naturemedicine

Article

https://doi.org/10.1038/s41591-022-02014-8

Influence of the microbiome, diet and genetics on inter-individual variation in the human plasma metabolome

Received: 5 July 2021

Accepted: 15 August 2022

Published online: 10 October 2022

Check for updates

Lianmin Chen[1,2,3,4], Daria V. Zhernakova[1,5], Alexander Kurilshikov[1], Sergio Andreu-Sánchez[1,2], Daoming Wang[1], Hannah E. Augustijn[1,6], Arnau Vich Vila[1,7], Lifelines Cohort Study*, Rinse K. Weersma[7], Marnix H. Medema[6], Mihai G. Netea[8,9], Folkert Kuipers[2,10], Cisca Wijmenga[1], Alexandra Zhernakova[1] and Jingyuan Fu[1,2] ✉

The levels of the thousands of metabolites in the human plasma metabolome are strongly influenced by an individual's genetics and the composition of their diet and gut microbiome. Here, by assessing 1,183 plasma metabolites in 1,368 extensively phenotyped individuals from the Lifelines DEEP and Genome of the Netherlands cohorts, we quantified the proportion of inter-individual variation in the plasma metabolome explained by different factors, characterizing 610, 85 and 38 metabolites as dominantly associated with diet, the gut microbiome and genetics, respectively. Moreover, a diet quality score derived from metabolite levels was significantly associated with diet quality, as assessed by a detailed food frequency questionnaire. Through Mendelian randomization and mediation analyses, we revealed putative causal relationships between diet, the gut microbiome and metabolites. For example, Mendelian randomization analyses support a potential causal effect of *Eubacterium rectale* in decreasing plasma levels of hydrogen sulfite—a toxin that affects cardiovascular function. Lastly, based on analysis of the plasma metabolome of 311 individuals at two time points separated by 4 years, we observed a positive correlation between the stability of metabolite levels and the amount of variance in the levels of that metabolite that could be explained in our analysis. Altogether, characterization of factors that explain inter-individual variation in the plasma metabolome can help design approaches for modulating diet or the gut microbiome to shape a healthy metabolome.

The plasma metabolome represents a functional readout of metabolic activities within different organs and tissues of the body. Levels of specific plasma metabolites may therefore reflect the presence of specific diseases or an individual's susceptibility to developing complex metabolic diseases such as cardiovascular and kidney disorders, diabetes, cancers and Crohn's disease[1]. Elucidating the genetic, dietary and microbial factors that shape human metabolism is crucial for understanding the origin and determinants of plasma metabolites, and hence for the eventual design of intervention strategies aimed at a healthy metabolome.

Inter-individual variations in the human plasma metabolome have already been linked to genetics, diet and the gut microbiome in several cohort-based studies[1–4]. For instance, a reference map of potential determinants of the human serum metabolome was established in

A full list of affiliations appears at the end of the paper. *A list of members and their affiliations appears in the Supplementary information.
✉e-mail: j.fu@umcg.nl

491 individuals from an Israeli cohort, and the authors reported 335 metabolites that were significantly explained by diet and 182 that were explained by the gut microbiome[3]. More recently, the Personalized Responses to Dietary Composition Trial assessed the impact of diet and the microbiome on host metabolism in 1,098 individuals from the United Kingdom and United States and observed that the microbial species associated with healthy dietary habits overlapped with those associated with favorable cardiometabolic and postprandial markers[4].

As diet, genetics and the gut microbiome are highly heterogeneous between different countries, we aimed to: (1) systematically identify dietary, genetic and microbial factors that are associated with plasma metabolites; (2) identify which of the three factors (diet, genetics or the microbiome) explains the most inter-individual variability in metabolites compared with the other two (the dominant factor); and (3) assess their causal relationships using in silico approaches. To do so, we quantified the plasma levels of 1,183 metabolites in 1,368 individuals from the population-based Lifelines DEEP (LLD)[5] and Genome of the Netherlands (GoNL)[6] cohorts, including LLD_1 ($n = 1{,}054$), LLD_2 ($n = 237$) and GoNL ($n = 77$). In addition, 311 LLD_1 individuals were followed up after 4 years[7]. For each participant, we had information on the gut microbiome, genetic background and dietary habits. In addition, we assessed whether diet-associated metabolites can be used to predict an individual's dietary quality score reflecting diet–disease relationships[8] and examined whether genetics-associated metabolites can pinpoint dysregulated molecular pathways in complex diseases. Importantly, as potential causal relationships among metabolites, diet and the microbiome remain largely unexplored, metabolites associated with multiple factors offered us an opportunity to infer their underlying causality using Mendelian randomization (MR) and mediation analyses[9].

Results

Untargeted plasma metabolites in Dutch cohorts

In this study, we examined plasma metabolomes in 1,679 fasting plasma samples from 1,368 individuals from two LLD[5] sub-cohorts (LLD_1 and LLD_2) and the GoNL[6] cohort (Extended Data Fig. 1 and Supplementary Table 1). The LLD_1 cohort was the discovery cohort, with information about genetics, diet and the gut microbiome available for 1,054 participants. Moreover, 311 LLD_1 subjects were followed up 4 years later (LLD_1 follow-up). We also included two independent replication cohorts: 237 LLD_2 participants for whom we had genetic and dietary data and 77 GoNL participants for whom only genetic data were available (Extended Data Fig. 1 and Supplementary Table 1). Untargeted metabolomics profiling was done using flow-injection time-of-flight mass spectrometry (FI-MS)[10,11], which yielded plasma levels of 1,183 metabolites (Supplementary Table 2). These metabolites covered a wide range of lipids, organic acids, phenylpropanoids, benzenoids and other metabolites (Extended Data Fig. 2a). As we observed weak (absolute $r_{Spearman} < 0.2$) correlations among the 1,183 metabolites (Extended Data Fig. 2b), data reduction was not required and, consequently, all metabolites were subjected to subsequent analyses. We validated the identification and quantification of some metabolites (for example, bile acids, creatinine, lactate, phenylalanine and isoleucine) by comparing their abundance levels from FI-MS with those previously determined by liquid chromatography with tandem mass spectrometry (LC-MS/MS)[12] or NMR[13] ($r_{Spearman} > 0.62$; Extended Data Fig. 2c,d).

Factors explaining inter-individual metabolome variations

To compare the relative importance of diet, genetics and the gut microbiome in explaining inter-individual plasma metabolome variability, we calculated the proportion of variance explained by these three factors for the whole plasma metabolome profile and for the individual metabolites separately. We have detailed information on 78 dietary habits (Supplementary Table 3), 5.3 million human genetic variants and the abundances of 156 species and 343 MetaCyc pathways for each individual of the LLD_1 cohort. Diet, genetics and the gut microbiome could explain 9.3, 3.3 and 12.8%, respectively, of inter-individual variations in the whole plasma metabolome, without adjusting for covariates (see the Methods section 'Distance matrix-based variance estimation'; false discovery rate (FDR) < 0.05; Fig. 1a and Supplementary Table 4), whereas intrinsic factors (age, sex and body mass index (BMI)) and smoking collectively explained 4.9% of the variance. Together, these factors explain 25.1% of the variance in the plasma metabolome (Fig. 1a).

Next, we tested for pairwise associations between each metabolite and the dietary variables, genetic variants and microbial taxa. We observed 2,854 associations with dietary habits (Supplementary Table 5), 48 associations with 40 unique genetic variants (metabolite quantitative trait loci (mQTLs); Supplementary Table 6), 1,373 associations with gut bacterial species (Supplementary Table 7) and 2,839 associations with bacterial MetaCyc pathways (Supplementary Table 8) (see the Methods sections 'Associations with dietary habits', 'QTL mapping' and 'Microbiome-wide associations'). In total, 769 metabolites were significantly associated with at least one factor (Fig. 1b and Supplementary Tables 5–8). We then performed interaction analysis to assess the role of diet–microbiome, genetics–microbiome and diet–genetics interactions in regulating the human metabolome using an interaction term in the linear model (see the Methods section 'Interaction analysis'). Among these, 185 metabolites were associated with multiple factors and seven were affected by either genetics–microbiome, genetics–diet or diet–microbiome interactions (Supplementary Table 9).

As interactions were limited, we further assessed the proportion of variance of each metabolite that was explained by these factors using an additive model with the least absolute shrinkage and selection operator (lasso) method (see the Methods section 'Estimating the variance of individual metabolites'). In general, the inter-individual variations in 733 metabolites could be explained by at least one of the three factors ($FDR_{F\text{-test}} < 0.05$; Supplementary Table 10). In detail, dietary habits contributed 0.4–35% of the variance in 684 metabolites; microbial abundances contributed 0.7–25% of the variance in 193 metabolites; and genetic variants contributed 3–28% of the variance in 44 metabolites (adjusted r^2; $FDR_{F\text{-test}} < 0.05$; Supplementary Table 10). We also estimated the explained variance of metabolites using Elastic Net[14], which is designed for highly correlated features, and found that the estimated explained variances were comparable between linear regression and the Elastic Net regression (Supplementary Fig. 1).

We further compared the variance explained by each type of factor (diet, genetics or the microbiome) and assigned the dominant factor for each metabolite if one factor explained more variance than the other two. Inter-individual variations in 610 metabolites were mostly explained by diet, 85 were explained by the gut microbiome and 38 were explained by genetics (Supplementary Table 10). Hereafter, we refer to these as diet-dominant, microbiome-dominant and genetics-dominant metabolites, respectively. The dominant factors of metabolites highlight their origin. For instance, ten out of the 21 diet-dominant metabolites for which diet explained >20% of the variance ($FDR_{F\text{-test}} < 0.05$; Supplementary Table 10) were food components based on their annotation in the Human Metabolome Database (HMDB)[15]. Similarly, of the 85 microbiome-dominant metabolites, 23 were annotated in the HMDB as microbiome-related metabolites (including 15 uremic toxins). Furthermore, out of the 38 genetics-dominant metabolites, ten were lipid species and eight were amino acids. Taken together, our analysis highlights that one factor—either dietary, genetic or microbial—can have a dominant effect over the other two in explaining the variances of plasma metabolites, with diet or the microbiome being particularly dominant. However, we also found that the variances in 185 metabolites were significantly attributable to more than one factor (Supplementary Table 10), including six metabolites associated with both genetics and the microbiome and 153 metabolites associated with both diet and the microbiome. For example, genetics and the microbiome explained 4 and 5%, respectively, of the variance in

https://doi.org/10.1038/s41591-022-02014-8

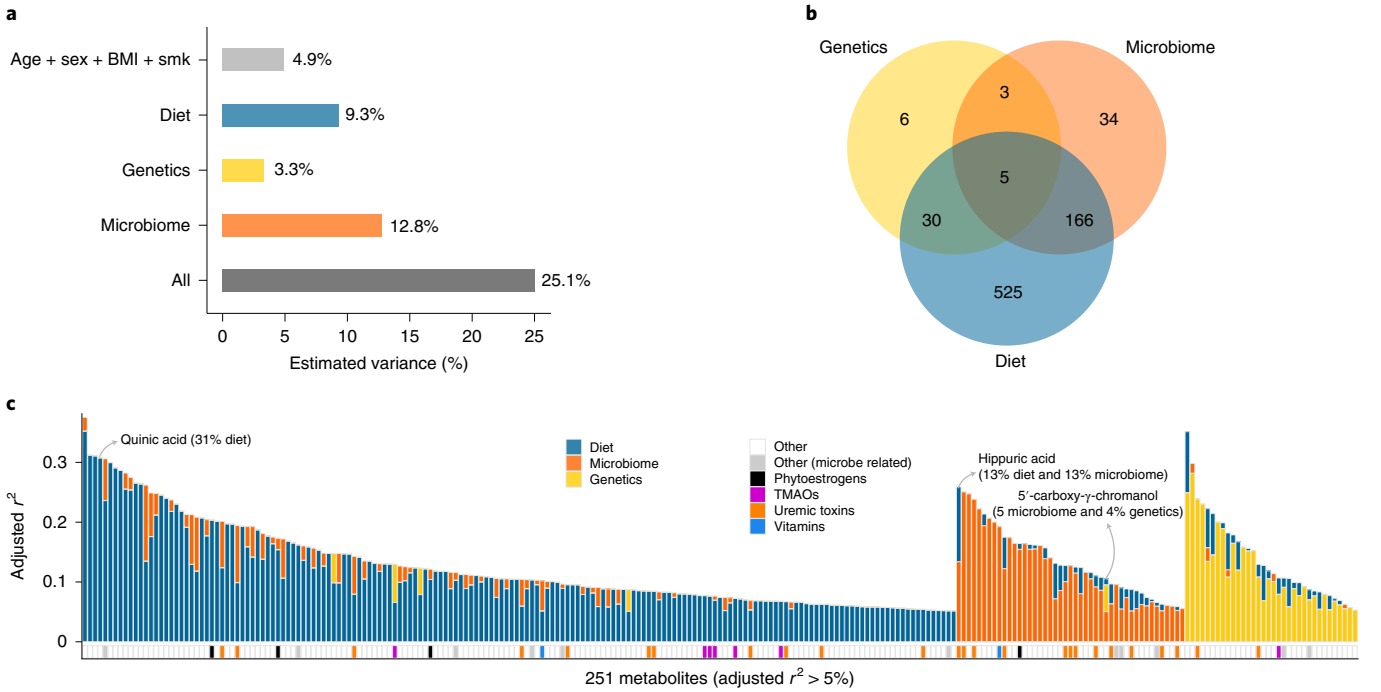


Fig. 1 | Contributions of genetics, diet and the microbiome to inter-individual variation in the plasma metabolome. a, Inter-individual variation in the whole plasma metabolome explained by the indicated factors, estimated using the PERMANOVA method. All, all of the indicated factors combined; smk, smoking status. **b**, Venn diagram indicating the number of metabolites whose inter-individual variation was significantly explained by diet, genetics or the gut microbiome, as estimated using the linear regression method ($FDR_{F\text{-test}}$ < 0.05). **c**, Inter-individual variations in metabolites explained by diet, genetics or the gut microbiome, as estimated using the linear regression method (the lasso regression method was applied for feature selection) with a significant estimated adjusted $r^2 > 5\%$ ($FDR_{F\text{-test}} < 0.05$). The blue bars represent dietary contributions to metabolite variations, the yellow bars indicate genetic contributions and the orange bars indicate microbial contributions. The other colors indicate the metabolic categories of metabolites (see legend). The *y* axis indicates the proportion of variation explained. TMAO, trimethylamine *N*-oxide.

plasma 5′-carboxy-γ-chromanol (Fig. 1c)—a dehydrogenated carboxylate product of 5′-hydroxy-γ-tocopherol[16] that may reduce cancer and cardiovascular risk[17]. Another example is hippuric acid—a uremic toxin that can be produced by bacterial conversion of dietary proteins[18], with 13% of its variance explained by diet and 13% explained by the microbiome (Fig. 1c).

Temporal variability of the metabolites over time

Temporal changes in plasma metabolites can reflect changes in an individual's diet, gut microbiome and health status. When assessing the plasma metabolome in the 311 LLD_1 follow-up samples, we indeed observed a significant shift in the plasma metabolome, with a significant difference in the second principal component ($P_{PC1\ paired\ Wilcoxon} = 0.1$ and $P_{PC2\ paired\ Wilcoxon} = 1.3 \times 10^{-5}$; Fig. 2a). Baseline genetics, diet and microbiome, together with age, sex and BMI, could explain 59.4% of the variance in the follow-up plasma metabolome ($P_{PERMANOVA} = 0.004$) (Supplementary Fig. 2). We also observed that temporal stability can vary substantially between different metabolites (see the Methods section 'Temporal consistency of individual metabolites'; Supplementary Table 11). Previously, we had assessed the changes in the gut microbiome in the LLD_1 follow-up cohort and linked these to changes in the plasma metabolome[7]. Here, we further checked the temporal variability of the plasma metabolome and assessed the stability of diet-, microbiome- and genetics-dominant metabolites over time. Interestingly, the temporal correlation of the microbiome-dominant metabolites was similar to that of the genetics-dominant metabolites ($P_{Wilcoxon} = 0.51$; Fig. 2b), whereas the temporal correlation between diet-dominant metabolites was significantly lower than between microbiome- and genetics-dominant metabolites ($P_{Wilcoxon} < 3.4 \times 10^{-5}$; Fig. 2b). However, the dominant dietary, microbial and genetic factors identified at baseline also explained similar variance in metabolic levels in the follow-up samples (Extended Data Fig. 3 and Supplementary Table 10). Our data also revealed a positive correlation between stability and the amount of variance that could be explained: the more variance explained, the more stable a metabolite is over time (Fig. 2c). For a few metabolites, we could not replicate the variance explained at baseline at the second time point, and these metabolites also showed weak or no correlation in their abundances between the two time points. For example, *N*-acetylgalactosamine showed very weak correlation between the two time points ($r = 0.13$; $P = 0.02$), and its genetic association was not replicated at the second time point.

Having established the variances in metabolites explained by diet, genetics and the gut microbiome and the dominant factors that explained most of this variance, we focused on detailing specific associations and on the potential implications of our findings for assessing diet quality and improving our understanding of the genetic risk of complex diseases and the interaction and causality relationships among diet, the microbiome, genetics and metabolism.

The metabolome reflects the diet quality score

We observed 2,854 significant associations ($FDR_{Spearman} < 0.05$) between 74 dietary factors and 726 metabolites (Fig. 3a and Supplementary Table 5; see the Methods section 'Lifelines diet quality score prediction'). Associations with food-specific metabolites can, in theory, be used to verify food questionnaire data. For instance, the strongest association we observed was between quinic acid levels and coffee intake ($r_{Spearman} = 0.54$; $P = 1.6 \times 10^{-80}$; Fig. 3b). Quinic acid is found in a wide variety of different plants but has a particularly high concentration in coffee. Another example is 2,6-dimethoxy-4-propylphenol, which was strongly associated with fish intake ($r_{Spearman} = 0.53$; $P = 1.5 \times 10^{-76}$; Fig. 3c). This association is expected as this compound is particularly present in smoked fish according to HMDB annotation[15]. In addition,

https://doi.org/10.1038/s41591-022-02014-8

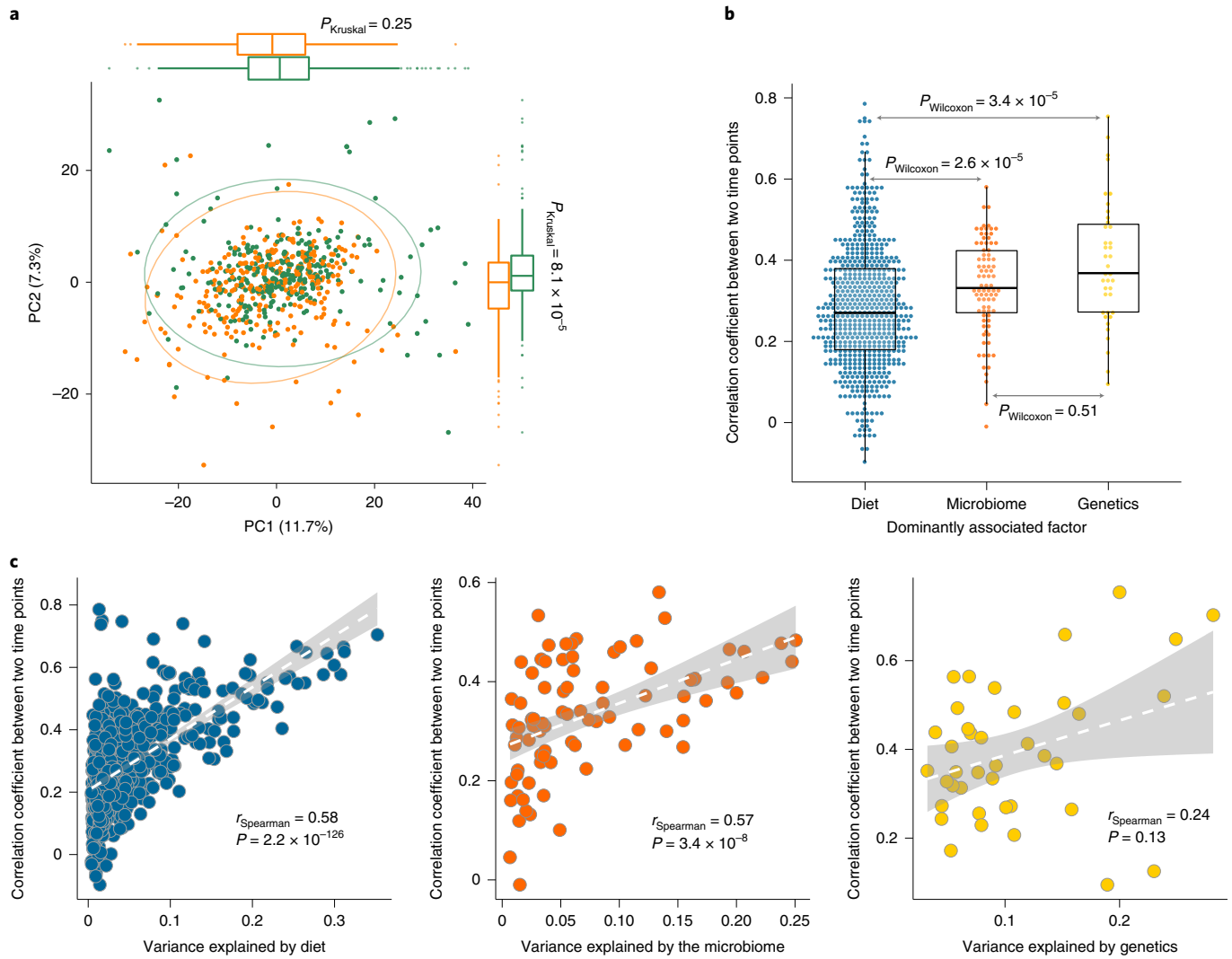


Fig. 2 | Temporal stability of plasma metabolites. a, Principal component analysis of metabolite levels at two time points (Euclidean dissimilarity). The green dots indicate baseline samples and the orange dots indicate follow-up samples (n = 311 biologically independent samples). The Kruskal–Wallis test (two sided) was used to check differences between baseline and follow-up. **b**, Temporal stability of metabolites stratified by the dominantly associated factor for each metabolite. The Wilcoxon test (two sided) was used to check the differences between groups. Each dot represents one metabolite. The y axis indicates the Spearman correlation coefficient of abundances of each metabolite between two time points (n = 311 biologically independent samples). In **a** and **b**, the box plots show the median and first and third quartiles (25th and 75th percentiles) of the first and second principal components (**a**) or correlation coefficients (**b**); the upper and lower whiskers extend to the largest and smallest value no further than 1.5× the interquartile range (IQR), respectively; and outliers are plotted individually. **c**, Correlation between metabolite stability and the metabolite variance explained by diet (left), genetics (middle) and the microbiome (right). The x axis indicates the inter-individual variation explained by each factor and the y axis indicates the Spearman correlation coefficient (two sided) of abundances of each metabolite between the two time points. The dashed white lines show the best fit and the gray shading represents the 95% confidence interval (CI) (n = 311 biologically independent samples).

we also detected associations between dietary factors and metabolic biomarkers of some diseases. For example, 1-methylhistidine is a biomarker for cardiometabolic diseases including heart failure[19] that is enriched in meat, and we observed significant associations between 1-methylhistidine and meat ($r_{Spearman}$ = 0.12; P = 7.2 × 10^{-5}) and fish intake ($r_{Spearman}$ = 0.11; P = 3.1 × 10^{-4}) as well as a lower level of 1-methylhistidine in vegetarians ($r_{Spearman}$ = −0.15; P = 9.7 × 10^{-7}; Fig. 3d).

Given the relationship between diet, metabolism and human health, we wondered whether the plasma metabolome could predict diet quality. For each of the Lifelines participants, we constructed a Lifelines Diet Score based on food frequency questionnaire (FFQ) data that reflected the relative diet quality based on diet–disease relationships[8]. To build a metabolic model to predict an individual's diet quality, we used LLD_1 as the training set and LLD_2 as the validation set. The resulting metabolic model included 76 metabolites, 51 of which were dominantly associated with diet. The diet score predicted by metabolites showed a significant association with the real diet score assessed by the FFQ in the validation set ($r^2_{adjusted}$ = 0.27; $P_{F\text{-test}}$ = 3.5 × 10^{-5}; Fig. 3e). We also tested four other dietary scores (the Alternate Mediterranean Diet Score[20], Healthy Eating Index (HEI)[21], Protein Score[22] and Modified Mediterranean Diet Score[23]) and found that the HEI predicted by plasma metabolites was also significantly associated with the FFQ-based HEI ($r^2_{adjusted}$ = 0.23; $P_{F\text{-test}}$ = 6.5 × 10^{-5}; Supplementary Table 12).

Genetic associations of plasma metabolites

Genetic associations of plasma metabolites may provide functional insights into the etiologies of complex diseases. After correcting for the first two genetic principal components, age, sex, BMI, smoking, 78 dietary habits, 40 diseases and 44 medications, QTL mapping in LLD_1 identified 48 study-wide, independent genetic associations between

https://doi.org/10.1038/s41591-022-02014-8

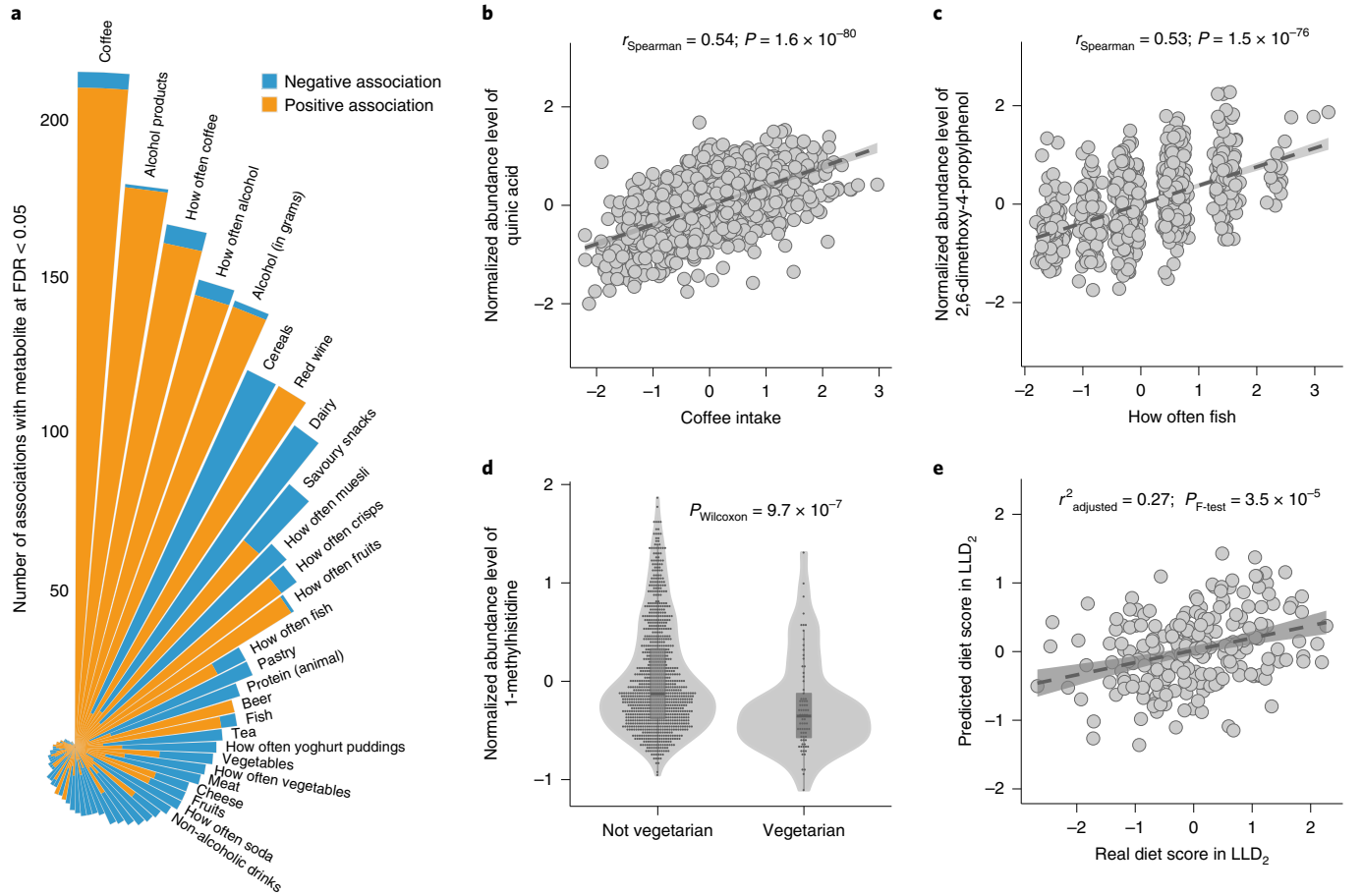


Fig. 3 | Associations between dietary habits and plasma metabolites. **a**, Summary of the associations between diet and metabolites. The bars represent dietary habits, with the bar order sorted by the number of significant associations. Association directions are colored differently: orange indicates a positive association, whereas blue indicates a negative association. The length of each bar indicates the number of significant associations at FDR < 0.05 (Spearman; two sided). **b**, Association between plasma quinic acid levels and coffee intake. The *x* and *y* axes indicate residuals of coffee intake and the metabolic abundance after correcting for covariates, respectively (n = 1,054 biologically independent samples). **c**, Association between plasma 2,6-dimethoxy-4-propylphenol levels and fish intake frequency (n = 1,054 biologically independent samples). The *x* and *y* axes refer to residuals of fish intake and metabolic abundance after correcting for covariates, respectively. **d**, Differential plasma levels of 1-methylhistidine between vegetarians and non-vegetarians (n = 1,054 biologically independent samples). The *y* axis indicates normalized residuals of metabolic abundance. The *P* value from the Wilcoxon test (two sided) is shown. The box plots show the median and first and third quartiles (25th and 75th percentiles) of the metabolite levels. The upper and lower whiskers extend to the largest and smallest value no further than 1.5× the IQR, respectively. Outliers are plotted individually. **e**, Association between the diet quality score predicted by the plasma metabolome (*y* axis) and the diet quality score assessed by the FFQ (*x* axis) (n = 237 biologically independent samples). In **b**, **c** and **e**, each gray dot represents one sample, the dark gray dashed line shows the linear regression line and the gray shading represents the 95% CI. In **b** and **c**, the association strength was assessed using Spearman correlation (two sided; the correlation coefficient and *P* value are reported) and in **e**, the prediction performance was assessed with linear regression (*F*-test; two sided; the adjusted r^2 value and *P* value are reported).

44 metabolites and 40 single-nucleotide polymorphisms (SNPs) ($P_{Spearman} < 4.2 \times 10^{-11}$; clumping $r^2 = 0.05$; clumping window = 500 kilobases (kb); Fig. 4a and Supplementary Table 6). All 48 genetic associations were replicated in either LLD_1 follow-up or the two independent replication datasets (LLD_2 and GoNL; Supplementary Fig. 3 and Supplementary Table 6). We also assessed the impact of physical activity, as assessed by questionnaires[24], on the genetics association of metabolism, but found its influence to be negligible (Supplementary Fig. 4). Functional mapping and annotation (FUMA) of genome-wide association studies (GWAS)[25] analysis revealed that the identified mQTLs were enriched in genes expressed in the liver and kidney (Extended Data Fig. 4) and related to metabolic phenotypes (Supplementary Table 6).

The strongest association we found was between the caffeine metabolite 5-acetylamino-6-formylamino-3-methyluracil (AFMU) and SNP rs1495741 near the *N*-acetyltransferase 2 (*NAT2*) gene ($r_{Spearman} = -0.52$; $P = 1.7 \times 10^{-66}$; Fig. 4b), which showed strong linkage disequilibrium ($r^2 = 0.98$) with a SNP, rs35246381, that was recently reported to be associated with urinary AFMU[26]. AFMU is a direct product of NAT2 activity and has been associated with bladder cancer risk[27]. Interestingly, the plasma level of AFMU was associated not only with coffee intake ($r_{Spearman} = 0.29$; $P = 9.2 \times 10^{-22}$; Supplementary Table 5) and the genotype of rs1495741, but also with their interactions (Supplementary Table 9). Individuals with a homologous AA genotype had a similar level of coffee intake, but their correlation between coffee intake and plasma AFMU level was significantly lower compared with individuals with GG and GA genotypes (Fig. 4e,f).

Pleotropic mQTL effects were also observed at several loci, including *SLCO1B1*, *FADS2*, *KLKB1* and *PYROXD2* (Supplementary Table 6). For example, three associations (related to three metabolites, two of them lipids) were observed for two SNPs (rs67981690 and rs4149067; linkage disequilibrium $r^2 = 0.72$ in Northern Europeans from Utah) in *SLCO1B1*, which encodes the solute carrier organic anion transporter family member 1B1. Expression of the SLCO1B1 protein is specific to the liver, where this transporter is involved in the transport

https://doi.org/10.1038/s41591-022-02014-8

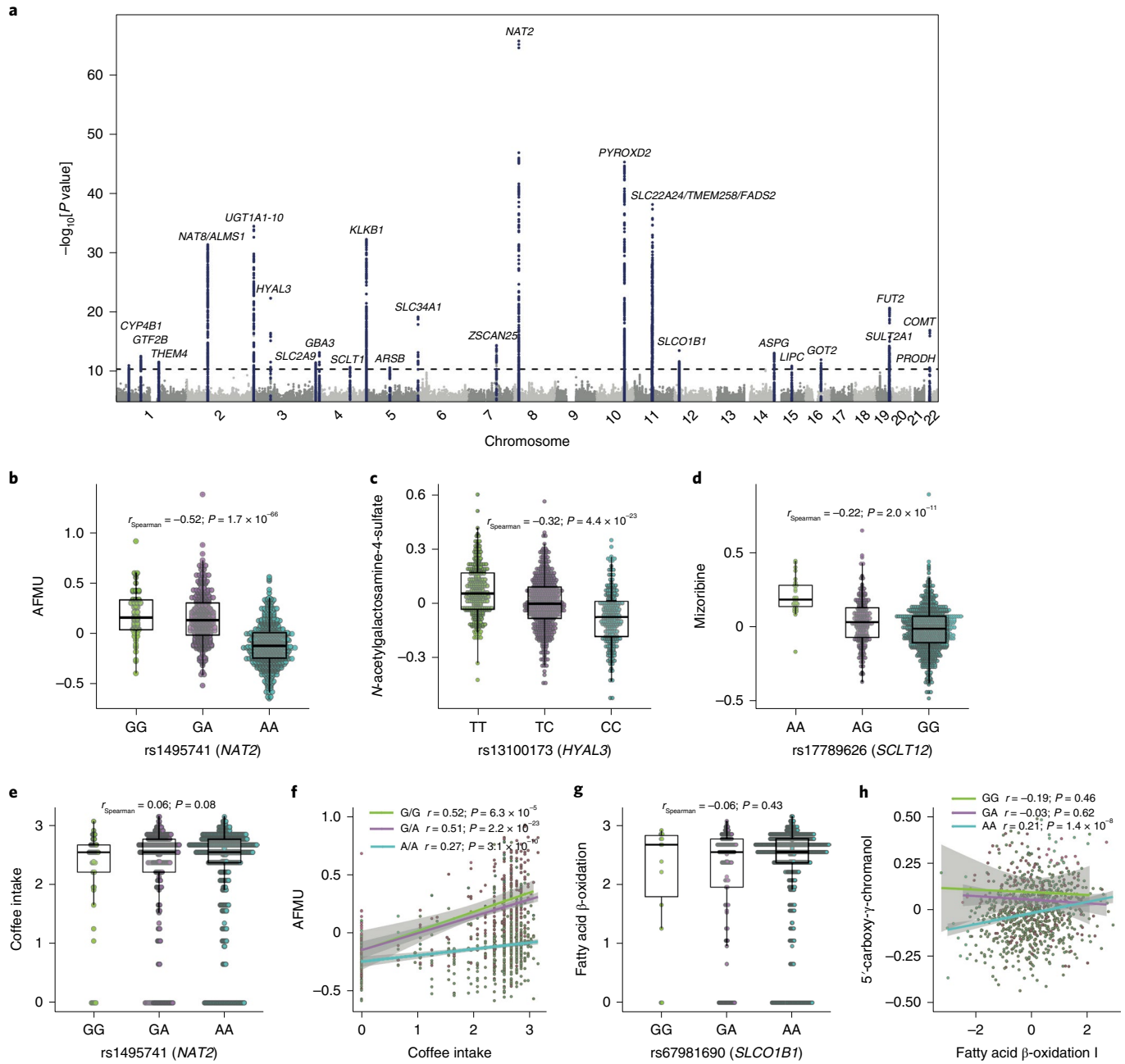


Fig. 4 | Genetic associations of plasma metabolites. a, Manhattan plot showing 48 independent mQTLs identified linking 44 metabolites and 40 genetic variants with $P < 4.2 \times 10^{-11}$ (Spearman; two sided). Representative genes for the SNPs with significant mQTLs are labeled. **b**, Association between a tag SNP (rs1495741) of the *NAT2* gene and plasma AFMU levels. **c**, Association between a SNP (rs13100173) within the *HYAL3* gene and plasma levels of *N*-acetylgalactosamine-4-sulfate. **d**, Association between a tag SNP (rs17789626) of the *SCLT1* gene and plasma mizoribine levels. **e**, Differences in coffee intake between participants with different genotypes at rs1495741. **f**, Correlations between coffee intake and AFMU in participants with different genotypes at rs1495741. **g**, Differences in bacterial fatty acid β-oxidation pathway abundance in participants with different genotypes at rs67981690. **h**, Correlations between bacterial fatty acid β-oxidation pathway abundance and 5′-carboxy-γ-chromanol in participants with different genotypes at rs67981690. In **b**–**e** and **g**, the *x* axis indicates the genotype of the corresponding SNP and the *y* axis indicates normalized residuals of the corresponding metabolic abundance ($n = 927$ biologically independent samples). Each dot represents one sample. The box plots show the median and first and third quartiles (25th and 75th percentiles) of the metabolite levels. The upper and lower whiskers extend to the largest and smallest value no further than 1.5× the IQR, respectively. Outliers are plotted individually. The association strength is shown by the Spearman correlation coefficient and corresponding *P* value (two sided). In **f** and **h**, the *x* axis indicates the normalized abundance of coffee intake (**f**) or the bacterial fatty acid β-oxidation pathway (**h**) and the *y* axis indicates the normalized residuals of the corresponding metabolic abundance. Each dot represents one sample ($n = 927$ biologically independent samples). The lines indicate linear regressions for each genotype group separately. Areas with light gray shading indicate the 95% CI of the linear regression lines. The association strength per genotype is shown by the Spearman correlation and the corresponding *P* value (two sided).

of various endogenous compounds and drugs, including statins[28], from blood into the liver. The SLCO1B1 locus has also been linked to plasma levels of fatty acids and to statin-induced myopathy[29]. Furthermore, we detected a genetics–microbiome interaction between rs67981690 and microbial fatty acid oxidation pathways in regulating plasma levels of 5′-carboxy-γ-chromanol ($P = 1.5 \times 10^{-3}$), where the

https://doi.org/10.1038/s41591-022-02014-8

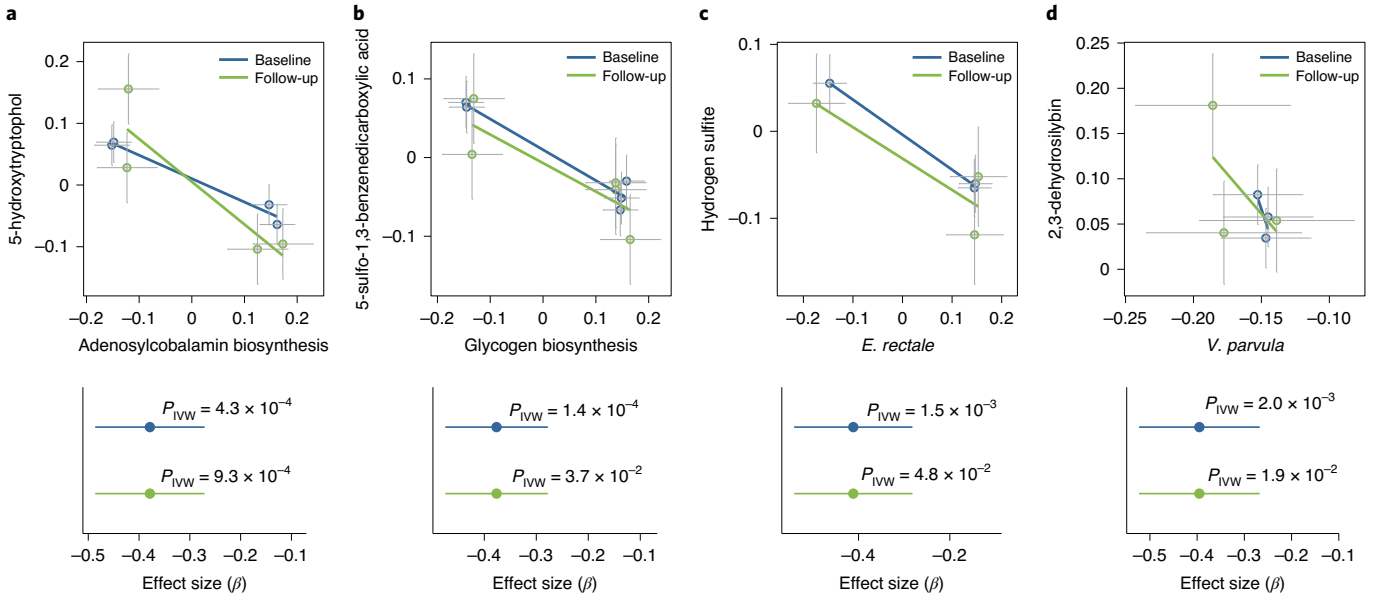


Fig. 5 | Causal relationships between microbiomes and plasma metabolites as assessed by MR analysis. a, Analysis of the association between adenosylcobalamin biosynthesis pathway abundance and 5-hydroxytryptophol levels. **b**, Glycogen biosynthesis pathway abundance versus 5-sulfo-1,3-benzenedicarboxylic acid levels. **c**, *E. rectale* abundance versus hydrogen sulfite levels. **d**, *Veillonella parvula* abundance versus 2,3-dehydrosilybin levels. In the top panels of **a**–**d**, the *x* axis shows the SNP exposure effect, and the *y* axis shows the SNP outcome effect and each dot represents a SNP. Error bars represent the s.e. of each effect size. The bottom panels of **a**–**d**, show the MR effect size (center dot) and 95% CI for the baseline (blue) and follow-up (green) datasets of the LLD_1 cohort, estimated with the IVW MR approach (two sided) ($n = 927$ biologically independent samples at baseline and $n = 311$ biologically independent samples at follow-up).

association of the bacterial fatty acid oxidation pathway with plasma levels of 5′-carboxy-γ-chromanol was dependent on the genotype of rs67981690 (Fig. 4g,h).

To identify novel mQTLs, we performed a systematic search of all published mQTL studies from 2008 onwards (Supplementary Table 13). This approach identified three novel mQTLs in our datasets (Supplementary Table 13) that were either not located close to previously reported mQTLs (distance > 1,000 kb) or not in linkage disequilibrium ($r^2 < 0.05$). The first two novel SNPs—rs13100173 at *HYAL3* and rs11741352 at *ARSB*—were associated with *N*-acetylgalactosamine-4-sulfate (Fig. 4c,d), which is associated with mucopolysaccharidosis[30]. Interestingly, *N*-acetylgalactosamine-4-sulfate can bind to HYAL proteins (HYAL1, HYAL2, HYAL3 and HYAL4), suggesting that mQTLs can also pinpoint potential metabolite–protein interactions. The third novel mQTL was rs17789626 at *SCLT1*, which was associated with mizoribine—a compound used to treat nephrotic syndrome[31].

A causal role for the microbiome in determining metabolites

We established 4,212 associations between 208 metabolites and 314 microbial factors (114 species and 200 MetaCyc pathways) ($FDR_{LLD1} < 0.05$; $P_{LLD1\,follow\text{-}up} < 0.05$; Supplementary Tables 7 and 8). Interestingly, many of the metabolites that were associated with microbial species and MetaCyc pathways are also known to be gut microbiome related based on their HMDB annotations[15]. For instance, we observed 919 associations with 25 uremic toxins, 142 associations with thiamine (vitamin B1) and 117 associations with five phytoestrogens (FDR < 0.05; Supplementary Tables 7 and 8). Uremic toxins and thiamine have been shown to be related to various diseases, including chronic kidney disease and cardiovascular diseases[32,33]. Phytoestrogens are a class of plant-derived polyphenolic compounds that can be transformed by gut microbiota into metabolites that promote the host's metabolism and immune system[33,34].

To assess whether gut microbiome composition causally contributes to plasma metabolite levels, we carried out bi-directional MR analyses (see the Methods section 'Bi-directional MR analysis'). Here, we focused on the 37 microbial features that were associated with at least three independent genetic variants at $P < 1 \times 10^{-5}$ and with 45 metabolites (Supplementary Table 14). At FDR < 0.05 (corresponding to $P = 2 \times 10^{-3}$ obtained from the inverse variance weighted (IVW) test)[35], we observed four potential causal relationships at baseline that could also be found in the follow-up in the microbiomes to metabolites direction (Fig. 5a–d and Supplementary Tables 15 and 16) but not in the opposite direction (Supplementary Table 17), and these outcomes were maintained following weighted median testing ($P < 0.03$; Supplementary Fig. 5). To ensure that the data followed MR assumptions, we performed several sensitivity analyses, including checking for horizontal pleiotropy (MR-Egger[36] intercept $P > 0.05$; Supplementary Table 15) and heterogeneity (Cochran's *Q* test $P > 0.05$; Supplementary Table 15) and leave-one-out analysis (Extended Data Fig. 5). We did not use causal estimates derived using the MR-Egger method to filter the results, as its power to detect causality is known to be low[36]. These sensitivity checks further confirmed the reliability of these four MR causal estimates.

We further found that increased abundance of microbial adenosylcobalamin biosynthesis (coenzyme B12) was associated with reduced plasma levels of 5-hydroxytryptophol (Fig. 5a)—a uremic toxin related to Parkinson's disease[37]. We also found that plasma hydrogen sulfite levels were related to *Eubacterium rectale* (Fig. 5c)—a core gut commensal species[38] that is highly prevalent (presence rate = 97%) and abundant (mean abundance = 8.5%) in both our cohorts and in other populations[39–41]. As a strict anaerobe, *E. rectale* promotes the host's intestinal health by producing butyrate and other short-chain fatty acids from non-digestible fibers[42], and a reduced abundance of this species has been observed in subjects with inflammatory bowel disease[39,43] and colorectal cancer[44] compared with healthy controls. As a toxin, hydrogen sulfite interferes with the nervous system, cardiovascular functions, inflammatory processes and the gastrointestinal and renal system[45]. Our results thus reveal a potential new beneficial effect of *E. rectale*.

To further investigate the metabolic potential of individual bacterial species, we applied newly developed pipelines to identify

microbial primary metabolic gene clusters (gutSMASH pathways)[46] and microbial genomic structural variants (SVs)[47]. These two tools profile microbial genomic entities that are implicated in metabolic functions. By associating 1,183 metabolites with 3,075 gutSMASH pathways and 6,044 SVs (1,782 variable SVs (vSVs) and 4,262 deletion SVs (dSVs); see Methods), we observed 23,662 associations with gutSMASH pathways and 790 associations with bacterial SVs ($FDR_{LLD1} < 0.05$; $P_{LLD1\ follow\text{-}up} < 0.05$; Supplementary Tables 18–20). These associations connect the genetically encoded functions of microbes with metabolites, thereby providing putative mechanistic information underlying the functional output of the gut microbiome. In one example, we observed that the microbial uremic toxin biosynthesis pathways, including the glycine cleavage pathway (in *Olsenella* and *Clostridium* species) and the hydroxybenzoate-to-phenol pathway (in *Clostridium* species) responsible for hippuric acid and phenol sulfate biosynthesis, were associated with the hippuric acid (*Olsenella* species: $r_{Spearman} = 0.15$; $P = 9.3 \times 10^{-7}$; *Clostridium* species: $r_{Spearman} = 0.18$; $P = 5.9 \times 10^{-9}$) and phenol sulfate ($r_{Spearman} = 0.17$; $P = 4.2 \times 10^{-8}$; Extended Data Fig. 6a) levels measured in plasma, respectively ($FDR_{LLD1} < 0.05$ and $P_{LLD1\ follow\text{-}up} < 0.05$; Extended Data Fig. 6b).

Diet–microbiome mediation effects in the control of metabolites

Next, we carried out a mediation analysis to investigate the links between diet, the microbiome and metabolites. For 675 microbial features that were associated with both dietary habits and metabolites (FDR < 0.05), we applied bi-directional mediation analysis to evaluate the effects of microbiome and metabolites for diet (see the Methods section 'Bi-directional mediation analysis'). This approach established 146 mediation linkages: 133 for the dietary impact on the microbiome through metabolites and 13 for the dietary impact on metabolites through the microbiome ($FDR_{mediation} < 0.05$ and $P_{inverse\text{-}mediation} > 0.05$; Fig. 6a,b and Supplementary Table 21). Most of these linkages were related to the impact of coffee and alcohol on microbial metabolic functionalities (Fig. 6a).

Coffee contains various phenolic compounds that can be converted to hippuric acid by colonic microflora[48]. Hippuric acid is an acyl glycine that is associated with phenylketonuria, propionic acidemia and tyrosinemia[49]. We observed that hippuric acid can mediate the impact of drinking coffee on *Methanobrevibacter smithii* abundance ($P_{mediation} = 2.2 \times 10^{-16}$; Fig. 6c). We also observed that hulupinic acid, which is commonly detected in alcoholic drinks, can mediate the impact of beer consumption on the *Clostridium methylpentosum* ferredoxin:NAD^+ oxidoreductase (Rnf) complex ($P_{mediation} = 2.2 \times 10^{-16}$; Fig. 6d)—an important membrane protein in driving the ATP synthesis essential for all bacterial metabolic activities[50].

Of the dietary impacts on metabolites through the microbiome (Fig. 6b and Supplementary Table 21), one interesting example is a *Ruminococcus* species vSV (300–305 kb) that encodes an ATPase responsible for transmembrane transport of various substrates[51]. This *Ruminococcus* species vSV mediated the effect of fruit consumption on plasma levels of urolithin B ($P_{mediation} = 2.2 \times 10^{-16}$; Fig. 6e). Urolithin B is a gut microbiota metabolite that protects against myocardial ischemia/reperfusion injury via the p62/Keap1/Nrf2 signaling pathway[52]. Taken together, our data provide potential mechanistic underpinnings for diet–metabolite and diet–microbiome relationships.

Discussion

By generating fasting plasma profiles of 1,183 metabolites in 1,679 samples from 1,368 individuals (311 with 4-year follow-up data) for whom we also have extensive dietary records, genetics and gut microbiome data, we carried out systematic diet, genetics and microbiome association analyses. Our results show that diet and the gut microbiome play a more dominant role than genetics in explaining inter-individual variability in metabolism, and the more variance that was explained in a metabolite, the more stable that metabolite was over time.

Dietary components are fundamental resources for the plasma metabolome, and a recent study illustrated that an individual's dietary habits can predict the levels of specific metabolites present in plasma[3], highlighting that the plasma metabolome mirrors personal dietary habits. Nevertheless, it remained to be established whether it was possible to assess an individual's diet quality score based on their plasma metabolome. Using a machine learning-based prediction model, we showed that diet quality estimated by an individual's plasma metabolome showed a significant correlation with diet quality estimated by the FFQ, suggesting that the plasma metabolome to some extent reflects diet quality.

Dietary components serve as substrates in gut microbial metabolic pathways, leading to the formation of a series of metabolites that can be absorbed from the intestine into the host's circulation. Although earlier studies had linked gut microbial taxonomic abundances to plasma metabolites[3,4,13,41,53], these investigations did not capture the specific microbial enzymes responsible for metabolite generation, even though this information is required to connect associated links to underlying molecular mechanisms[54]. Using gutSMASH and microbial SVs, we identified putative metabolic functionalities for previously unannotated microbial genetic sequences. In addition, through bi-directional mediation analysis, we identified hundreds of mediation linkages that provide insight into diet–microbiome interactions in human metabolic health, as illustrated by several metabolites (for example, phenol and pipecolic acid) that have previously been related to cardiometabolic and kidney diseases[32]. Notably, these mediation linkages mainly show that the impact of diet composition on the microbiome can be mediated by metabolites, highlighting the pronounced selective power of dietary habits in shaping the gut microbiome. Nevertheless, as these results are mainly based on observational data, interpretation of such associations should be made with caution, and future intervention and experimental studies that focus on specific diet and microbial genomic capacities are essential to confirm causality.

Apart from diet and the gut microbiome, human genetics also acts as a potential determinant of the plasma metabolome. With this metabolome dataset, we not only replicated previously reported mQTLs, but we also identified three mQTLs involving three loci not previously known to be associated with any metabolites. The mQTLs we characterized could be linked to cardiometabolic and chronic kidney diseases, as illustrated by the tissue-specific gene expression analysis and pleotropic mQTL effects. We also used genetic variants as instruments in MR to infer causal relationships between the gut microbiome and metabolites. This analysis showed that the microbiome may causally contribute to the levels of toxins (hydrogen sulfite and 5-hydroxytryptophol) related to chronic kidney disease and cardiometabolic diseases[32]. The causal relationships between microbiomes and metabolites that we have established thus reveal potential metabolic functionalities of gut microbes that impact on human health.

We acknowledge several limitations in our study. Untargeted plasma metabolome was profiled using FI-MS without compound separation using liquid chromatography columns, and no genuine standards were used. Although the abundances of a few metabolites were well validated using the LC-MS/MS or NMR platforms, identification and quantification of mass peaks using the FI-MS approach is still generally less accurate than in the classical LC-MS/MS platform. We systematically investigated the contributions of genetics, diet and the microbiome to inter-individual metabolome variations and replicated the explained variance in two independent metabolic profile assessments of the same cohort performed 4 years apart. However, overfitting may still have been an issue and could potentially have biased the conclusions. Our findings should be further replicated

https://doi.org/10.1038/s41591-022-02014-8

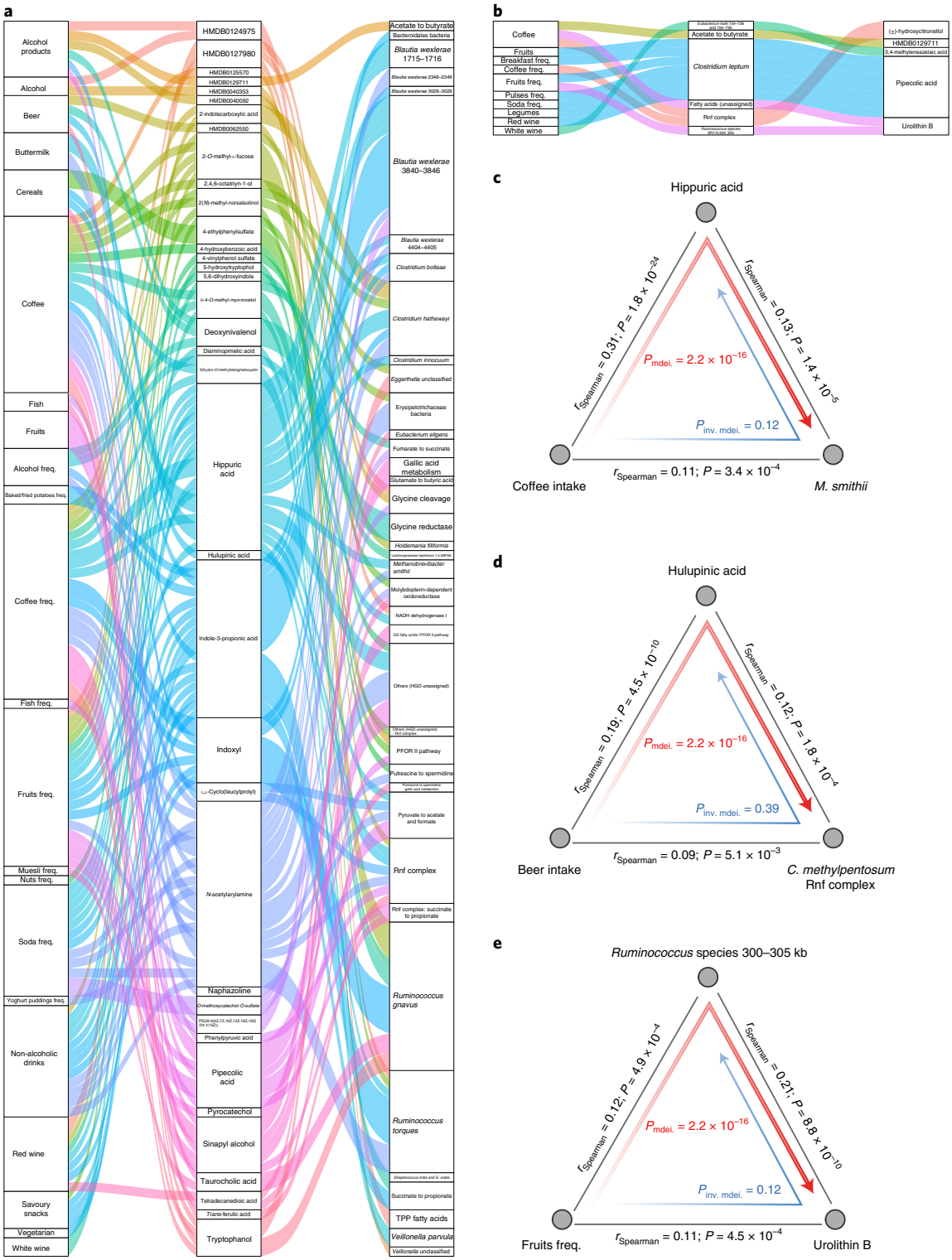


Fig. 6 | Mediation analysis identifies linkages between the gut microbiome, metabolites and dietary habits. a, Parallel coordinates chart showing the 133 mediation effects of plasma metabolites that were significant at FDR < 0.05. Shown are dietary habits (left), plasma metabolites (middle) and microbial factors (right). The curved lines connecting the panels indicate the mediation effects, with colors corresponding to different metabolites. freq., frequency; PFOR, pyruvate:ferredoxin oxidoreductase; OD, oxidative decarboxylation; HGD, 2-hydroxyglutaryl-CoA dehydratase; TPP, thiamine pyrophosphate. **b**, Parallel coordinates chart showing the 13 mediation effects of the microbiome that were significant at FDR < 0.05. Shown are dietary habits (left), microbial factors (middle) and plasma metabolites (right). For the microbial factors column, number ranges represent the genomic location of microbial structure variations (SVs) in kilobyte unit, and colons represent the detailed annotation of certain gutSMASH pathway. **c**, Analysis of the effect of coffee intake on the abundance of *M. smithii* as mediated by hippuric acid. **d**, Analysis of the effect of beer intake on the *C. methylpentosum* Rnf complex pathway as mediated by hulupinic acid. **e**, Analysis of the effect of fruit intake on urolithin B in plasma as mediated by a vSV in *Ruminococcus* species (300–305 kb). In **c**–**e**, the gray lines indicate the associations between the two factors, with corresponding Spearman coefficients and *P* values (two sided). Direct mediation is shown by a red arrow and reverse mediation is shown by a blue arrow. Corresponding *P* values from mediation analysis (two sided) are shown. inv., inverse; mdei., mediation.

using an independent cohort for which similarly extensive datasets on the metabolome, genetics, the microbiome, diet, disease and medication are also available. However, it is challenging to obtain such cohorts. We included as many participants as possible for the replication, including two independent sets of individuals from the LLD_2 and GoNL cohorts. Nonetheless, our study was still underpowered. At the observed effect size in the discovery set and at the $P < 0.05$ level, we have 80% power to replicate the findings for 100% of the genetic associations, but only 60% power for the microbial associations. In addition, causal relationships between the microbiome and metabolites were based on one-sample MR; replication in independent cohorts with larger sample sizes and two-sample MR analysis may further strengthen those observations and better establish possible biological significance. Finally, the LLD cohort was comprised of Dutch participants of Caucasian ethnicity from the northern region of the Netherlands. It is thus possible that the LLD results are biased towards a region-specific microbial background constrained by host genetics and local environmental exposures. We primarily focused on biologically plausible mechanisms by integrating different layers of omics to provide mechanistic hypotheses, and experimental validation is thus warranted.

Taken together, the dietary, genetic and microbial associations with plasma metabolites and the causal and mediation linkages that we report here provide a comprehensive resource that can guide follow-up studies aimed at designing preventive and therapeutic strategies for human metabolic health.

Online content

Any methods, additional references, Nature Research 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/s41591-022-02014-8.

References

1. Suhre, K. et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54–60 (2011).
2. Shin, S. Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
3. Bar, N. et al. A reference map of potential determinants for the human serum metabolome. *Nature* **588**, 135–140 (2020).
4. Asnicar, F. et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).
5. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
6. Boomsma, D. I. et al. The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22**, 221–227 (2014).
7. Chen, L. et al. The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **184**, 2302–2315 (2021).
8. Vinke, P. C. et al. Development of the food-based Lifelines Diet Score (LLDS) and its application in 129,369 Lifelines participants. *Eur. J. Clin. Nutr.* **72**, 1111–1119 (2018).
9. Sanna, S. et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).
10. Fuhrer, T., Zampieri, M., Sevin, D. C., Sauer, U. & Zamboni, N. Genomewide landscape of gene–metabolome associations in *Escherichia coli*. *Mol. Syst. Biol.* **13**, 907 (2017).
11. Fuhrer, T., Heer, D., Begemann, B. & Zamboni, N. High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection–time-of-flight mass spectrometry. *Anal. Chem.* **83**, 7074–7080 (2011).
12. Wang, D. M. et al. Characterization of gut microbial structural variations as determinants of human bile acid metabolism. *Cell Host Microbe* **29**, 1802–1814 (2021).
13. Kurilshikov, A. et al. Gut microbial associations to plasma metabolites linked to cardiovascular phenotypes and risk: a cross-sectional study. *Circ. Res.* **124**, 1808–1820 (2019).
14. Zou, H. & Hastie, T. Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. B* **67**, 301–320 (2005).
15. Wishart, D. S. et al. HMDB 4.0: the Human Metabolome Database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
16. Zhao, Y. et al. Analysis of multiple metabolites of tocopherols and tocotrienols in mice and humans. *J. Agric. Food Chem.* **58**, 4844–4852 (2010).
17. Jiang, Q., Christen, S., Shigenaga, M. K. & Ames, B. N. γ-Tocopherol, the major form of vitamin E in the US diet, deserves more attention. *Am. J. Clin. Nutr.* **74**, 714–722 (2001).
18. Pallister, T. et al. Hippurate as a metabolomic marker of gut microbiome diversity: modulation by diet and relationship to metabolic syndrome. *Sci. Rep.* **7**, 13670 (2017).
19. Razavi, A. C. et al. Novel findings from a metabolomics study of left ventricular diastolic function: the Bogalusa Heart Study. *J. Am. Heart Assoc.* **9**, e015118 (2020).
20. Trichopoulou, A. & Vasilopoulou, E. Mediterranean diet and longevity. *Br. J. Nutr.* **84**, S205–S209 (2000).
21. Krebs-Smith, S. M. et al. Update of the Healthy Eating Index: HEI-2015. *J. Acad. Nutr. Diet.* **118**, 1591–1602 (2018).
22. Moller, G. et al. A protein diet score, including plant and animal protein, investigating the association with HbA1c and eGFR—the PREVIEW project. *Nutrients* **9**, 763 (2017).
23. Khalili, H. et al. Adherence to a Mediterranean diet is associated with a lower risk of later-onset Crohn's disease: results from two large prospective cohort studies. *Gut* **69**, 1637–1644 (2020).
24. Wendel-Vos, G. C., Schuit, A. J., Saris, W. H. & Kromhout, D. Reproducibility and relative validity of the short questionnaire to assess health-enhancing physical activity. *J. Clin. Epidemiol.* **56**, 1163–1169 (2003).
25. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
26. Schlosser, P. et al. Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nat. Genet.* **52**, 167–176 (2020).
27. Selinski, S., Blaszkewicz, M., Ickstadt, K., Hengstler, J. G. & Golka, K. Refinement of the prediction of *N*-acetyltransferase 2 (NAT2) phenotypes with respect to enzyme activity and urinary bladder cancer risk. *Arch. Toxicol.* **87**, 2129–2139 (2013).
28. Lee, H. H. & Ho, R. H. Interindividual and interethnic variability in drug disposition: polymorphisms in organic anion transporting polypeptide 1B1 (OATP1B1; *SLCO1B1*). *Br. J. Clin. Pharm.* **83**, 1176–1184 (2017).
29. Group, S. C. et al. *SLCO1B1* variants and statin-induced myopathy—a genomewide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
30. Yogalingam, G., Muller, V., Hopwood, J. J. & Anson, D. S. Regulation of *N*-acetylgalactosamine 4-sulfatase expression in retrovirus-transduced feline mucopolysaccharidosis type VI muscle cells. *DNA Cell Biol.* **18**, 187–195 (1999).
31. Ohtomo, Y. et al. High-dose mizoribine therapy for childhood-onset frequently relapsing steroid-dependent nephrotic syndrome with cyclosporin nephrotoxicity. *Pediatr. Nephrol.* **20**, 1744–1749 (2005).

32. Wang, Z. & Zhao, Y. Gut microbiota derived metabolites in cardiovascular health and disease. *Protein Cell* **9**, 416–431 (2018).
33. Wang, Y. et al. Decoding microbial genomes to understand their functional roles in human complex diseases. *iMeta* **1**, e14 (2022).
34. Seyed Hameed, A. S., Rawat, P. S., Meng, X. & Liu, W. Biotransformation of dietary phytoestrogens by gut microbes: a review on bidirectional interaction between phytoestrogen metabolism and gut microbiota. *Biotechnol. Adv.* **43**, 107576 (2020).
35. Bowden, J. et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).
36. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
37. Levin, J., Bötzel, K., Giese, A., Vogeser, M. & Lorenzl, S. Elevated levels of methylmalonate and homocysteine in Parkinson's disease, progressive supranuclear palsy and amyotrophic lateral sclerosis. *Dement. Geriatr. Cogn. Disord.* **29**, 553–559 (2010).
38. Gacesa, R. et al. Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732–739 (2022).
39. Chen, L. et al. Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat. Commun.* **11**, 4018 (2020).
40. Karcher, N. et al. Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. *Genome Biol.* **21**, 138 (2020).
41. Chen, L. et al. Genetic and microbial associations to plasma and fecal bile acids in obesity relate to plasma lipids and liver fat content. *Cell Rep.* **33**, 108212 (2020).
42. Ríos-Covián, D. et al. Intestinal short chain fatty acids and their link with diet and human health. *Front. Microbiol.* **7**, 185 (2016).
43. Vich Vila, A. et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* **10**, eaap8914 (2018).
44. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
45. Beltowski, J. Hydrogen sulfide in pharmacology and medicine—an update. *Pharm. Rep.* **67**, 647–658 (2015).
46. Andreu, V. P. et al. A systematic analysis of metabolic pathways in the human gut microbiota. Preprint at *bioRxiv* https://doi.org/10.1101/2021.02.25.432841 (2021).
47. Zeevi, D. et al. Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).
48. Ogawa, M., Suzuki, Y., Endo, Y., Kawamoto, T. & Kayama, F. Influence of coffee intake on urinary hippuric acid concentration. *Ind. Health* **49**, 195–202 (2011).
49. Duranton, F. et al. Normal and pathologic concentrations of uremic toxins. *J. Am. Soc. Nephrol.* **23**, 1258–1270 (2012).
50. Biegel, E. & Muller, V. Bacterial Na^+-translocating ferredoxin:NAD^+ oxidoreductase. *Proc. Natl Acad. Sci. USA* **107**, 18138–18142 (2010).
51. Holland, I. B. & Blight, M. A. ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. *J. Mol. Biol.* **293**, 381–399 (1999).
52. Zheng, D. et al. Urolithin B, a gut microbiota metabolite, protects against myocardial ischemia/reperfusion injury via p62/Keap1/Nrf2 signaling pathway. *Pharm. Res.* **153**, 104655 (2020).
53. Visconti, A. et al. Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* **10**, 4505 (2019).
54. Bradley, P. H. & Pollard, K. S. Building a chemical blueprint for human blood. *Nature* **588**, 36–37 (2020).

[1]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [2]Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [3]Department of Cardiology, Nanjing Medical University, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China. [4]Cardiovascular Research Center, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School, Nanjing Medical University, Suzhou, China. [5]Laboratory of Genomic Diversity, Center for Computer Technologies, ITMO University, St. Petersburg, Russia. [6]Bioinformatics Group, Wageningen University, Wageningen, the Netherlands. [7]Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [8]Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, the Netherlands. [9]Department of Immunology and Metabolism, Life and Medical Sciences Institute, University of Bonn, Bonn, Germany. [10]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. ✉e-mail: j.fu@umcg.nl

Lifelines Cohort Study

Cisca Wijmenga[1], Alexandra Zhernakova[1], Jingyuan Fu[1,2] and Rinse K. Weersma[7]

Methods

Study cohorts

The LLD cohort (n = 1,500) is a sub-cohort of the large prospective Lifelines cohort study from the north of the Netherlands[5,55]. The cohort is 58% female and 42% male, with a mean age of 45.04 years (s.d. = 13.60). The mean BMI is 25.26 (s.d. = 4.18) and 12% of participants are obese (BMI > 30)[5]. All Lifelines participants signed an informed consent form before sample collection. The ethics review board of the University Medical Center Groningen has approved the study with reference number M12.113965. For this study, we involved 1,054 out of 1,500 LLD participants (LLD_1) for whom detailed dietary habit, stool microbiome and plasma untargeted metabolomics information is available. After removing relatives and genetic outliers, 927 individuals were subjected to genetics analysis, and for 311 of them we also have 4-year follow-up metabolome data (Supplementary Table 1). We further included several replication cohorts: the GoNL cohort (n = 77)[56] and the remaining set of LLD participants for whom we have genetics and plasma metabolome data but not microbiome data (n = 237; LLD_2) (Fig. 1a and Supplementary Table 1).

Data generation and preprocessing

Plasma metabolome. Plasma samples of study participants were collected and frozen at −80 °C with ethylenediaminetetraacetic acid. During extraction, plasma samples were thawed on ice, vortexed and spun down. Then, 20 µl plasma was combined with 180 µl 80% methanol and vortexed for 15 s. The samples were then incubated at 4 °C for 1 h to precipitate the proteins and then spun for 30 min at 3,200 RCF.

Untargeted metabolic profiling of fasting plasma samples was conducted at General Metabolics using FI-MS on an Agilent 6550 Q-TOF system[10,11]. In brief, the instrument was set to scan in full mass spectrometry at 1.4 Hz in negative ionization, 4 GHz High Res Mode and from 50–1,000 m/z. The solvent was 60:40 isopropanol:water supplemented with 1 mM NH_4F at pH 9.0, as well as 10 nM hexakis(1H, 1H,3H-tetrafluoropropoxy)phosphazene and 80 nM taurocholic acid for online mass calibration. Then, 100 µl of samples were injected into the ionization source in a random order. Data were acquired in profile mode, centroided and analyzed using MATLAB (MathWorks). Missing values were filled by recursion in the raw data. Upon identification of consensus centroids across all samples, ions were putatively annotated by accurate mass and isotopic patterns. Based on the HMDB (version 4.0)[15], a list of the expected ions found under these conditions was generated, including deprotonated, fluorinated and all major adducts. As this method does not employ chromatographic separation or in-depth MS2 characterization, it is not possible to distinguish between compounds with identical molecular formulas. The confidence of annotation reflects level 4 but is higher in practice in the case of intermediates of primary metabolism because they are the most abundant metabolites in cells[10,11]. Ion intensities were normalized by quantile normalization to compensate for slight variations in the sample amount. In this way, 1,183 peaks were annotated based on accurate mass using a 1-mDa tolerance (Supplementary Table 2). The annotated metabolites cover 18 chemical categories based on the HMDB[15], including 341 lipids and lipid-like molecules, 218 organic acids and derivatives, 196 organoheterocyclic compounds, 118 phenylpropanoids and polyketides, 109 benzenoids, 104 organic oxygen compounds and 97 additional metabolites belonging to another 12 categories (Supplementary Table 2). Finally, we estimated the effect of sample plate batch on metabolite level and detected no batch effects.

To investigate potential factors that may influence the human plasma metabolome, we correlated the first 100 principal components of the 1,183 metabolites (accounting for 73% of the total metabolome variance) with age, sex, BMI, smoking, 78 dietary habits, 39 diseases and the use of 44 medications (Supplementary Table 22). As we were interested in the impact of diet, genetics and the gut microbiome on metabolites, we decided to correct for age, sex, smoking and oral contraceptive use, based on the correlation results. To adjust for confounding factors, we first log-transformed the metabolite abundances, then applied a linear regression model that included all of the confounding factors as covariates, taking the residuals for the subsequent analysis.

Stool microbiome. Fecal samples were collected by participants at home and placed in the freezer (−20 °C) within 15 min of production. Subsequently, a nurse visited the participant to pick up the fecal samples on dry ice and transfer them to the laboratory. Aliquots were then made and stored at −80 °C until further processing (fecal samples of the GoNL cohort were stored in RNAlater). The same protocol for fecal DNA isolation and metagenomics sequencing was used in all four cohorts. Fecal DNA isolation was performed using the AllPrep DNA/RNA Mini Kit (80204; Qiagen). After DNA extraction, fecal DNA was sent to the Broad Institute of MIT and Harvard, where library preparation and whole-genome shotgun sequencing were performed on an Illumina HiSeq platform. From the raw metagenomics sequencing data, low-quality reads were discarded by the sequencing facility and reads belonging to the human genome were removed by mapping the data to the human reference genome (version NCBI37) using KneadData (version 0.4.6.1) and Bowtie 2 (version 2.1.0)[57,58].

Microbial taxonomic profiles were generated using MetaPhlAn2 (version 2.7.2)[59]. Microbial general pathways were determined using HUMAnN2 (ref. [60]), which maps DNA/RNA reads to a customized database of functionally annotated pan-genomes. HUMAnN2 reported the abundances of gene families from the UniProt Reference Clusters[61] (UniRef90), which were further mapped to microbial pathways from the MetaCyc metabolic pathway database[62,63]. In total, we detected 156 species and 343 pathways that were present in at least 10% of samples, retaining 98% of the original species composition and 100% of the original functional composition. The relative abundances of both species and pathway datasets were centered-log-ratio transformed, followed by inverse-rank transformation, before subsequent analysis[64].

We applied the SGV-Finder pipeline[47] to classify SVs that were either completely absent from the microbial genome of some samples (dSVs) or whose coverage was highly variable across samples (vSVs). Before SV classification, we applied an iterative coverage-based read assignment algorithm that resolves ambiguous read assignments to regions that are similar between different bacteria, using information on bacterial relative abundances in the microbiome, their genomic sequencing coverage and sequencing and alignment qualities[47]. In total, we classified 4,262 dSVs and 1,782 vSVs from 41 microbial species that were present in at least 10% of samples. The vSV data were inverse-rank transformed for subsequent analysis.

Metabolite-specific pathways were generated using the gutSMASH algorithm[46]. In total, we generated 3,075 microbial strain-level metabolite-specific pathways that were present in at least 10% of samples. The abundance of these pathways was recorded as reads per kilobase of transcript per million reads mapped, and inverse-rank transformation was applied before subsequent analysis.

Genotype data. Microarray genotype data for the LLD cohort were generated using the CytoSNP-12 Beadchip and Immunochip assays, as previously described[65]. Quality control checks on the LLD cohort were performed using the Haplotype Reference Consortium (version 1.0) preparation checking tool (version 4.2.3). We then uploaded the resulting VCF files to the Michigan Imputation Server[66]. Phasing and imputation were performed using the option SHAPEIT for phasing, population EUR and the mode Quality Control and Imputation. For all steps, we used version R1 as a reference[67]. We further excluded SNPs that had an imputation quality $r^2 < 0.5$, failed the Hardy–Weinberg equilibrium test ($P < 1 \times 10^{-6}$), had a call rate of <95% or had a minor allele frequency of <5%. In total, we obtained genotype data for 5.3 million SNPs (genome build hg19) for 927 individuals after removing relatives

and outliers. The genotypes of the GoNL samples were obtained by whole-genome sequencing.

Statistical analysis

Distance matrix-based variance estimation. We applied feature selection based on the permutational multivariate analysis of variance using distance matrices (PERMANOVA) procedure to estimate the contributions of different factors to inter-individual variations of the whole plasma metabolome. Phenotypic, genetic and microbial contributions were estimated based on the 927 participants for whom plasma metabolites, a stool microbiome, phenotypic data and genotype were available. First, we used each raw phenotypic and microbial feature to estimate inter-individual metabolic variations using the adonis function from vegan (version 2.5.5) with 1,000 times permutation. Only phenotypic and microbial features that could estimate inter-individual metabolic variations at a permutational FDR of <0.05 were kept. For genetic variants, we used SNPs with significant mQTLs. To deal with the colinearity of selected features, we applied hierarchical clustering analysis based on a feature inter-correlation distance matrix ($1 - r^2$). Features were assigned to different clusters based on 70% dissimilarity and the central feature in each cluster was selected as representative. All representative features were further included in PERMANOVA to estimate the combined contribution to inter-individual metabolome variation.

Associations with dietary habits. To assess associations between diet and metabolites, continuous dietary habits were inverse-rank transformed and corrected for age, sex and smoking. Spearman correlation was applied to assess the correlation between 78 dietary habits and 1,183 metabolites (residuals after regressing out age, sex, smoking and oral contraceptive use; Supplementary Table 2). The FDR was calculated using the Benjamini–Hochberg procedure[68].

QTL mapping. This analysis involved the 927 participants for whom there were genotype, plasma metabolome and phenotypic data. After adjusting for the first two genetic principal components, age, sex, BMI, smoking, 78 dietary habits, 40 diseases and 44 medications, as described above, a Java-based pipeline (version 1.4nZ)[69] was applied for QTL mapping by calculating the Spearman correlation between SNP dosage and metabolite residuals after regressing out the above covariates. We considered metabolite associations with $P < 4.2 \times 10^{-11}$ as significant—a threshold corresponding to a genome-wide significance cut-off of 5.0×10^{-8} divided by 1,183 tests. All independent mQTLs (clumping variants with linkage disequilibrium $r^2 < 0.05$ and a 500-kb window[25]) above this threshold were reported. We also tested the impact of physical activities on QTL effect by re-running the QTL analysis in 855 Lifelines participants for whom we had SQUASH (the short questionnaire to assess health-enhancing physical activity) physical activity scores[24].

Microbiome-wide associations. We previously reported that several medications can alter the gut microbiome significantly, including proton pump inhibitors, antibiotics and laxatives[70,71]. We therefore adjusted all microbial datasets for these confounding factors, together with age, sex and smoking. For microbial changes 4 years apart, we regressed out age and sex. Next, Spearman correlation was applied to check the associations between metabolites and microbial features, and *P* values were adjusted using the Benjamini–Hochberg procedure.

Interaction analysis. For metabolites associated with at least two types of factor (from genetics, the microbiome and diet), we further performed interaction analysis by assessing pairwise interactions between the two factors using a linear model ($y = a + b + a \times b$). *P* values were adjusted using the Benjamini–Hochberg method.

Estimating the variance of individual metabolites. To estimate the variance of each metabolite that was contributed by dietary, genetic and microbial features, we applied machine learning-based lasso regression from the glmnet package (version 2.0.16). While ensemble machine learning methods have previously been shown to outperform the predictive capabilities of linear methods such as lasso[3], lasso's interpretability and capacity to integrate highly correlated data layers (microbiome taxa and dietary habits) made it an attractive methodology for our analysis. We believe that, while the overall variance explained might be an underestimation of the predictive power of the available data layers, the relative variability explained by each data layer should be representative of the dominant factor that explains most variance in each metabolite.

All of the dietary, microbial (general species and pathway relative abundance) and genetic features that were significantly associated with a specific metabolite at FDR < 0.05 were involved in the model. These features were further selected using lasso with a lambda that gave a minimum mean error from a tenfold cross-validation in order to control for overfitting and to provide a conservative estimate of model performance. Finally, features selected by lasso were included in the linear model to estimate the variance contributed by different factors, and the adjusted r^2 and *F*-test *P* value were recorded. The FDR was calculated based on the Benjamini–Hochberg procedure. We also applied Elastic Net from the glmnet package (version 2.0.16) to estimate the variance of each metabolite contributed by different factors and compared the results with the above method.

Principal component analysis. The levels of all plasma metabolites were included in the principal component analysis. We applied the vegdist() function from the R package (version 2.5.5) vegan to calculate the Euclidean dissimilarity matrix based on the metabolite levels. Subsequently, classical metric multidimensional scaling was carried out based on the Euclidean distance matrix to obtain different principal coordinates.

Temporal consistency of individual metabolites. We used the Spearman correlation to check how consistent the levels of individual metabolites were between the baseline and 4-year follow-up. Metabolites with larger correlation coefficients were assumed to be more stable.

Lifelines diet quality score prediction. We first checked the Lifelines Diet Score[8] associations with each metabolite in 1,054 LLD samples and selected the significant metabolite features ($P < 0.05$) for lasso regression, as described above. Metabolites selected by lasso were used to build the linear model, and 230 LLD_2 samples with both diet score and plasma metabolome information available were used as validation. Adjusted r^2 and the *P* value from the *F*-test were reported to reflect the performance of the prediction model. We also carried out these analyses for four other dietary scores: the alternate Mediterranean Diet Score[20], HEI[21], Protein Score[22] and Modified Mediterranean Diet Score[23].

Tissue-specific gene expression analysis. Summary statistics of independent mQTLs were used for tissue-specific gene expression analysis with FUMA[25].

Bi-directional MR analysis. To evaluate whether the microbiome can causally contribute to plasma metabolites, we applied bi-directional MR analyses. A microbiome GWAS was performed using the same approach we applied for the metabolite GWAS, after correcting for age, sex, smoking and the use of proton pump inhibitors, antibiotics and laxatives. We focused on the 37 microbial features associated with at least three independent genetic variants at $P < 1 \times 10^{-5}$ in the baseline that could also be found in the follow-up samples with the same direction of association ($P < 0.05$) and on the 45 metabolites with significant associations with the microbiome ($FDR_{LLD1} < 0.05$ and $P_{LLD1\,follow\text{-}up} < 0.05$). The

relaxed significance threshold for choosing microbiome-associated SNPs as genetic instruments was used to increase the number of SNPs available for analyses, as described previously[9].

MR analysis was done using the TwoSampleMR (version 0.5.5) R package. While this package was developed for two-sample MR analysis, a recent paper showed that it is possible to use most two-sample MR methods in the one-sample setting, including IVW[35] and weighted median[72]. Therefore, MR estimates were calculated using Wald ratios and these Wald ratios were meta-analyzed using the IVW method[35]. In addition, we report MR estimates calculated using the weighted median test. We kept only the results based on three or more SNPs. To ensure the validity of the results, several sensitivity analyses were performed. We excluded MR estimates potentially driven by horizontal pleiotropy (removing results with MR-Egger[36] intercept $P < 0.05$) and heterogeneity (removing results with Cohran's Q test[35] $P < 0.05$). In addition, we carried out leave-one-out analysis[73] to check whether the MR estimates were possibly driven by a single SNP (removing the estimates where all but one leave-one-out configuration had $P < 0.05$). Multiple testing correction was performed using the Benjamini–Hochberg approach based on IVW P values. To avoid complex causality relationships, we excluded the results that showed a nominally significant MR estimate in the other direction ($P < 0.05$). For this analysis, metabolite-associated SNPs at a P value cut-off of 1×10^{-5} in the baseline group and $P < 0.1$ in the follow-up group were used as genetic instruments in IVW-based MR.

Bi-directional mediation analysis. For microbial features associated with both metabolites ($FDR_{LLD1} < 0.05$ and $P_{LLD1\ follow\text{-}up} < 0.05$) and dietary habits (FDR < 0.05), we first checked whether the dietary habits were associated with the metabolite using Spearman correlation (FDR < 0.05). Next, we carried out bi-directional mediation analysis with interactions ($y = x + m + x \times m$, where y is the outcome, x is the variable and m represents the mediator) between mediator and outcome taken into account using the mediate function from mediation (version 4.5.0) to infer the mediation effect of metabolites and microbiome for dietary impacts[74]. The FDR was calculated based on the Benjamini–Hochberg procedure.

Reporting summary

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

Data availability

All processed microbiome abundance data and full summary statistics of mQTLs are freely available via the MOLGENIS cloud (https://geneticsresearch.molgeniscloud.org/menu/main/home), with an interactive browser of the top 100,000 mQTLs. The annotation of metabolites is based on the HMDB (https://hmdb.ca; version 5). The raw metagenomics sequence, metabolomics and basic phenotype data (age, sex and BMI) are deposited in the European Genome-phenome Archive (EGA) database with the study ID EGAS00001001704, which includes dataset IDs EGAD00001001991 (raw metagenomics sequencing) and EGAD00001006953 (raw metabolomics data). However, the use of Lifelines data and materials must comply with the informed consent signed by Lifelines participants specifying that their collected data will not be used for commercial purposes. There is a minimal access procedure for access to the EGA dataset that includes the provision of a contact address and completion of an online data access form (https://goo.gl/forms/TWHlrmbXaXNqWnnl2), which is very simple and is only intended to ensure that the data are being requested for research/scientific purposes only. Submitted data access forms will be evaluated by the data manager and Lifelines. For requests from verified academic parties, access will be granted within 2 weeks. There are no restrictions on downstream data re-use or authorship requirements. For requests from commercial parties, Lifelines will perform a pre-data protection impact assessment (pre-DIPA) to assess the risks of the proposed processing of personal data (for example, purpose, storage, access, archiving and so on) with respect to the General Data Protection Regulation (GDPR) subject rights. Based on the outcome of the pre-DPIA, Lifelines will decide whether sharing data with the commercial entity is allowed and/or whether additional measures have to be taken. Genotype and metadata, including disease, medication and other clinical and lifestyle information, are however privacy sensitive. To ensure adherence to participant's privacy and informed consent, the rights of participants as described in the GDPR (EU privacy laws) and Lifelines biobank regulations, the complete genotype and phenotype data cannot be provided as open access and are only available from Lifelines under controlled access in a secure Lifelines Workspace or High-Performance Cluster (HPC) environment. As Lifelines is a non-profit organization dependent on (governmental) subsidies, a fee is required to cover the costs of controlled data access and supporting infrastructure. In brief, the step-by-step data access procedure is as follows: (1) data are requested by filling in the application form to request 'Available Lifelines-data' at https://www.lifelines.nl/researcher/how-to-apply/apply-here; (2) Lifelines will evaluate project proposals to ensure compliance with the Lifelines data access policy, the informed consent of Lifelines participants and the GDPR, and that the data are being requested for non-commercial research; (3) upon approval, Lifelines will send Data and Material Transfer Agreement contracts to the applicants; and (4) after the required contracts are signed, Lifelines will provide access to data via the Workspace or HPC and link the raw and processed Lifelines sequencing data to the Lifelines phenotypes. Lifelines strives to accomplish steps 2–4 at 2 weeks per step, assuming that no extra actions by the applicant or Lifelines are required. The fee for data access on the HPC is €3,500 for 1 year and the fee for the Lifelines Workspace environment is €4,500 for 1 year, or less for shorter periods of time. There are no restrictions on the downstream re-use of aggregated, non-identifiable results (as approved by Lifelines), nor are there authorship requirements, but Lifelines does request that it is acknowledged in publications using these data. The data access policy, data access fees and an example Data and Material Transfer Agreement (which includes details on how to acknowledge the use of Lifelines data in publications) are described in detail at https://www.lifelines.nl/researcher/how-to-apply. Note that data access for replication can be arranged via Lifelines. Lifelines will not charge an access fee for controlled access to the full dataset used in the manuscript (including phenotype and sequencing data) for a period of 3 months for the specific purpose of replication of the results presented in the current manuscript. Researchers interested in such a replication study can contact Lifelines at research@lifelines.nl. Further information can be obtained from Lifelines at https://www.lifelines.nl/researcher/how-to-apply/information-request or by contacting Lifelines at research@lifelines.nl.

Code availability

Analysis codes are available via https://github.com/GRONINGEN-MICROBIOME-CENTRE/Groningen-Microbiome/tree/master/Projects/LLDeep_plasma_GeneralMeta.

References

55. Wijmenga, C. & Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat. Genet.* **50**, 322–328 (2018).
56. Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
57. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
58. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421–432 (2019).

59. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
60. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
61. Bateman, A. et al. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
62. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
63. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2018).
64. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc.* **44**, 139–177 (1982).
65. Zhernakova, D. V. et al. Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nat. Genet.* **50**, 1524–1532 (2018).
66. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
67. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
68. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.* **57**, 289–300 (1995).
69. Fehrmann, R. S. N. et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
70. Imhann, F. et al. Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740–748 (2016).
71. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
72. Minelli, C. et al. The use of two-sample methods for Mendelian randomization analyses on single large datasets. *Int J. Epidemiol.* **50**, 1651–1659 (2021).
73. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
74. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R package for causal mediation analysis. *J. Stat. Softw.* **56**, 11 (2014).

Acknowledgements

We thank the participants and staff of the Lifelines cohort for their collaboration. We thank J. Dekens and J. Arends for management and technical support and K. Mc Intyre for English editing. We also thank the Genomics Coordination Center for providing data infrastructure and access to high-performance computing clusters, and M. van der Geest, M. Swertz and the MOLGENIS server for providing the platform for data access. This project was funded by the Netherlands Heart Foundation (IN-CONTROL CVON grants 2012-03 and 2018-27 to F.K., M.G.N., A.Z. and J.F.), Netherlands Organization for Scientific Research (NWO) (NWO Gravitation Exposome-NL (024.004.017) to J.F., A.K. and A.Z., NWO-VIDI 864.13.013 and NWO-VICI VI.C.202.022 to J.F., NWO-VIDI 016.178.056 to A.Z., NWO-VIDI 016.136.308 to R.K.W., NWO-VENI 194.006 to D.V.Z. and NWO Spinoza Prize SPI 92-266 to C.W.), European Research Council (ERC) (ERC Advanced Grant 2012-322698 to C.W., ERC Advanced Grant 2019-833247 to M.G.N., ERC Consolidator Grant 101001678 to J.F. and ERC Starting Grant 715772 to A.Z.), ZONMW MENORABEL Grant 733050814 to A.Z. and the RuG Investment Agenda grant Personalized Health to C.W. F.K. is also supported by the Noaber Foundation. J.F. and C.W. are also supported by the Netherlands Organ-on-Chip Initiative—an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of the Netherlands. L.C. is supported by the Natural Science Foundation of China (32270077), the De Cock-Hadders Foundation (20:20-13), the Natural Science Foundation of Jiangsu (BK20220709) and an NJMU starting grant (303073572NC21). D.W. holds a fellowship from the China Scholarship Council (CSC201904910478). R.K.W. is supported by the Seerave Foundation and the Dutch Digestive Foundation (16-14). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

F.K., C.W., A.Z. and J.F. conceptualized and managed the study. L.C., S.A.-S., D.W., H.E.A., D.V.Z., A.K. and A.V.V. generated the data. L.C., D.V.Z. and A.K. analyzed the data. L.C., F.K., A.Z. and J.F. drafted the manuscript. L.C., S.A.-S., D.W., H.E.A., D.V.Z., A.K., A.V.V., R.K.W., M.H.M., M.G.N., F.K., C.W., A.Z. and J.F. reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

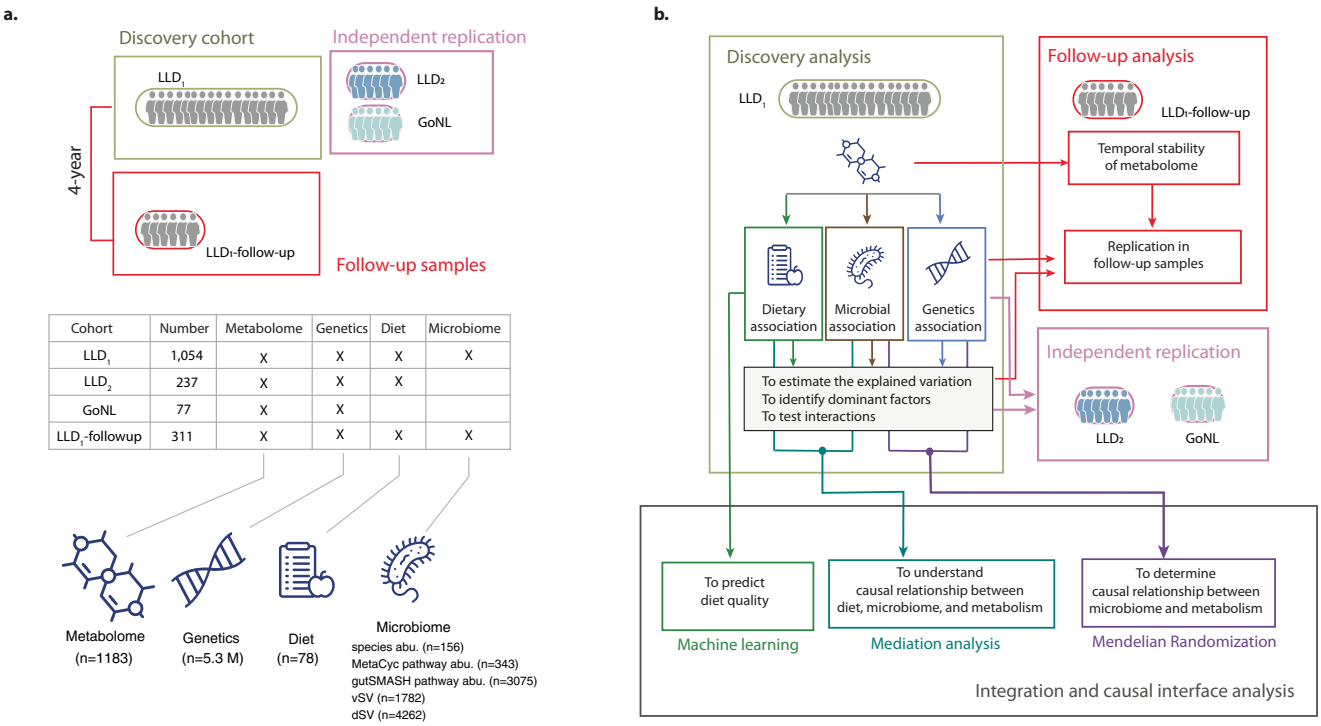
Extended data is available for this paper at https://doi.org/10.1038/s41591-022-02014-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-02014-8.

Correspondence and requests for materials should be addressed to Jingyuan Fu.

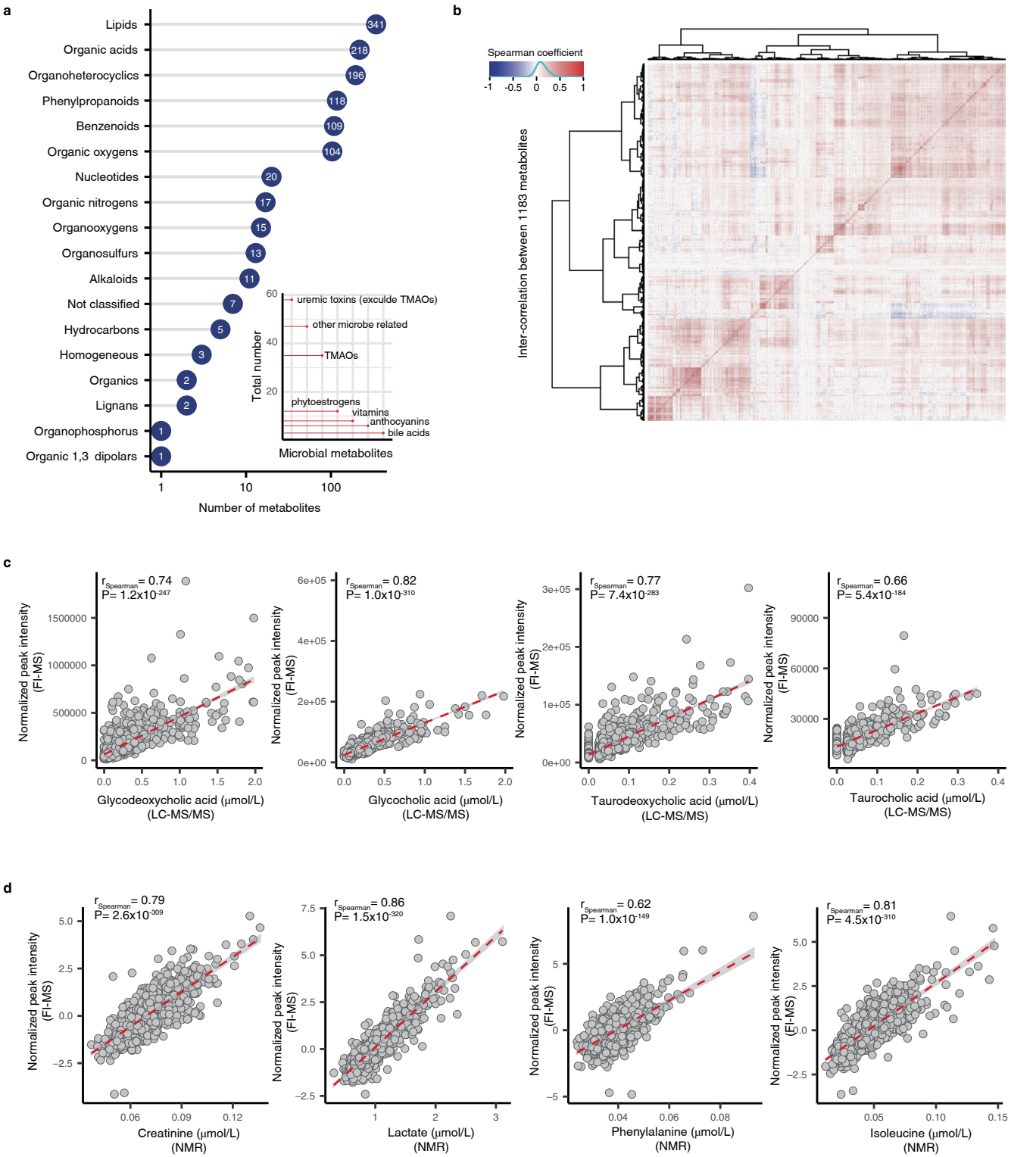
Peer review information *Nature Medicine* thanks Julian Griffin, Liming Liang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling editor: Michael Basson, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-022-02014-8



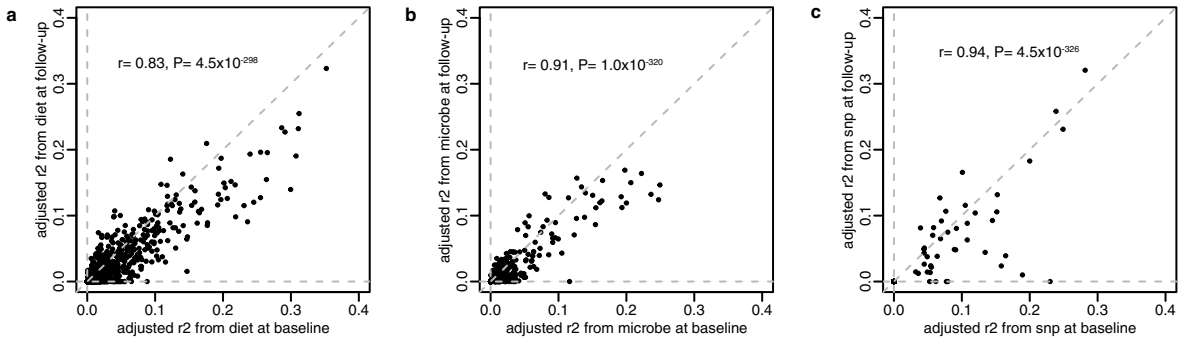
Extended Data Fig. 1 | Overview of study cohorts and analysis workflow. a. Summary of cohorts and datasets involved in the analyses. **b.** Detailed analysis workflow of the present study.

https://doi.org/10.1038/s41591-022-02014-8



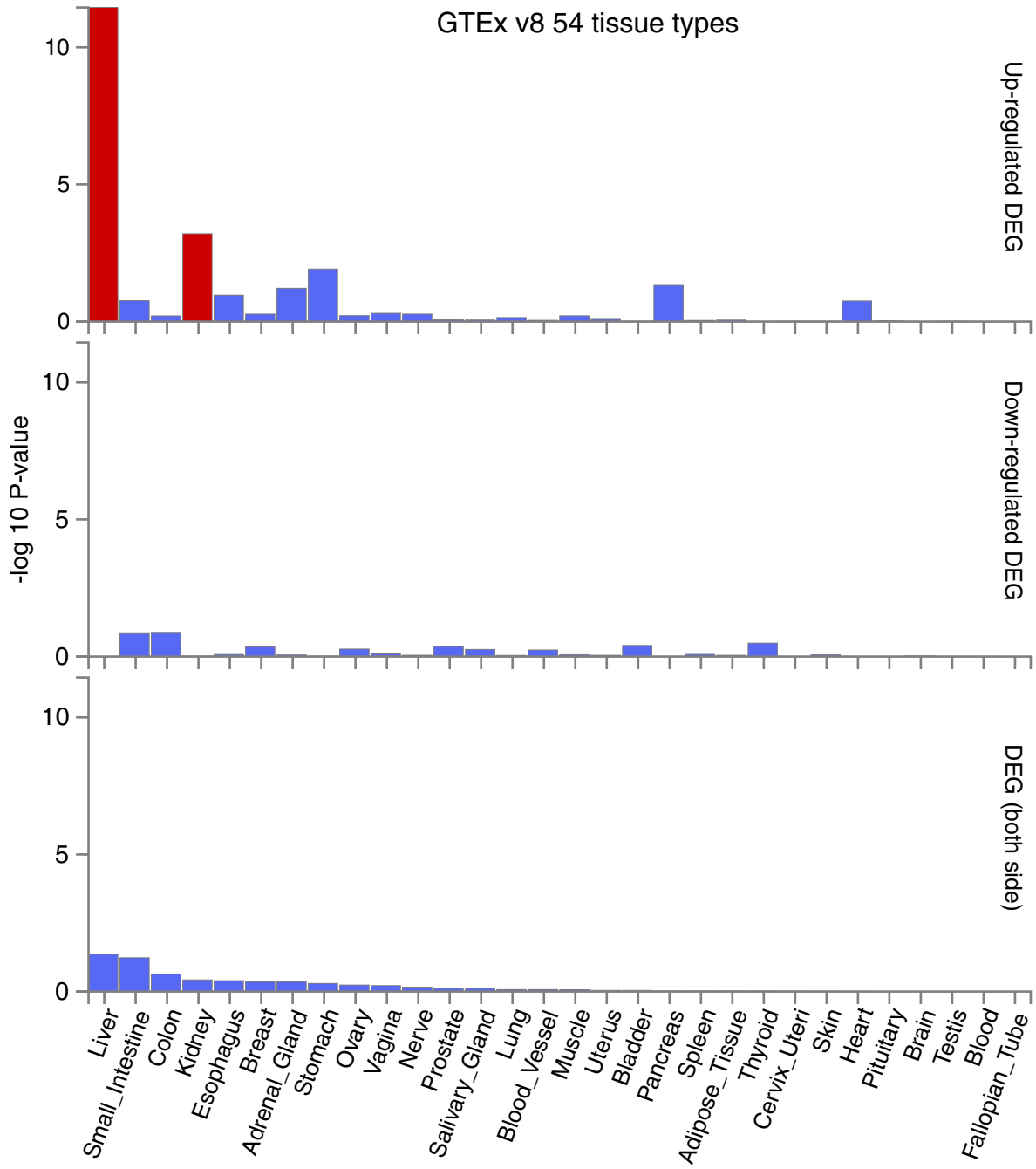
Extended Data Fig. 2 | Summary of plasma metabolites. a. Number of metabolites per metabolic categories based on annotation in the HMDB database. The x-axis indicates number of metabolites per category. The number of metabolites is also shown. The y-axis indicates metabolic categories. **b**. Inter-correlation between metabolites in the LLD baseline samples. Rows and columns represent metabolites. Color scheme represents the Spearman correlation coefficient. **c**. Comparison of metabolite concentrations between un-targeted FI-MS and LC-MS/MS platforms. The x-axis indicates metabolic abundance determined by LC-MS/MS. The y-axis indicates metabolic abundance determined by FI-MS. Each gray dot represents one sample. The red dashed line is the best fit line of linear regression with 95% confidence interval (CI). Spearman correlation coefficient between two measurements and the corresponding P value (two-sided) are shown. **d**. Comparison of metabolite concentrations between un-targeted FI-MS and NMR platforms. The x-axis indicates metabolic abundance determined by NMR and the y-axis indicates metabolic abundance determined by FI-MS. Each gray dot represents one sample. The red dashed line is the best fit line of linear regression with 95% CI. Spearman correlation coefficients between two measurements and the corresponding P value (two-sided) are shown.

https://doi.org/10.1038/s41591-022-02014-8



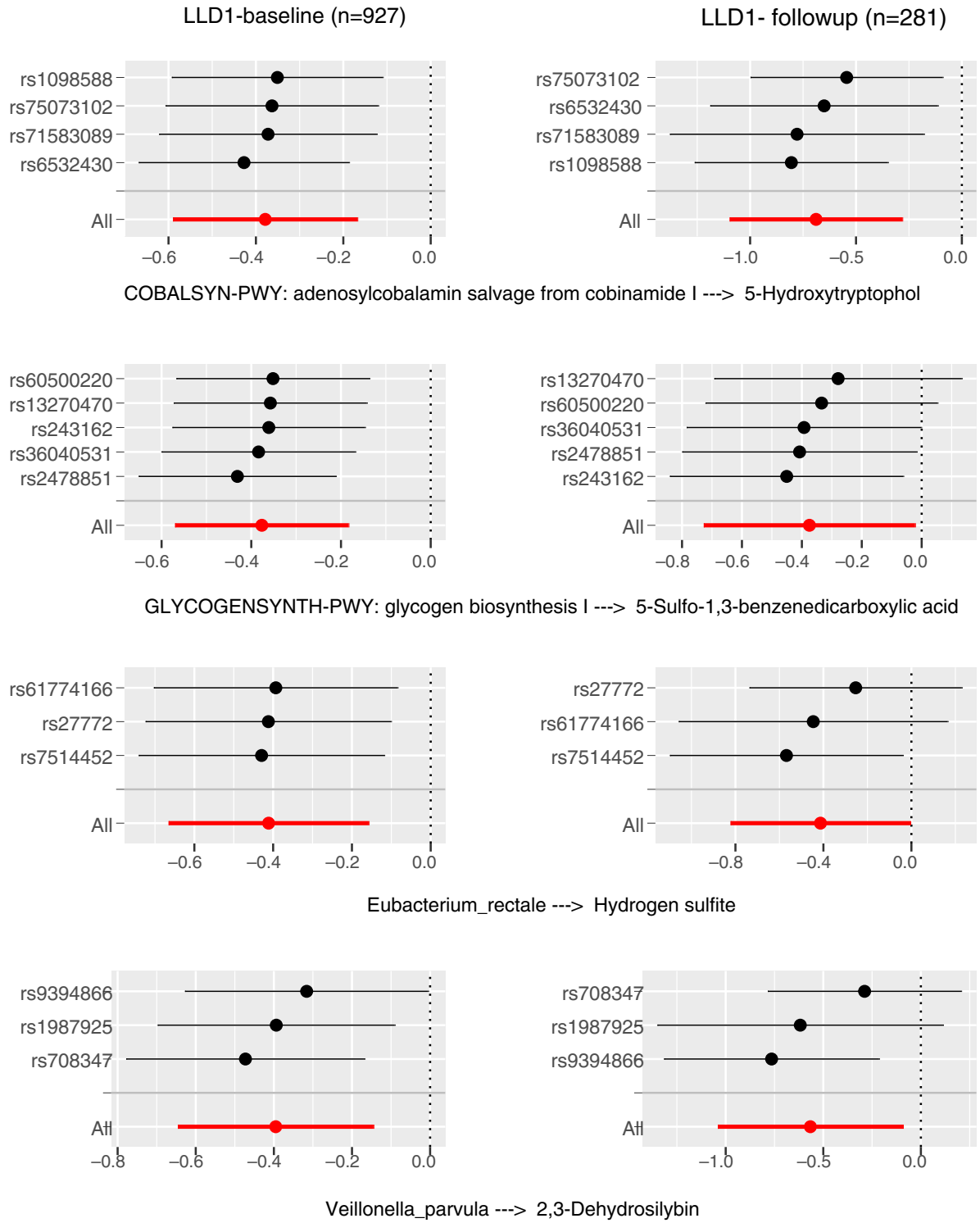
Extended Data Fig. 3 | Comparison of the proportion of metabolite variation explained by diet, microbiome and genetics at two time points. a. Proportion of the variation explained by diet. **b**. Proportion of the variation explained by microbiome. **c**. Proportion of the variation explained by genetics. Each dot represents a metabolite. The x-axis indicates explained variation at baseline and the y-axis indicates explained variation at follow-up. Similarity between the two time points was assessed using Spearman correlation (two-sided).

https://doi.org/10.1038/s41591-022-02014-8



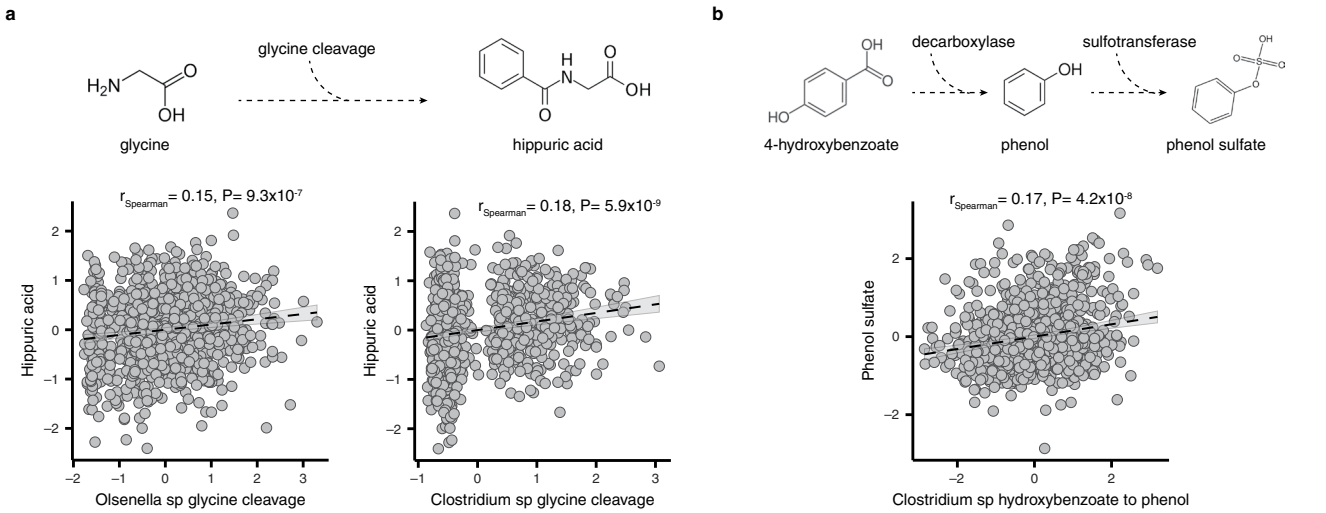
Extended Data Fig. 4 | Tissue-specific gene expression analysis with FUMA. All significant mQTLs observed in LLD_1 were used to check the enrichment of differentially expressed gene (DEG) sets in a specific tissue compared to all other tissue types based on GTEx (version 8). Red bars represent liver and kidney. Blue bars represent other tissues. The x-axis is the different tissue type and the y-axis indicates the significance of the differential abundance of the gene in one tissue compared to other tissue types in terms of -log10 transformed P value (Fisher's exact test, two-sided).

https://doi.org/10.1038/s41591-022-02014-8



Extended Data Fig. 5 | Leave-one-out sensitivity analysis for significant MR linkages. Forest plots of MR leave-one-out sensitivity results (IVW method) for four significant bi-directional MR linkages in LLD_1 and LLD_1-follow-up. Dots represent the estimated effect size. Bars represent 95% confidence intervals. The x-axis indicates effect size of MR.

https://doi.org/10.1038/s41591-022-02014-8



Extended Data Fig. 6 | Metabolites associated with gutSMASH pathways. **a**. Correlation between the microbial glycine cleavage pathway and plasma hippuric acid levels. **b**. Correlation between the microbial hydroxybenzoate to phenol pathway and plasma phenol sulfate levels. The upper part of each panel indicates the chemical transformation of the gutSMASH pathways. The lower part indicates the correlation between the pathway abundance (x-axis) and the plasma level of the corresponding metabolite (y-axis). Each dot represents one sample. The dark dashed line is the best fitted line of linear regression with 95% CI. Spearman correlation coefficient and the corresponding P value (two-sided) are also shown.

nature portfolio

Corresponding author(s): Jingyuan Fu

Last updated by author(s): Aug 11, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specific software was used for data collection.
Data analysis	All data analyses were conducted using publicly available tools. For this study the following software was used: kneadData (v0.4.6.1), Bowtie2 (v2.1.0), MetaPhlAn2 (v2.7.2), HUMAnN2 (v0.10.0), gutSMASH*, SGV-Finder*, vegan (v2.5.5), glmnet (v2.0.16), TwoSampleMR (v0.5.5), mediation (v4.5.0). Code used for data analyses is publicly available at: [https://github.com/GRONINGEN-MICROBIOME-CENTRE/Groningen-Microbiome/tree/master/Projects/LLDeep_plasma_GeneralMeta]. * No version number available.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All processed microbiome abundance data and full summary statistics of mQTLs are freely available via the MolGenis cloud (https://

geneticsresearch.molgeniscloud.org/menu/main/home), with an interactive browser of the top 100,000 mQTLs. The annotation of metabolites is based on the Human Metabolome Database (https://hmdb.ca, version 5). Tissue specific expression of genes is based on the Genotype-Tissue Expression (GTEx) database (https://gtexportal.org/home, version 8). The raw metagenomics sequence, metabolomics, and basic phenotype data (age, sex and BMI) are deposited at the EGA database with the study ID EGAS00001001704 (https://ega-archive.org/studies/EGAS00001001704), which include Dataset ID EGAD00001001991 for raw metagenomics sequencing (https://ega-archive.org/datasets/EGAD00001001991 and dataset ID EGAD00001006953 for raw metabolomics data (https://ega-archive.org/datasets/EGAD00001006953). However, the use of Lifelines data and materials must comply with the informed consent signed by Lifelines participants specifying that their collected data will not be used for commercial purposes. There is a minimal access procedure for the access of the EGA dataset that includes a contact address and a online data access form https://goo.gl/forms/TWHlrmbXaXNqWnnl2, which is very simple and is only intended to ensure that the data is being requested for research/scientific purposes only. Submitted data access forms will be evaluated by the data manager and Lifelines. For requests from verified academic parties, access will be granted within two weeks. There are no restrictions on downstream data re-use or authorship requirements. For requests from commercial parties, Lifelines will perform a pre-DPIA (Data Privacy Impact Assessment) to assess the risks of the proposed processing of personal data (e.g. purpose, storage, access, archiving, etc.) with respect to the GDPR subject rights. Based on the outcome of the pre-DPIA, Lifelines will decide whether sharing data with the commercial entity is allowed and/or whether additional measures have to be taken.
Genotype and metadata, including disease, medication and other clinical and lifestyle information, are however privacy sensitive. To ensure adherence to participant's privacy and informed consent, the rights of participants as described in the GDPR (EU privacy laws) and Lifelines biobank regulations, the complete genotype and phenotype data cannot be provided open-access and is only available from the Lifelines under controlled-access in a secure Lifelines Workspace or High Performance Cluster (HPC) environment. As Lifelines is a non-profit organization dependent on (governmental) subsidies, a fee is required to cover the costs of controlled data access and supporting infrastructure.
In brief, the step-by-step data access procedure is as follows: 1) Data is requested by filling the application form to request "Available Lifelines-data" at https://www.lifelines.nl/researcher/how-to-apply/apply-here; 2) Lifelines will evaluate project proposals to ensure compliance with the Lifelines data access policy, informed consent of Lifelines participants and the GDPR and that the data is being requested for non-commercial research; 3) Upon approval, Lifelines will send Data and Material Transfer Agreement (DMTA) contracts to the applicants; and 4) After the required contracts are signed, Lifelines will provide access to data via the Workspace or HPC and link the raw and processed DMP sequencing data to the Lifelines phenotypes. Lifelines strives to accomplish steps 2–4 at 2-weeks per step, assuming that no extra actions by the applicant or Lifelines are required.
The fee for data access on the HPC is €3,500 for one year and the fee for the Lifelines Workspace environment is €4,500 for one year, or less for shorter periods of time. There are no restrictions on downstream re-use of aggregated, non-identifiable results (as approved by Lifelines), nor are there authorship requirements, but Lifelines does request that it is acknowledged in publications using these data.
The data access policy, data access fees and an example DMTA (which includes details on how to acknowledge the use of Lifelines data in publications) are described in detail at https://www.lifelines.nl/researcher/how-to-apply. Note, data access for replication can be arranged via the Lifelines. Lifelines will not charge an access fee for controlled access to the full dataset used in the manuscript (including phenotype and sequencing data) for a period of 3 months, for the specific purpose of replication of the results presented in the current manuscript. Researchers interested in such a replication study can contact Lifelines at research@lifelines.nl. Further information can be obtained from Lifelines at https://www.lifelines.nl/researcher/how-to-apply/information-request or by contacting Lifelines at research@lifelines.nl.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The study includes both sexes, which is balanced as much as possible and with the male to female ratio of 41.8 to 58.2. Sex was included as a covariant for association analyses, and no sex-specific analysis was performed.
Population characteristics	The study has included the population-based cohorts (LLD1, LLD2 and GoNL) from the Netherlands, n=1138, 58.20% female, the mean age (SD) of participants is 45.04 (SD 13.60) years and their mean BMI is 25.26 (SD 4.18).
Recruitment	All participants cohorts (LLD1, LLD2, GoNL) are part of the prospective, population-based LifeLines cohort. From April to August 2013, all participants registered at the LifeLines Research Site in Groningen were invited to participate in the study, a study with deep omics profiling in addition to the regular LifeLines programme. All participants were collected without any specific selection. Most participants were from the north of the Netherlands and thus the reported results could be region-specific.
Ethics oversight	All participants signed an informed consent form prior to sample collection. The Medical Ethical Commitee of the University Medical Center Groningen (UMCG) has approved the study with reference number M12.113965. The Medical Ethical Committee (in Dutch: Medisch Ethische Toetsingsingscommissie or METc) of UMCG evaluates protocols for scientific research involving human beings. Such evaluation was given a legal context when on 1 December 1999, the Law on Medical Scientific Research involving Human Beings (WMO) took effect. The METc is authorized to evaluate research that is conducted by the UMCG. For any questions, please contact: +31 50 – 361 4204

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In total, this study includes 1,053 individuals of the Lifelines-DEEP (LLD1), 237 individuals from the LLD2 and 77 individuals from the GoNL cohorts, for whom we had collected extensive phenotypic datasets. Moreover, 311 individuals from the LLD1 cohort were followed up 4 years later. In order to ensure the analysis power, the study includes as much as subjects as possible. Thus no sample size calculation was performed. Detailed cohort description and sample size are shown in the Extended Data Fig 1.
Data exclusions	Only samples with missing metagenomics, metabolomics, genetics or dietary habits were excluded in analyses.
Replication	The LLD1 cohort served as a discovery cohort. Two independent replications were done using two independent cohorts (LLD2 and GoNL), respectively. One non-independent replication was done using the follow-up samples of the LLD1 cohort, which were profiled again 4 years later. Detailed replication scheme is shown in Extended Data Fig 1.
Randomization	This is human cohort-based analysis. The sample collection and sequencing were performed in a random order. No extra randomization was done for this study.
Blinding	This study is a human cohort based, observational study. Thus no blinding was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging