Dominant drivers of the human plasma metabolome

Jingyuan Fu (j.fu@umcg.nl)
University Medical Center Groningen https://orcid.org/0000-0001-5578-1236

Lianmin Chen
Nanjing Medical University & University Medical Center Groningen https://orcid.org/0000-0003-0660-3518

Sergio Andreu-Sánchez
University Medical Center Groningen

Daoming Wang
University Medical Center Groningen https://orcid.org/0000-0003-4623-8527

Hannah E. Augustijn
University Medical Center Groningen

Daria Zhernakova
University Medical Center Groningen

Alexander Kurilshikov
University of Groningen, University Medical Center Groningen https://orcid.org/0000-0003-2541-5627

Arnau Vich Vila
University of Groningen and University Medical Center Groningen

Rinse Weersma
University of Groningen and University Medical Center Groningen

Marnix Medema
Wageningen University https://orcid.org/0000-0002-2191-2821

Mihai Netea
Radboud University Nijmegen Medical Centre https://orcid.org/0000-0003-2421-6052

Folkert Kuipers
University Medical Center Groningen https://orcid.org/0000-0003-2518-737X

Cisca Wijmenga
University Medical Centre Groningen https://orcid.org/0000-0002-5635-1614

Alexandra Zhernakova
University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands

Article

Keywords: population cohort study, plasma metabolites, gut microbiome, genetics, diet, host-microbiome interactions

Posted Date: July 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-688716/v1

Version of Record: A version of this preprint was published at Nature Medicine on October 10th, 2022. See the published version at https://doi.org/10.1038/s41591-022-02014-8.

Abstract

The human plasma metabolome contains thousands of metabolites that are dependent on an individual's diet, genetics and gut microbiome, and on their interactions. By assessing 1,183 plasma metabolites in 1,368 extensively phenotyped individuals from the LifeLines-DEEP and GoNL cohorts, we quantified the proportion of inter-individual variation of the plasma metabolome explained by different factors and characterized 605 metabolites that are dominantly driven by diet, 85 driven by the gut microbiome, and 56 driven by genetics. We also developed a model that reflects the quality of an individual's diet using 76 mostly diet-driven metabolites. By applying Mendelian randomization (MR) and mediation analyses, we reveal putative causal relationships between diet, gut microbiome and metabolites. For example, MR supports the causal effect of *Eubacterium rectale* in decreasing plasma levels of toxic p-cresol and p-cresol sulfate, two metabolites related to cardiometabolic risk and chronic kidney disease. We further followed-up the plasma metabolome profile in 311 individuals after 4 years and observe that the metabolites with more variance explained by genetics, microbiome or diet are also more temporarily stable. Altogether, our characterization of the dominant drivers of plasma metabolites can help in the design of therapeutic approaches that target the diet, human genome, or gut microbiome to drive a healthy metabolome.

Introduction

The plasma metabolome represents a functional readout of the 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 the human metabolism is crucial for understanding the origin and determinants of plasma metabolites and for the eventual design of intervention strategies for a healthy metabolome.

Inter-individual variations in the human plasma metabolome have already been linked to genetics, diet and the gut microbiome in several human cohort–based studies [1–4]. For instance, a reference map of potential determinants of the human serum metabolome was established in 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 Response to Dietary Composition Trial (PREDICT1) assessed the impact of diet and microbiome on host metabolism in 1,098 individuals from the UK and the USA 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 gut microbiome are highly heterogeneous between different countries, we aimed to identify dietary, genetic and microbial determinants of plasma metabolites in the Dutch population and to characterize the main drivers of inter-individual metabolic variability. 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 311 LLD individuals who were followed-up after 4 years [7] (**Figure S1A**). For each participant, we had information on the gut microbiome, genetic background and dietary habits (**Figure S1A**). Notably, our metabolic data contained more than 800 metabolites that have not been studied previously [1–4], which can expand our understanding of plasma metabolome. In addition, we have assessed whether diet-associated metabolites can be used predict an individual dietary quality score reflecting diet–disease relationships [8] and examined what kinds of novel insights genetics-associated metabolites provide into the molecular pathways of complex diseases. Importantly, as potential causal relationships among metabolites, diet and microbiome remain largely unexplored, metabolites associated to multiple factors offered us an opportunity to infer their underlying causality using Mendelian randomization (MR) and mediation analyses [9,10]. Finally, we also assessed the temporal stability of the metabolites that are driven by genetics, diet and the gut microbiome.

Results

Untargeted plasma metabolites in Dutch cohorts

We applied liquid chromatography mass spectrometry (LCMS) to profile 1,679 fasting plasma samples from 1,368 individuals of the LLD [5] and GoNL [6] cohorts for whom we had collected extensive phenotypic datasets (**Figure S1A**, **Table S1**). In detail, dietary habits, genetics, the gut microbiome and plasma metabolome information were available for 1,054 LLD participants ($LLD_{baseline}$), with 331 of them also followed-up 4 years later ($LLD_{follow-up}$). In addition, we included two replication cohorts: 237 LLD participants for whom we had dietary habits, genetics and plasma metabolome information ($LLD2_{replication}$) and 77 GoNL participants for whom only genetics and plasma metabolome information were available ($GoNL_{replication}$). Our untargeted metabolomics approach quantified plasma levels of 1,183 metabolites (**Figure S1B**, **Table S2**), covering a wide range of lipids, organic acids, phenylpropanoids, benzenoids and other metabolites. Notably, our metabolome dataset contains more than 800 metabolites that were not assessed by previous studies [1–3, 11], including numerous lipids and lipid-like molecules that were not covered sufficiently in the previous blood metabolite blueprint [3,12] (**Table S2**). We validated the levels of some metabolites, such as creatinine, lactate, phenylalanine and isoleucine, whose abundance levels had been profiled previously using nuclear magnetic resonance [13] ($r_{Spearman}$>0.62, **Figure S1C**). We also observed weak (absolute $r_{Spearman}$<0.2) correlations among the 1,183 metabolites (**Figure S1D**), indicating that the levels of these metabolites were largely independent of each other. This meant that data reduction was not essential, and all metabolites were subjected to subsequent analyses. The microbiome and dietary analyses are further explained in the Methods section.

Inter-individual metabolome variations are dominantly driven by diet and the gut microbiome

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 that can be explained by these factors for the whole plasma metabolome profile and for the individual metabolites separately. We have

detailed information on 78 dietary habits (**Table S3**), 5.3 million human genetic variants and the abundances of 156 species and 343 MetaCyc pathways for each individual of the $LLD_{baseline}$ cohort. Diet, genetics and gut microbiome could explain 8.7%, 4.5%, and 11.6%, respectively, of the inter-individual variations in the whole plasma metabolome (False discovery rate $(FDR)_{PMAV}$<0.05, Fig. 1A, **Table S4**), whereas intrinsic factors (age, sex, BMI and smoking) could explain 4.9% of the variance. Together, these factors contributed 26.2% of the variation in the plasma metabolome (Fig. 1A).

We further assessed the association of diet, genetics and microbiome with individual metabolites and observed 2,854 associations with dietary habits (**Table S5**), 73 with genetic variants (metabolite quantitative trait loci (mQTLs), **Table S6**), 1,373 with gut bacterial species (**Table S7**) and 2,839 with bacterial MetaCyc pathways (**Table S8**) (Methods). 774 metabolites were significantly associated with at least one factor (**Tables S5-8**). We further assessed the proportion of variance of each metabolite that was explained by these factors. In general, inter-individual variations in 746 metabolites could be significantly explained by at least one of the three factors (FDR_{F-test}<0.05, **Table S9**). In detail, dietary habits contributed 0.4%–34% of the variance in 690 metabolites, microbial abundances contributed 0.6%–26% of the variance in 198 metabolites and genetic variants contributed 3%–28% of the variance in 62 metabolites (adjusted r^2, FDR_{F-test}<0.05, **Table S9**).

The inter-individual variations in 605 metabolites were dominantly explained by diet, 85 by the gut microbiome and 56 by genetics (Fig. 1B, **Table S9**). For instance, diet explained more than 20% of the variance in 13 metabolites (FDR_{F-test}<0.05, **Table S9)**, of which 10 are food components based on their HMDB annotation [14]. Similarly, of the 85 metabolites whose variation was dominantly explained by the gut microbiome, 27 were defined as microbial-related metabolites (including 16 uremic toxins). Furthermore, out of 56 metabolites whose inter-individual variations were dominantly driven by genetic variants, 12 are amino acids and 18 are lipids.

Even though the inter-individual variations of most metabolites seemed to be dominantly driven by one type of factor, variations in 200 metabolites were significantly attributed to more than one factor (Fig. 1B), including seven metabolites associated with both genetics and microbiome and 156 metabolites associated with both diet and microbiome. For example, genetics and microbiome explained 6% and 5%, respectively, of the variation of plasma 5'-carboxy-gamma-chromanol (Fig. 1C), a dehydrogenated carboxylate product of 5'-hydroxy-r-tocopherol [15] that reduces cancer and cardiovascular risk [16]. Another example is hippuric acid, a uremic toxin that can be produced as a product of bacterial conversion of dietary proteins [17], with 12% of its variation was explained by diet and 12% explained by microbiome (Fig. 1C). Taken together, our analysis highlights that variations in most plasma metabolites were mainly explained by one type of factor, particularly diet or microbiome, but the variations in 200 metabolites were significantly attributed to multiple factors, with 156 (78%) associated with both diet and microbiome (Fig. 1B), suggesting the importance of diet–microbiome interactions in regulating the human metabolome.

Temporal stability of the metabolites over time

Unlike our genetics, the gut microbiome and individual dietary habits can change over time. If these two driver factors are unstable, this can alter metabolite levels as well. We compared the metabolome data from 311 samples ($LLD_{follow-up}$) collected 4 years after the baseline measurements in the same individuals. Here we saw that the human plasma metabolome can shift over time, observing a significant difference in the 2nd PC ($P_{PC1\ paired\ Wilcoxon}$=0.1 and $P_{PC2\ paired\ Wilcoxon}$=1.3x10^{-5}, Fig. 2A). We also observed that temporal stability can vary substantially between different metabolites (**Table S10**). Interestingly, the temporal correlation of the microbiome-driven metabolites was similar to that of genetics-driven metabolites ($P_{Wilcoxon}$=0.51, Fig. 2B), which is in line with our recent observation that the gut microbiome of these individuals is relatively stable between baseline and 4-year follow up [9]. We further show that the microbial and genetic contributions to inter-individual plasma metabolite variations were highly comparable between baseline and follow-up for microbiome and genetics, respectively (**Figure S2**, **Table S9**). The temporal correlation of diet-driven metabolites was significantly lower than that of the microbiome- and genetics-driven metabolites ($P_{Wilcoxon}$<3.4x10^{-5}, Fig. 2B). Our data also showed a positive correlation between stability and explained variation, i.e. the more variation explained, the more stable a metabolite was over time (Fig. 2C).

Metabolome reflects diet quality score

We observed 2,854 significant associations ($FDR_{Spearman}$<0.05) between 74 dietary factors and 726 metabolites (Fig. 3A, **Table S5**, see Methods). Associations to 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 = 1.6x10^{-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 = 1.5x10^{-76}, Fig. 3C). This association is expected as this compound is particularly present in smoked fish according to HMDB annotation [14]. Additionally, we detected associations between dietary factors and metabolic biomarkers of some diseases. For example, 1-methylhistidine is a biomarker for cardiometabolic diseases, including heart failure [18], that is enriched in meat, and we observed significant associations of 1-methylhistidine with meat ($r_{Spearman}$=0.12, P = 7.2x10^{-5}) and fish intake ($r_{Spearman}$=0.11, P = 3.1x10^{-4}) and a lower level of L-methylhistidine in vegetarians ($r_{Spearman}$=-0.15, P = 9.7x10^{-7}, Fig. 3D).

Given the relationship between diet, metabolism and human health, we hypothesized that the plasma metabolome could predict the diet quality. For each of the Lifelines participants, we have a Lifelines Diet Score (DS) that is constructed based on food frequency questionnaire (FFQ) data and reflects relative diet quality based on diet–disease relationships [8]. To build a metabolic model to predict individual diet quality, we used the $LLD_{baseline}$ (n = 1,019) as the training set and the additional LLD2 (n = 230) cohort as the validation set. The resulting metabolic model included 76 metabolites (51 primarily driven by dietary habits) and significantly predicted the DS in the validation set ($r^2_{adjusted}$ = 0.27, P_{F-test}=3.5x10^{-5}, Fig. 3E).

Novel human genetic determinants of plasma metabolites

Genetic determinants of plasma metabolites may provide functional insights into the etiologies of complex diseases. After correcting for age, sex, smoking and oral contraceptive use, QTL mapping in $LLD_{baseline}$ identified 63 study-wide independent mQTLs ($P_{Spearman}<4.2x10^{-11}$, clumping $r^2 = 0.05$, clumping window = 500kb, Fig. 4A, **Table S6**). Other factors, i.e. dietary habits and gut microbiome, can interfere with mQTL identification by inducing non-genetic variation. Regressing out these factors may therefore increase QTL mapping power, but this has rarely been done because most cohorts lack this information [19]. We performed a stepwise-regression to correct for dietary habits, diseases, medication and gut microbiome in our QTL mapping, and this yielded an additional 10 independent mQTLs ($P < 4.2x10^{-11}$, Fig. 4A, **Table S6**). Importantly, all 73 mQTLs between 63 metabolites and 57 mQTLs could be repeatedly observed in $LLD_{follow-up}$ or replicated in the two replication datasets ($LLD2_{replication}$ and $GoNL_{replication}$, **Figure S3**, **Table S6**), indicating that these mQTLs are robust and the genetic impact on metabolites was stable over time. Using FUMA (Functional Mapping and Annotation of GWAS) [20], we found that the identified mQTLs are enriched for genes expressed in liver and kidney (**Figure S4**) and related to metabolic phenotypes (**Table S6**).

Notably, upon a systematic search in the GWAS catalog [21] and PubMed, 67 of the 73 mQTLs appeared to be novel (**Table S6**). These mQTLs could point to disease susceptibility, including cardiometabolic and chronic kidney disease driver genes, as can be seen in the pleotropic mQTL effects observed at *SLCO1B1, FADS2, KLKB1* and *PYROXD2* (**Table S6**). For example, eight mQTLs (related to eight metabolites, six of them lipids) were observed for four independent SNPs (rs4149056, rs77289848, rs4149067 and rs67981690) in *SLCO1B1*, which encodes the solute carrier organic anion transporter family member 1B1. Expression of SLCO1B1 is specific to the liver, where this transporter is involved in transport of various endogenous compounds and drugs, including statins [22], from blood into the liver. SLCO1B1 has also been linked to plasma levels of fatty acids and to statin-induced myopathy [23], and *SLCO1B1* variant rs4149056 has been linked to statin response and risk of heart failure [24]. We observed that rs4149056 was associated with plasma 5'-carboxy-gamma-chromanol ($r_{Spearman}=0.24$, $P = 2.8x10^{-13}$, Fig. 4B), a biomarker for cardiovascular risk [15], and with plasma 3-O-protocatechuoylceanothic acid ($r_{Spearman}=0.22$, $P = 8.7x10^{-12}$, Fig. 4C), which is also relevant for cardiovascular risk [25].

The mQTLs we identified enrich our understanding of genetic control of human metabolism. For instance, the strongest association we found was between the caffeine metabolite 5-acetylamino-6-formylamino-3-methyluracil (AFMU) and the N-acetyltransferase 2 gene (*NAT2*) tag SNP (rs1495741) ($r_{Spearman}=0.56$, $P = 3.8x10^{-77}$, Fig. 4D), and it showed strong linkage disequilibrium ($r^2 = 0.98$) with recently reported SNP rs35246381 that is associated with urine AFMU [26]. AFMU is a direct product of NAT2 activity and has been associated with bladder cancer risk [27]. As a key enzyme in the conjugation of xenobiotics and drugs with various hydrazine or arylamine structures, NAT2 is also able to bioactivate many known carcinogens [28]. Notably, rs1495741 has been recognized as a risk-locus for bladder cancer [29], which suggests that the interaction between this genetic variant and caffeine metabolism may be implicated in bladder cancer risk.

Causal role for the microbiome in determining the plasma metabolome

Through their distinct metabolic activities, gut microbes can contribute to variations in the host's plasma metabolome. We established 4,212 associations between 208 metabolites and 314 microbial factors (114 species and 200 MetaCyc pathways) ($FDR_{baseline}$<0.05, $P_{follow-up}$<0.05, **Table S7-8**). Interestingly, many of the metabolites that were associated with microbial species and MetaCyc pathways are also known to be gut microbe—related, based on their HMDB annotations [14]. For instance, we observed 919 associations with 25 uremic toxins, 142 associations with thiamine (vitamin B1) and 117 associations with 5 phytoestrogens (FDR < 0.05, **Table S7-8**). Uremic toxins and thiamine derived by gut microbes have been shown to be related to various diseases, including chronic kidney and cardiovascular diseases [30]. 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 [31]. Our data thus provides functional insights into the role of gut microbial composition in host health and disease, but experimental confirmation is still needed to substantiate these findings.

To assess whether gut microbiome composition causally contributes to plasma metabolite levels, we carried out bi-directional MR analyses (see Methods). Here we focused on the 37 microbial features that were associated with at least three independent genetic variants at $P < 1x10^{-5}$ and on 45 metabolites with significant associations to the microbiome. At FDR < 0.05 (corresponding to a $P = 5.2x10^{-3}$ obtained from the inverse variance weighted (IVW) MR test), we observed seven causal relationships at baseline between five microbial abundances and six metabolites in the microbiomes to metabolites direction, but not in the opposite direction (**Table S11-13**). Notably, five of these seven causal relationships were related to four uremic toxins (p-cresol, p-cresol sulfate, 5-hydroxytryptophol and 4-ethyl-2-methoxyphenyl-oxidanesulfonic acid) that can be produced by gut microbes. Furthermore, two of these uremic toxins (p-cresol and p-cresol sulfate) were related to *Eubacterium rectale* (P < 0.05, Fig. 5A-B), a core gut commensal species [32] that is highly prevalent (presence rate = 97%) and abundant (mean abundance = 8.5%) in both our cohort and other populations [33–35]. 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 [36], and a reduced abundance of this species has been observed in subjects with inflammatory bowel disease [33,37] and colorectal cancer [38] compared to healthy controls. As uremic toxins, p-cresol and its sulfate are involved in the etiologies of chronic kidney disease and cardiometabolic diseases [30]. Importantly, by further annotating the genomes of *E. rectale*, we characterized genes that encode a sulfatase and 4-hydroxybenzoyl-CoA reductase that are responsible for consecutive solvolysis of p-cresol sulfate into p-cresol and for subsequent reduction of p-cresol into benzoyl-CoA (Fig. 5C). Our results thus reveal a potential new beneficial effect of *E. rectale* through degradation of uremic toxins.

To further zoom in on the metabolic potential of individual species, we applied newly developed pipelines to identify microbial primary metabolic gene clusters (gutSMASH pathways) [39] and microbial genomic structural variants (SVs) [40]. These two tools profile microbial genomic entities that are implicated in metabolic functions. By associating 1,183 metabolites to 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 ($FDR_{baseline}$<0.05, $P_{follow-up}$<0.05, **Table S14-16**). 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 microbial uremic toxin biosynthesis pathways, including the glycine cleavage pathway (in species *Olsenella sp* and *Clostridium sp*) and the hydroxybenzoate to phenol pathway (in *Clostridium sp*), responsible for hippuric acid and phenol sulfate biosynthesis were associated with the hippuric acid (*Olsenella sp*: $r_{Spearman}$=0.15, P = $9.3x10^{-7}$; *Clostridium sp*: $r_{Spearman}$=0.18, P = $5.9x10^{-9}$) and phenol sulfate ($r_{Spearman}$=0.17, P = $4.2x10^{-8}$, **Figure S5A**) levels measured in plasma, respectively ($FDR_{baseline}$<0.05 and $P_{follow-up}$<0.05, **Figure S5B**).

Diet–microbiome interactions involved in control of the plasma metabolome

Next, we carried out a mediation analysis to investigate the links between diet, 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 mediation effects of microbiome and metabolites for diet (see Methods). This established 195 mediation linkages: 185 for dietary impact on the microbiome through metabolites and 10 for dietary impact on metabolites through the microbiome ($FDR_{mediation}$<0.05 and $P_{inverse\ mediation}$>0.05, Fig. 6A-B, **Table S17**). Most of these linkages were related to the impact of coffee and alcohol on microbial metabolic functionalities (Fig. 6A). Coffee contains various hydroxycinnamic acids, including ferulic acid, that can be catabolized by gut microbes and may play a beneficial role in alleviating features of diabetes [41,42]. We observed that ferulic acid can mediate the impact of drinking coffee on a vSV of *Ruminococcus sp.* (300–305 kb) ($P_{mediation}$=$2.2x10^{-16}$, Fig. 6C) that encodes an ATPase component that can be activated by ferulic acid [43]. 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 ($P_{mediation}$=$2.2x10^{-16}$, Fig. 6D), an important membrane protein in driving the ATP synthesis essential for all bacterial metabolic activities [44]. Of the dietary impacts on metabolites through the microbiome (Fig. 6B, **Table S17**), one interesting example is two *Eubacterium hallii* vSVs (734–736 and 754–756 kb) that encode an ATPase responsible for transmembrane transport of various substrates [45]. These two *E. hallii* vSVs mediated the effect of white wine consumption on plasma levels of pipecolic acid ($P_{mediation}$=$2.0x10^{-3}$, Fig. 6E). Pipecolic acid present in human plasma is mainly derived from the catabolism of dietary lysine by intestinal bacteria rather than by direct intake from diet [46], and pipecolic acid levels are evaluated in chronic liver disease [47]. Notably, lysine is one of the most abundant amino acids in white wine [48]. Taken together, our data provide potential mechanistic underpinnings for diet–metabolite and diet–microbiome interactions.

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 identified a series of dietary, genetic and microbial determinants of plasma metabolite levels. Our results clearly demonstrate that most metabolites were driven by just one of these three factors, but the inter-individual variations of 200 metabolites were attributed to multiple factors, particularly diet and gut microbiome. Metabolites that had more variation explained by a specific factor were also more stable over time. We then built a metabolite model that can robustly indicate an individual's diet quality. We also applied MR analysis to infer potential causal relationships between the gut microbiome and specific metabolites present in plasma and mediation analysis to reveal the directionality of impact among diet, metabolites and gut microbiome. Finally, our metabolome data enriches the human plasma metabolome with over 800 metabolites that have not been measured in previous genetics–diet–microbiome analyses using different LC-MS platforms.

Dietary compounds are the fundamental resources for the plasma metabolome, and a recent study illustrated that individual dietary habits can accurately predict the levels of some metabolites present in plasma [3], thereby 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 could use an individual's plasma metabolome to robustly predict their diet quality score, an index that reflects diet quality based on solid contemporary evidence on diet–disease relationships [8]. For now, our findings might be useful for guiding people towards better metabolic health through changes in their dietary habits.

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 [3,4,35,49,50], 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 [12]. 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 insights into diet–microbiome interactions in human metabolic health, as illustrated by several metabolites (e.g. phenol and pipecolic acid) that have previously been related to cardiometabolic and kidney diseases [30]. Notably, these mediation linkages mainly showed that the dietary impact on the microbiome can be mediated by metabolites, highlighting the pronounced selective power of dietary habits in shaping the gut microbiome through metabolites. 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 capacity are essential for confirming causality.

Apart from diet and the gut microbiome, human genetics also acts as a determinant of the plasma metabolome. With this unique metabolome dataset, we not only replicated six previously reported mQTLs, we also identified 67 mQTLs involving 22 loci not previously known to be associated with any trait. Many of these novel mQTLs could be identified as cardiometabolic and chronic kidney disease driver genes, as illustrated by the tissue-specific gene expression analysis and pleotropic mQTLs effects observed for *SLCO1B1*, *FADS2*, *KLKB1* and *PYROXD2*. Importantly, 10 of the 67 novel mQTLs were only discovered after adjustment for diet and microbial factors, which highlights the need to take these factors into account. We further used genetic variants as instruments in MR to infer causal relationships between the gut microbiome and metabolites. This showed that the microbiome may causally contribute to the levels of uremic toxins (p-cresol, p-cresol sulfate, 5-hydroxytryptophol and 4-ethyl-2-methoxyphenyl-oxidanesulfonic acid) that act as risk factors for chronic kidney disease and cardiometabolic diseases [30]. Interestingly, we further confirmed this by annotating the genome of *E. rectale*, where we found genes that encode sulfatase and 4-hydroxybenzoyl-CoA reductase, which are responsible for degrading p-cresol and its sulfate. The causal relationships between microbiomes and metabolites that we have established thus reveal potential metabolic functionalities of gut microbes.

Taken together, we have integrated multiple layers of omics information in a large prospective cohort in order to investigate the roles of diet–genetics–microbiome interactions in controlling human metabolism. The dietary, genetic and microbial determinants of plasma metabolites and the causal and mediation linkages we report are a comprehensive resource that can guide follow-up studies aimed at designing preventive and therapeutic strategies for human metabolic health.

Methods

Study cohort

The LifeLines-DEEP cohort (LLD, n=1,500) is a sub-cohort of the large prospective Lifelines cohort study from the north of the Netherlands [5,19]. The cohort is 58% female and 42% male, with a mean age of 45.04 years (sd=13.60). The mean BMI is 25.26 (sd=4.18), and 12% of participants are obese (BMI>30) [5]. All LifeLines participants signed an informed consent form prior to sample collection. Institutional ethics review board approval is available for LLD under reference number M12.113965. For this study, we involved 1,054 participants for whom detailed dietary habits, stool microbiome and plasma untargeted metabolomics are available. For 933 of them, genetic information is also available (**Table S1**), and 331 also have a 4-year follow-up data (**Table S1**).

Data generation and preprocessing

Plasma metabolome

Metabolite levels of fasting plasma samples were measured at General Metabolics, Inc., Boston, USA, using an untargeted platform, as described previously [51,52]. Plasma samples of study participants were collected and frozen at -80°C with EDTA. During extraction, plasma samples were thawed on ice, vortexed

and spun down. 20μL of plasma was combined with 180μL of 80% methanol and vortexed for 15 seconds. The samples were then incubated at 4°C for 1 hour to precipitate proteins and then spun for 30 minutes at 3,200 RCF. 100μL of supernatant was removed and used for Flow-Injection Time-of-Flight Mass Spectrometry (FIA-TOF) analysis.

In brief, FIA-TOF analysis was performed on a platform consisting of an Agilent 1260 Infinity II LC pump coupled to a Gerstel MPS autosampler (CTC Analytics, Zwingen, Switzerland) and an Agilent 6550 Series Quadrupole TOF mass spectrometer (Agilent, Santa Clara, CA, USA) with a Dual AJS ESI source operating in negative mode, as described previously [52]. The flow rate was 150 μL/min of mobile phase consisting of isopropanol and water (60:40, v/v) with 1 mM ammonium fluoride. For online mass axis correction, two ions in Agilent's ESI-L Low Concentration Tuning Mix (G1969-85000) were used. Mass spectra were recorded in profile mode from molecular weight (m/z) 50 to 1,050 with a frequency of 1.4 s for 2×0.48 min (i.e. each sample was measured twice and high consistency between the two injections was observed, r>0.9) using the highest resolving power (4 GHz HiRes).

All the mass spectrometry data processing and analysis was performed with MATLAB (Mathworks, Natick, MA, USA) using functions embedded in the bioinformatics, statistics, database and parallel computing toolboxes, as described previously [52]. The resulting data included the intensity of each mass peak in each analyzed sample. Peak-picking was carried out once for each sample from the total profile spectrum obtained by summing all single scans recorded over time and using wavelet decomposition as provided by the bioinformatics toolbox. In this procedure, a cutoff was applied to filter out peaks of less than 5,000 ion counts in the summed spectrum in order to avoid detection of features too low to deliver meaningful insights. Centroid lists from samples were then merged into a single matrix by binning the accurate centroid masses within the tolerance given by the instrument resolution. The resulting matrix lists the intensity of each mass peak in each sample analyzed. An accurate common m/z was recalculated with a weighted average of the values obtained from independent centroiding. To correct for temporal drift of metabolite intensity, we used a moving-median normalization: each ion intensity was adjusted based on the median intensity value of that ion in the 90 samples run before and after. After merging, 31,302 common ions were obtained. Of these, 1,183 were annotated (**Table S2**) based on accurate mass using a 1 mDa tolerance. Annotation was based on assumption that -H(-) and .F(-) are the possible ionization options. The annotated metabolites cover 18 chemical categories based on the Human Metabolome Database (HMDB) [14], 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 (**Table S2**). Finally, we estimated the effect of sample plate batch on metabolite level and detected no batch effects.

As we were interested in the impact of diet, genetics and gut microbiome on metabolites, we decided to adjust for these confounding factors because they can potentially influence metabolites but were not relevant to our research focus. To investigate potential confounding factors, we correlated the first 100 principal components of the 1,183 metabolites (accounting for 73% of the total metabolome variations)

with age, sex, BMI, smoking, 78 dietary habits, 39 diseases and use of 44 medications. Based on the correlation results (**Table S18**), we decided to correct for age, sex, smoking and oral contraceptive use. To adjust for confounding factors, we first log-transformed the metabolite abundances and then applied a linear regression model that included all 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 minutes after 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 (Qiagen; cat. 80204). After DNA extraction, fecal DNA was sent to the Broad Institute of Harvard and MIT in Cambridge, Massachusetts, USA, where library preparation and whole genome shotgun sequencing were performed on the Illumina HiSeq platform. From the raw metagenomic 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 Bowtie2 (v.2.1.0) [53,54].

Microbial taxonomic profiles were generated using MetaPhlAn2 (v.2.7.2) [55]. Microbial general pathways were determined using HUMAnN2 [56], 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 [57] (UniRef90), which were further mapped to microbial pathways from the MetaCyc metabolic pathway database [58,59]. 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 [60].

We applied the SGV-Finder pipeline [40] to classify structural variants (SVs) that are either completely absent from the microbial genome of some samples (deletion SVs, dSVs) or those whose coverage is highly variable across samples (variable SVs, vSVs). Prior to 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 [40]. 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 was inverse rank–transformed for subsequent analysis.

Metabolite-specific pathways were generated by using the gutSMASH algorithm [39]. 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 RPKMs (reads per kilobase of transcript per million reads mapped), and inverse-rank transformation was applied before subsequent analysis.

Genotype

Microarray genotype data for the LLD cohort was generated using the CytoSNP and ImmunoChip assays, as previously described [61]. Quality control checks on the LLD cohort were performed using the Haplotype Reference Consortium (v1.0) preparation checking tool (v4.2.3). We then uploaded the resulting VCF files to the Michigan Imputation Server [62]. 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 reference [63]. We further excluded SNPs that had imputation quality r^2<0.5, failed the Hardy-Weinberg equilibrium test ($P<1\times10^{-6}$), or had a call rate <95% or a minor allele frequency <5%. In total, we obtained genotype data for 7 million SNPs (Genome Build hg19) for all individuals.

Statistical analysis

Associations to dietary habits

To assess dietary associations to 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 (**Table S2**). The false discovery rate (FDR) was calculated using the Benjamini-Hochberg (BH) procedure [64].

Microbiome-wide associations

We previously reported that several medications can alter gut microbiome significantly, including proton pump inhibitors, antibiotics and laxatives [65,66]. 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 BH procedure.

Quantitative trait locus (QTL) mapping

This analysis involved 933 participants for whom there is genotype, plasma metabolome and phenotypic data. After adjusting for age, sex, BMI, smoking and oral contraceptive use, as described above, a Java-based pipeline (version 1.4nZ) [67] was applied for QTL mapping by calculating the Spearman correlation between SNP dosage and metabolite abundances. To investigate whether adjustment for other covariates could uncover more of these metabolite QTLs (mQTLs), we stepwise-adjusted for dietary habits, diseases, medications and gut microbiome composition. We considered metabolite associations with a P-value < 4.2×10^{-11} as significant, a threshold corresponding to a genome-wide significance cutoff of 5.0×10^{-8} corrected for 1,183 tested metabolites. We report all independent mQTLs (clumping variants with linkage disequilibrium r^2 <0.05 and a 500kb window [20]) with a P-value<4.2×10^{-11}. For genetics

associations to metabolite changes, we regressed out age and sex from metabolites and reported all independent mQTLs at a P-value<5×10^{-8}.

Tissue-specific gene expression analysis

Summary statistics of independent mQTLs were used for tissue-specific gene expression analysis using FUMA (Functional Mapping and Annotation of GWAS) [20].

Distance matrix–based variance estimation

We applied feature selection based on the permutational multivariate analysis of variance using distance matrices (PMAV) procedure to estimate the contributions of different factors to inter-individual variations of the whole plasma metabolome. Phenotypic and microbial contributions were estimated based on the 1,054 participants for whom plasma metabolites, stool microbiome and phenotypic data were available, while the genetic contribution was estimated based on the 933 participants for whom genotype was also available. We first used each phenotypic and microbial feature to estimate inter-individual metabolic variations using the *adonis* function from vegan (version 2.5.5) with 1000 times permutation. Only phenotypic and microbial features that can estimate inter-individual metabolic variations at a permutational FDR<0.05 were kept. For genetic variants, we used SNPs with significant mQTLs. To deal with the collinearity of selected features, we applied hierarchical clustering analysis based on feature inter-correlation distance matrix ($1-r^2$). Features were assigned into different clusters based on 70% dissimilarity, and the central feature in each cluster was selected as representative. All representative features were further included in PMAV to estimate the combined contribution to inter-individual metabolome variation.

Estimating variance of individual metabolites

To estimate the variance of each metabolite that is contributed by dietary, genetic and microbial features, we applied machine learning–based least absolute shrinkage and selection operator (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 variation 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 major drivers of each metabolite variation.

All 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 gives a minimum mean error from a 10-fold cross-validation in order to control for overfitting and to provide a conservative estimate of model performance. Finally, features selected by lasso are included in the linear model to estimate the variance

contributed by different factors, and the adjusted r^2 and the F-test P-value were recorded. FDR was calculated based on the BH procedure.

Lifelines diet quality score prediction

We first checked the Lifelines dietary score (DS) associations to each metabolite in 1,054 LLD samples and selected the significant metabolite features (P-value<0.05) for lasso regression, as described above. Metabolites selected by lasso were used to build the linear model, and 230 LLD2 samples with both DS and plasma metabolome available were used as validation. Adjusted r^2 and the P-value from F-test were reported to reflect the performance of prediction model.

Bi-directional Mendelian randomization analysis

To evaluate causal relationships between the gut microbiome and metabolites, we applied bi-directional Mendelian randomization (MR) analyses. We focused on the 37 microbial features associated with at least three independent genetic variants at $P < 1x10^{-5}$ in the baseline that could also be replicated in the follow-up samples with the same direction of association (P<0.05) and on the 45 metabolites with significant associations to the microbiome ($FDR_{baseline}$<0.05, $P_{follow-up}$<0.05 and P_{delta}<0.05). MR analysis was done using the TwoSampleMR (version 0.5.5) package with the inverse variance weighting test. We also report results from the Egger test. Leave-one-out sensitivity analysis was carried out to check whether the MR results were being driven by a single SNP (**Figure S6**).

Genomic annotation of Eubacterium rectale

We used Prokka [68] with default settings to annotate the eight isolated reference genomes of *E. rectale* [34], which resulted in a list of genes with functional annotations. We characterized the genes encoding sulfatase and 4-hydroxybenzoyl-CoA reductase that are responsible for p-cresol sulfate hydrolyzating into p-cresol and reduction of p-cresol, which is widely distributed in these genomes.

Bi-directional mediation analysis

For microbial features associated with both metabolites ($FDR_{baseline}$<0.05, $P_{follow-up}$<0.05 and P_{delta}<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, bi-directional mediation analysis was carried out using the *mediate* function from mediation (version 4.5.0) to infer the mediation effect of metabolites and microbiome for dietary impacts. FDR was calculated based on the BH procedure.

Principal component analysis (PCA)

The levels of all plasma metabolites were included in the PCA. 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.

Replication

To validate the robustness of our results, we replicated the significant metabolites associations to genetics and microbiome, our metabolite variance estimation and our MR results in corresponding omics datasets of the Genome of the Netherlands (GoNL) cohort [69], the LLD 4-year follow-up cohort and the LLD2 cohort (**Figure 1A**, **Table S1**).

Declarations

ACKNOWLEDGEMENTS

We thank the participants and staff of the LifeLines cohort for their collaboration, particularly B. Bolmer and S. Gerritsma for coordinating the LifeLines data. 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. This project was funded by the Netherlands Heart Foundation (IN-CONTROL CVON grant 2012-03 and 2018-27 to F.K., M.N., A.Z. and J.F.), the 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.), the 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.), the RuG Investment Agenda Grant Personalized Health to C.W. F.K. is supported also by the Noaber Foundation, Lunteren, the Netherlands. 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. also holds a joint fellowship from the University Medical Centre Groningen and China Scholarship Council (CSC201708320268) and a Foundation De Cock-Hadders grant (20:20-13). 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., D.W., H.A., D.Z., A.K. and A.V.V. generated the data. L.C. analyzed the data. L.C., F.K., A.Z. and J.F. drafted the manuscript. L.C., S.A., D.W., H.A., D.Z., A.K., A.V.V., R.W., M.M., M.N., F.K., C.W., A.Z. and J.F. reviewed and edited the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

References

1 Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54-60, doi:10.1038/nature10354 (2011).

2 Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550, doi:10.1038/ng.2982 (2014).

3 Bar, N. *et al.* A reference map of potential determinants for the human serum metabolome. *Nature* **588**, 135-140, doi:10.1038/s41586-020-2896-2 (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, doi:10.1038/s41591-020-01183-8 (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, doi:10.1136/bmjopen-2014-006772 (2015).

6 Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221-227, doi:10.1038/ejhg.2013.118 (2014).

7 Chen, L. *et al.* The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **184**, 2302-2315 e2312, doi:10.1016/j.cell.2021.03.024 (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, doi:10.1038/s41430-018-0205-z (2018).

9 Lianmin, C. *et al.* The long-term genetic stability and individual specificity of the human gut microbiome. *Cell,* **185** (2021).

10 Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat Genet* **51**, 600-605, doi:10.1038/s41588-019-0350-x (2019).

11 Yousri, N. A. *et al.* Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nature Communications* **9**, doi:ARTN 33310.1038/s41467-017-01972-9 (2018).

12 Bradley, P. H. & Pollard, K. S. Building a chemical blueprint for human blood. *Nature* **588**, 36-37, doi:10.1038/d41586-020-03122-6 (2020).

13 Kurilshikov, A. *et al.* Gut Microbial Associations to Plasma Metabolites Linked to Cardiovascular Phenotypes and Risk: A Cross-Sectional Study. *Circ Res* **124**, doi:10.1161/CIRCRESAHA.118.314642 (2019).

14 Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **46**, D608-D617, doi:10.1093/nar/gkx1089 (2018).

15 Zhao, Y. *et al.* Analysis of multiple metabolites of tocopherols and tocotrienols in mice and humans. *J Agric Food Chem* **58**, 4844-4852, doi:10.1021/jf904464u (2010).

16 Jiang, Q., Christen, S., Shigenaga, M. K. & Ames, B. N. gamma-tocopherol, the major form of vitamin E in the US diet, deserves more attention. *Am J Clin Nutr* **74**, 714-722, doi:10.1093/ajcn/74.6.714 (2001).

17 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, doi:10.1038/s41598-017-13722-4 (2017).

18 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, doi:10.1161/JAHA.119.015118 (2020).

19 Wijmenga, C. & Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat Genet*, 322-328, doi:10.1038/s41588-018-0066-3 (2018).

20 Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826, doi:10.1038/s41467-017-01261-5 (2017).

21 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012, doi:10.1093/nar/gky1120 (2019).

22 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 Pharmacol* **83**, 1176-1184, doi:10.1111/bcp.13207 (2017).

23 Group, S. C. *et al.* SLCO1B1 variants and statin-induced myopathy–a genomewide study. *N Engl J Med* **359**, 789-799, doi:10.1056/NEJMoa0801936 (2008).

24 Yu, B. *et al.* Loss-of-function variants influence the human serum metabolome. *Sci Adv* **2**, e1600800, doi:10.1126/sciadv.1600800 (2016).

25 Chang, Y. *et al.* Fragment-based discovery of novel pentacyclic triterpenoid derivatives as cholesteryl ester transfer protein inhibitors. *European journal of medicinal chemistry* **126**, 143-153 (2017).

26 Schlosser, P. *et al.* Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nat Genet* **52**, 167-176, doi:10.1038/s41588-019-0567-8 (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, doi:10.1007/s00204-013-1157-7 (2013).

28 Suhre, K. *et al.* A genome-wide association study of metabolic traits in human urine. *Nat Genet* **43**, 565-569, doi:10.1038/ng.837 (2011).

29 Rothman, N. *et al.* A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet* **42**, 978-984, doi:10.1038/ng.687 (2010).

30 Wang, Z. & Zhao, Y. Gut microbiota derived metabolites in cardiovascular health and disease. *Protein Cell* **9**, 416-431, doi:10.1007/s13238-018-0549-0 (2018).

31 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, doi:10.1016/j.biotechadv.2020.107576 (2020).

32 Gacesa, R. *et al.* The Dutch Microbiome Project defines factors that shape the healthy gut microbiome. *bioRxiv*, 2020.2011.2027.401125, doi:10.1101/2020.11.27.401125 (2020).

33 Chen, L. *et al.* Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nature Communications* **11**, 4018, doi:10.1038/s41467-020-17840-y (2020).

34 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, doi:10.1186/s13059-020-02042-y (2020).

35 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, doi:10.1016/j.celrep.2020.108212 (2020).

36 Ríos-Covián, D. *et al.* Intestinal short chain fatty acids and their link with diet and human health. *Frontiers in microbiology* **7**, 185 (2016).

37 Vich Vila, A. *et al.* Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci Transl Med* **10**, eaap8914, doi:10.1126/scitranslmed.aap8914 (2018).

38 Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**, 766, doi:10.15252/msb.20145645 (2014).

39 Andreu, V. P. *et al.* A systematic analysis of metabolic pathways in the human gut microbiota. *bioRxiv* (2021).

40 Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43-48, doi:10.1038/s41586-019-1065-y (2019).

41 Natella, F., Nardini, M., Belelli, F. & Scaccini, C. Coffee drinking induces incorporation of phenolic acids into LDL and increases the resistance of LDL to ex vivo oxidation in humans. *Am J Clin Nutr* **86**,

604-609, doi:10.1093/ajcn/86.3.604 (2007).

42 Song, Y. *et al.* Feruloylated oligosaccharides and ferulic acid alter gut microbiome to alleviate diabetic syndrome. *Food Res Int* **137**, 109410, doi:10.1016/j.foodres.2020.109410 (2020).

43 Salau, V. F. *et al.* Ferulic Acid Modulates Dysfunctional Metabolic Pathways and Purinergic Activities, While Stalling Redox Imbalance and Cholinergic Activities in Oxidative Brain Injury. *Neurotox Res* **37**, 944-955, doi:10.1007/s12640-019-00099-7 (2020).

44 Biegel, E. & Muller, V. Bacterial Na+-translocating ferredoxin:NAD+ oxidoreductase. *Proc Natl Acad Sci U S A* **107**, 18138-18142, doi:10.1073/pnas.1010318107 (2010).

45 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, doi:10.1006/jmbi.1999.2993 (1999).

46 Fujita, T., Fujita, M., Kodama, T., Hada, T. & Higashino, K. Determination of D- and L-pipecolic acid in food samples including processed foods. *Ann Nutr Metab* **47**, 165-169, doi:10.1159/000070040 (2003).

47 Kawasaki, H., Hori, T., Nakajima, M. & Takeshita, K. Plasma levels of pipecolic acid in patients with chronic liver disease. *Hepatology* **8**, 286-289, doi:10.1002/hep.1840080216 (1988).

48 Miras-Avalos, J. M., Bouzas-Cid, Y., Trigo-Cordoba, E., Orriols, I. & Falque, E. Amino Acid Profiles to Differentiate White Wines from Three Autochtonous Galician Varieties. *Foods* **9**, doi:10.3390/foods9020114 (2020).

49 Alexander Kurilshikov, I. C. v. d. M., Lianmin Chen , Marc Jan Bonder , Kiki Schraa , Joost Rutten , Niels P Riksen , Jacqueline de Graaf , Marije Oosting , Serena Sanna , Leo AB Joosten , Marinette van der Graaf , Tessa Brand , Debby PY Koonen , Martijn JR van Faassen , P Eline Slagboom , Ramnik J Xavier , Folkert Kuipers , Marten Hofker , Cisca Wijmenga , Mihai G Netea , Alexandra Zhernakova , Jingyuan Fu. Gut Microbial Associations to Plasma Metabolites Linked to Cardiovascular Phenotypes and Risk: A Cross-Sectional Study. *Circulation Research* **124**, https://doi.org/10.1161/CIRCRESAHA.1118.314642, doi:https://doi.org/10.1161/CIRCRESAHA.118.314642 (2019).

50 Visconti, A. *et al.* Interplay between the human gut microbiome and host metabolism. *Nature Communications* **10**, doi:ARTN 450510.1038/s41467-019-12476-z (2019).

51 Fuhrer, T., Zampieri, M., Sevin, D. C., Sauer, U. & Zamboni, N. Genomewide landscape of gene-metabolome associations in Escherichia coli. *Molecular Systems Biology* **13**, doi:ARTN 90710.15252/msb.20167150 (2017).

52 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. *Analytical Chemistry*

83, 7074-7080, doi:10.1021/ac201267k (2011).

53 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, doi:10.1186/gb-2009-10-3-r25 (2009).

54 Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421-432, doi:10.1093/bioinformatics/bty648 (2019).

55 Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**, 902-903, doi:10.1038/nmeth.3589 (2015).

56 Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* **15**, 962-968 (2018).

57 Bateman, A. *et al.* UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-D212 (2015).

58 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).

59 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* **46**, D633-D639 (2018).

60 Aitchison, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44**, 139-160 (1982).

61 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, doi:10.1038/s41588-018-0224-7 (2018).

62 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287, doi:10.1038/ng.3656 (2016).

63 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

64 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).

65 Imhann, F. *et al.* Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740-748, doi:10.1136/gutjnl-2015-310376 (2016).

66 Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565-569, doi:10.1126/science.aad3369 (2016).

67 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 Genetics* **7**, doi:ARTN e100219710.1371/journal.pgen.1002197 (2011).

68 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

69 Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818-825, doi:10.1038/ng.3021 (2014).

Figures

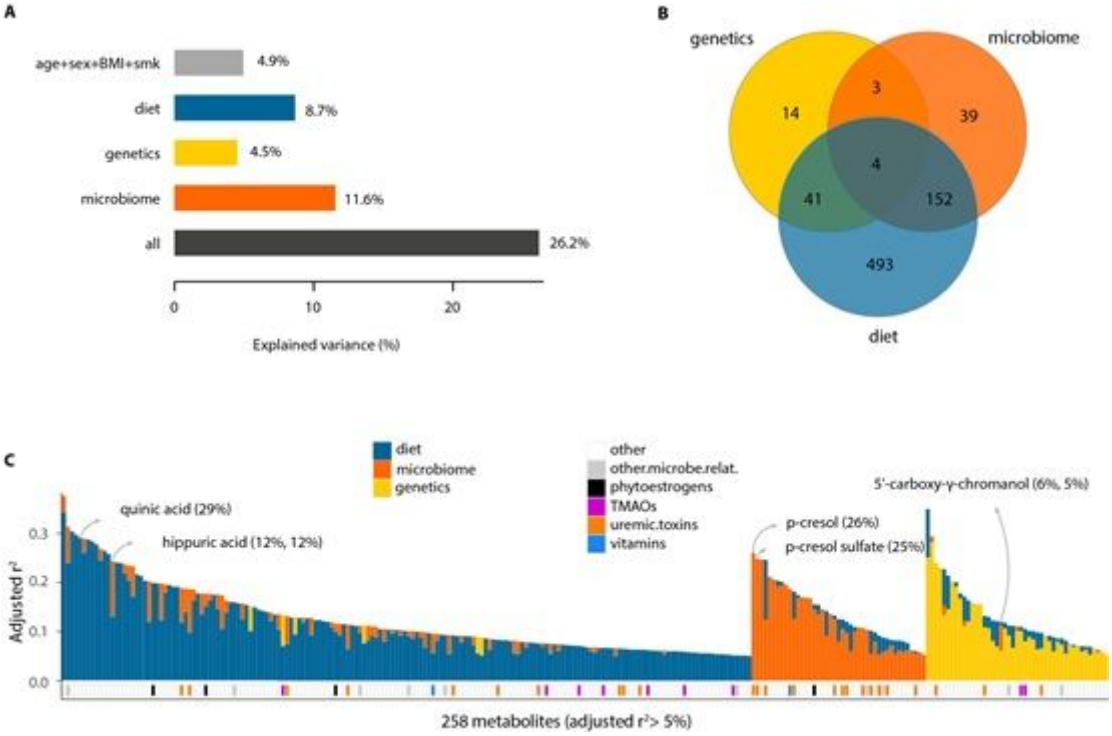

Figure 1

Diet and microbiome contribute more than genetics to inter-individual plasma metabolome variations. A. Dietary, genetic and gut microbiome contributions to the inter-individual variation in the whole plasma metabolome, estimated using the PMAV method. B. Overlap of metabolites involved in inter-individual metabolite variations that can be significantly explained by diet, genetics or the gut microbiome, estimated using the lasso regression method (FDRF-test<0.05). C. Inter-individual variations in metabolites that can be significantly explained by diet, genetics or the gut microbiome, estimated using the lasso regression method with a significant estimated adjusted r2 more than 5% (FDRF-test<0.05). Blue bars represent dietary contributions to metabolite variations. Yellow bars indicate genetic contributions. Orange bars indicate microbial contributions. Other colors (see legend) indicate the metabolic categories of metabolites.

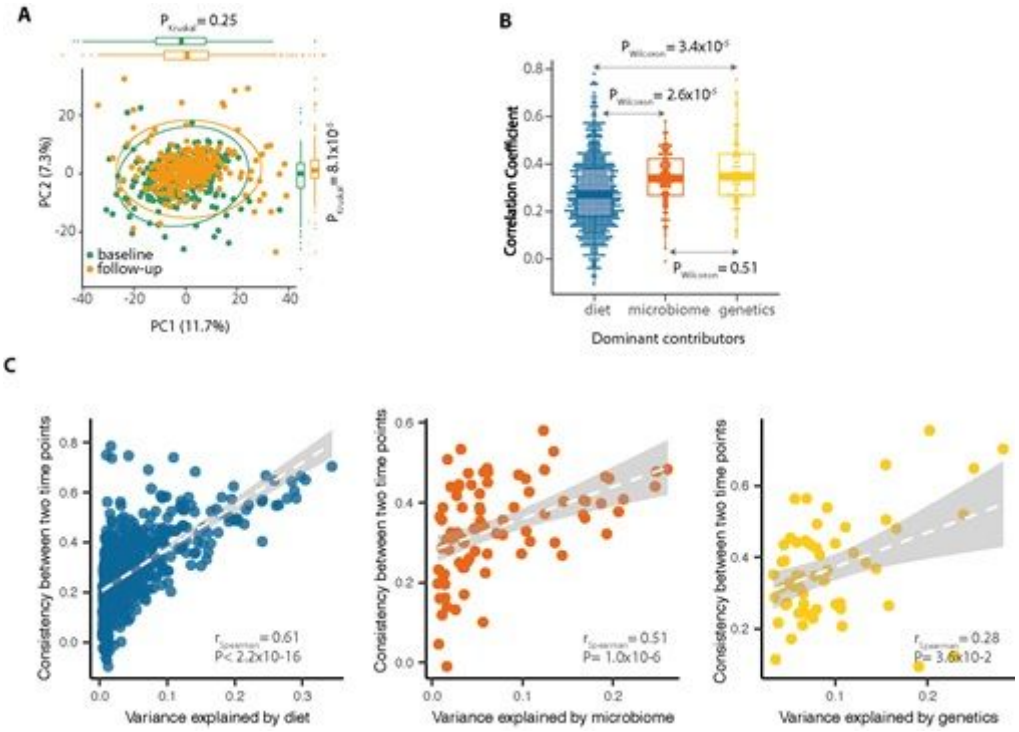

Figure 2

Stability of plasma metabolites. A. Principal component analysis of metabolite levels (Euclidean dissimilarity). Green dots indicate baseline subjects. Yellow dots indicate follow-up samples. B. Stability of metabolites that were dominated by diet, genetics and microbiome. C. Consistency between metabolite stability and the metabolite variance explained by diet, genetics and microbiome.

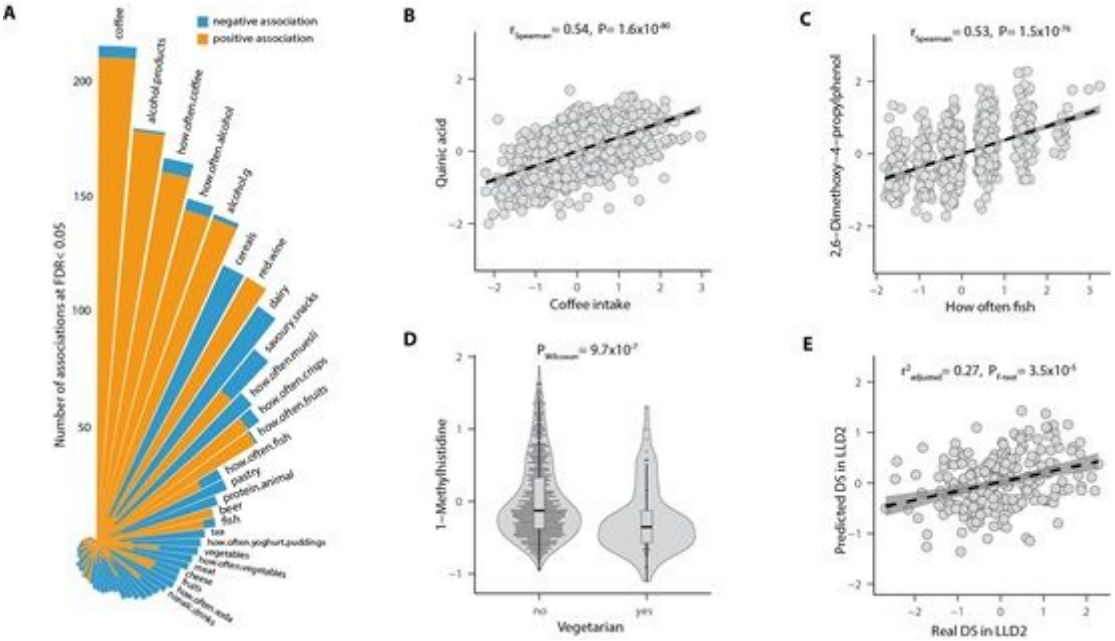

Figure 3

Associations between dietary habits and plasma metabolites. A. Summary of dietary associations to metabolites. Bars represent dietary habits, with bar order sorted by the number of significant

associations. Association directions are colored differently: yellow indicates a positive association, blue indicates a negative association. B. Positive association between plasma quinic acid and coffee intake. C. Positive association between plasma 2,6-dimethoxy-4-propylphenol and fish intake frequency. Association strength was assessed using Spearman correlation, and both correlation coefficient and P-value are reported. D. Differential plasma levels of 1-methylhistidine between vegetarians and non-vegetarians. The P-value from the Wilcoxon test is shown. E. Plasma metabolome significantly predicts dietary quality score. X-axis shows is the LLD dietary score. Y-axis shows the dietary score predicted based on the prediction model. Prediction performance is assessed with linear regression, and both adjusted r2 and the P-value from the F test are reported.

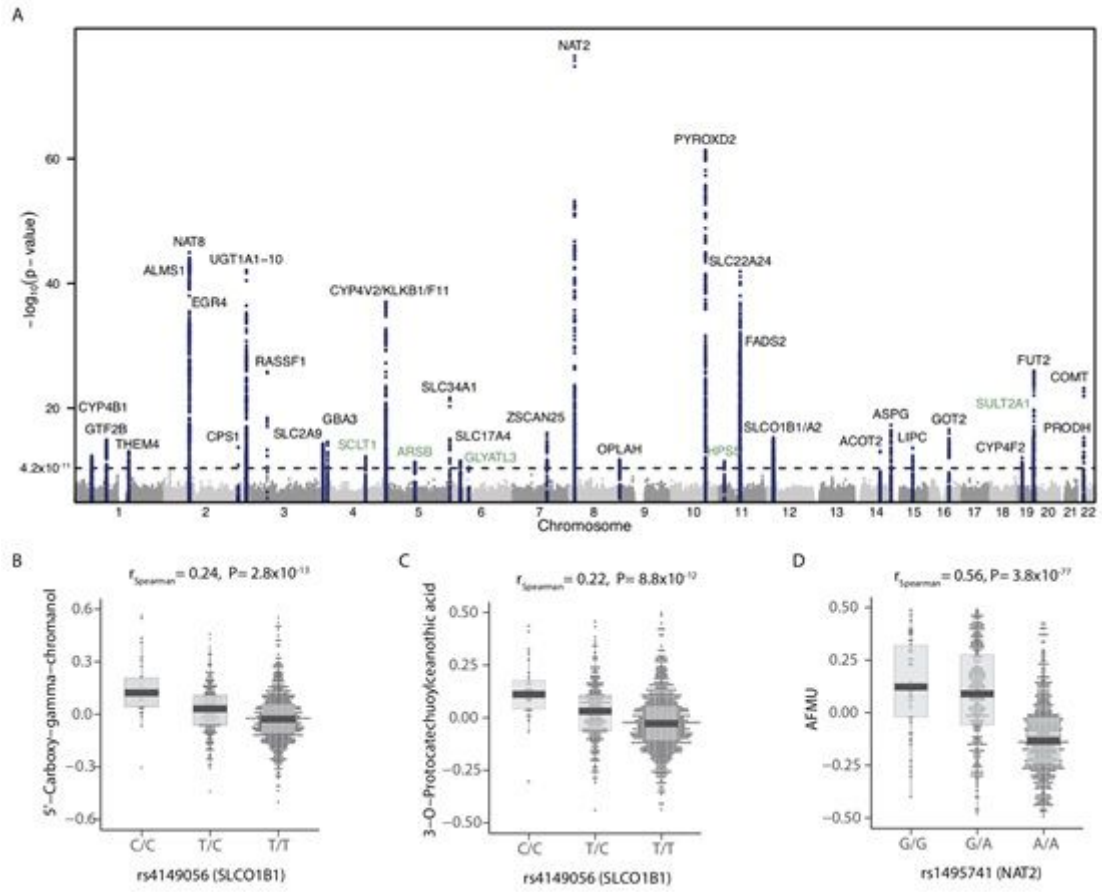

Figure 4

Novel genetic determinants of plasma metabolites. A. Manhattan plot showing the 73 independent metabolite quantitative trait loci (mQTLs) between 63 metabolites and 57 genetic variants with P < 4.2x10-11. 67 of these 73 mQTLs are novel. Representative genes for the SNPs with significant mQTLs are listed. Gene names marked in green are additional mQTLs observed after step-wise correction for dietary habits, diseases, medications and gut microbial composition. Independent mQTLs are defined by clumping r2 of 0.05 and a window size of 500kb. The linkage disequilibrium (LD) block (100 kb) of each mQTL is highlighted in blue. B-C. A SNP (rs4149056) within gene SLCO1B1 associates with both plasma levels of 5'-carboxy-gamma-chromanol (B) and 3-o-protocatechuoylceanothic acid (C). D. A tag loci (rs1495741) of the NAT2 gene associates with plasma AFMU levels.

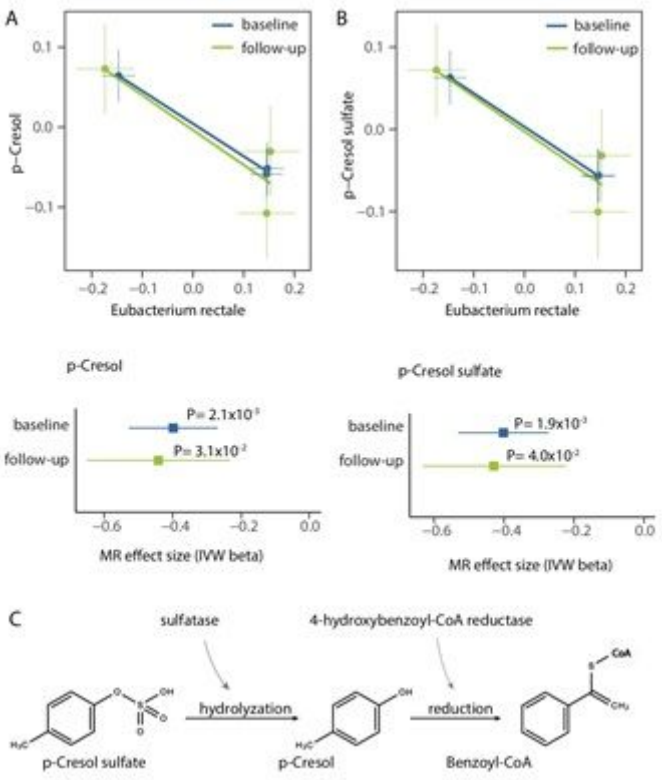

Figure 5

Causal relationships between the abundance of Eubacterium rectale and plasma levels of p-cresol and its sulfate. A-B. Mendelian randomization (MR) analysis of E. rectale abundance with p-cresol (A) and p-cresol sulfate (B). In the upper panel, the X-axis shows the SNP—exposure effect and the Y-axis shows the SNP—outcome effect. Each dot represents a SNP. The lower panel shows the combined MR results for the discovery and replication datasets. Blue dots represent discovery. Green dots refer to replication. C. Annotations of E. rectale genome. We characterized genes encoding sulfatase and 4-hydroxybenzoyl-CoA reductase that are responsible for p-cresol sulfate hydrolyzating into p-cresol and the reduction of p-cresol.

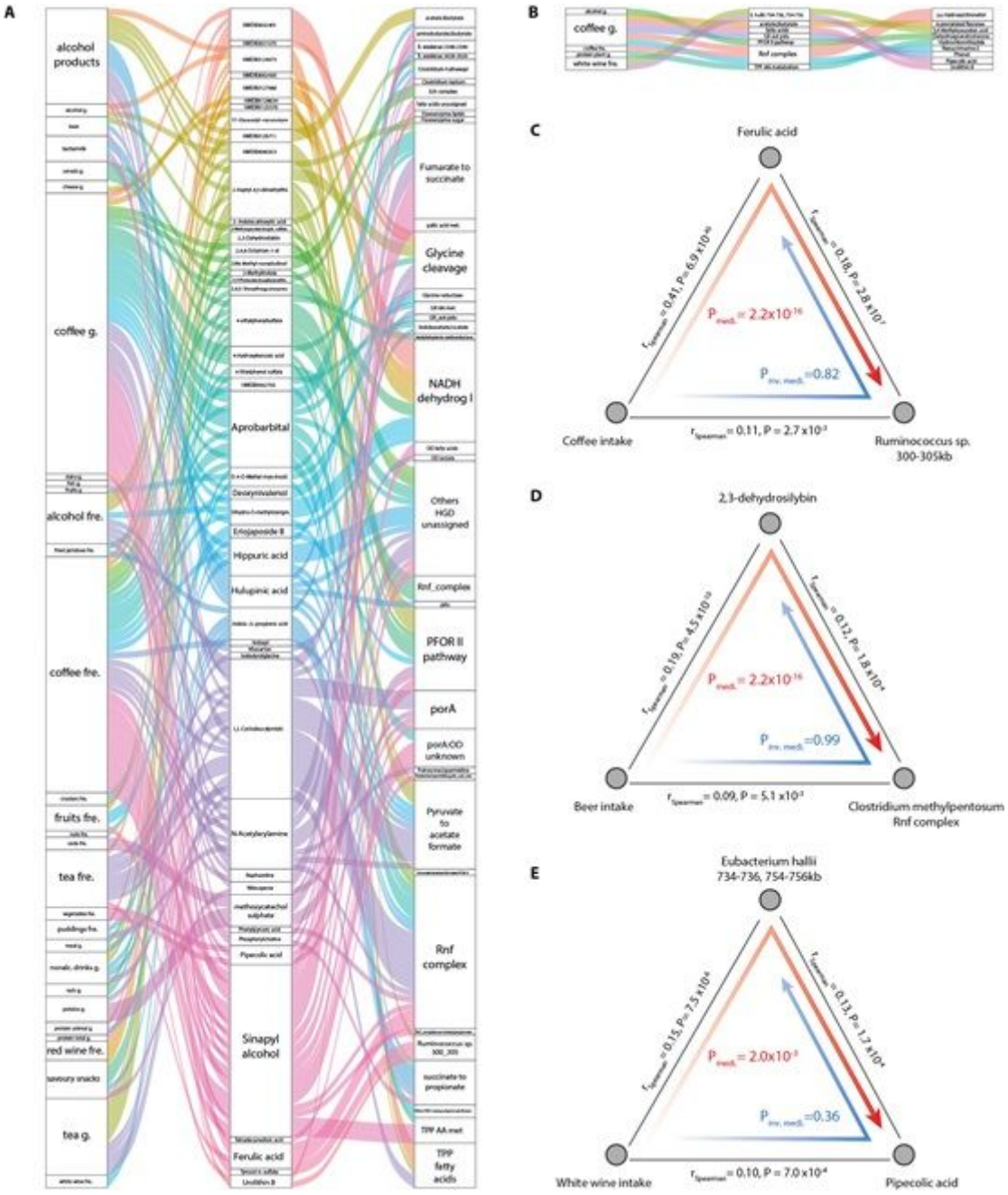

Figure 6

Mediation linkages between the gut microbiome, metabolites and dietary habits. A. Parallel coordinates chart showing the 185 mediation effects of plasma metabolites that were significant at FDR < 0.05. Left panel shows dietary habits. Middle panel shows plasma metabolites. Right panel shows microbial factors. The curved lines connecting the panels indicate the mediation effects, with the colors corresponding to different metabolites. B. Parallel coordinates chart showing the 10 mediation effects of the microbiome that were significant at FDR < 0.05. Left panel shows dietary habits. Middle panel shows microbial factors. Right panel shows plasma metabolites. The curved lines connecting the panels indicate the mediation effects, with the colors corresponding to the different microbial factors. C. Coffee intake influences the variability of a vSV in Ruminococcus sp. (300—305 kb) through ferulic acid (Pmediation=2.2x10-16). D. Beer intake influences the Clostridium methylpentosum Rnf complex

pathway through 2,3-dehydrosilybin (Pmediation=2.2x10-16). E. White wine intake influences pipecolic acid in plasma through a vSV in Eubacterium hallii (734–736, 754–756 kb) (Pmediation=2.0x10-3). Grey lines indicate the associations between two factors with corresponding Spearman coefficients and P-values. Direct mediation is shown by a red arrow. Reverse mediation is shown by a blue arrow.

Supplementary Files

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

- SupFigsubmission.docx
- Supptablessubmission.xlsx
- flatFuepc.pdf
- flatFurs.pdf