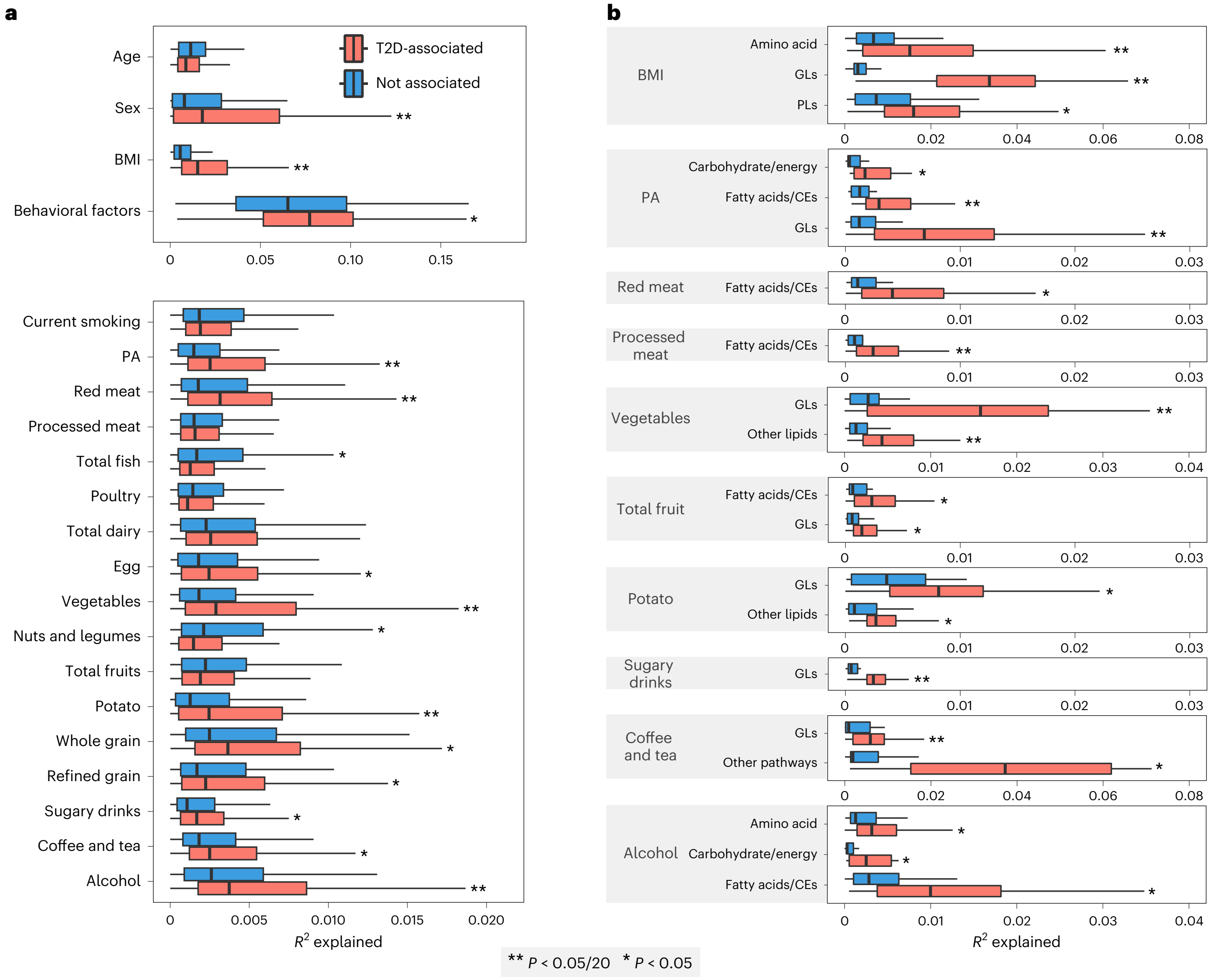


Fig. 4 | Variance of metabolites explained by modifiable risk factors. **a**, Boxplots comparing variance explained by age, sex and modifiable risk factors (including smoking, PA and intakes of 15 main food groups), for T2D-associated metabolites versus non-associated metabolites. **b**, Boxplots showing several specific biochemical categories of metabolites that drove the differential R^2. Each box shows the IQR, line in box indicates median and whiskers extend from the box to the smallest and largest value within 1.5 IQR from the lower and upper quartiles. Wilcoxon test was used to compare R^2 of the T2D-associated versus that of other metabolites; **two-sided $P < 0.0025$ (Bonferroni correction for 20 examined factors); *two-sided $P < 0.05$. For each metabolite, we first fitted a linear regression to regress inverse normal transformed metabolite on age, sex, BMI (standardized), PA (METs hours per week; standardized), all 15 main food groups (red meat, processed meat, poultry, fish and seafood, egg, total dairy, total vegetables, total fruits, potato, nuts and legume, whole grain, refined grain, sugary drinks, coffee and tea and alcohol; servings per day), fasting status and other cohort-specific variables simultaneously. We then calculated R^2 of the metabolites explained by each of the risk factors based on association coefficients and the variance of metabolite and risk factors. The analyses were conducted in NHS, NHS2, HPFS, SOL and WHI separately ($n = 16,883$) by main racial/ethnic groups and R^2 were averaged for the comparison.

A metabolomic signature to reflect the complex metabolic states predictive of T2D risk

Finally, we developed a multi-metabolite signature reflecting the complex metabolic states predictive of future T2D risk using elastic net regression, focusing on T2D-associated metabolites shared between the two metabolomic platforms to facilitate translational applicability of our findings. A leave-one-cohort-out cross-validation approach was applied to avoid overfitting (Methods and Supplementary Fig. 18a). In independent testing cohorts, the metabolomic signature alone demonstrated decent prediction performance for incident T2D risk, with an area under the receiver operating characteristic (ROC) curve (AUC) ranging from 0.62 to 0.86. Compared to a conventional model with traditional risk factors, the model that additionally included the metabolomic signature substantially improved T2D risk prediction with the AUC ranging from 0.69 to 0.92 (AUC increment $P < 0.05$ in all cohorts, except $P = 0.054$ in SOL) (Fig. 6a–c, Extended Data Fig. 8, Supplementary Fig. 12 and Supplementary Table 18b). In secondary analyses of five datasets with available fasting glucose, the addition of the metabolomic signature improved the model AUC significantly ($P < 0.05$ in three datasets) to marginally ($P = 0.06$ in SOL) beyond traditional risk factors and fasting glucose, except for PREDIMED ($P = 0.18$) (Extended Data Fig. 9).

Across cohorts, crude incidence of T2D increased from 7.7% in the lowest to 37.7% in the highest decile of the metabolomic signature (Fig. 6d). In a multivariable-adjusted analysis combining all cohorts, participants in the highest decile had a 5.1-fold higher risk of T2D compared to those in the lowest decile (RR = 5.07; 95 CI%, 4.02–6.39) (Fig. 6e and Supplementary Table 18c). Further assessing associations with

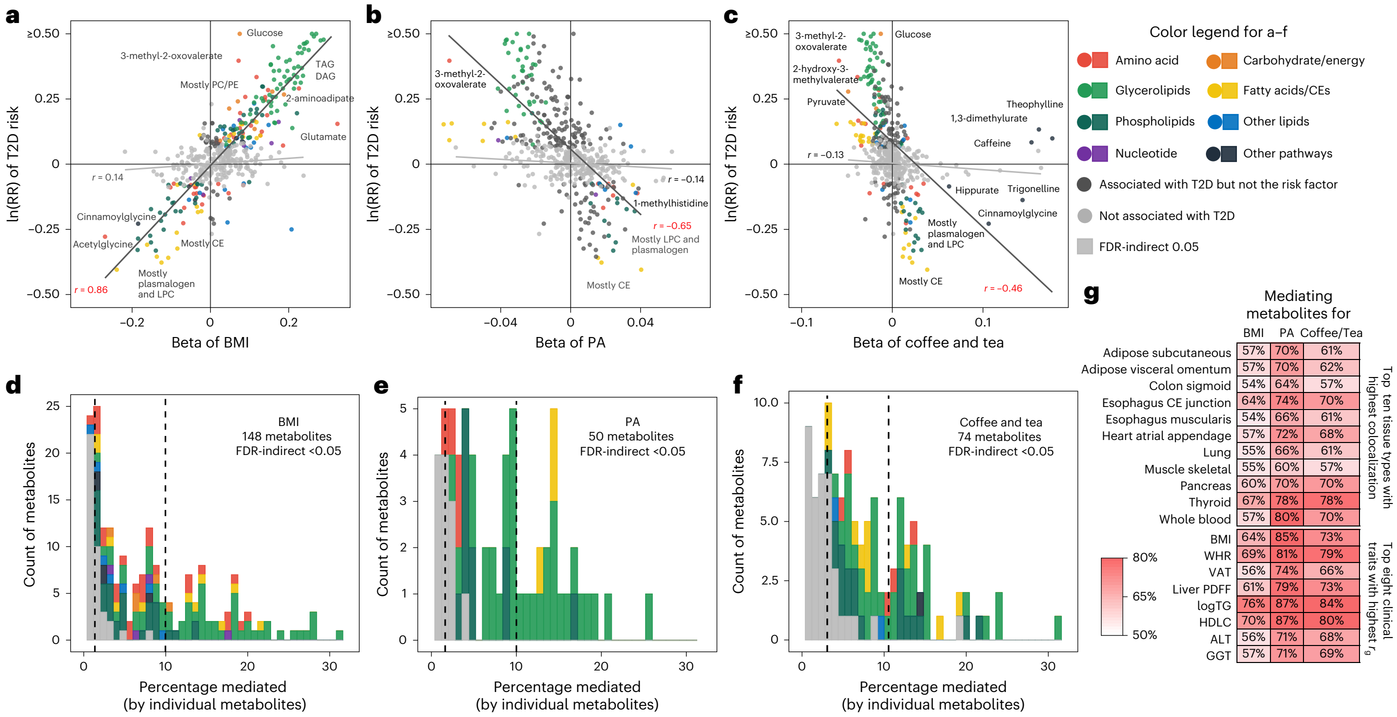


Fig. 5 | Metabolites that potentially mediate associations between modifiable risk factors and T2D risk. a–**c**, Scatterplots compare the associations of metabolites with the risk factors BMI (**a**), PA (**b**) and coffee and/or tea consumption (**c**) versus their associations with T2D risk. Each dot represents a metabolite (colored: associated with the risk factor and incident T2D at FDR < 0.05 by biochemical category; dark gray: associated with incident T2D but not the risk factor; light gray: not associated with incident T2D); and the two trend lines are for T2D-associated (dark gray) and non-associated metabolites (light gray) separately. Association coefficients (betas) for risk factors are from MWASs in which all risk factors were mutually adjusted (including age, sex, BMI, PA, consumption of 15 main food groups, fasting status and other cohort-specific variables). For metabolites associated with a risk factor and incident T2D in an epidemiologically expected direction, we conducted mediation analysis testing the indirect effect (risk factor − T2D association via a metabolite). **d**–**f**, For metabolites whose indirect effects were in the same direction as the total effect, we present the distribution of proportion mediated (indirect effect/total effect) for BMI (**d**), PA (**e**) and coffee and/or tea consumption (**f**). All analyses were conducted separately in NHS, NHS2, HPFS, SOL and WHI (*n* up to 16,883 for individual metabolites) and results were combined using meta-analysis. **g**, For metabolites showing significant mediating effects between risk factors and incident T2D, we highlighted the top tissue types where these metabolites showed the most genetic colocalizations with tissue-specific gene expression, and the top clinical traits with which these metabolites have most genetic correlation.

modifiable diet/lifestyle factors, we found that greater BMI and higher consumption of red meat and sugary drinks were associated with a higher metabolomic signature score, whereas more PA and higher intakes of whole grain, coffee/tea and wine were associated with a lower signature score (Fig. 6f and Supplementary Table 18d).

The final metabolomic signature model, derived based on all study cohorts, comprised 44 metabolites (including 20 amino acids, 19 involved in lipid/energy metabolism and five others), with many potentially linking modifiable risk factors to T2D risk (Fig. 6f and Supplementary Table 18a). For instance, alanine, which connected higher BMI and intakes of red meat and sugary drinks with higher T2D risk, was found as a potential mediator between BMI and T2D risk by our mediation and two-step MR analyses (Supplementary Tables 14a and 17a,b). Several metabolites, including trigonelline, hippuric acid, isoleucine and glycine, connected higher coffee/tea intake to lower T2D risk (Fig. 6f). Taking together, this metabolomic signature may serve as a predicting/monitoring biomarker to facilitate risk prediction, risk stratification and evaluation of effects of diet/lifestyle interventions on T2D prevention.

Discussion

This is one of the largest and most comprehensive investigations of metabolomic profiles associated with T2D risk, integrating blood metabolomic, genomic and diet/lifestyle data across racially and ethnically diverse cohorts. Collectively, our study identified a profile of 235 metabolites reflecting a dysregulated metabolism driven by both genetics and modifiable risk factors and predicts future T2D risk.

A key strength of this study is the harmonized analysis of individual-level data from ten prospective cohort studies using standardized protocols. This design provided high statistical power, enabling the identification of 235 metabolites prospectively associated with T2D risk, offering a comprehensive view of the metabolic landscape underlying T2D pathogenesis and substantially expanding upon the 123 metabolites reported in a recent literature-review-based meta-analysis of more than 60 studies[4]. Our identified significant associations include 34 that were only nominally significant in previous studies and 33 never linked to T2D risk. The use of individual-level data also allowed consistent adjustments of covariates and result comparisons across population groups and metabolomic platforms—which are not feasible in literature-review-based meta-analyses. Notably, associations between the identified metabolites and T2D risk remain robust after adjustments for obesity/adiposity, blood lipids, blood pressures, lifestyle factors or kidney function, and were generally consistent across popular liquid chromatography–tandem mass spectroscopy (LC–MS) platforms and major racial and ethnic groups.

Previous mQTL studies have advanced our understanding of genetic regulation of metabolic homeostasis[7,8,24,42,43]. Our study offers additional insights into the shared genetic architectures between metabolites and T2D. First, genetic determinants of T2D-associated metabolites were enriched in pathways central to T2D pathogenesis, including regulatory signaling of glucose response, insulin resistance and lipid homeostasis, despite their modest contributions to the overall metabolite variation. In addition, many of these metabolites were

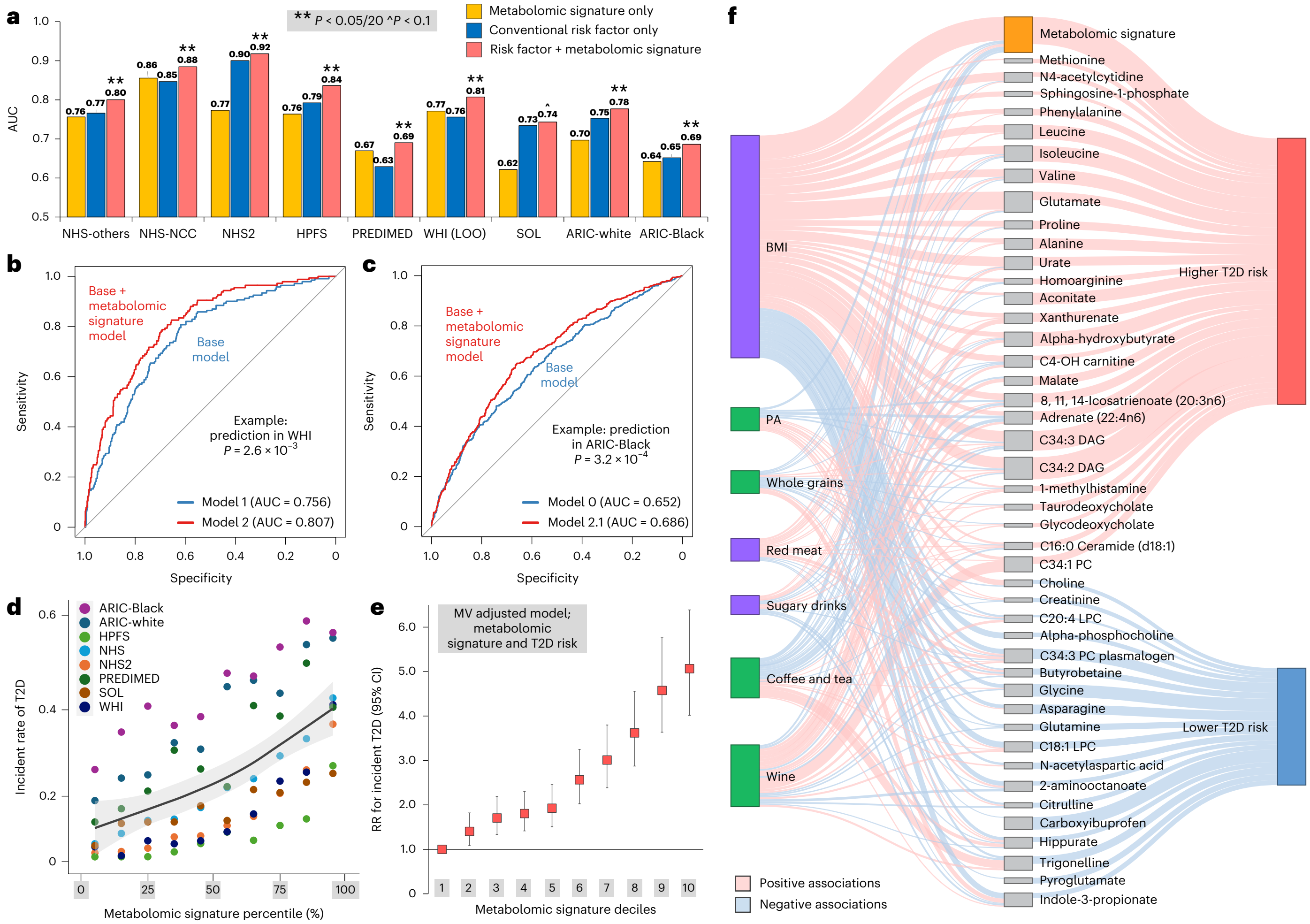


Fig. 6 | A multi-metabolite signature for T2D risk prediction. a, AUC for T2D risk prediction in each cohort. Yellow: the model with metabolomic signature only, acquired using a leave-one cohort-out cross-validation approach to avoid overfitting (within WHI, the signature was acquired using a leave-one-out cross-validation); blue: the model with conventional risk factors including age, sex, smoking, BMI, dyslipidemia, hypertension, lipid-lowering medication use, anti-hypertensive medication use and family history of T2D; red: the model with conventional risk factors plus the metabolomic signature. For cohorts analyzed with Cox model, we plotted AUC estimated at the median follow-up time. We compared the AUC of the conventional plus metabolomic signature model to that of the conventional model; **two-sided $P < 0.01$, ^two-sided $P < 0.1$. **b**,**c**, Two examples of ROC curves and two-sided P values from WHI (**b**) and Black participants from ARIC (**c**). **d**, Crude incident rate of T2D by cohort, across deciles of the metabolomic signature, with a smooth trendline and 95% CI (gray band) from locally estimated scatterplot smoothing (LOESS). **e**, Relative risk ratio (points) and 95% CI (lines) for incident T2D, comparing participants in higher versus the lowest deciles of the metabolomic signature. Analyses were conducted separately in NHS, NHS2, HPFS, SOL, WHI, PREDIMED and Black and white participants from ARIC, separately, adjusting for age, sex, smoking, alcohol consumption, fasting status, hypertension, dyslipidemia, lipid-lowering medication use, anti-hypertensive medication use, BMI, WHR, family history of T2D and cohort-specific variables. We plotted relative risk ratios from the meta-analysis ($n = 20{,}930$). **f**, In multivariable analysis, BMI, red meat intake and sugary drink consumption (purple) were associated positively with the metabolomic signature, whereas PA, and intakes of coffee/tea, whole grains and wine (green), were associated inversely with the metabolomic signature (FDR < 0.05). A Sankey plot was used to demonstrate the associations between each of the 44 metabolites constituting the final metabolomic signature with these risk factors and with T2D risk (band-width proportional to the association coefficients).

genetically correlated with traits reflecting T2D pathophysiology, such as insulin secretion, insulin resistance, obesity, ectopic fat deposition and liver function. Furthermore, circulating levels of T2D-associated metabolites may reflect biological regulations within specific tissue types relevant to nutrient metabolism (digestive track, pancreas and liver), endocrine/metabolic regulation (thyroid, pancreas and adipose tissues), and inflammation (whole blood and visceral fat). Mapping metabolites—particularly those with strong genetic regulation—to relevant tissues and physiological functions can facilitate mechanistic interpretation. For example, TAGs 46:1 and 46:2 were linked to visceral but not subcutaneous fat, gene expression in pancreas, and insulin secretion and sensitivity indices, suggesting a role in visceral adiposity-related insulin resistance[44]. Notably, although dyslipidemia is often viewed as a consequence of diabetes[45], our findings and recent evidence[4,28,29] indicate a complex interplay between lipid and amino acid metabolism and glucose homeostasis. Future studies may leverage our results to further explore mechanisms linking circulating metabolites to T2D risk.

Obesity, diet and lifestyle can directly influence circulating metabolome[9–11]. We showed that obesity, PA and diet may impose substantial impacts on the subset of metabolites associated with T2D risk, which is consistent with the notion that environmental factors need to disturb causal pathways to affect T2D risk[46]. We also identified specific metabolites probably mediating risk factor–T2D associations. These findings, together with our genetic results, highlight potential causal pathways underlying T2D that deserve further mechanistic investigations. For instance, several metabolites mediating the inverse association between PA and T2D risk seem to be involved in ectopic fat-related insulin resistance and liver function impairment, whereas metabolites mediating the association between coffee/tea consumption and T2D risk were linked to polyphenol metabolism, glucose response, insulin resistance, ectopic fat deposition and liver function. Future clinical trials and functional studies could prioritize these pathways when investigating the causal effects of PA and coffee (or tea) consumption on metabolic health.

The blood metabolome reflects overall biological states and may serve as a prediction or monitoring tool in T2D prevention and therapeutic interventions. In the final step, we developed a multi-metabolite signature that robustly predicted future T2D risk, either used alone or in combination with conventional risk factors, and could identify people with extremely high risk of T2D before T2D diagnosis. The metabolomic signature is also associated with key modifiable risk factors and comprises metabolites that may mediate the associations between various diet/lifestyle factors and T2D risk. Collectively, this metabolomic signature captures the complex metabolic states associated with T2D risk, and is applicable in future clinical and research settings, as either a prediction tool to identify people with high risk of T2D for early prevention, or an intermediate biomarker to evaluate the efficacy of dietary and lifestyle interventions.

We acknowledge several limitations. First, although metabolomic data were harmonized between two LC–MS platforms, some metabolites were unique to one platform, limiting their sample sizes to specific cohorts. Second, although MR analysis is used frequently to infer causality between metabolites and diseases[47–49], its results should be interpreted cautiously, because some metabolites have weak genetic instruments and many molecules within the same pathways share genetic loci. To minimize false positives, we used the conservative mode-based estimate as our primary method, and confirmed findings with another three MR methods. We note that the lack of significant MR results does not preclude potential biological connections between a metabolite and T2D. Third, due to the observational design, our study cannot establish causality. Randomized trials are warranted to assess how diet/lifestyle affect T2D-associated metabolites and T2D risk. Finally, although our study included people with racially and ethnically diverse backgrounds, and associations were generally consistent across groups, 77% of our participants were non-Hispanic white individuals, highlighting the need for further replication and additional investigations in more diverse populations.

In summary, we identified 235 metabolites associated with incident T2D, potentially reflecting the influence of genetic and modifiable factors (especially diet, PA and adiposity) on metabolic pathways underlying T2D risk. This included 67 significant associations not previously reported encompassing BA, lipid, carnitine, urea cycle and arginine/proline, glycine and histidine metabolic pathways. As a resource, our findings may aid mechanistic and clinical research to investigate pathways underlying T2D pathophysiology. Our metabolomic signature may serve as a powerful tool for risk stratification and as a monitoring biomarker to inform precision T2D prevention and early intervention.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-025-04105-8.

References

1. *IDF Diabetes Atlas 2025*, 11th edn (International Diabetes Federation, 2025).
2. Galicia-Garcia, U. et al. Pathophysiology of type 2 diabetes mellitus. *Int. J. Mol. Sci.* **21**, 6275 (2020).
3. Roden, M. & Shulman, G. I. The integrative biology of type 2 diabetes. *Nature* **576**, 51–60 (2019).
4. Morze, J. et al. Metabolomics and type 2 diabetes risk: an updated systematic review and meta-analysis of prospective cohort studies. *Diabetes Care* **45**, 1013–1024 (2022).
5. Guasch-Ferre, M. et al. Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis. *Diabetes Care* **39**, 833–846 (2016).
6. Qi, Q. et al. Host and gut microbial tryptophan metabolism and type 2 diabetes: an integrative analysis of host genetics, diet, gut microbiome and circulating metabolites in cohort studies. *Gut* **71**, 1095–1105 (2022).
7. Feofanova, E. V. et al. A genome-wide association study discovers 46 loci of the human metabolome in the Hispanic Community Health Study/Study of Latinos. *Am. J. Hum. Genet.* **107**, 849–863 (2020).
8. Feofanova, E. V. et al. Whole-genome sequencing analysis of human metabolome in multi-ethnic populations. *Nat. Commun.* **14**, 3111 (2023).
9. Li, J. et al. The Mediterranean diet, plasma metabolome, and cardiovascular disease risk. *Eur. Heart J.* **41**, 2645–2656 (2020).
10. Chen, L. et al. Influence of the microbiome, diet and genetics on inter-individual variation in the human plasma metabolome. *Nat. Med.* **28**, 2333–2343 (2022).
11. Bar, N. et al. A reference map of potential determinants for the human serum metabolome. *Nature* **588**, 135–140 (2020).
12. Bender, D. A. & Mayes, P. A. in *Harper's Illustrated Biochemistry* 31st edn (eds Rodwell, V. W. et al.) Ch. 14 (McGraw-Hill Education, 2018).
13. Boden, G. & Laakso, M. Lipids and glucose in type 2 diabetes: what is the cause and effect? *Diabetes Care* **27**, 2253–2259 (2004).
14. Bellini, L. et al. Targeting sphingolipid metabolism in the treatment of obesity/type 2 diabetes. *Expert Opin. Ther. Targets* **19**, 1037–1050 (2015).
15. Raubenheimer, P. J., Nyirenda, M. J. & Walker, B. R. A choline-deficient diet exacerbates fatty liver but attenuates insulin resistance and glucose intolerance in mice fed a high-fat diet. *Diabetes* **55**, 2015–2020 (2006).
16. van der Veen, J. N., Lingrell, S., da Silva, R. P., Jacobs, R. L. & Vance, D. E. The concentration of phosphatidylethanolamine in mitochondria can modulate ATP production and glucose metabolism in mice. *Diabetes* **63**, 2620–2630 (2014).
17. Messias, M. C. F., Mecatti, G. C., Priolli, D. G. & de Oliveira Carvalho, P. Plasmalogen lipids: functional mechanism and their involvement in gastrointestinal cancer. *Lipids Health Dis.* **17**, 41 (2018).
18. Yu, J. et al. Update on glycerol-3-phosphate acyltransferases: the roles in the development of insulin resistance. *Nutr. Diabetes* **8**, 34 (2018).
19. Lee, J. M., Lee, H., Kang, S. & Park, W. J. Fatty acid desaturases, polyunsaturated fatty acid regulation, and biotechnological advances. *Nutrients* **8**, 23 (2016).
20. Ferrell, J. M. & Chiang, J. Y. L. Understanding bile acid signaling in diabetes: from pathophysiology to therapeutic targets. *Diabetes Metab. J.* **43**, 257–272 (2019).
21. Vangipurapu, J., Fernandes Silva, L., Kuulasmaa, T., Smith, U. & Laakso, M. Microbiota-related metabolites and the risk of type 2 diabetes. *Diabetes Care* **43**, 1319–1325 (2020).
22. Virmani, M. A. & Cirulli, M. The role of L-carnitine in mitochondria, prevention of metabolic inflexibility and disease initiation. *Int. J. Mol. Sci.* **23**, 2717 (2022).
23. Huxley, R. et al. Coffee, decaffeinated coffee, and tea consumption in relation to incident type 2 diabetes mellitus: a systematic review with meta-analysis. *Arch. Intern. Med.* **169**, 2053–2063 (2009).

24. Surendran, P. et al. Rare and common genetic determinants of metabolic individuality and their effects on human health. *Nat. Med.* **28**, 2321–2332 (2022).
25. Lotta, L. A. et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat. Genet.* **53**, 54–64 (2021).
26. Chen, Y. et al. Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. *Nat. Genet.* **55**, 44–53 (2023).
27. Mahajan, A. et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet.* **54**, 560–572 (2022).
28. Soremekun, O. et al. Lipid traits and type 2 diabetes risk in African ancestry individuals: a Mendelian randomization study. *EBioMedicine* **78**, 103953 (2022).
29. Yuan, S. & Larsson, S. C. An atlas on risk factors for type 2 diabetes: a wide-angled Mendelian randomisation study. *Diabetologia* **63**, 2359–2371 (2020).
30. Achari, A. E. & Jain, S. K. Adiponectin, a therapeutic target for obesity, diabetes, and endothelial dysfunction. *Int. J. Mol. Sci.* **18**, 1321 (2017).
31. Chavez-Talavera, O., Tailleux, A., Lefebvre, P. & Staels, B. Bile acid control of metabolism and inflammation in obesity, type 2 diabetes, dyslipidemia, and nonalcoholic fatty liver disease. *Gastroenterology* **152**, 1679–1694 (2017).
32. Chiu, T. T., Jensen, T. E., Sylow, L., Richter, E. A. & Klip, A. Rac1 signalling towards GLUT4/glucose uptake in skeletal muscle. *Cell Signal.* **23**, 1546–1554 (2011).
33. Holm, L. J. & Buschard, K. L-Serine: a neglected amino acid with a potential therapeutic role in diabetes. *APMIS* **127**, 655–659 (2019).
34. Razquin, C. et al. Lysine pathway metabolites and the risk of type 2 diabetes and cardiovascular disease in the PREDIMED study: results from two case-cohort studies. *Cardiovasc. Diabetol.* **18**, 151 (2019).
35. Nakatsu, D. et al. L-Cysteine reversibly inhibits glucose-induced biphasic insulin secretion and ATP production by inactivating PKM2. *Proc. Natl Acad. Sci. USA* **112**, E1067–E1076 (2015).
36. GTEx Consortium The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
37. Lichtenstein, A. H. et al. 2021 Dietary guidance to improve cardiovascular health: a scientific statement from the American Heart Association. *Circulation* **144**, e472–e487 (2021).
38. Ardisson Korat, A. V., Willett, W. C. & Hu, F. B. Diet, lifestyle, and genetic risk factors for type 2 diabetes: a review from the Nurses' Health Study, Nurses' Health Study 2, and Health Professionals' Follow-up Study. *Curr. Nutr. Rep.* **3**, 345–354 (2014).
39. Hu, F. B. et al. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N. Engl. J. Med.* **345**, 790–797 (2001).
40. Liang, N. & Kitts, D. D. Role of chlorogenic acids in controlling oxidative and inflammatory stress conditions. *Nutrients* **8**, 16 (2015).
41. Musial, C., Kuban-Jankowska, A. & Gorska-Ponikowska, M. Beneficial properties of green tea catechins. *Int. J. Mol. Sci.* **21**, 1744 (2020).
42. Han, X. et al. Integrating genetics and metabolomics from multi-ethnic and multi-fluid data reveals putative mechanisms for age-related macular degeneration. *Cell Rep. Med.* **4**, 101085 (2023).
43. Rhee, E. P. et al. A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* **18**, 130–143 (2013).
44. Neeland, I. J. et al. Visceral and ectopic fat, atherosclerosis, and cardiometabolic disease: a position statement. *Lancet Diabetes Endocrinol.* **7**, 715–725 (2019).
45. Athyros, V. G. et al. Diabetes and lipid metabolism. *Hormones (Athens)* **17**, 61–67 (2018).
46. Kolb, H. & Martin, S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC Med.* **15**, 131 (2017).
47. Liu, J. et al. A Mendelian randomization study of metabolite profiles, fasting glucose, and type 2 diabetes. *Diabetes* **66**, 2915–2926 (2017).
48. Yuan, S., Merino, J. & Larsson, S. C. Causal factors underlying diabetes risk informed by Mendelian randomisation analysis: evidence, opportunities and challenges. *Diabetologia* **66**, 800–812 (2023).
49. Lotta, L. A. et al. Genetic predisposition to an impaired metabolism of the branched-chain amino acids and risk of type 2 diabetes: a Mendelian randomisation analysis. *PLoS Med.* **13**, e1002179 (2016).

Jun Li[1,2] ✉, Jie Hu[3,4,5], Huan Yun[5], Zhendong Mei[1], Xingyan Wang[5], Kai Luo[6], Marta Guasch-Ferré[2,7,8], Xikun Han[5], Buu Truong[5], Jordi Merino[8], Chengyong Jia[6], Miguel Ruiz-Canela[9,10,11], Casey M. Rebholz[12], Eun Hye Moon[13], Taryn Alkis[13], Guning Liu[14], Jie Yao[15], Xiyuan Zhang[16], Bianca C. Porneala[17], Jordi Salas-Salvadó[11,18], Thomas J. Wang[19], Josée Dupuis[20,21], Elizabeth Selvin[12], Xiuqing Guo[15], Shilpa N. Bhupathiraju[22], Jennifer A. Brody[23], Yongmei Liu[24], Alexis C. Wood[25], Kari E. North[26], Su Yon Jung[27], Ching-Ti Liu[20], Nona Sotoodehnia[28], Simin Liu[29], Lesley F. Tinker[30], A. Heather Eliassen[2,5,22], JoAnn E. Manson[1,5], Jose C. Florez[31,32,33], Robert E. Gerszten[32,34], Clary B. Clish[35], Liming Liang[5,36], Rozenn N. Lemaitre[23], Katherine L. Tucker[16], Stephen S. Rich[37], Jerome I. Rotter[15], Miguel Angel Martínez-González[2,9,10,11], Kathryn M. Rexrode[3], James B. Meigs[32,33,38], Eric Boerwinkle[13], Robert C. Kaplan[6,30], Frank B. Hu[2,5,22], Bing Yu[13] & Qibin Qi[6] ✉

[1]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [2]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [3]Division of Women's Health, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [4]Center for Genomic Medicine and Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [5]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [6]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA. [7]Department of Public Health, University of Copenhagen, Copenhagen, Denmark. [8]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [9]Department of Preventive Medicine and Public Health, University of Navarra, Pamplona, Spain. [10]Instituto de Investigacion Sanitaria de Navarra (IdiSNA), Edificio LUNA-Navarrabiomed, Pamplona, Spain. [11]CIBER Fisiopatologıa de la Obesidad y Nutricion (CIBERObn), Instituto de Salud Carlos III, Madrid, Spain. [12]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [13]Department of Epidemiology, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [14]Department of Environmental and Occupational Health Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [15]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA. [16]Department of Biomedical and Nutritional Sciences, College of Health Sciences, University of Massachusetts Lowell, Lowell, MA, USA. [17]Heller School for Social Policy and Management, Brandeis University, Waltham, MA, USA. [18]Human Nutrition Unit, Faculty of Medicine and Health Sciences, Institut d'Investigacio Sanitaria Pere Virgili, Universitat Rovira i Virgili, Reus, Spain. [19]University of Michigan Medical School, Ann Arbor, MI, USA. [20]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [21]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada. [22]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [23]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [24]Department of Medicine, Divisions of Cardiology and Neurology, Duke University Medical Center, Durham, NC, USA. [25]USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA. [26]Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA. [27]Department of Epidemiology, Fielding School of Public Health, Translational Sciences Section, Jonsson Comprehensive Cancer Center, School of Nursing, University of California, Los Angeles, CA, USA. [28]Cardiovascular Health Research Unit, Division of Cardiology, University of Washington, Seattle, WA, USA. [29]Department of Epidemiology and Biostatistics, The Joe C. Wen School of Population and Public Health; Mary and Steve Wen Cardiovascular Division, Department of Medicine, School of Medicine; and The Center for Global Cardiometabolic Health and Nutrition, The University California Irvine (UCI), Irvine, CA, USA. [30]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [31]Diabetes Unit, Department of Medicine and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [32]Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [33]Department of Medicine, Harvard Medical School, Boston, MA, USA. [34]Division of Cardiovascular Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [35]Metabolomics Platform, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [36]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [37]Department of Genome Sciences, University of Virginia, Charlottesville, VA, USA. [38]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. ✉e-mail: juli@bwh.harvard.edu; qibin.qi@einsteinmed.edu

Methods

Study participants and ethics approval

Our MWAS for incident T2D involves the use of data from ten prospective cohorts, including the Nurses' Health Study (NHS; initiated in 1976 with 121,701 female nurses aged 30–55 years[9,50]), NHS2 (started in 1989 with 116,429 female nurses aged 25–42 years[9,50]), Health Professionals Follow-Up Study (HPFS; started in 1986 with 51,529 male health professions aged 40–75 years[9]), Hispanic Community Health Study/Study of Latinos (SOL; enrolled 16,415 Hispanic/Latino adults aged 18–74 years during 2008–2011[51,52]), Women's Health Initiative (WHI; initiated in 1993 enrolling 68,132 women aged 50–79 years to one of three clinical trials or an observational study[53]), Atherosclerosis Risk in Communities (ARIC) study (enrolled 15,792 mostly Black and white US adults aged 45–64 years during 1987–1989[54]), Framingham Heart Study Offspring cohort (FHS; enrolled 5,124 adults; we focused on those attended the fifth examination during 1991–1995), Multi-Ethnic Study of Atherosclerosis (MESA; initiated in 2000 with 6,814 adults aged 45–84 years[55,56]), the Boston Puerto Rican Health Study (BPRHS; enrolled 1,500 self-identified Puerto Rican adults aged 45–75 years) and the Prevención con Dieta Mediterránea Study (PREDIMED; a 5-year dietary trial with 7,447 adults aged 55–80 years[57]). In each cohort, comprehensive data on demographics, medical and family history, diet, lifestyle and other health information were collected at baseline and were updated during longitudinal follow-ups. Blood samples were collected at baseline and/or during follow-ups. Our MWAS for incident T2D included participants with qualified metabolomics data, and were free of diabetes, cardiovascular disease and cancer at study baseline. The final analysis included 6,890 participants from NHS; 3,692 from NHS2 and 2,529 from HPFS; 2,821 from SOL; 1,392 from WHI; 1,288 white and 1,433 Black participants from ARIC; 1,424 from FHS; 902 from MESA; 378 from BPRHS and 885 from PREDIMED (Extended Data Table 1). Each study was approved by Institutional Review Boards at respective institutions or study centers, and all participants provided informed consent. Our GWAS for metabolites included participants from eight cohorts comprising NHS, NHS2, HPFS, SOL, WHI, ARIC, FHS and, in addition, the Cardiovascular Health Study (CHS; enrolled 5,201 adults during 1989–1990 and 678 predominantly Black participants in 1992–1993[58,59]) (Supplementary Table 7). The detailed descriptions of the design, data collection, ethical review of each cohort, and our inclusion and exclusion criteria are provided in Supplementary Methods.

Ascertainment of T2D

In all cohorts, incident T2D was defined when a participant was free of diabetes at baseline but was identified as having T2D during longitudinal follow-up. Detailed information on diagnosis criteria in each cohort is included in Supplementary Methods, and follow-up years and numbers of incident cases are listed in Extended Data Table 1. Briefly, in NHS/HPFS, T2D were identified by follow-up questionnaires, and confirmed through a supplementary questionnaire based on diagnostic criteria from the National Diabetes Data Group before 1998[60] and the American Diabetes Association (ADA) criteria after 1998[61,62]. In SOL, T2D was defined if a participant had fasting glucose ≥7.0 mmol l^{-1}, fasting ≤8 h and nonfasting glucose ≥11.1 mmol l^{-1}, post oral glucose tolerance test glucose ≥11.1 mmol l^{-1}, HbA1c ≥ 6.5%, current use of antidiabetic medications or self-reported physician-diagnosed diabetes[63]. In WHI, T2D was determined based on self-reported history of diabetes or using antidiabetic medications (pills or shots) in any visits/interviews. In ARIC and FHS, T2D was diagnosed if a person had fasting glucose ≥7.0 mmol l^{-1}, fasting ≤8 h and nonfasting glucose ≥11.1 mmol l^{-1}, or current use of antidiabetic medications with ARIC further considering self-reported physician-diagnosed diabetes[64,65]. T2D cases in MESA and BPRHS were determined according to the ADA criteria[66], which included fasting plasma glucose level ≥7.0 mmol l^{-1} or the use of antidiabetic medications or insulin[56,67]. In PREDIMED, T2D was adjudicated through blind assessment by a Clinical Endpoint and Adjudication of Events Committee, based on the ADA criteria[68].

Assessment of diet, lifestyle factors and covariates

Detailed information on data collection in each cohort is in Supplementary Methods. Briefly, demographic factors (for example, self-reported sex, and race and ethnicity), socioeconomic status, health information (for example, medical conditions and family history) and lifestyle (for example, smoking history and PAs), anthropometrics and blood pressure, were collected at baseline and follow-up visits, through self-administrated questionnaires, or in-person or telephone-based interviews by trained staff. PA was quantified as metabolic equivalent (MET) in hours per week. We calculated BMI based on baseline weight and height, and WHR based on waist and hip circumferences. Blood clinical biomarkers were measured using standard assays. Among participants with serum creatinine data, eGFR was estimated using the CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) formula, based on age, sex and race in NHS/HPFS, WHI, ARIC and PREDIMED[69], and standard reference equations for Hispanics adjusting for age and sex in SOL. In PREDIMED, two propensity scores were estimated to account for the probability of assignment to intervention groups[57].

In NHS/HPFS, diet was assessed using a semi-quantitative food frequency questionnaire (FFQ) every 4 years; in our analysis we averaged the intakes from the two FFQs closest to the time of blood draw (NHS: 1986 and 1990; NHS2: 1995 and 1999; HPFS: 1994 and 1998). In WHI, ARIC, FHS, MESA and BPRHS, diet was similarly assessed by FFQs designed and validated for application to their targeted populations (for example, multiethnic and geographically diverse populations in WHI[70–72] and Puerto Rican population in BPRHS[73]). In SOL, diet was assessed using two 24-h dietary recalls and a food propensity questionnaire[74]. The overall dietary quality was assessed by the Alternate Healthy Eating Index-2010 (AHEI-2010)[75] in all cohorts except for the PREDIMED trial, in which it was assessed by a 14-item Mediterranean Diet Adherence Screener score[57]. In NHS/HPFS, SOL and WHI, we also calculated baseline consumptions of 15 main food groups in the unit of servings per day.

Metabolomic profiling, quality control and data harmonization

Metabolomic profiling in NHS/HPFS, WHI, MESA, PREDIMED, FHS and CHS was conducted with the Metabolomics Platforms at the Broad Institute of MIT and Harvard University, using three to four complementary LC–MS methods[9,65,76]. Metabolomic profiling in SOL and ARIC (serum samples) and BPRHS (plasma samples) was conducted using LC–MS based methods by the Metabolon DiscoveryHD4 Panel at the Metabolon Inc.[63,77,78]. Detailed protocols for both platforms have been described previously[53,79].

Data processing was conducted within each study and, if applicable, separately within each batch (or substudy) if several batches/substudies were conducted within a cohort. Samples were removed if their metabolite detection rate was <80%, or were identified as outliers by multidimensional scaling analysis within a specific race/ethnic group. Metabolites were filtered if their detection rate across samples was <80% and, if applicable, had a coefficient of variation >20% for quality control (QC) samples. After quality filtering, missingness of each metabolite were imputed using the half minimum value, and the data were then standardized for analysis. Across all cohorts, we matched metabolites by their HMDB ID and/or PubChem ID, provided by the corresponding metabolomic laboratories. A total of 1,273 named metabolites were initially qualified for analysis in at least one cohort. To reduce single-study bias, we limited our analyses to 469 metabolites that were available in at least four independent cohorts, or available in at least three independent cohorts if the three cohorts covered both Metabolomic platforms. Finally, 407 metabolites from NHS, 363 from NHS2, 291 from HPFS, 364 from WHI, 327 from MESA, 274 from PREDIMED, 188 from FHS, 283 from SOL, 139 from ARIC and 231 from BPRHS were harmonized for our analysis (Extended Data Table 1). In CHS, 411 metabolites were included in genetic analyses (Supplementary Table 7). Details of the metabolomic profiling, QC and data processing are in the Supplementary Methods.

Metabolome-wide association analysis for incident T2D

Details of analytical approaches and models are provided in Supplementary Methods and Supplementary Table 1. Briefly, all association analyses were conducted separately for each cohort, stratified by major racial/ethnic groups when sample sizes permitted. Metabolites were inversely normal transformed by each substudy and racial/ethnic group (if applicable) in each cohort. To analyze the association between each metabolite and T2D risk, we applied Cox regression for studies of longitudinal cohort design (NHS excluding the T2D nested case–control substudy, NHS2, HPFS, SOL, ARIC, WHI, FHS, MESA and BPRHS); logistic regression for the NHS T2D nested case–control substudy; and Cox regression with Barlow weights[80] and robust estimators for the PREDIMED T2D nested case–cohort study. The basic multivariate model (model 1) was adjusted for age, sex, smoking status, alcohol consumption and, if applicable, education, family income, fasting status, lipid-lowering medications, anti-hypertensive medications, family history of diabetes, self-reported physician-diagnosed hypertension, self-reported physician-diagnosed dyslipidemia and study-specific covariates. The main model was further adjusted for BMI and WHR (model 2). In sensitivity analyses, model 1 was further adjusted for PA and dietary quality index (model 3); high-density lipoprotein (HDL)-cholesterol, low-density lipoprotein (LDL)-cholesterol and triglycerides (model 4), or systolic and diastolic blood pressures (model 5). In another sensitivity analysis, model 2 was further adjusted for eGFR in NHS, NHS2, HPFS, SOL, ARIC, WHI and PREDIMED. For each metabolite, association results from all available cohorts and racial/ethnic groups were combined using a fixed-effect, inverse-variance-weighted (IVW) meta-analysis, and a meta-analyzed FDR < 0.05 was considered statistically significant. In secondary analyses, meta-analysis was conducted combining results from the same racial/ethnic groups, or cohorts using the same platforms.

To annotate the novelty of the identified associations, we reviewed previous prospective cohort studies linking circulating metabolites to T2D risk. We used a literature-review-based meta-analysis[4] that included all studies published before 6 March 2021 as an anchor, and searched for additional studies published from 2021 to 2024[21,81–94]. We considered an association as 'previously reported,' if the association was statistically significant in a published study after multiple testing correction based on the study's prespecified analysis plan.

GWAS of metabolites

Detailed information on genotyping arrays, imputation methods, sample size and GWAS and meta-analysis methods, is provided in Supplementary Methods and Supplementary Table 7. Briefly, genotyping were conducted using several types of array by previous studies in NHS/HPFS[95], SOL[96], ARIC[7], WHI[97], CHS[98] and FHS[43]. Imputation was conducted based on the HRC reference panel in NHS/HPFS and CHS; 1000 Genomes Project phase 3 worldwide reference panel in SOL, 1000 Genomes Project phase 3 v.5 in WHI and HapMap CEU population release v.22 in FHS with comprehensive pre- and postimputation QC. GWAS of metabolites were conducted previously in the NHS/HPFS (median $n = 6{,}610$, range 971–8,054) and WHI ($n = 1{,}256$) using the RVTESTS tool[6,42,99], in SOL ($n = 3{,}933$) using a linear mixed-effect model in GMMAT[7] and in ARIC ($n = 1{,}772$ and $n = 1509$ for African American and non-Hispanic white participants, respectively)[7], CHS ($n = 263$) and FHS ($n = 1{,}802$)[43], with detailed analysis procedures described in previous publications[7,42,43].

GWAS summary statistics from each cohort were lifted over to Genome Build v.37 and filtered, retaining single nucleotide polymorphisms with a minor allele frequency ≥ 0.01 and imputation ratio ≥0.3. For each metabolite, an IVW fixed-effect meta-analysis, implemented in METAL[100], was used to combine GWAS results from the cohorts in which the metabolite was available. Genomic control was implemented before and after meta-analysis[100]. The final GWAS were available for 458 out of 469 harmonized metabolites, with the total sample size ranging from 1,074 to 18,590 (median $n = 8{,}611$). We compared significant mQTLs identified at $P < 5 \times 10^{-8}$ and 1.09×10^{-10} (that is, 5×10^{-8} further correcting for 458 metabolites) levels. Manhattan plots were derived using R package CMplot and regional plots were draw with LocusZoom[101]. In a secondary analysis, we compared genetic effect heterogeneity between racial/ethnic groups at the identified mQTLs for T2D-associated metabolites (Supplementary Methods).

We annotate the novelty of our significant mQTLs for the 165 T2D-associated metabolites at $P < 1.09 \times 10^{-10}$, by comparing our results to eight previous studies (with $N \geq 4{,}000$ and used LC–MS based metabolomic platforms)[8,24–26,102–105]. We considered a locus for a specific metabolite as 'previously reported' if the reported lead genetic variant was the same lead variant, or not the same lead variant but was significant in our study; or not in our study but within the clumping range of our identified locus. We considered a locus for a metabolite as potentially new if our locus was not previously reported for this metabolite, or this metabolite was not previously reported in these studies.

Lead variants for metabolites, pathway analysis and proportion of variance explained

We used the PLINK clumping function ($P < 5 \times 10^{-8}$ and $r^2 < 0.01$ in a 1,000-kb window) to identify independent genetic variants associated with each metabolite. For metabolite with no variant at $P < 5 \times 10^{-8}$, a single lead variant with the smallest P was selected. Gene annotation for top variants was conducted using the SNPNexus web tool[106]. Canonical pathway enrichment analyses was conducted using the MetaCore software with the default background[107]; and we compared top enriched pathways for genes annotated to mQTLs of T2D-related metabolites versus those of non-associated metabolites. We calculated the R^2 of each metabolite explained by independent lead genetic variants using the formula $\sum_{i=1}^{k} \beta \times \beta \times 2 \times \mathrm{MAF} \times (1 - \mathrm{MAF})$, in which k is the number of independent lead variants, and β is the association coefficient between the variant and the metabolite. We compared the R^2 distribution for the T2D-associated versus non-associated metabolites using Wilcoxon test.

Genetic correlation r_g between metabolites and T2D-related traits

We acquired publicly available GWAS summary statistics from large consortium studies for T2D (180,834 cases and 1,159,055 controls)[27], fasting insulin ($N = 98{,}210$)[108], proinsulin ($N = 45{,}861$)[109], HOMA-IR and HOMA-B ($N = 51{,}750$)[110], BMI-adjusted insulin sensitivity index (ISI, $N = 53{,}657$) and insulin fold-change (IFC; $N = 55{,}124$)[111], BMI and WHR ($N = {\sim}700{,}000$)[112] and lipids ($N = {\sim}1{,}500{,}000$)[113]. We conducted GWAS for HBA1c ($N = 390{,}982$), subcutaneous fat volume ($N = 37{,}912$), visceral fat volume ($N = 37{,}912$), liver proton density fat fraction (PDFF; $N = 29{,}512$), pancreas PDFF ($N = 28{,}624$) and liver enzymes ($N = {\sim}390{,}000$) in the UK Biobank using BOLT-LMM (Supplementary Methods). We calculated r_g between each metabolite and each clinical trait using linkage disequilibrium score regression, based on their GWAS summary data overlapping with the 1.2 M HapMap3 variants after excluding the major histocompatibility complex region in the European population[114]. For each trait, we compared the distribution of its r_g with T2D-associated versus non-associated metabolites, using chi-squared test, and considered FDR < 0.05 (correcting for numbers of comparisons tested) as statistically significant.

Genetic colocalization

We obtained tissue-specific *cis*-eQTLs summary statistics from the GTEx project v.8[115,116]. The shared causal variants between each metabolite and tissue-specific transcriptome from 47 tissue types, were examined using colocalization analysis implemented in the coloc.abf() function in R package 'coloc' v.5[117]. For each metabolite, we input the GWAS summary statistics for all variants within ±500 kb of its independent lead variants (Supplementary Methods). A posterior probability of H4 (PPH4) > 0.8 was considered as strong evidence for genetic colocalization. Within each tissue type, we used univariant logistic regression to test whether the proportions of mQTL–eQTL colocalizations are higher for the T2D-associated versus non-associated metabolites, and a one-sided FDR < 0.05 (correcting for 47 tissue types) was considered as statistically significant. We applied a similar coloc approach to examine genetic colocalizations between circulating metabolites and T2D[27]. We then aligned mQTL–T2D colocalizations with tissue-specific eQTL–mQTL colocalizations by metabolites and shared causal variants, to interpret the potential functionality of metabolites in T2D pathogenesis.

MR analysis

To infer the potential causal relationships between 233 T2D-associated metabolites (with genetic data) and T2D risk, we applied four MR methods implemented in the MendelianRandomization R package[118]: we used mode-based estimate (MBE) as the main method as it is generally conservative and robust to outliers; we further applied weighted-median, IVW and MR-egger to indicate result consistency[119]. When testing the direction from metabolites to T2D, we used independent variants from clumping ($P < 5 \times 10^{-8}$ and $r^2 < 0.01$ in a 1,000-kb window) excluding the *HLA* region as genetic instrumental variables. If fewer than three variants were identified, we reduced the clumping P threshold until at least three variants were identified. We considered a potential causal relationship when MBE–FDR < 0.05 and at least two other MR methods showed the same effect directions as those from MBE. Sensitivity analyses were conducted, either to remove variants mapped to the top 3 recurrent loci (*GCKR*, *ZNF259*, *FADS* cluster) from the instrumental variables, or to use only independent variants clumped at $P < 1.09 \times 10^{-10}$ as the instrumental variables of metabolites, using the IVW MR method (due to fewer variants retained). When testing the direction from T2D to metabolites, we used independent lead variants associated with T2D at $P < 5 \times 10^{-8}$ as the instrumental variables. For the 148 metabolites that are potential mediators between BMI and T2D risk, we applied MR analysis to test the direction from BMI to metabolites. Details are provided in Supplementary Methods.

MWASs for modifiable risk factors

We fitted linear models to regress inversely normal transformed metabolite levels on age, sex (only in SOL), current smoking status, BMI, PA, intakes of 15 main food groups and fasting status, simultaneously together with cohort-specific covariates. Analyses were conducted in NHS/HPFS, SOL and WHI, separately, further stratified by substudies or racial groups (Supplementary Methods). Association coefficients between metabolites and each particular risk factor were then combined across analytical sets using a fixed-effect IVW meta-analysis. The R^2 of each metabolite explained by specific risk factors were first calculated in each analytical set using the formula $\beta \times \beta \times \text{variance}(\text{risk factor})/\text{variance}(\text{metabolite})$, with the β being the association coefficients between the metabolite and the risk factor; and then averaged across all analytical sets. We compared the distributions of R^2 for T2D-associated versus non-associated metabolites using the Wilcoxon test.

Mediation analysis between risk factors, metabolites and T2D risk

Details for mediation analysis are described in Supplementary Methods. Briefly, our analysis focused on BMI, PA, coffee/tea consumption and red/processed meat intake. For each risk factor, metabolites (1) that were associated with both the risk factor and T2D risk and (2) whose association directions with the risk factor and T2D risk were consistent with the pre-assumed epidemiological relationships between the risk factor and T2D risk, were considered. We tested whether, and to what degree, each metabolite mediated the association between a risk factor and T2D risk using the CMAverse R package[120], adjusting age, sex, smoking, BMI and PA (if not the tested risk factor), calorie intake and other cohort-specific covariates, separately in NHS/HPFS, SOL and WHI. We combined total, indirect and direct effects, respectively, from each analytical set using a fixed-effect meta-analysis. The mediated proportion was calculated by dividing indirect effect to total effect. Metabolites with an indirect effect FDR < 0.05 and a consistent effect direction between the indirect and total effects, was considered as a potential mediator between a risk factor and T2D risk.

A multimetabolite signature for incident T2D prediction

We used metabolites shared between the Broad Institute and the Metabolon platforms (excluding glucose) to develop the signature to increase its generalizability to future studies. To avoid overfitting in model development and testing, we employed a leave-one-cohort-out cross-validation approach, in which we set aside one cohort as the testing set each time, and trained a prediction model for the set-aside cohort using data from all other cohorts (Extended Data Fig. 8). Given the heterogeneity of our cohorts, we did not pool individual-level data for model training. Instead, we applied a two-step approach to train the prediction model in a representable cohort (that is, WHI, which assessed the most shared metabolites for all its participants) but also leveraged association data from several other cohorts. In each iteration (that is, for each held-out testing cohort), we first conducted a metabolome-wide meta-analysis for T2D risk using all cohorts except WHI and the held-out cohort. Then, metabolites associated with T2D risk at FDR < 0.05 in the first step and shared between the two metabolomic platforms, were used as input in a Cox regression with elastic net regularization, implemented using the glmnet R package[121], to construct a metabolomic signature model for T2D prediction in WHI. The derived model was further applied to the held-out cohort to calculate a metabolomic signature score. Within WHI, a leave-one-out cross-validation approach was used to acquire the unbiased metabolomic signature score. For details, please see Supplementary Methods.

The metabolomic signature scores, calculated in each held-out cohort, were then standardized. To evaluate whether the signature improved the T2D risk prediction, we fitted three sets of logistic (in SOL, and T2D nested case–control substudy in NHS) or Cox models (all other datasets): one model including only the metabolomic signature; a conventional risk factor model including age, sex, smoking, lipid-lowering medication use, anti-hypertensive medications, family history of diabetes, hypertension, dyslipidemia and BMI; and a third model including all conventional risk factors and the metabolomic signature. We compared the AUC between the conventional model versus the conventional plus metabolomic signature model. In a secondary analysis, we further included blood glucose (from metabolomic assays) in the conventional model to evaluate the added value of the metabolomic signatures beyond blood glucose.

In each cohort, we calculated the crude incident rate of T2D across deciles of the signature score. We fitted logistic or Cox models to analyze the relative risk of T2D, comparing higher versus lowest deciles of the metabolomic signature, adjusting for the same covariates in the main analysis model 2. In NHS/HPFS, SOL and WHI, we examined associations between the metabolomic signature with baseline risk factors, by regressing the signature score on age, sex (if appropriate), current smoking status, BMI, PA, intakes of 15 main food groups and fasting status simultaneously, together with cohort-specific covariates, using linear regression. All analysis was conducted separately in each cohort, and results were combined using a meta-analysis. FDR < 0.05 was considered as statistically significant.

We conducted two sensitivity analyses during model development. One was to use SOL (measured the most metabolites using the Metabolon platform) as the representative training cohort instead of WHI, which showed a similar, albeit slightly weaker, model performance in held-out cohorts (Extended Data Fig. 8). The other was to compare between elastic net versus lasso regularizations[121], which reaffirmed that elastic net regression had compatible but a slightly better performance versus lasso regression (Supplementary Fig. 13). Separately from the leave-one-cohort-out cross-validation, we presented a final metabolomic signature model for future studies, developed using data from all study cohorts. For this model, we first conducted a metabolome-wide meta-analysis for T2D risk in all cohorts except WHI, and then used significant metabolites (FDR < 0.05) as input in a Cox regression with elastic net regularization for T2D prediction in WHI. The selected metabolites and their coefficients of this final model are highly consistent with those of models applied to each held-out cohort (Supplementary Table 18a).

Reporting summary

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

Data availability

This study did not generate any new/raw data but used existing data from several population-based cohorts, including the NHS, NHS2, HPFS, SOL, WHI, ARIC, FHS, MESA and BPRHS cohorts and the PREDIMED trial. Because of participant confidentiality and privacy concerns, these datasets are each governed by an approved data access policy, and are available upon request with formal applications submitted to the respective cohort committees, to adhere to data security and ethical considerations. Data for NHS/NHS2 (detailed policies and access procedures https://nurseshealthstudy.org/; email to nhsaccess@channing.harvard.edu) HPFS (https://www.hsph.harvard.edu/hpfs/) are available upon written request; applications to use resources will be reviewed by an External Collaborators Committee for evaluation of the fit of the data for the proposed methodology, and verification that the proposed use meets the guidelines of the Ethics and Governance Framework and the consent that was provided by the participants. HCHS/SOL has established a process for the scientific community to apply for access to participant data and materials, with requests reviewed by the SOL Steering Committee (https://sites.cscc.unc.edu/hchs/). WHI metabolomic, genomic and clinical data are available upon reasonable request to the WHI Publications and Presentations (P&P) Committee. Upon approval, requesters will be provided with details to access to the data (https://www.whi.org/propose-a-paper). Data access for FHS (detailed data policy at https://www.framinghamheartstudy.org/), MESA (https://www.mesa-nhlbi.org/), and ARIC (https://aric.cscc.unc.edu/aric9/) in the current study was approved by the TOPMed Publications and Presentations Steering Committees with data access provided by an approved project (10065). GWAS summary statistics for metabolites from NHS/HPFS (doi:10.1016/j.xcrm.2023.101085), SOL and ARIC (doi:10.1016/j.ajhg.2020.09.003) and FHS (doi: 10.1016/j.cmet.2013.06.013) were each acquired from prior publications. For the PREDIMED trial (http://www.predimed.es/), due to the restrictions imposed by the Informed Consent and the Institutional Review Board, bona fide investigators interested in analyzing the PREDIMED dataset used for the present article may submit a brief proposal and statistical analysis plan to the corresponding author. Upon approval from the PREDIMED Steering Committee and Institutional Review Boards, the data will be made available to them using an onsite secure access data enclave. BPRHS data are available upon reasonable request, and information on data request can be found at https://www.uml.edu/research/uml-cph/research/bprhs/. Source data are provided with this paper.

Code availability

The main code used to conduct this study is available on GitHub at https://github.com/JL-BWHlab/TOPMed_MWAS.

References

50. Colditz, G. A., Manson, J. E. & Hankinson, S. E. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J. Womens Health* **6**, 49–62 (1997).
51. Lavange, L. M. et al. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* **20**, 642–649 (2010).
52. Sorlie, P. D. et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* **20**, 629–641 (2010).
53. Paynter, N. P. et al. Metabolic predictors of incident coronary heart disease in women. *Circulation* **137**, 841–853 (2018).
54. Rebholz, C. M. et al. Serum metabolomic profile of incident diabetes. *Diabetologia* **61**, 1046–1054 (2018).
55. Bild, D. E. et al. Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156**, 871–881 (2002).
56. Christine, P. J. et al. Longitudinal associations between neighborhood physical and social environments and incident type 2 diabetes mellitus: the Multi-Ethnic Study of Atherosclerosis (MESA). *JAMA Intern. Med.* **175**, 1311–1320 (2015).
57. Estruch, R. et al. Primary prevention of cardiovascular disease with a Mediterranean diet supplemented with extra-virgin olive oil or nuts. *N. Engl. J. Med.* **378**, e34 (2018).
58. Kuller, L. H. et al. 10-year follow-up of subclinical cardiovascular disease and risk of coronary heart disease in the Cardiovascular Health Study. *Arch. Intern. Med.* **166**, 71–78 (2006).
59. Fried, L. P. et al. The Cardiovascular Health Study: design and rationale. *Ann. Epidemiol.* **1**, 263–276 (1991).
60. National Diabetes Data Group Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *Diabetes* **28**, 1039–1057 (1979).
61. Manson, J. E. et al. Physical activity and incidence of non-insulin-dependent diabetes mellitus in women. *Lancet* **338**, 774–778 (1991).
62. American Diabetes Association Standards of medical care in diabetes—2010. *Diabetes Care* **33**, S11–S61 (2010).
63. Chen, G. C. et al. Serum sphingolipids and incident diabetes in a US population with high diabetes burden: the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Am. J. Clin. Nutr.* **112**, 57–65 (2020).
64. Selvin, E. et al. Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. *N. Engl. J. Med.* **362**, 800–811 (2010).
65. Merino, J. et al. Metabolomics insights into early type 2 diabetes pathogenesis and detection in individuals with normal fasting glucose. *Diabetologia* **61**, 1315–1324 (2018).
66. Genuth, S. et al. Follow-up report on the diagnosis of diabetes mellitus. *Diabetes Care* **26**, 3160–3167 (2003).
67. Bertoni, A. G. et al. Inflammation and the incidence of type 2 diabetes: the Multi-Ethnic Study of Atherosclerosis (MESA). *Diabetes Care* **33**, 804–810 (2010).
68. Salas-Salvado, J. et al. Reduction in the incidence of type 2 diabetes with the Mediterranean diet: results of the PREDIMED-Reus nutrition intervention randomized trial. *Diabetes Care* **34**, 14–19 (2011).
69. Valente, M. A. et al. The Chronic Kidney Disease Epidemiology Collaboration equation outperforms the Modification of Diet in Renal Disease equation for estimating glomerular filtration rate in chronic systolic heart failure. *Eur. J. Heart Fail.* **16**, 86–94 (2014).
70. Cespedes, E. M. et al. Multiple healthful dietary patterns and type 2 diabetes in the Women's Health Initiative. *Am. J. Epidemiol.* **183**, 622–633 (2016).
71. Patterson, R. E. et al. Measurement characteristics of the Women's Health Initiative food frequency questionnaire. *Ann. Epidemiol.* **9**, 178–187 (1999).

72. Block, G. et al. A data-based approach to diet questionnaire design and testing. *Am. J. Epidemiol.* **124**, 453–469 (1986).
73. Tucker, K. L., Bianchi, L. A., Maras, J. & Bermudez, O. I. Adaptation of a food frequency questionnaire to assess diets of Puerto Rican and non-Hispanic adults. *Am. J. Epidemiol.* **148**, 507–518 (1998).
74. Siega-Riz, A. M. et al. Food-group and nutrient-density intakes by Hispanic and Latino backgrounds in the Hispanic Community Health Study/Study of Latinos. *Am. J. Clin. Nutr.* **99**, 1487–1498 (2014).
75. McCullough, M. L. & Willett, W. C. Evaluating adherence to recommended diets in adults: the Alternate Healthy Eating Index. *Public Health Nutr.* **9**, 152–157 (2006).
76. Hu, J. et al. Differences in metabolomic profiles between Black and white women and risk of coronary heart disease: an observational study of women from four US cohorts. *Circ. Res.* **131**, 601–615 (2022).
77. Zheng, Y. et al. Associations between metabolomic compounds and incident heart failure among African Americans: the ARIC Study. *Am. J. Epidemiol.* **178**, 534–542 (2013).
78. Rivas-Tumanyan, S. et al. Novel plasma metabolomic markers associated with diabetes progression in older Puerto Ricans. *Metabolites* **12**, 513 (2022).
79. Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal. Chem.* **81**, 6656–6667 (2009).
80. Barlow, W. E., Ichikawa, L., Rosner, D. & Izumi, S. Analysis of case-cohort designs. *J. Clin. Epidemiol.* **52**, 1165–1172 (1999).
81. Wang, G. et al. Gestational diabetes mellitus, postpartum lipidomic signatures, and subsequent risk of type 2 diabetes: a lipidome-wide association study. *Diabetes Care* **46**, 1223–1230 (2023).
82. Wang, F. et al. Integration of epidemiological and blood biomarker analysis links haem iron intake to increased type 2 diabetes risk. *Nat. Metab.* **6**, 1807–1818 (2024).
83. Parnell, L. D. et al. Metabolite patterns link diet, obesity, and type 2 diabetes in a Hispanic population. *Metabolomics* **17**, 88 (2021).
84. Papandreou, C. et al. Plasma trimethylamine-N-oxide and related metabolites are associated with type 2 diabetes risk in the Prevención con Dieta Mediterranea (PREDIMED) trial. *Am. J. Clin. Nutr.* **108**, 163–173 (2018).
85. Miao, G. et al. Longitudinal plasma lipidome and risk of type 2 diabetes in a large sample of American Indians with normal fasting glucose: the Strong Heart Family Study. *Diabetes Care* **44**, 2664–2672 (2021).
86. Liu, J. et al. Metabolic and genetic markers improve prediction of incident type 2 diabetes: a nested case-control study in Chinese. *J. Clin. Endocrinol. Metab.* **107**, 3120–3127 (2022).
87. Lee, K. S. et al. Association of circulating metabolites with incident type 2 diabetes in an obese population from a national cohort. *Diabetes Res Clin. Pr.* **180**, 109077 (2021).
88. Lee, D. H. et al. The metabolic potential of inflammatory and insulinaemic dietary patterns and risk of type 2 diabetes. *Diabetologia* **67**, 88–101 (2024).
89. Jiang, X. et al. Serum metabolomic profiling of incident type 2 diabetes mellitus in the Multi-ethnic Study of Atherosclerosis and Rotterdam Study. *J. Clin. Endocrinol. Metab.* **110**, e2700–e2710 (2025).
90. Haslam, D. E. et al. Discovery and validation of plasma, saliva and multi-fluid plasma-saliva metabolomic scores predicting insulin resistance and diabetes progression or regression among Puerto Rican adults. *Diabetologia* **67**, 1838–1852 (2024).
91. Floegel, A. et al. Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* **62**, 639–648 (2013).
92. Chai, J. C. et al. Serum metabolomics of incident diabetes and glycemic changes in a population with high diabetes burden: the Hispanic Community Health Study/Study of Latinos. *Diabetes* **71**, 1338–1349 (2022).
93. Bragg, F. et al. Predictive value of circulating NMR metabolic biomarkers for type 2 diabetes risk in the UK Biobank study. *BMC Med.* **20**, 159 (2022).
94. Bragg, F. et al. Circulating metabolites and the development of type 2 diabetes in Chinese adults. *Diabetes Care* **45**, 477–480 (2022).
95. Lindstrom, S. et al. A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts. *PLoS ONE* **12**, e0173997 (2017).
96. Conomos, M. P. et al. Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos. *Am. J. Hum. Genet.* **98**, 165–184 (2016).
97. Baldassari, A. R. et al. Multi-ethnic genome-wide association study of decomposed cardioelectric phenotypes illustrates strategies to identify and characterize evidence of shared genetic effects for complex traits. *Circ. Genom. Precis. Med.* **13**, e002680 (2020).
98. Roberts, J. D. et al. Genetic investigation into the differential risk of atrial fibrillation among Black and white individuals. *JAMA Cardiol.* **1**, 442–450 (2016).
99. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423–1426 (2016).
100. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
101. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
102. Yin, X. et al. Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. *Nat. Commun.* **13**, 1644 (2022).
103. Shin, S. Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
104. Schlosser, P. et al. Genetic studies of paired metabolomes reveal enzymatic and transport processes at the interface of plasma and urine. *Nat. Genet.* **55**, 995–1008 (2023).
105. Hysi, P. G. et al. Metabolome genome-wide association study identifies 74 novel genomic regions influencing plasma metabolites levels. *Metabolites* **12**, 61 (2022).
106. Oscanoa, J. et al. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res.* **48**, W185–W192 (2020).
107. Cirillo, E., Parnell, L. D. & Evelo, C. T. A review of pathway-based analysis tools that visualize genetic variants. *Front Genet* **8**, 174 (2017).
108. Lagou, V. et al. Sex-dimorphic genetic effects and novel loci for fasting glucose and insulin variability. *Nat. Commun.* **12**, 24 (2021).
109. Broadaway, K. A. et al. Loci for insulin processing and secretion provide insight into type 2 diabetes risk. *Am. J. Hum. Genet.* **110**, 284–299 (2023).
110. Manning, A. K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).

111. Williamson, A. et al. Genome-wide association study and functional characterization identifies candidate genes for insulin-stimulated glucose uptake. *Nat. Genet.* **55**, 973–983 (2023).
112. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
113. Graham, S. E. et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
114. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
115. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
116. *GTEx Portal* (Broad Institute of MIT and Harvard, accessed 1 September 2023); https://gtexportal.org/home
117. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, e1008720 (2020).
118. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J. Epidemiol.* **46**, 1734–1739 (2017).
119. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* **44**, 313–329 (2020).
120. Shi, B., Choirat, C., Coull, B. A., VanderWeele, T. J. & Valeri, L. CMAverse: a suite of functions for reproducible causal mediation analyses. *Epidemiology* **32**, e20–e22 (2021).
121. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

Acknowledgements

We thank the investigators, staff and participants involved in each cohort. The current study is supported by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) including R01DK119268 (Q.Q.) and R00DK122128 (J.L.). Other funding sources for this study include R01HL060712 (F.B.H. and Q.Q.), R01HL170904 (Q.Q.), R01HL136266 (R.C.K.) and HHSN268201600034I (contract to the Broad Institute) from the National Heart, Lung and Blood Institute (NHLBI); R01DK126698 (Q.Q.), R01DK081572 (R.E.G., C.B.C. and T.J.W.), U01DK140761 (A.C.W., J.I.R. and Q.Q.), R01DK134672 (R.C.K.), R01DK120870 (Q.Q.) and the New York Regional Center for Diabetes Translation Research (P30DK111022) from NIDDK; R01AG085320 (R.C.K.) from National Institute on Aging; and a cooperative agreement UM1HG008898 from the National Human Genome Research Institute (NHGRI). K.L. was supported by the American Heart Association (AHA) postdoctoral fellowship (23POST1020455). E.S. was supported by NHLBI K24HL152440 from and an AHA Merit Award. C.M.R. was supported by NHLBI R01HL153178. M.G.-F. is supported by a Novo Nordisk Foundation grant NNF24OC0095435. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Funding sources and acknowledgement of participating cohorts are listed in detail in Supplementary Information.

Author contributions

J.L., J.H., B.Y. and Q.Q. conceived and designed the study. J.L., M.R.-C., J.S.-S., J.D., E.S., S.N.B., A.H.E., J.E.M., J.C.F., R.N.L., K.L.T., S.S.R., J.I.R., M.A.M.-G., K.M.R., J.B.M., E.B., R.C.K., F.B.H., B.Y. and Q.Q. acquired funding and curated data. C.B.C. and R.E.G. were involved in metabolomic profiling and data generation. J.L., J.H., H.Y., Z.M., X.W., K.L., M.G.-F., X.H., B.T., J.M., C.J., C.M.R., E.H.M., T.A., G.L., J.Y., X.Z., B.C.P., T.J.W., X.G., J.A.B., Y.L., A.C.W., K.E.N., S.Y.J., C.-T.L., N.S., S.L. and L.F.T. processed data. J.L., J.H., H.Y., Z.M., X.W., K.L., X.H., B.T., J.M., C.J., C.M.R., E.H.M., T.A., G.L., J.Y., X.Z. and B.C.P. conducted analyses. J.L., J.H., L.L., F.B.H., B.Y. and Q.Q. provided feedback on analyses. J.L. and Q.Q. wrote the initial manuscript draft. J.L., J.H. and Q.Q. critically revised the paper. All authors reviewed the paper, provided important revision and feedback, and approved the final version of the manuscript.

Competing interests

S.S.R. is a consultant to Westat, the Administrative Coordinating Center for the NHLBI Trans-Omics for Precision Medicine (TOPMed) program. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-025-04105-8.

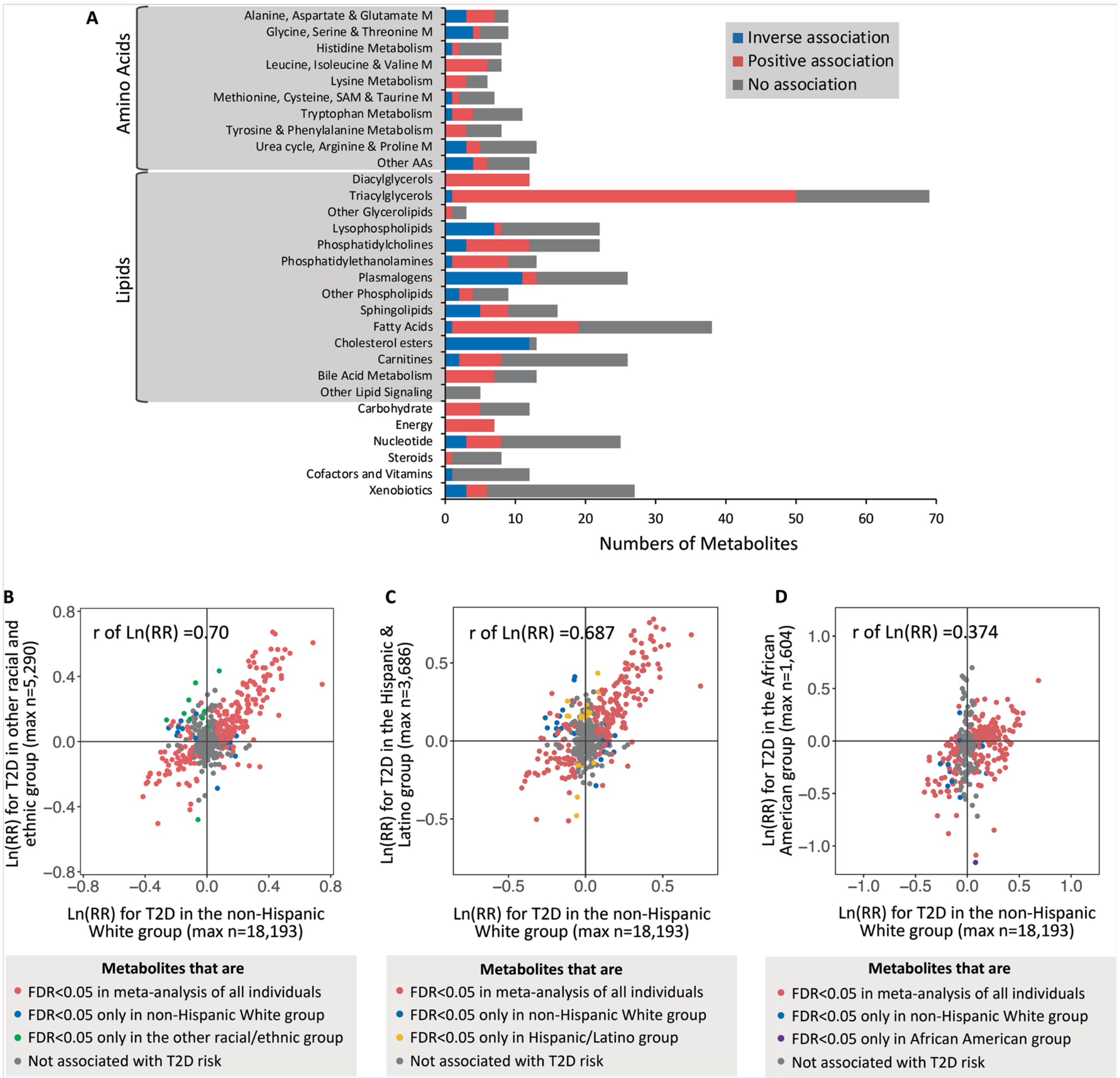
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-025-04105-8.

Correspondence and requests for materials should be addressed to Jun Li or Qibin Qi.

Peer review information *Nature Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Ashley Castellanos-Jankiewicz, in collaboration with the *Nature Medicine* team.

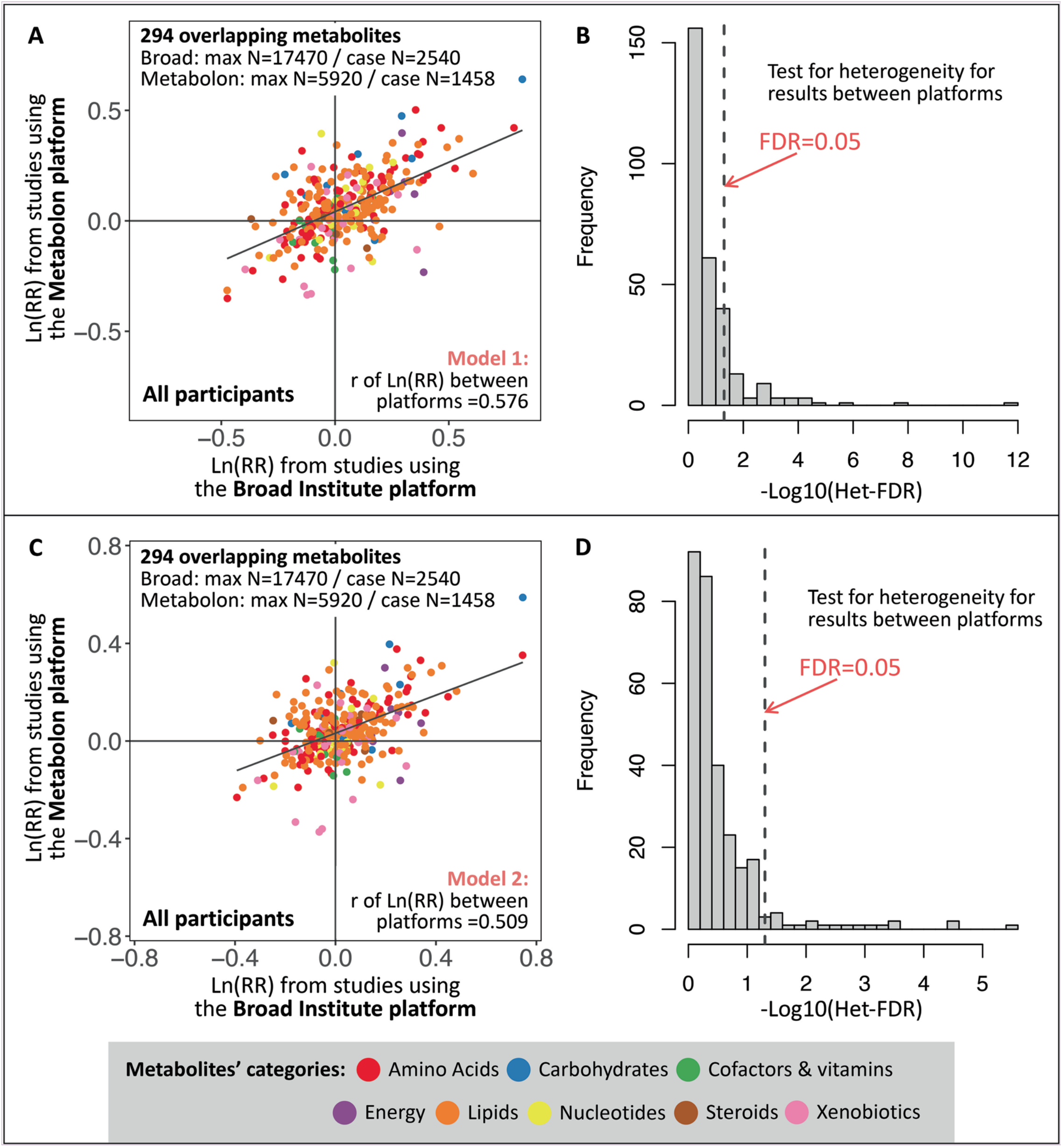
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41591-025-04105-8

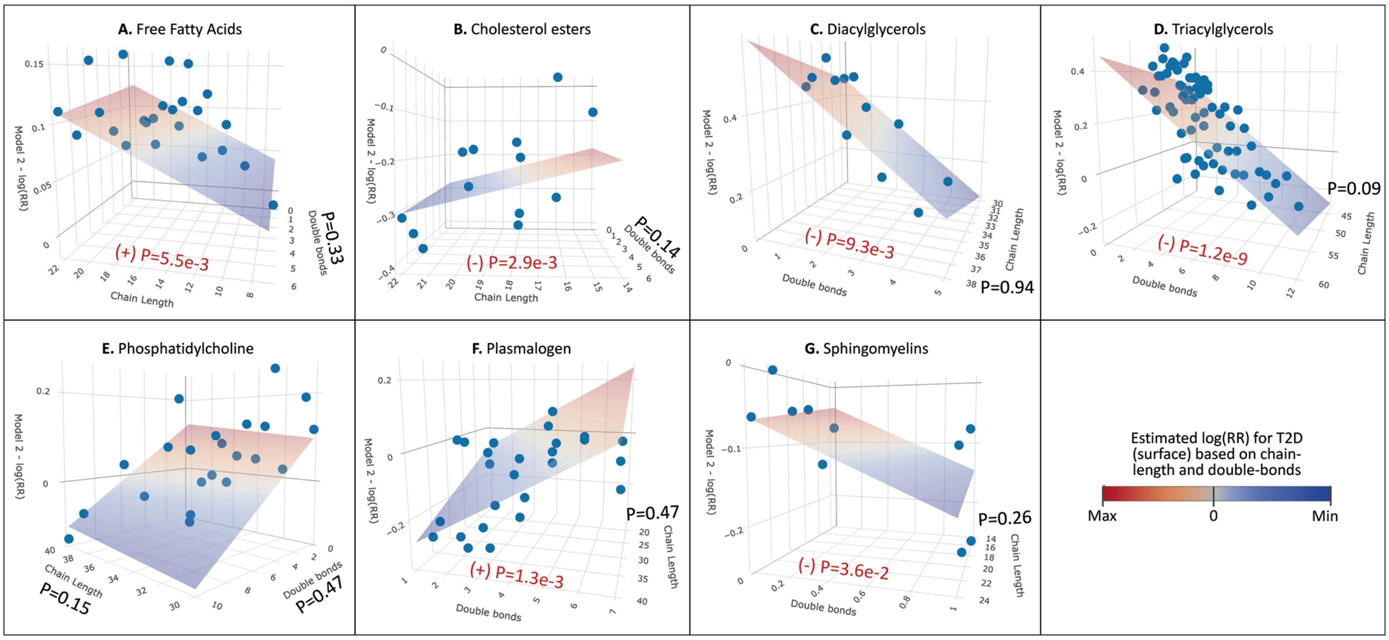


Extended Data Fig. 1 | Biochemical categories of the 469 analyzed metabolites, and their associations with incident T2D comparing non-Hispanic White individuals vs. individuals of other races and ethnicities. (**A**) Numbers of metabolites with positive, inverse, or null associations with T2D risk by biochemical category. We compared the association coefficients of each metabolite with T2D risk in the non-Hispanic White group to those from all individuals of other races and ethnics (**B**), Hispanic/Latino participants (**C**), and African American participants (**D**). Sample sizes for individual metabolites vary, depending on their availability in each cohort; the maximum sample sizes are 18,193 for non-Hispanic White individuals, 3,686 for Hispanic/Latino individuals, and 1,604 for African American individuals (see Supplementary Table S4). Association coefficients were presented as natural log of relative risk (RR) per SD increment in metabolites. In each cohort, we first conducted MWAS for incident T2D stratified by major racial/ethnic groups (that is, non-Hispanic White, African American, Hispanic/Latino, or mixed non-White individuals depending on sample size). The main model was adjusted for age, sex, smoking, alcohol consumption, fasting status, lipid-lowering mediation use, anti-hypertensive medication use, hypertension, dyslipidemia, body mass index, waist-hip ratio, family history of T2D, and other cohort-specific variables. Results presented in **A** were from meta-analysis of all participants. When comparing between racial/ethnic groups in panel **B**-**D**, we meta-analyzed the results within each group.

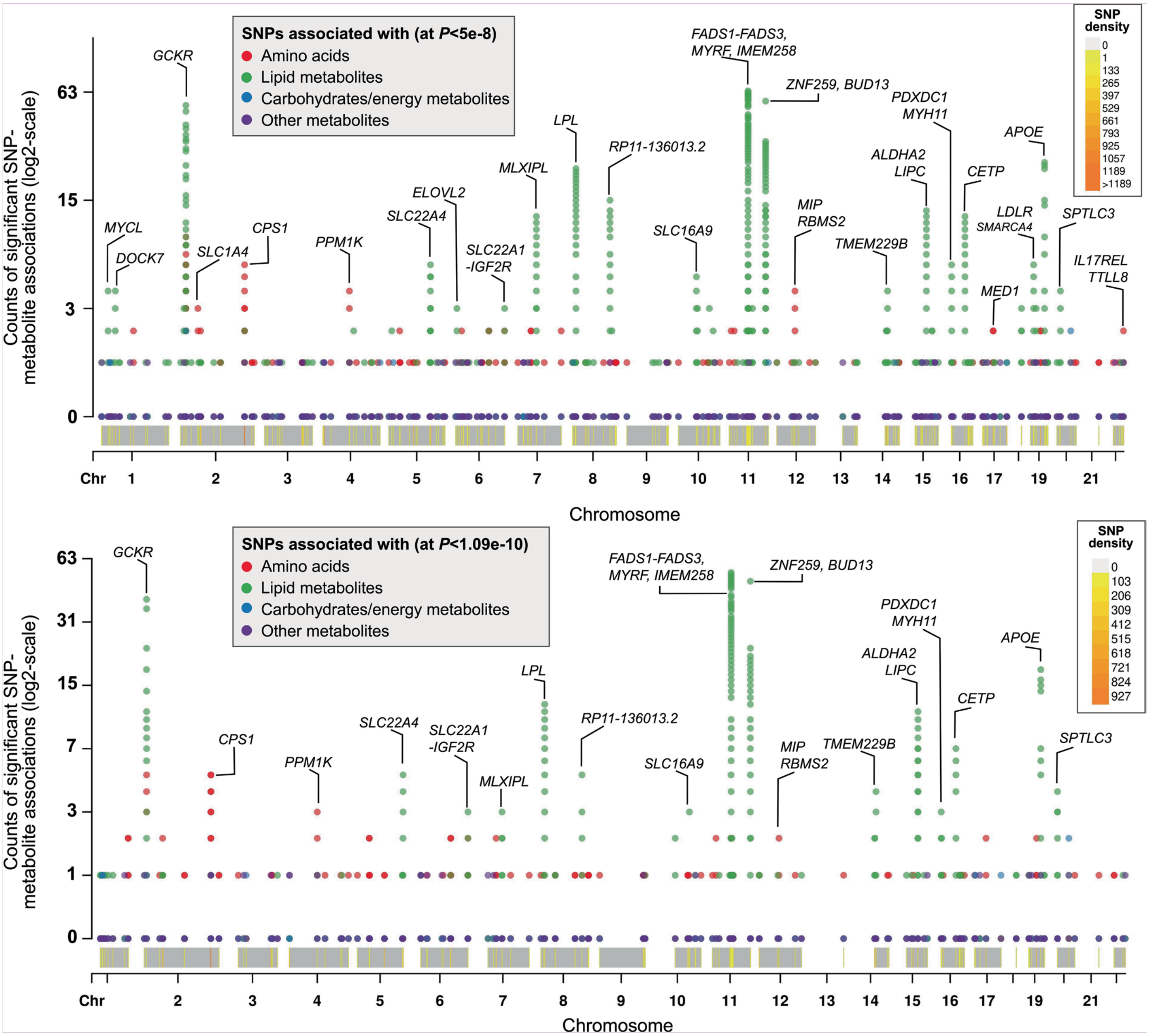
https://doi.org/10.1038/s41591-025-04105-8



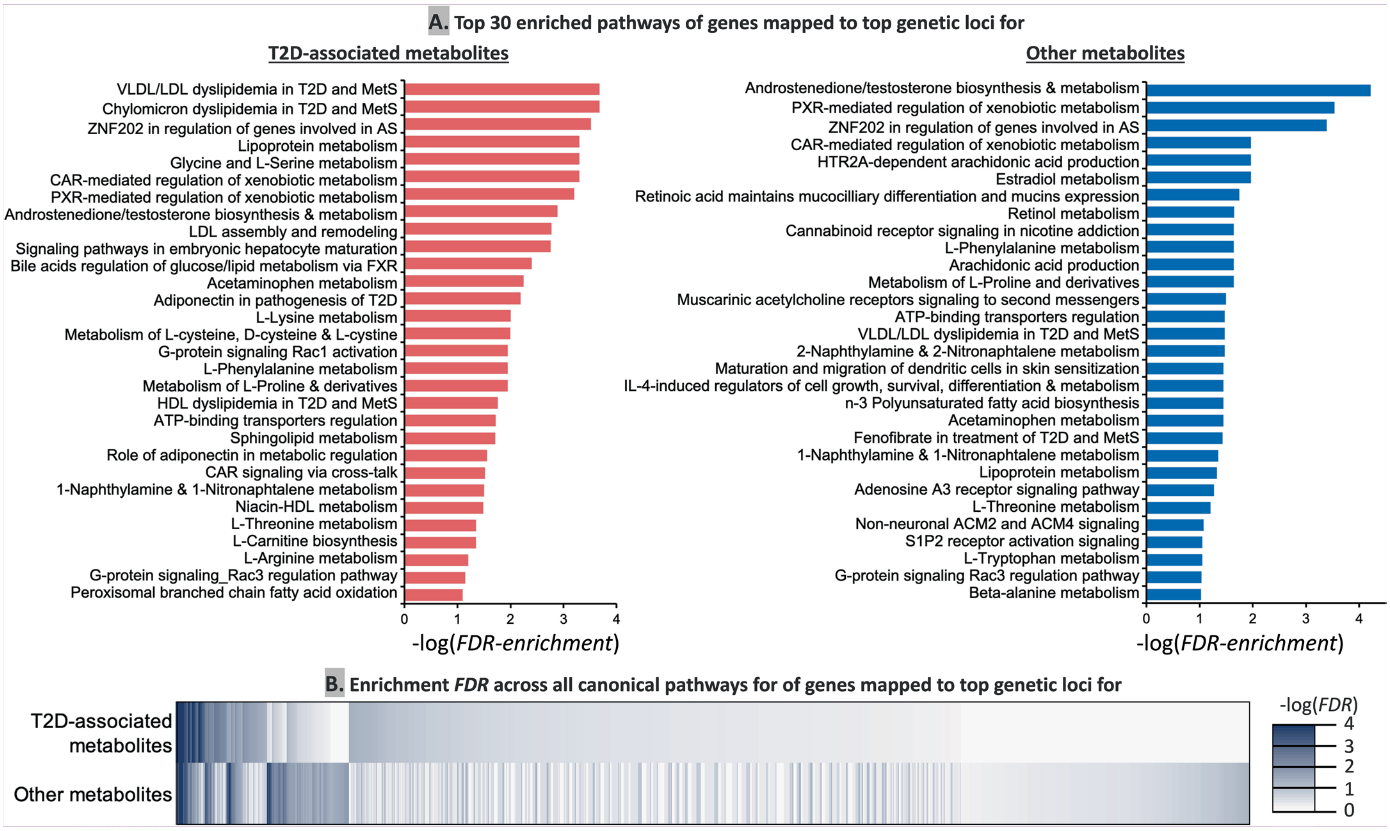
Extended Data Fig. 2 | Comparison of associations between metabolites and T2D risk across the two metabolomic platforms. In each cohort and stratified by major racial/ethnic groups, associations between inversely normal transformed metabolites and T2D risk were analyzed using Cox or logistic regressions. Results were then meta-analyzed separately for cohorts profiled at the Broad Institute *vs*. those profiled at the Metabolome Inc. A total of 294 overlapping metabolites were included in the comparison. **A** and **C** compare the association coefficients (that is, natural log-transformed relative risk ratio [RR] of T2D risk per standard deviation increase in metabolite levels) between the two platforms from Model 1 and Model 2, respectively. **B** and **D** show distributions of FDR testing for association heterogeneity between the two platforms, for Model 1 and Model 2, respectively.



Extended Data Fig. 3 | Association with T2D risk for complex lipids and fatty acids by carbon chain length and double-bond numbers. For complex lipid metabolites and fatty acids, we tested the correlation between their association coefficients (with T2D, from Model 2) with carbon chain length and double-bond numbers. Correlations with $P < 0.05$ were demonstrated, including for free fatty acids (**A**), cholesterol esters (**B**), diacylglycerols (**C**), triacylglycerols (**D**), phosphatidylcholines (**E**), plasmalogens (**F**), and sphingomyelins (**G**). In each sub-figure, x- and y-axis each represents carbon chain length and double-bond numbers, respectively; and the z-axis represents the natural log-transformed relative risk (RR) for T2D per standard deviation increase in the levels of metabolites. Significant correlations and P values were highlighted in red (+ and − indicate positive and negative correlations, respectively).

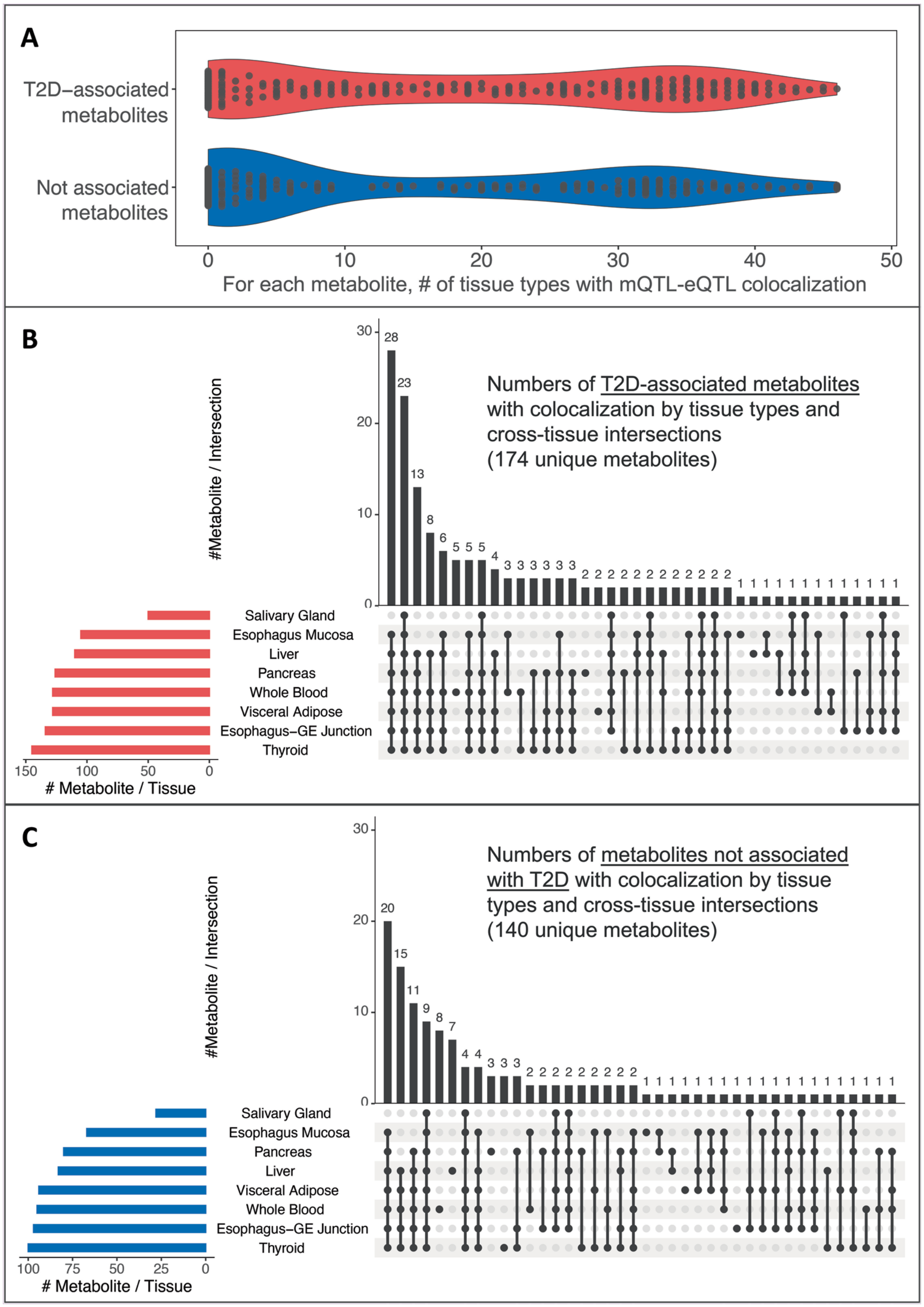


Extended Data Fig. 4 | Genetic determinants of T2D-associated metabolites. The Manhattan-like plots show significant genetic variants associated with any of the T2D-associated metabolites, at the standard genome-wide significant level ($P < 5\times10^{-8}$; upper panel) and after Bonferroni corrections for 458 metabolites with genetic data ($P < 1.09\times10^{-10}$; lower panel). The x-axis demonstrates chromosomal positions; y-axis shows the numbers of T2D-related metabolites associated with each variants; and the color depicts the major biochemical categories of the metabolite (amino acids, lipids, carbohydrates and energy metabolism, and others). Genome-association study was conducted in each of the 8 cohorts by major racial/ethnic groups, and meta-analyzed using fixed effect meta-analysis in METAL. Among the 235 T2D-associated metabolites, 233 had GWAS summary data and were included in the analyses.

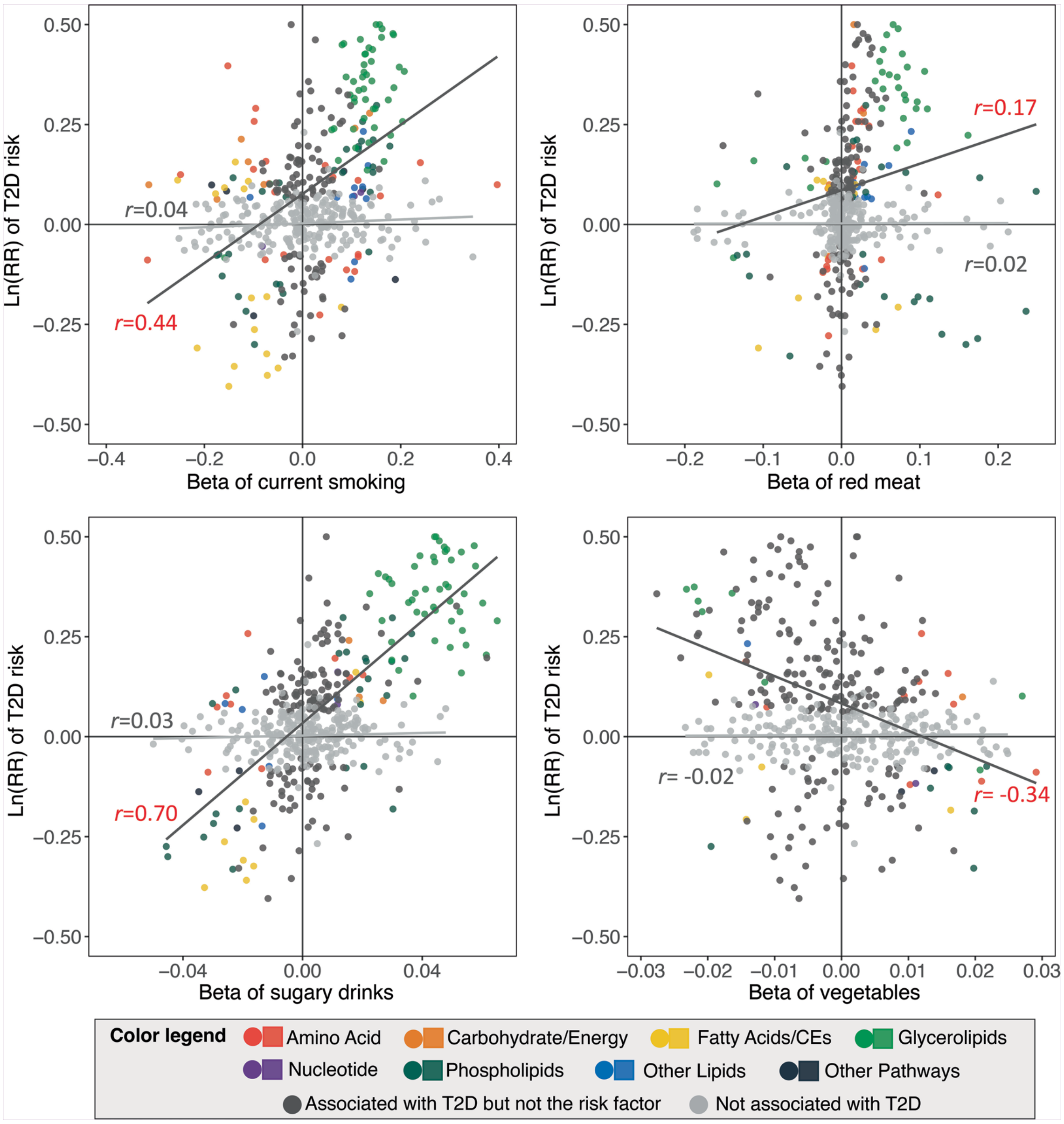


Extended Data Fig. 5 | Comparison of top enriched canonical pathways for genes mapped to mQTLs of T2D-associated metabolites vs. those mapped to mQTLs of non-associated metabolites. A. The top 30 enriched pathways identified for genes mapped to mQTLs of T2D-associated metabolites (left) *vs*. those for non-associated metabolites (right). **B**. We also observed a clear difference in the overall enrichment pattern of canonical pathways, when comparing the enrichment-*FDR* for genes mapped to mQTLs of T2D-associated metabolites *vs*. those of non-associated metabolites across all 1,140 tested canonical pathways.

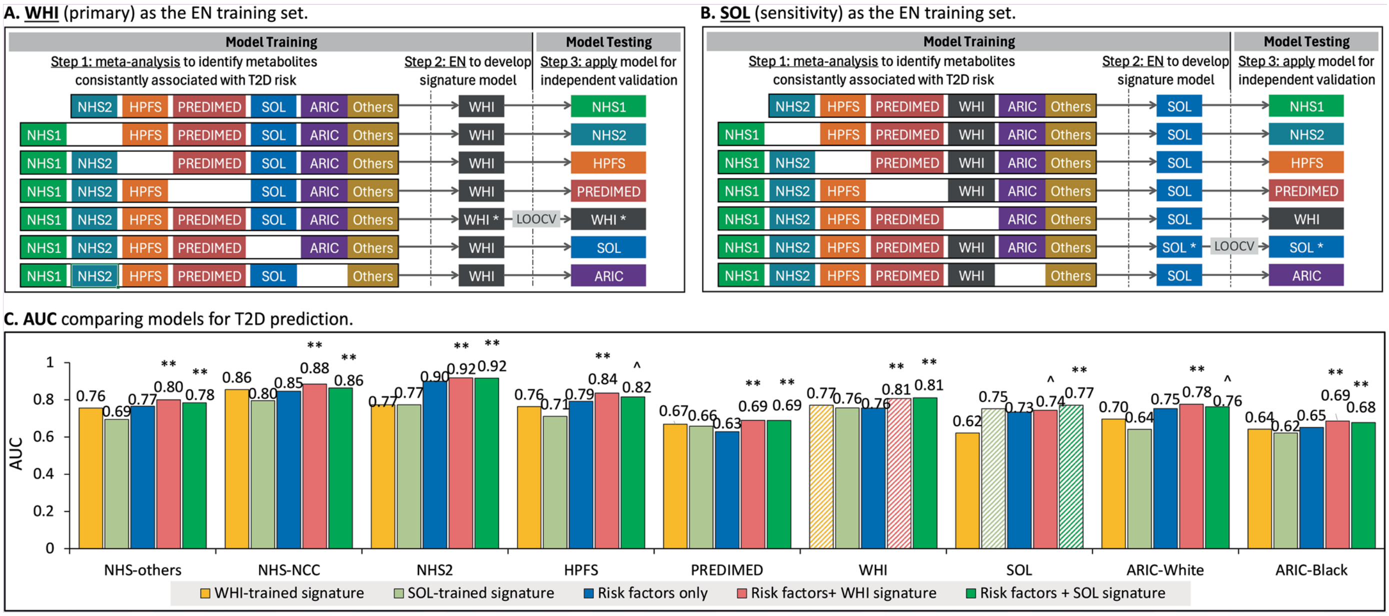
https://doi.org/10.1038/s41591-025-04105-8



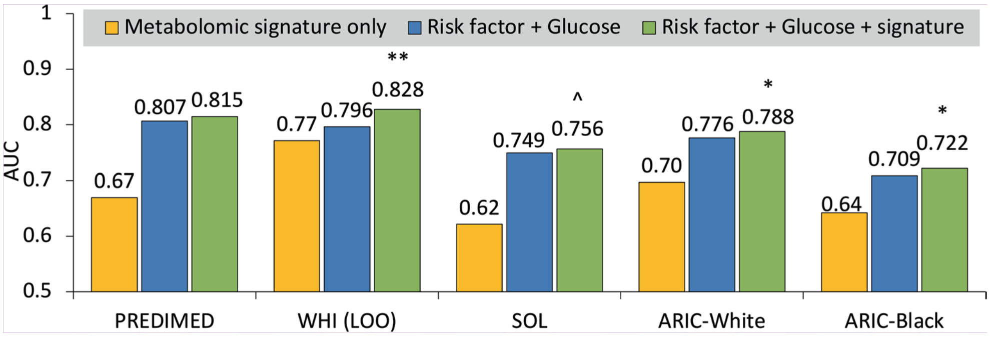
Extended Data Fig. 6 | Numbers of tissue-specific eQTL-mQTL colocalizations by metabolite's association with T2D and key tissue types. (**A**) We calculated the numbers of tissue types that each metabolite had significant mQTL-eQTL colocalizations with, and then compared numbers of colocalized tissue types across all T2D-associated metabolites *vs*. non-associated metabolites. Further, for the 8 selected tissue types (7 with significant enrichment of mQTL-eQTL colocalizations among T2D-associated metabolites plus liver), we used upset plots to depict the numbers of metabolites with mQTL-eQTL colocalizations, stratified by tissue types (left horizontal bars) and cross-tissue intersections (vertical bars), separately for T2D-associated metabolites (**B**) and non-associated metabolites (**C**).



Extended Data Fig. 7 | Associations of each circulating metabolites with baseline modifiable risk factors, and with incident T2D. Here we presented results for current smoking, red meat intake, sugary beverage intake, and vegetable intake. In the scatter plots, we compared the associations between metabolites with a risk factor *vs*. their association with incident T2D. Each dot represents a metabolite (colored: associated with the risk factor and incident T2D at *FDR* < 0.05 by biochemical category, dark grey: associated with incident T2D but not with the risk factor; light grey: not associated with incident T2D), and we presented the trend lines (and correlation coefficients) separately for T2D-associated metabolites (dark grey) and non-associated metabolites (light grey). Association coefficients (beta) for risk factors are from metabolome-wide association analysis with all risk factors mutually adjusted simultaneously (including age, sex, and BMI, physical activity, 15 major food groups, fasting status, and other cohort specific variables). This analysis was conducted separately in NHS, NHSII, HPFS, SOL, and WHI (n = 16,883) and results were combined using a meta-analysis. Association coefficients (ln[RR]) for T2D risk are from Model 2 (the main analysis model).

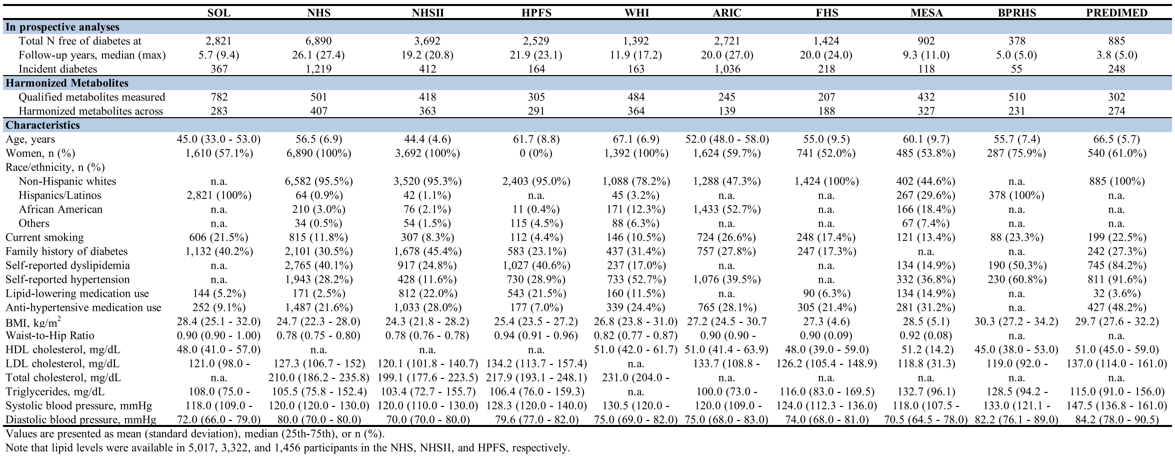


Extended Data Fig. 8 | Schematic plot and results for metabolomic signature development and testing. A. We primarily used WHI, which assessed the most metabolites shared between the two platforms for all its participants, as a representable training cohort. For each of held-out testing cohort, we first conducted a metabolome-wide meta-analysis for T2D risk including all cohorts except WHI and the held-out cohort. Metabolites associated with T2D risk at FDR < 0.05 and shared between the two platforms were then used as the input, in an elastic net Cox regression to construct a metabolomic signature model for T2D risk prediction in WHI. We next applied the derived model to the held-out cohort to calculate a metabolomic signature score. In WHI, a leave-one-out cross-validation (LOOCV) approach was used to acquire an unbiased metabolomic signature score for each individual without overfitting. **B.** We conducted a sensitivity analysis using SOL, which measured the most metabolites on the Metabolon platform for all its participants, as the training cohort. **C.** The AUC for T2D risk prediction in each cohort, comparing models with *vs*. without (blue) the metabolomic signatures, beyond traditional risk factors (age, sex, smoking, lipid-lowering medication use, anti-hypertensive medication use, family history of diabetes, hypertension, dyslipidemia, and BMI). ** Two-sided $P < 0.01$; * $P < 0.05$, ^ $P < 0.1$; slash: signature scores were calculated using LOOCV.



Extended Data Fig. 9 | Metabolomic signature for T2D prediction with the conventional model additionally adjusting for fasting glucose in cohorts with available data. We compared AUC for T2D risk prediction across three models in a secondary analysis. Model 1 (yellow) included only the metabolomic signature. Model 2 (blue) included traditional T2D risk factors, comprising age, sex, smoking, lipid-lowering medication use, anti-hypertensive medications, family history of diabetes, hypertension, dyslipidemia, and BMI, and a T2D diagnostic biomarker, blood glucose, assessed by the metabolomic assays. Model 3 (green) additionally included the metabolomic signature score on the basis of Model 2. We compared Model 3 *vs*. Model 2 to evaluate if the metabolomic signatures demonstrated added value beyond traditional risk factors and blood glucose. ** Two-sided $P < 0.01$, * $P < 0.05$, and ^ $P < 0.1$.

Extended Data Table 1 | Characteristics of study participants included in the prospective analyses



	SOL	NHS	NHSII	HPFS	WHI	ARIC	FHS	MESA	BPRHS	PREDIMED
In prospective analyses										
Total N free of diabetes at	2,821	6,890	3,692	2,529	1,392	2,721	1,424	902	378	885
Follow-up years, median (max)	5.7 (9.4)	26.1 (27.4)	19.2 (20.8)	21.9 (23.1)	11.9 (17.2)	20.0 (27.0)	20.0 (24.0)	9.3 (11.0)	5.0 (5.0)	3.8 (5.0)
Incident diabetes	367	1,219	412	164	163	1,036	218	118	55	248
Harmonized Metabolites										
Qualified metabolites measured	782	501	418	305	484	245	207	432	510	302
Harmonized metabolites across	283	407	363	291	364	139	188	327	231	274
Characteristics										
Age, years	45.0 (33.0 - 53.0)	56.5 (6.9)	44.4 (4.6)	61.7 (8.8)	67.1 (6.9)	52.0 (48.0 - 58.0)	55.0 (9.5)	60.1 (9.7)	55.7 (7.4)	66.5 (5.7)
Women, n (%)	1,610 (57.1%)	6,890 (100%)	3,692 (100%)	0 (0%)	1,392 (100%)	1,624 (59.7%)	741 (52.0%)	485 (53.8%)	287 (75.9%)	540 (61.0%)
Race/ethnicity, n (%)										
Non-Hispanic whites	n.a.	6,582 (95.5%)	3,520 (95.3%)	2,403 (95.0%)	1,088 (78.2%)	1,288 (47.3%)	1,424 (100%)	402 (44.6%)	n.a.	885 (100%)
Hispanics/Latinos	2,821 (100%)	64 (0.9%)	42 (1.1%)	n.a.	45 (3.2%)	n.a.	n.a.	267 (29.6%)	378 (100%)	n.a.
African American	n.a.	210 (3.0%)	76 (2.1%)	11 (0.4%)	171 (12.3%)	1,433 (52.7%)	n.a.	166 (18.4%)	n.a.	n.a.
Others	n.a.	34 (0.5%)	54 (1.5%)	115 (4.5%)	88 (6.3%)	n.a.	n.a.	67 (7.4%)	n.a.	n.a.
Current smoking	606 (21.5%)	815 (11.8%)	307 (8.3%)	112 (4.4%)	146 (10.5%)	724 (26.6%)	248 (17.4%)	121 (13.4%)	88 (23.3%)	199 (22.5%)
Family history of diabetes	1,132 (40.2%)	2,101 (30.5%)	1,678 (45.4%)	583 (23.1%)	437 (31.4%)	757 (27.8%)	247 (17.3%)	n.a.	n.a.	242 (27.3%)
Self-reported dyslipidemia	n.a.	2,765 (40.1%)	917 (24.8%)	1,027 (40.6%)	237 (17.0%)	n.a.	n.a.	134 (14.9%)	190 (50.3%)	745 (84.2%)
Self-reported hypertension	n.a.	1,943 (28.2%)	428 (11.6%)	730 (28.9%)	733 (52.7%)	1,076 (39.5%)	n.a.	332 (36.8%)	230 (60.8%)	811 (91.6%)
Lipid-lowering medication use	144 (5.2%)	171 (2.5%)	812 (22.0%)	543 (21.5%)	160 (11.5%)	n.a.	90 (6.3%)	134 (14.9%)	n.a.	32 (3.6%)
Anti-hypertensive medication use	252 (9.1%)	1,487 (21.6%)	1,033 (28.0%)	177 (7.0%)	339 (24.4%)	765 (28.1%)	305 (21.4%)	281 (31.2%)	n.a.	427 (48.2%)
BMI, kg/m^2	28.4 (25.1 - 32.0)	24.7 (22.3 - 28.0)	24.3 (21.8 - 28.2)	25.4 (23.5 - 27.2)	26.8 (23.8 - 31.0)	27.2 (24.5 - 30.7	27.3 (4.6)	28.5 (5.1)	30.3 (27.2 - 34.2)	29.7 (27.6 - 32.2)
Waist-to-Hip Ratio	0.90 (0.90 - 1.00)	0.78 (0.75 - 0.80)	0.78 (0.76 - 0.78)	0.94 (0.91 - 0.96)	0.82 (0.77 - 0.87)	0.90 (0.90 -	0.90 (0.09)	0.92 (0.08)	n.a.	n.a.
HDL cholesterol, mg/dL	48.0 (41.0 - 57.0)	n.a.	n.a.	n.a.	51.0 (42.0 - 61.7)	51.0 (41.4 - 63.9)	48.0 (39.0 - 59.0)	51.2 (14.2)	45.0 (38.0 - 53.0)	51.0 (45.0 - 59.0)
LDL cholesterol, mg/dL	121.0 (98.0 -	127.3 (106.7 - 152)	120.1 (101.8 - 140.7)	134.2 (113.7 - 157.4)	n.a.	133.7 (108.8 -	126.2 (105.4 - 148.9)	118.8 (31.3)	119.0 (92.0 -	137.0 (114.0 - 161.0)
Total cholesterol, mg/dL	n.a.	210.0 (186.2 - 235.8)	199.1 (177.6 - 223.5)	217.9 (193.1 - 248.1)	231.0 (204.0 -	n.a.	n.a.	n.a.	n.a.	n.a.
Triglycerides, mg/dL	108.0 (75.0 -	105.5 (75.8 - 152.4)	103.4 (72.7 - 155.7)	106.4 (76.0 - 159.3)	n.a.	100.0 (73.0 -	116.0 (83.0 - 169.5)	132.7 (96.1)	128.5 (94.2 -	115.0 (91.0 - 156.0)
Systolic blood pressure, mmHg	118.0 (109.0 -	120.0 (120.0 - 130.0)	120.0 (110.0 - 130.0)	128.3 (120.0 - 140.0)	130.5 (120.0 -	120.0 (109.0 -	124.0 (112.3 - 136.0)	118.0 (107.5 -	133.0 (121.1 -	147.5 (136.8 - 161.0)
Diastolic blood pressure, mmHg	72.0 (66.0 - 79.0)	80.0 (70.0 - 80.0)	70.0 (70.0 - 80.0)	79.6 (77.0 - 82.0)	75.0 (69.0 - 82.0)	75.0 (68.0 - 83.0)	74.0 (68.0 - 81.0)	70.5 (64.5 - 78.0)	82.2 (76.1 - 89.0)	84.2 (78.0 - 90.5)

Values are presented as mean (standard deviation), median (25th-75th), or n (%).

Note that lipid levels were available in 5,017, 3,322, and 1,456 participants in the NHS, NHSII, and HPFS, respectively.