A reference map of potential determinants for the human serum metabolome

https://doi.org/10.1038/s41586-020-2896-2

Received: 23 January 2019

Accepted: 29 September 2020

Published online: 11 November 2020

Check for updates

Noam Bar[1,2,81], Tal Korem[1,2,3,4,5,81], Omer Weissbrod[1,2,6], David Zeevi[1,2,7], Daphna Rothschild[1,2], Sigal Leviatan[1,2], Noa Kosower[1,2], Maya Lotan-Pompan[1,2], Adina Weinberger[1,2], Caroline I. Le Roy[8], Cristina Menni[8], Alessia Visconti[8], Mario Falchi[8], Tim D. Spector[8], The IMI DIRECT consortium*, Jerzy Adamski[9,10,11], Paul W. Franks[12,13], Oluf Pedersen[14] & Eran Segal[1,2]✉

The serum metabolome contains a plethora of biomarkers and causative agents of various diseases, some of which are endogenously produced and some that have been taken up from the environment[1]. The origins of specific compounds are known, including metabolites that are highly heritable[2,3], or those that are influenced by the gut microbiome[4], by lifestyle choices such as smoking[5], or by diet[6]. However, the key determinants of most metabolites are still poorly understood. Here we measured the levels of 1,251 metabolites in serum samples from a unique and deeply phenotyped healthy human cohort of 491 individuals. We applied machine-learning algorithms to predict metabolite levels in held-out individuals on the basis of host genetics, gut microbiome, clinical parameters, diet, lifestyle and anthropometric measurements, and obtained statistically significant predictions for more than 76% of the profiled metabolites. Diet and microbiome had the strongest predictive power, and each explained hundreds of metabolites—in some cases, explaining more than 50% of the observed variance. We further validated microbiome-related predictions by showing a high replication rate in two geographically independent cohorts[7,8] that were not available to us when we trained the algorithms. We used feature attribution analysis[9] to reveal specific dietary and bacterial interactions. We further demonstrate that some of these interactions might be causal, as some metabolites that we predicted to be positively associated with bread were found to increase after a randomized clinical trial of bread intervention. Overall, our results reveal potential determinants of more than 800 metabolites, paving the way towards a mechanistic understanding of alterations in metabolites under different conditions and to designing interventions for manipulating the levels of circulating metabolites.

We used mass spectrometry to profile serum samples from 491 healthy individuals for whom we had previously collected extensive clinical, lifestyle, dietary, genetics and gut microbiome data[10] (Methods, Extended Data Table 1). Our untargeted metabolomics analysis measured the levels of 1,251 metabolites, covering a wide range of biochemicals including lipids, amino acids, xenobiotics, carbohydrates, peptides and nucleotides, and approximately 30% unidentified compounds (Extended Data Fig. 1a, Methods, Supplementary Table 1). To classify unidentified metabolites and aid in biomarker discovery, we designed models that accurately predict the candidate biological pathway of the metabolites (Extended Data Fig. 2, Supplementary Note 1, Supplementary Table 2–5). Most metabolites we measured were prevalent across the cohort, including 498 metabolites that were detected in all samples and 1,104 metabolites that were detected in more than 50% of the samples (Extended Data Fig. 1b). After quality control (Methods), 475 individuals with high-quality data were included in the subsequent analyses.

To validate the accuracy of our metabolomic measurements, we compared the levels of creatinine and cholesterol to those obtained using standardized laboratory tests (Methods) that were performed independently on a second blood sample taken from the participants on the same visit, and found good agreement (creatinine, Pearson's $R = 0.87$; cholesterol, $R = 0.79$; Extended Data Fig. 1c, d). We further

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [3]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA. [4]Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY, USA. [5]CIFAR Azrieli Global Scholars Program, CIFAR, Toronto, Ontario, Canada. [6]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [7]Center for Studies in Physics and Biology, The Rockefeller University, New York, NY, USA. [8]Department for Twin Research & Genetic Epidemiology, King's College London, London, UK. [9]Research Unit Molecular Endocrinology and Metabolism, Genome Analysis Center, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [10]Lehrstuhl für Experimentelle Genetik, Technische Universität München, Freising-Weihenstephan, Germany. [11]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [12]Lund University Diabetes Center, Department of Clinical Sciences, Lund University, Malmö, Sweden. [13]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [14]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [81]These authors contributed equally: Noam Bar, Tal Korem. *A list of members and their affiliations appears at the end of the paper. ✉e-mail: eran.segal@weizmann.ac.il

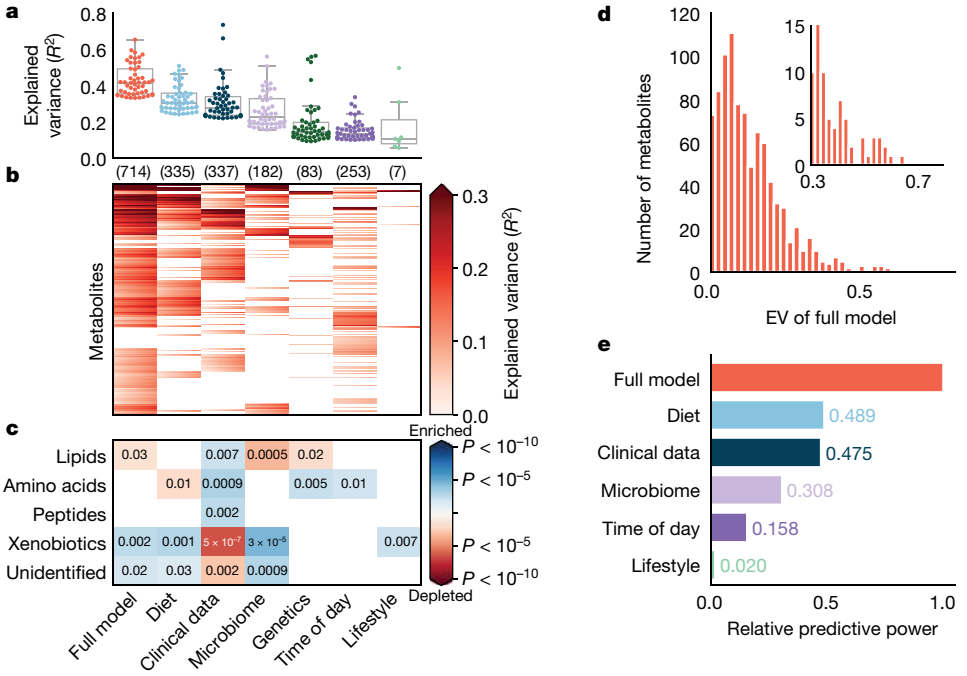


Fig. 1 | Diet, gut microbiome, genetics and clinical data predict the levels of most serum metabolites. This dataset arises from fivefold cross-validation predictions of metabolite levels based on separate models for each feature group. **a**, Box and swarm plots (centre, median; box, interquartile range (IQR); whiskers, 1.5 × IQR) showing the explained variance (EV; R^2) of the top 50 significantly predicted metabolites (when available) of each feature group (group names are listed in **c**). **b**, Heat map with colour gradient from left to right corresponding to the 95% confidence interval for the EV, for each metabolite by every feature group. Only metabolites with significant predictions (FDR < 10%) are shown, and the number of metabolites per group is shown in brackets at the top. *P* values and confidence intervals were estimated using bootstrapping (Methods). **c**, Enrichment of metabolite types in the predictions by each feature group (two-sided Mann–Whitney *U*-test; Methods). Only significant enrichments are shown ($P < 0.05$ after 10% FDR correction). Exact *P* values are written in each cell. **d**, Histogram of the number of metabolites with each value of EV obtained using the full model. The inset shows the EV range of 0.3–0.8. **e**, The fraction of total EV of each feature group compared to the total EV of a model with all feature groups excluding genetics (full model). The total EV is the sum of the EV of the first 15 metabolite principal components weighted by the EV of each principal component (Methods).

found that, for 20 participants for which samples were taken one week apart, the two profiles were significantly correlated (Spearman $\rho = 0.68 \pm 0.06$, median ± s.d.); this was in contrast to samples from different participants, which showed no correlation (Spearman $\rho = 0.05 \pm 0.12$; Methods, Extended Data Fig. 1e). These results validate the reproducibility and accuracy of our data, are consistent with previous work showing long-term stability in the human metabolome[11], and confirm that this metabolic profile is a unique, person-specific signature.

Robust predictions of serum metabolites

We trained gradient-boosted decision trees[12] (GBDT) algorithms that predict metabolite levels in held-out individuals (Methods, Supplementary Note 2). GBDT systematically outperformed linear models (Lasso; Methods), with a median and maximum explained variance gain of 8.3% and 43.2%, respectively, for prediction with diet data, and 4.6% and 14.9%, respectively, for microbiome data (Extended Data Fig. 3). Notably, our predictions for more than 76% of the metabolite groups tested were statistically significant with at least one feature group after multiple hypothesis correction (Methods). In total, 335 metabolites were significantly explained by diet-related features, and 182 by the microbiome (Fig. 1a, b). Our models explained more than 10% of the variance for 543 metabolite groups (median 10.2%; range 0–73.5%; Fig. 1d, Supplementary Table 6), and more than 50% of the variance for 17 metabolites.

We next checked, for each feature group, whether any type of metabolite was enriched with superior predictions (Fig. 1c, Methods). We found that clinical data better predicted metabolites that were classified as blood lipids, amino acids and peptides, as opposed to xenobiotics and unidentified compounds, on which it performed worse than on other metabolites. By contrast, microbiome data better explained levels of xenobiotics ($P < 10^{-4}$) and unidentified compounds ($P < 0.001$), highlighting its potential for explaining the origins of the large number of unidentified compounds. We further found that predictions based on clinical data were significantly correlated with those based on diet (Spearman's $\rho = 0.30$, $P < 10^{-20}$), and had a weaker correlation with predictions based on the microbiome ($R = 0.21$, $P < 10^{-11}$). Predictions based on microbiome data had the highest correlation with those based on diet ($R = 0.44$, $P < 10^{-20}$). Finally, we found that metabolites associated with genetics could not be predicted by other feature groups, and there was a weak correlation between the prediction accuracy of a model containing all other features ('full model', Methods) and the heritability of metabolites ($R = 0.09$, $P < 0.005$). Altogether, each feature group was particularly informative with respect to a different set of metabolites (Extended Data Figs. 4, 5a).

To estimate the relative predictive power of each feature group across all metabolites, we built models to predict the principal metabolomic components (Extended Data Fig. 5b). Diet had the largest predictive power, inferring 48.9% of the variance explained by a model containing all features (Methods), whereas lifestyle factors explained only 1.9% (Fig. 1e). Notably, the predictive power of microbiome data was 30.8% that of the full model. As a large portion of these predictions did not overlap with the predictions of other data, these results highlight the importance of microbiome data in predicting and potentially determining serum metabolite levels.

Replication in external cohorts

To test the robustness and reproducibility of the models based on gut microbiome data, we validated their accuracy in two geographically

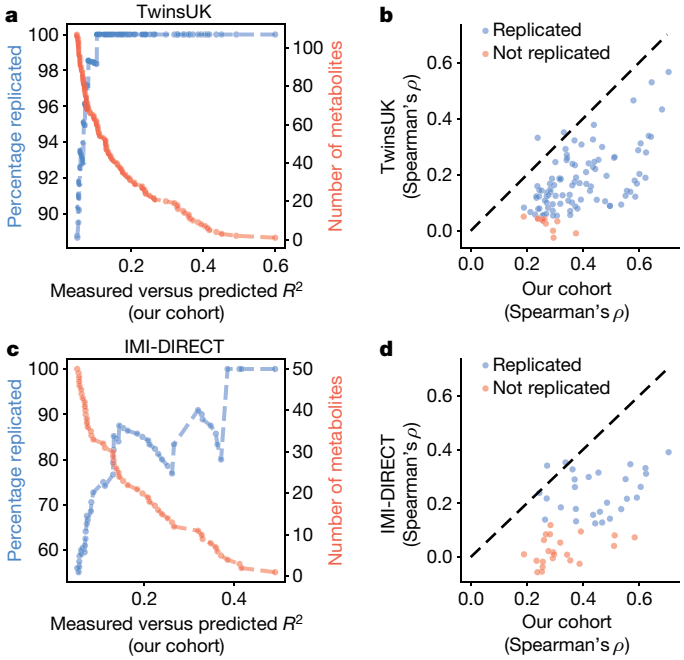


Fig. 2 | Validation of microbiome-based predictions of metabolites in two independent cohorts. a, **c**, R^2 of predicted metabolites in our cohort plotted against the rate of replicated associations in the replication cohorts (TwinsUK, **a**; IMI DIRECT, **c**), computed as the fraction of significant replications out of all predictions with equal or higher predicted R^2 in our cohort (left *y* axis; blue; FDR <10%), and the cumulative number of metabolites (right *y* axis; red). **b**, **d**, Spearman correlation between true and predicted levels of metabolites in our cohort plotted against the same correlations in the replication cohorts (TwinsUK, **b**; IMI DIRECT, **d**). Metabolites are coloured by the replication success (replicated, blue ($n = 95$ (**b**), $n = 28$ (**d**)); not replicated, red ($n = 12$ (**b**), $n = 22$ (**d**)); FDR < 10%).

independent cohorts (Methods): 1,004 samples from healthy older participants from the UK (from the TwinsUK Registry[7]), and 245 samples from Northern European individuals with type 2 diabetes (from the IMI DIRECT cohort[8]; Extended Data Table 1). Validation data were not available to us while developing the prediction models, which were trained only on samples from the initial Israeli cohort. We obtained predictions for metabolites that had statistically significant predictions (false discovery rate (FDR) < 0.1) with $R^2 > 5\%$ in the Israeli cohort (107 metabolites in TwinsUK, 50 in IMI DIRECT), using only microbiome data from the validation cohorts. Notably, 95 out of 107 and 28 out of 50 predictions were replicated (FDR < 0.1) in the healthy TwinsUK cohort and in the IMI DIRECT cohort of patients with type 2 diabetes, respectively, including all top 60 predictions in the TwinsUK cohort (Fig. 2, Supplementary Tables 7, 8). We note that most of the replicated associations are accompanied by a reduction in effect size; this is expected, particularly as a result of study-specific biases. These results indicate that our models reveal robust associations between serum metabolites and the gut microbiome, despite differences between both the populations and the protocols and staff used to assemble these cohorts. Finally, most significant associations between metabolite levels and body mass index were also replicated in the TwinsUK cohort[13] with high accuracy ($R = 0.85$, $P < 10^{-10}$; Extended Data Fig. 5c, Supplementary Table 9).

Diet and microbiome models are independent

Because the diet modulates the gut microbiome[14], we compared the explained variance of metabolites obtained by models based on either. Although some metabolites—mostly related to coffee consumption—were significantly predicted by both diet and microbiome data, many were not (Supplementary Table 10). Furthermore, adding microbiome data to a diet-based prediction model improved its accuracy in 66% of cases (median and maximum gain of 2.1% and 62.6% respectively; Supplementary Table 11), whereas adding permuted data reduced the performance in 82% of cases (median and maximum gain of −1.7% and 7.4% respectively; Extended Data Fig. 5d–f). Finally, 34 metabolites were significantly predicted only by microbiome data. Altogether, these results suggest that the gut microbiome may be modulating the production of many circulating metabolites, independent of diet.

We next used feature attribution analysis (SHAP[9]; Methods) to interpret these models, infer the drivers of each prediction, and examine interactions between different predictive factors (Extended Data Fig. 6, Supplementary Note 3). We found dozens of diet and bacterial features that were strongly predictive of blood metabolites in our models (Fig. 3a, Extended Data Fig. 7). Notably, the reported consumption of coffee (both long-term and short-term; Methods) was a stronger predictor than other dietary features for the levels of a large number of xenobiotics and unidentified compounds. These included metabolites from the xanthine metabolism pathway such as paraxanthine (diet prediction Pearson $R = 0.64$, $P < 10^{-20}$) and caffeine ($R = 0.68$, $P < 10^{-20}$), as previously reported[15]. These metabolites were also significantly predicted using microbiome data, with a Clostridiaceae species being the main predictor. Another strong feature was long-term fish consumption, which accurately predicted the levels of several blood lipids including 3-carboxy-4-methyl-5-propyl-2-furanpropionic acid (diet $R = 0.71$, $P < 10^{-20}$), a uraemic toxin that accumulates in the serum of patients with chronic kidney disease[16] and has also been suggested to prevent and reverses steatosis[17]. X-16124 (microbiome $R = 0.77$, $P < 10^{-20}$) and X-11850 ($R = 0.7$, $P < 10^{-20}$) are two unidentified metabolites that were accurately predicted by microbiome data, and specifically by bacteria from the Eggerthellaceae family and *Clostridium* genus, respectively. Microbiome data was also highly predictive of the uraemic toxins phenylacetylglutamine ($R = 0.63$, $P < 10^{-20}$) and indoxyl sulfate ($R = 0.37$, $P < 10^{-20}$), which have been previously reported in association with cardiovascular disease[18] and chronic kidney disease[19]; these predictions were driven by a species from the Lachnospiraceae family.

To assess whether a few important taxa are sufficient for accurate prediction, we defined the 'main predictor' of each metabolite as the taxon with the maximal mean absolute SHAP value. Nineteen bacterial taxa were the main predictors for the top 50 microbiome-predicted metabolites (prediction $R > 0.4$; Supplementary Table 12). One Clostridiceae species was the main predictor of 22 of these, which are also strongly associated with coffee consumption in diet-based models. *Clostridium* sp. CAG:138 was the main predictor of five metabolites, including phenylacetylcarnitine ($R = 0.47$, $P < 10^{-20}$) and *p*-cresol-glucuronide ($R = 0.64$, $P < 10^{-20}$) as previously reported[20]. Other taxa, however, were the main predictors of only one or two top metabolites, demonstrating that many different bacteria are required to accurately predict the levels of different metabolites. Among the main bacterial predictors of the top 100 metabolites, 89 belonged to Firmicutes, highlighting the strong predictive power of this phylum. It is notable that, although Bacteroidetes is the second most abundant phylum in our cohort (Extended Data Fig. 8a), none of its species were among these main predictors.

To check whether main predictors are sufficient for accurate prediction, for each metabolite we compared the accuracy of a full microbiome model to the accuracy of a model based only on its main predictor (Fig. 3b). We found that a model based on the main predictor could explain only a median of 36% of the explained variance of the full microbiome-based model. Cinnamoylglycine, for example, is significantly predicted using microbiome data ($R = 0.49$, $P < 10^{-20}$); however, a model based on its main predictor fails to provide a significant prediction. By contrast, some metabolites are exclusively predicted by a single bacterial species, such as the unidentified metabolite X-16124, for which a model based on an Eggerthellaceae species explained 93%

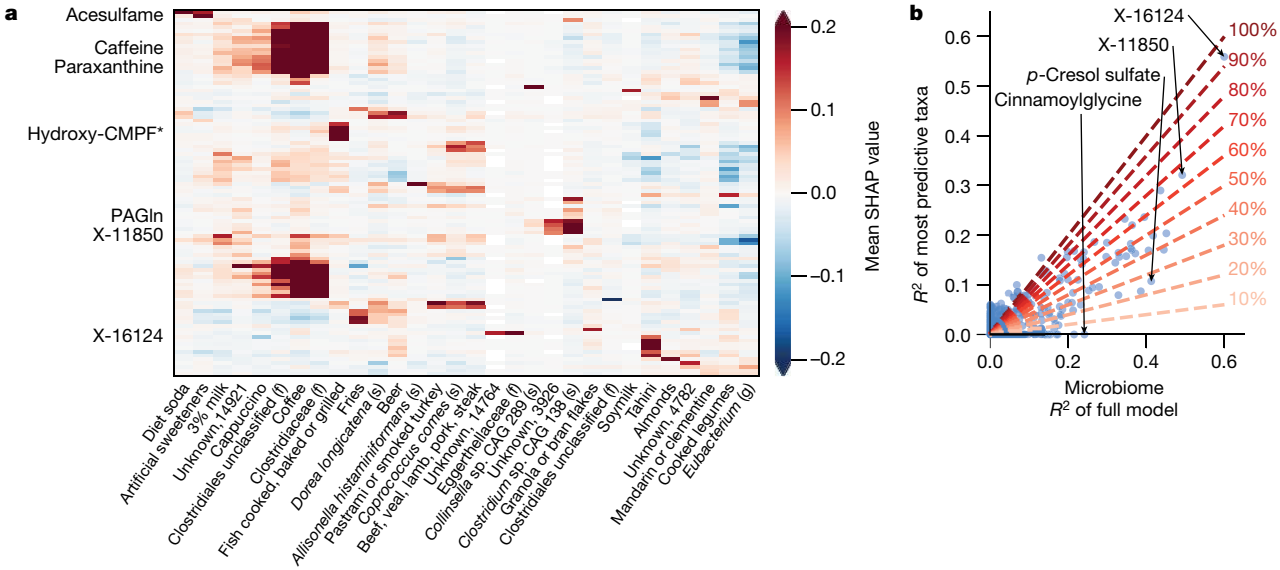


Fig. 3 | Diet and gut microbiome data independently explain a wide range of biochemicals. a, Subset of a heat map showing the directional mean absolute SHAP values (Methods) of various features (*x* axis) computed from fivefold cross-validation models that predict metabolite levels (*y* axis) using two separate models, one based on diet and another on gut microbiome data. Positive (negative) SHAP values indicate that higher (lower) feature values lead, on average, to higher predicted values. Shown are the top 100 predicted metabolites using diet and gut microbiome, and the top 30 features by maximum mean absolute SHAP value across all metabolites. See extended heat map in Extended Data Fig. 7g. **b**, The EV of every metabolite from microbiome-based prediction models (*x* axis) compared to using only the top predictor of that metabolite, selected as the feature with the largest mean absolute SHAP value (*y* axis). Dashed red lines mark different *y*:*x* ratios. PAGln, phenylacetylglutamine; hydroxy-CMPF*, hydroxy-3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid.

of the variance of a full microbiome-based model. Indeed, in 95% of the individuals in which this bacteria was detectable, X-16124 was also detectable in serum, compared to only 23% of individuals for which this bacteria was not detected (Mann–Whitney *U*-test, $P < 10^{-20}$; Extended Data Fig. 8b).

New genetic–metabolomics associations

Several genome-wide association studies have found that human genetics influences serum metabolites[2,3,21–24]. The median serum metabolite ACE-heritability, using the traditional twin model, was estimated to be 25%, whereas the median narrow-sense heritability, based only on discovered genetic loci, was estimated to be 2.1%[2]. Because we measured several molecules that were not yet identified in these studies, we searched for associations between levels of these molecules and single nucleotide polymorphisms (SNPs; Supplementary Note 4). Notably, we found 68 statistically significant associations ($P < 5 \times 10^{-11}$ for all), of which—to the best of our knowledge (Methods)—22 have not been previously reported (Supplementary Table 13). These include ethylmalonate, a branched fatty acid that has been reported in association with anorexia nervosa[25] and that was associated with rs2066938, which explained 50% of its variance. This SNP is a variant of the 3′-untranslated region of the gene *UNC119B*, which we also found to be associated with butyrylcarnitine, in line with previous reports[2]. Other examples include 2′-*O*-methyluridine and 2′-*O*-methylcytidine—both of which are nucleotides involved in pyrimidine metabolism—which we found to associate with a missense variant in the *PHYHD1* gene and have been previously reported to be negatively correlated with *PHYHD1* expression[26]. We further found that X-21441—which we predicted as an androgenic steroid (Supplementary Note 1)—was associated with rs8187710, a missense variant in the *ABCC2* gene, explaining 11% of its variance (Extended Data Fig. 9). rs8187710 was previously demonstrated to be associated with non-alcoholic fatty liver disease[27]. Notably, X-21441 was also negatively correlated with age in our cohort ($R = -0.3$, $P < 10^{-7}$), independent of the genotype (Extended Data Fig. 9c). This suggests that X-21441 might be an independent metabolic risk factor that mediates the genetic susceptibility to non-alcoholic fatty liver disease and age, a known risk factor for the disease[28].

Proof-of-concept clinical validation

As a proof-of-concept analysis, we examined whether some of the feature–metabolite interactions we uncovered may be causal. We used our diet-based models to select the top 5% of metabolites that were either positively or negatively associated with normal consumption of white or whole wheat bread (Fig. 4a, b, Methods). We then analysed the serum metabolome from the beginning and the end of a previously conducted week-long intervention[29], in which two randomized groups of ten healthy individuals increased their consumption of either wholegrain sourdough bread or industrial white bread (Fig. 4a, Methods). Notably, we found that metabolites that were positively associated with the consumption of whole wheat bread in our discovery cohort increased significantly more after the sourdough bread intervention (median fold change 1.62) than metabolites that were negatively associated with it (median fold change 0.66; Mann–Whitney *U*-test, $P < 10^{-10}$; Fig. 4c). Moreover, we found no statistically significant differences when comparing the mean fold change of these metabolites under the white bread intervention ($P > 0.1$; Fig. 4c).

Some metabolites for which levels increased after the sourdough bread intervention were previously linked to the consumption of wholegrain wheat flour. A notable example is betaine, an amino acid that has been shown to improve vascular risk factors[30] and is also highly abundant in wheat bran and germ[31]. We found that the mean fold change in betaine levels in the sourdough bread group was 6.16, as opposed to 0.82 in the white bread group (Mann–Whitney *U*-test, $P < 0.004$; Fig. 4d). Another example is cytosine, for which the mean fold change was far greater in the sourdough bread group (78.5) compared to the white bread group (0.53) ($P < 0.002$; Fig. 4e) To the best of our knowledge, unlike betaine, cytosine levels have not previously been linked to bread consumption.

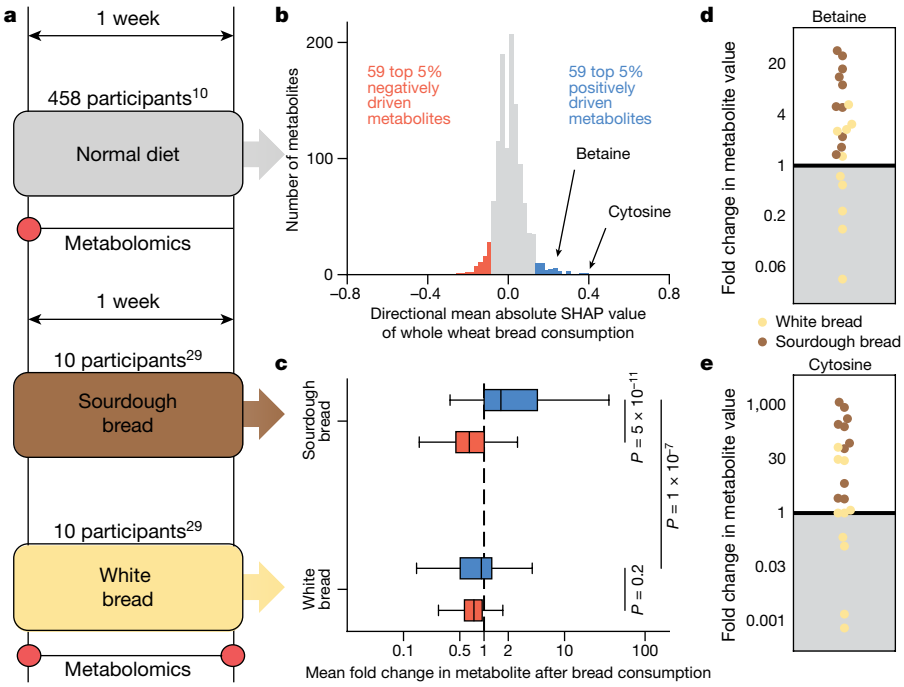


Fig. 4 | Increases in levels of metabolites explained by bread after an intervention of increased bread consumption. a, Schematic of the measurement of metabolites and of white bread and whole wheat bread consumption. The prediction models were constructed using samples from distinct participants who consumed and recorded their normal diet for one week[10] (n = 458 participants). We analysed samples from the first week of a randomized controlled trial[29], in which 10 participants increased their consumption of wholegrain sourdough bread and 10 others increased their consumption of industrial white bread. **b**, Histogram of directional mean absolute SHAP values of whole wheat bread consumption for metabolites computed on the basis of held-out samples from our cohort. The top 5% (n = 59; blue) positively associated metabolites and the top 5% (n = 59; red) negatively associated metabolites are used for further analysis. **c**, Box plots (centre, median; box, IQR; whiskers, 1.5 × IQR) showing the mean fold change in the top 5% positively (blue) and negatively (red) associated metabolites, separated by intervention group. They show a significantly higher mean fold change for the top 5% positively associated metabolites compared with the top 5% negatively associated metabolites under the sourdough bread intervention (two-sided Mann–Whitney U-test, $P = 5 \times 10^{-11}$). **d**, **e**, The fold changes of metabolite levels for both betaine (**d**; two-sided Mann–Whitney U-test, P = 0.0036) and cytosine (**e**; P = 0.0014) were higher in the sourdough bread group compared to the white bread group.

A similar analysis of metabolites that were associated with white bread consumption in our cohort did not find significant changes in their levels after intervention, potentially due to high baseline white wheat consumption in the typical diet of the study population. Overall, these results suggest that some of the associations that we found here might be causal.

Discussion

Although our cohort is not the largest in which serum metabolomics were measured, it is—to our knowledge—the only one in which these measurements were coupled with such a diverse array of potential determinants. Still, it has several limitations. First, although drug intake has been shown to have a large effect on the serum metabolome profile[32], our cohort was healthy and had limited drug intake. We are therefore likely to be underestimating the potential effect of drug intake on blood metabolites. Second, replication of results is still required for predictions by most factors other than the microbiome. Third, owing to the lack of reliable annotations, we have not associated metabolites with specific enzymes; this could be addressed in subsequent experimental studies by focusing on strongly predictive taxa. Finally, because this study is mainly based on observational data, interpretation of interactions should be made with caution, and the associations cannot be considered as causal.

Taken together, our results reveal a comprehensive list of potential determinants for circulating blood metabolites. Many of the associations and interactions detected here replicate previously reported findings, supporting the validity of our results. The majority of them, however, are new, making them a useful resource for future studies, either for improving molecular understanding of health and disease, or for forming the basis of interventional studies aimed at altering the levels of blood metabolites.

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/s41586-020-2896-2.

1. Psychogios, N. et al. The human serum metabolome. *PLoS ONE* **6**, e16957 (2011).
2. Shin, S.-Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
3. Long, T. et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.* **49**, 568–578 (2017).
4. Wikoff, W. R. et al. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl Acad. Sci. USA* **106**, 3698–3703 (2009).
5. Xu, T. et al. Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC Med.* **11**, 60 (2013).
6. Playdon, M. C. et al. Comparing metabolite profiles of habitual diet in serum and urine. *Am. J. Clin. Nutr.* **104**, 776–789 (2016).
7. Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **16**, 144–149 (2013).
8. Koivula, R. W. et al. Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: rationale and design of the epidemiological studies within the IMI DIRECT Consortium. *Diabetologia* **57**, 1132–1142 (2014).
9. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. Preprint at https://arxiv.org/abs/1802.03888v3 (2018).
10. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).

11. Yousri, N. A. et al. Long term conservation of human metabolic phenotypes and link to heritability. *Metabolomics* **10**, 1005–1017 (2014).
12. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) (Neural Information Processing Systems Foundation, 2017).
13. Cirulli, E. T. et al. Profound perturbation of the metabolome in obesity is associated with health risk. *Cell Metab*. **29**, 488–500.e2 (2019).
14. David, L. A. et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
15. Azam, S., Hadi, N., Khan, N. U. & Hadi, S. M. Antioxidant and prooxidant properties of caffeine, theobromine and xanthine. *Med. Sci. Monit*. **9**, BR325–BR330 (2003).
16. Tsutsumi, Y. et al. Renal disposition of a furan dicarboxylic acid and other uremic toxins in the rat. *J. Pharmacol. Exp. Ther*. **303**, 880–887 (2002).
17. Prentice, K. J. et al. CMPF, a metabolite formed upon prescription omega-3-acid ethyl ester supplementation, prevents and reverses steatosis. *EBioMedicine* **27**, 200–213 (2018).
18. Nemet, I. et al. A cardiovascular disease-linked gut microbial metabolite acts via adrenergic receptors. *Cell* **180**, 862–877.e22 (2020).
19. Hung, S.-C., Kuo, K.-L., Wu, C.-C. & Tarng, D.-C. Indoxyl sulfate: A novel cardiovascular risk factor in chronic kidney disease. *J. Am. Heart Assoc*. https://doi.org/10.1161/JAHA.116.005022 (2017).
20. Evenepoel, P., Meijers, B. K. I., Bammens, B. R. M. & Verbeke, K. Uremic toxins originating from colonic microbial metabolism. *Kidney Int*. **76**, S12–S19 (2009).
21. Yousri, N. A. et al. Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nat. Commun*. **9**, 333 (2018).
22. Suhre, K. et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54–60 (2011).
23. Gieger, C. et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*. **4**, e1000282 (2008).
24. Kettunen, J. et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet*. **44**, 269–276 (2012).
25. Capo-chichi, C. D. et al. Riboflavin and riboflavin-derived cofactors in adolescent girls with anorexia nervosa. *Am. J. Clin. Nutr*. **69**, 672–678 (1999).
26. Darst, B. F., Lu, Q., Johnson, S. C. & Engelman, C. D. Integrated analysis of genomics, longitudinal metabolomics, and Alzheimer's risk factors among 1,111 cohort participants. *Genet. Epidemiol*. **43**, 657–674 (2019).
27. Sookoian, S., Castaño, G., Gianotti, T. F., Gemma, C. & Pirola, C. J. Polymorphisms of MRP2 (*ABCC2*) are associated with susceptibility to nonalcoholic fatty liver disease. *J. Nutr. Biochem*. **20**, 765–770 (2009).
28. Hamaguchi, M. et al. Aging is a risk factor of nonalcoholic fatty liver disease in premenopausal women. *World J. Gastroenterol*. **18**, 237–243 (2012).
29. Korem, T. et al. Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. *Cell Metab*. **25**, 1243–1253.e5 (2017).
30. Olthof, M. R., van Vliet, T., Boelsma, E. & Verhoef, P. Low dose betaine supplementation leads to immediate and long term lowering of plasma homocysteine in healthy men and women. *J. Nutr*. **133**, 4135–4138 (2003).
31. Craig, S. A. S. Betaine in human nutrition. *Am. J. Clin. Nutr*. **80**, 539–549 (2004).
32. Liu, J. et al. Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug-metabolite atlas. *Nat. Med*. **26**, 110–117 (2020).

The IMI DIRECT consortium

Henrik Vestergaard[14,15], Manimozhiyan Arumugam[14], Torben Hansen[14], Kristine Allin[14], Tue Hansen[14], Mun-Gwan Hong[16], Jochen Schwenk[16], Ragna Haussler[16], Matilda Dale[16], Toni Giorgino[17,18], Marianne Rodriquez[19], Mandy Perry[20], Rachel Nice[20], Timothy McDonald[20,21], Andrew Hattersley[21], Angus Jones[21], Ulrike Graefe-Mody[22], Patrick Baum[23], Rolf Grempler[23], Cecilia Engel Thomas[24,25,26], Federico De Masi[25,26], Caroline Anna Brorsson[25,26], Gianluca Mazzoni[25,26], Rosa Allesøe[25,26], Simon Rasmussen[25,26], Valborg Gudmundsdóttir[25,26], Agnes Martine Nielsen[25,26], Karina Banasik[25,26], Konstantinos Tsirigos[25,26], Birgitte Nilsson[25,26], Helle Pedersen[25,26], Søren Brunak[25,26], Tugce Karaderi[25,26], Agnete Troen Lundgaard[25,26], Joachim Johansen[25,26], Ramneek Gupta[26], Peter Wad Sackett[26], Joachim Tillner[27], Thorsten Lehr[28], Nina Scherer[28], Christiane Dings[28], Iryna Sihinevich[28], Heather Loftus[29], Louise Cabrelli[29], Donna McEvoy[30], Andrea Mari[31], Roberto Bizzotto[31], Andrea Tura[31], Leen 't Hart[32,33,34], Koen Dekkers[33], Nienke van Leeuwen[33], Roderick Slieker[33,34], Femke Rutters[34], Joline Beulens[34], Giel Nijpels[34], Anitra Koopman[34], Sabine van Oort[34], Lenka Groeneveld[34], Leif Groop[35], Petra Elders[36], Ana Viñuela[37,79], Anna Ramisch[37], Emmanouil Dermitzakis[37], Beate Ehrhardt[38], Christopher Jennison[38], Philippe Froguel[39,40], Mickaël Canouil[40], Amélie Boneford[40], Ian McVittie[41], Dianne Wake[41], Francesca Frau[42], Hans-Henrik Staerfeldt[43], Kofi Adragni[44], Melissa Thomas[44], Han Wu[44], Imre Pavo[45], Birgit Steckel-Hamann[45], Henrik Thomsen[46], Giuseppe Nicola Giordano[47], Hugo Fitipaldi[47], Martin Ridderstråle[47], Azra Kurbasic[47], Naeimeh Atabaki Pasdar[47], Hugo Pomares-Millan[47], Pascal Mutie[47], Robert Koivula[47,48], Nicky McRobert[48], Mark McCarthy[48,49,50,80], Agata Wesolowska-Andersen[48], Anubha Mahajan[50,80], Moustafa Abdalla[50], Juan Fernandez[50], Reinhard Holl[51], Alison Heggie[52], Harshal Deshmukh[52], Anita Hennige[53], Susanna Bianzano[53], Barbara Thorand[54,55], Sapna Sharma[55,56], Harald Grallert[55,56], Jonathan Adam[56], Martina Troll[56], Andreas Fritsche[57], Anita Hill[58], Claire Thorne[58], Michelle Hudson[58], Teemu Kuulasmaa[59], Jagadish Vangipurapu[59], Markku Laakso[59], Henna Cederberg[59], Tarja Kokkola[59], Yunlong Jiao[60], Stephen Gough[60], Neil Robertson[60], Helene Verkindt[61], Violeta Raverdi[61], Robert Caiazzo[61], Francois Pattou[61], Margaret White[62], Louise Donnelly[62], Andrew Brown[62], Colin Palmer[62], David Davtian[62], Adem Dawed[62], Ian Forgie[62], Ewan Pearson[62], Hartmut Ruetten[53], Petra Musholt[63], Jimmy Bell[64], Elizabeth Louise Thomas[64], Brandon Whitcher[64], Mark Haid[65], Claudia Nicolay[66], Miranda Mourby[67], Jane Kaye[67,68], Nisha Shah[67], Harriet Teare[67], Gary Frost[69], Bernd Jablonka[70], Mathias Uhlen[71], Rebeca Eriksen[72], Josef Vogt[73], Avirup Dutta[73], Anna Jonsson[73], Line Engelbrechtsen[73], Annemette Forman[73], Nadja Sondertoft[73], Nathalie de Preville[74], Tania Baltauss[74], Mark Walker[75], Johann Gassenhuber[76], Maria Klintenberg[77], Margit Bergstrom[77] & Jorge Ferrer[78]

[15]Bornholms Hospital, Rønne, Denmark. [16]Affinity Proteomics, Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Solna, Sweden. [17]Biophysics Institute (IBF-CNR), National Research Council of Italy, Milan, Italy. [18]Department of Biosciences, University of Milan, Milan, Italy. [19]Biotech & Biomarkers Research Department, Institut de Recherches Internationales Servier, Croissy sur Seine, France. [20]Blood Sciences, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. [21]Institute of Clinical and Biological Sciences, University of Exeter Medical School, Exeter, UK. [22]Therapeutic Area CNS, Retinopathies and Emerging Areas, Boehringer Ingelheim International GmbH, Ingelheim am Rhein, Germany. [23]Translational Medicine & Clinical Pharmacology, Boehringer Ingelheim International GmbH, Biberach an der Riss, Germany. [24]Affinity Proteomics, Science for Life Laboratory, School of Biotechnology, KTH - Royal Institute of Technology, Solna, Sweden. [25]Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [26]Section for Bioinformatics, Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark. [27]Clinical Operations, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany. [28]Clinical Pharmacy, Saarland University, Saarbrücken, Germany. [29]Clinical Research Centre, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [30]Clinical Research Facility, Royal Victoria Infirmary, Newcastle upon Tyne, UK. [31]CNR Institute of Neuroscience, Padova, Italy. [32]Department of Biomedical Data Sciences, Molecular Epidemiology Section, Leiden University Medical Center, Leiden, The Netherlands. [33]Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, The Netherlands. [34]Department of Epidemiology and Biostatistics, Amsterdam UMC- location VUmc, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands. [35]Department of Clinical Sciences, Diabetes and Endocrinology Unit, Lund University, Skåne University Hospital Malmö, CRC, Malmö, Sweden. [36]Department of General Practice, Amsterdam UMC- location VUmc, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands. [37]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [38]Department of Mathematical Sciences, University of Bath, Bath, UK. [39]Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK. [40]Université de Lille, INSERM UMR 1283, CNRS UMR 8199, Institut Pasteur de Lille, EGID, Lille, France. [41]Diabetes Research Network, Royal Victoria Infirmary, Newcastle upon Tyne, UK. [42]Digital and Data Sciences, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany. [43]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark. [44]Eli Lilly and Company, Indianapolis, IN, USA. [45]Eli Lilly Regional Operations GmbH, Vienna, Austria. [46]Faculty of Medical and Health Sciences, University of Copenhagen, Copenhagen, Denmark. [47]Genetic and Molecular Epidemiology Unit, Lund University Diabetes Centre, Department of Clinical Sciences, CRC, Lund University, SUS, Malmö, Sweden. [48]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [49]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [50]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [51]Institute for Epidemiology and Medical Biometry, ZIBMT, University of Ulm, Ulm, Germany. [52]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK. [53]Therapeutic Area CardioMetabolism and Respiratory Medicine, Boehringer Ingelheim International GmbH, Ingelheim am Rhein, Germany. [54]Institute of Epidemiology II, Research Unit of Diabetes Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany. [55]German Center for Diabetes Research (DZD), Neuherberg, Germany. [56]Research Unit of Molecular Epidemiology, Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany. [57]Medizinische Universitätsklinik Tübingen, Eberhard Karls Universität Tübingen, Tübingen, Germany. [58]NIHR Exeter Clinical Research Facility, University of Exeter Medical School, Exeter, UK. [59]Internal Medicine, Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland. [60]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK. [61]Inserm, Université de Lille, CHU Lille, Lille Pasteur Institute, EGID, Lille, France. [62]Population Health and Genomics, School of Medicine, University of Dundee, Dundee, UK. [63]R&D Global Development, Translational Medicine & Clinical Pharmacology (TMCP), Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany. [64]Research Centre for Optimal Health, Department of Life Sciences, University of Westminster, London, UK. [65]Research Unit of Molecular Endocrinology and Metabolism, Helmholtz Zentrum München, Neuherberg, Germany. [66]Lilly Deutschland GmbH, Bad Homburg, Germany. [67]Centre for Health, Law and Emerging Technologies (HeLEX), Faculty of Law, University of Oxford, Oxford, UK. [68]Technologies (HeLEX), Melbourne Law School, University of Melbourne, Carlton, Victoria, Australia. [69]Section for Nutrition Research, Faculty of Medicine, Imperial College London, London, UK. [70]Strategy and Innovation, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany. [71]Systems Biology, Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Solna, Sweden. [72]Section for Nutrition Research, Division of Digestive Diseases, Department of Metabolism, Digestion and Reproduction, Faculty of Medicine, Imperial College London, London, UK. [73]The Novo Nordisk Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Science, University of Copenhagen, Copenhagen, Denmark. [74]Translational and Clinical Research, Metabolism Innovation Pole, Institut de Recherches Internationales Servier, Suresnes, France. [75]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. [76]Diabetes Division, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany. [77]VO Endokrinologi, Enheten för Diabetesstudier, Skånes Universitetssjukhus i Lund, Lund, Sweden. [78]Institut d'Investigacions Biomediques August Pi i Sunye, Centre Esther Koplowitz, Barcelona, Spain. [79]Present address: Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. [80]Present address: Human Genetics, Genentech, South San Francisco, CA, USA.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Description of cohorts

We analysed banked samples from two previously collected cohorts[10,29], for a total of 491 Israeli individuals. Studies were approved by Tel Aviv Sourasky Medical Center Institutional Review Board (IRB), approval numbers TLV-0658-12, TLV-0050-13 and TLV-0522-10; Kfar Shaul Hospital IRB, approval number 0–73. All participants signed written informed consent forms. Full study designs, including inclusion and exclusion criteria are described elsewhere[10,29]. In brief, participants in both studies were healthy individuals aged between 18 and 70. The participants answered detailed medical, lifestyle and nutritional questionnaires, provided stool and serum samples for metagenomic sequencing and metabolomics, were genotyped, underwent a comprehensive blood test, and for a period of at least one week, recorded all of their daily activities and nutritional intake in real-time using their smartphones with a specialized app provided to them[29]. Both blood and stool samples were not taken under strict fasting conditions. Sixteen samples of participants for which microbiome data was not available to us were excluded from all analyses. Meetings in which participants provided blood samples took place in two different centres, Weizmann (45% of participants) and Tel-Aviv (55% of participants). All meetings in Weizmann took place within the first half of the day, whereas most meetings in Tel-Aviv took place during the second half of the day (82% of the participants).

Feature groups

The 'diet' feature group includes both answers for a detailed food frequency questionnaire (FFQ) aimed at capturing long term dietary habits, and the daily mean consumption of different food types, computed over a week based on real-time logging. In both cases we kept only items that were reported to be consumed at least once by at least 1% of our participants, resulting in 670 different food types from logging, and 141 different items from the FFQ.

The 'macronutrients' feature group includes the daily mean consumption of macronutrients (lipids, proteins, carbohydrates), calories and water, calculated from real-time logging.

The 'anthropometrics' feature group includes weight, BMI, waist and hips circumference, and waist-to-hips ratio.

The 'cardiometabolic' feature group includes systolic and diastolic blood pressure, heart rate in beats per minute and a glycemic status as previously described[33].

The 'drugs' feature group includes 30 binary features representing the intake of 20 common medications as reported in questionnaires, in addition to 10 medication groups as previously described[33]. We included only drugs reported to be used by at least 1% of our participants.

The 'clinical data' feature group includes the age and sex of the participants, and the following feature groups described above: anthropometrics, cardiometabolic and drugs.

The 'lifestyle' feature group includes smoking status (current, past), stress levels obtained from questionnaires, and the daily mean sleeping time, exercise time and midday sleep time based on real time logging.

The 'time of day' feature is a binary feature indicating whether the sample was taken during the first half of the day.

The 'seasonal effects' feature is the month in which the sample was taken. In some analyses we also grouped months by season (Winter: December–February; Spring: March–May; Summer: June–August; Autumn: September–November).

The 'microbiome' feature group includes bacterial relative abundance calculated both by considering coverage (see the following section 'Microbiome preprocessing'), and by MetaPhlAn2[34], as well as the first 10 principal components computed over the log transformed relative abundance of a bacterial gene catalogue[35] as previously described[33,36]. Preprocessing steps are described in the following section.

We further defined a full model that included all of the above.

Metabolomics profiling and preprocessing

Metabolite concentrations were measured in serum samples by Metabolon, by using an untargeted liquid chromatography coupled to mass spectrometry (LC–MS) platform as previously described[2,37,38]. A total of 540 serum samples were profiled, 19 of which were control samples (technical replicate) pooled from several individuals. The other 521 serum samples belonged to 491 participants.

We removed from further analysis 27 metabolites with fewer than 10 measurements across our cohort, and 54 metabolites that we found to have significantly different distributions in samples collected in two different recruitment centres (Mann–Whitney U-test, $P < 0.05/1{,}251$; Bonferroni corrected; Supplementary Table 14). For the remaining 1,170 metabolites, we performed robust standardization (subtracting the median and dividing by the standard deviation) over the log (base 10)-transformed levels, followed by clipping outlier samples which were further than 5 standard deviations. We next used two separate normalization schemes: one for single metabolites, which we subsequently used in the feature attribution analysis, and the second for metabolite groups, which we used for global and enrichment analyses.

For single metabolites, we regressed metabolite levels against storage times (only for metabolites present in at least 50 samples), and finally, imputed missing values as the minimum value per metabolite. For the second scheme, metabolites were grouped by correlation with a Spearman's ρ threshold of 0.85. This is done in order to handle possible bias resulting from uncertainty of metabolite assignments and a high rate of highly correlated mass spectrometry peaks, and resulted in 1,067 metabolite groups, 982 of which are singletons. The value of the metabolite group was set to the mean. The category of each metabolite group was assigned based on majority vote, where unidentified compounds were excluded from the vote unless all metabolites in the group were unidentified.

Microbiome preprocessing

Sample collection, DNA extraction, and sequencing of the samples in this study was previously described[10,29,33]. In brief, we used only samples that were collected using swabs, filtered metagenomic reads containing Illumina adapters, filtered low-quality reads and trimmed low-quality read edges. We detected host DNA by mapping with GEM[39] to the human genome (hg19) with inclusive parameters, and removed human reads. We subsampled all samples to have 10 million reads.

Bacterial relative abundance estimation was performed by mapping bacterial reads to species-level genome bins (SGB) representative genomes[40] (Supplementary Table 15). We selected all SGB representatives from groups with at least five genomes, and for these representatives genomes kept only unique regions as a reference dataset. Mapping was performed using Bowtie2[41] and abundance was estimated by calculating the mean coverage of unique genomic regions across the 50% most densely covered areas as previously described[36,42]. Feature names include the lowest taxonomy level identified.

Comparing metabolomics to laboratory tests

We compared the levels of both creatinine and cholesterol, which we previously obtained via standard laboratory tests[10] with their metabolomic levels. Because the tests were performed by two different laboratories, we centred the tests by reducing from the value of each sample the mean of all tests taken in the laboratory in which it was performed. We then performed a standardization of the resulting measurements. The metabolomic profiling and the laboratory tests were performed on two samples taken at the same blood draw.

Correlation of metabolic profiles within and between individuals

We compared the Spearman's correlations between standardized metabolomic profiles of the same participant taken one week apart ($n = 20$) to correlations between standardized metabolomic profiles of different individuals ($n = 475$). Each pair of samples taken from the same participant was run in the same metabolomic batch. In the group of different individuals, only pairs of individuals from the same batch were included (resulting in a total of 3,835 such pairs), and were further stratified by sex.

Predictive models of metabolite groups

We used gradient boosting decision trees from the LightGBM (v.2.1.2) package[12], in order to predict the levels of 1,067 metabolite groups on the basis of 7 feature groups in held-out individuals. In order to estimate the EV of each metabolite group we ran a fivefold cross-validation (CV) model using each feature group as input, and evaluated the results using the coefficient of determination (R^2). For all prediction results except those based on human genetics (Methods) we computed 95% confidence intervals and P values via 1,000 iterations of bootstrapping[43]. In each bootstrap iteration, we performed a random fivefold cross validation, in which in each fold we randomly sampled (with replacement) a group of participants from the training set to have the same size as the current training set. We next used this set to train our model and evaluated the performance of the model on the set of participants in the remaining fold. Finally, we computed the coefficient of determination between the measured values of the metabolite and the concatenation of the CV's predicted values as obtained from the bootstrapping iteration. We applied the Fisher transformation to the estimations of the explained variance we got from bootstrapping in order to induce normality[44], and then computed a standard error, and estimated the P values via the normal cumulative distribution function using the Wald test[45], such that our null hypothesis is that the explained variance should distribute normally with zero mean. Confidence intervals were computed empirically from the bootstrapping results. We corrected P values of predictions for multiple hypotheses using the Benjamini–Hochberg procedure over all feature groups (10% FDR). In all CV and bootstrapping runs we used a fixed and predetermined set of hyperparameters: for the microbiome and diet feature groups: learning_rate = 0.005, max_depth = default, feature_fraction = 0.2, num_leaves = default, min_data_in_leaf = 15, metric = L2, early_stopping_rounds = None, n_estimators = 2000, bagging_fraction = 0.8, bagging_freq = 1; for other feature groups: learning_rate = 0.01, max_depth = 5, feature_fraction = 0.8, num_leaves = 25, min_data_in_leaf = 15, metric = L2, early_stopping_rounds = None, n_estimators = 200, bagging_fraction = 0.9, bagging_freq = 5.

Human genetics based prediction models

To obtain the predictions based on human genetics, we used a similar fivefold CV scheme, in which in every fold we calculated the associations between SNPs and metabolite levels within the training fold, and then trained a model on only the top 10 SNPs that reached genome-wide significance (Bonferroni-adjusted). For folds in which no SNP reached the significance level, we assigned every sample in the test fold with the mean metabolite level of the training fold. Owing to high complexity and running time issues, P values and confidence intervals were not computed based on bootstrapping; rather, we estimated the P values of the Pearson's correlation between the true and predicted metabolite levels. Metabolites for which the R^2 was negative were assigned a P value of 1.

Testing for SNP associations with metabolites

Genotype processing and imputation of 413 individuals were described previously[33]. We performed genome-wide associations for single metabolites ($n = 1,170$) and calculated the P value and the estimated effect sizes using plink (v.1.07). When declaring a genome-wide significance for the SNP–metabolite associations we used a conservative Bonferroni adjustment procedure to control for the false discovery rate due to the large number of SNPs tested ($P < (5 \times 10^{-8})/1,170$). We performed all genome wide associations using imputed genotypes.

For named molecules, their chemical identification, super and sub pathways are presented as provided by Metabolon. For unidentified molecules, super and sub pathways are estimated on the basis of our biological pathway classifier. We did our best to scan the available literature for known associations between genetic loci and metabolites before reporting an association as novel. The main resources included the GWAS Catalog[46] and the GWAS server[2,22].

Pathway category enrichment analysis

For each pathway category we used a Mann–Whitney U-test comparing the prediction accuracy of metabolites from that category compared to prediction accuracy of metabolites from other categories. Direction of enrichment was determined by the sign of the Mann–Whitney U-test statistic. We considered only metabolite groups for which at least one feature group had a significant prediction (after correcting for multiple hypotheses), resulting in 819 metabolite groups.

Validation of metabolite predictions based on microbiome

We validated the robustness of the associations between the gut microbiome composition and the levels of circulating metabolites in two independent cohorts in which we had access to both metagenomics sequencing. Serum metabolomics in these cohorts were performed using the same Metabolon platform that we used for the discovery cohort. The first validation cohort included 1,004 samples of healthy participants from the TwinsUK cohort[7], for which there was an average of 0.9 ± 1.3 years gap between the collection of faecal and blood samples. The second validation cohort included 245 samples of participants of European ancestry with type 2 diabetes (T2D) from the IMI DIRECT consortium[8]. Data from both these validation cohorts were not available to us while developing the prediction models. The metagenomics samples from both cohorts went through the exact same analysis pipeline as our discovery cohort to extract the bacterial features that our prediction models were based on. We then applied our models on these data to obtain the metabolite predictions for both cohorts. Only metabolites that were significantly predicted based only on microbiome data with $R^2 > 5\%$ (FDR <0.1) in our discovery cohort were considered for further analysis (107 metabolites out of 678 in TwinsUK, 50 metabolites out of 261 in IMI DIRECT). Within every validation cohort, we performed robust standardization (subtracting the median and dividing by the standard deviation) over the log (base 10) transformed levels, followed by clipping outlier samples which were further than 5 standard deviations, and finally, imputed missing values as the minimum value per metabolite. The analysis of these geographically distinct cohorts holds multiple potential sources of noise, including different methods, centres and staff involved in assembling these cohorts, as well as different cohort demographics, clinical manifestations, different genetic background and dietary and lifestyle preferences. Therefore, we defined a successful replication as one that restores the original ranking of the participants as dictated by the true levels of the metabolite in hand. Hence, for every metabolite, in each validation cohort, we computed the Spearman's correlation between its true levels and its predicted levels. A replication was considered significant if the FDR adjusted P value of the Spearman's correlation was lower than 0.1 and the correlation coefficient was strictly positive.

Feature attribution analysis

To explain the output of our machine learning models and find specific associations between features and metabolite levels, we used SHAP (Shapley additive explanations)[47], a recently introduced framework for interpreting predictions, which assigns each feature an importance value

for a particular prediction. In brief, for a specific prediction, the SHAP value of a feature is defined as the change in the expected value of the output of the model when this feature is observed versus when it is missing. It is computed using a sum that represents the effect of each feature being added to the model averaged over all possible orderings of features being introduced. Shapley-value-based analysis in gut microbiome data was recently demonstrated to be useful, as it enabled the estimation of complex contributions of gut microbiome taxa to functional shifts, while maintaining global community composition properties[48].

Individual SHAP values were computed for held-out individuals in fivefold CV using the module TreeExplainer (v.0.24.0)[9,49], based on models trained only on features from the respective feature group. Before training, we standardized the levels of target metabolites, so that SHAP values from different models would be comparable (they are measured in the same units as the target). In each CV fold we ran a random hyperparameter search consistent of 10 iterations using the module RandomizedSearchCV from sklearn (v.0.20.4), and chose the best model for predicting the held-out individuals and computing SHAP values. In all feature attribution analyses we used the ungrouped list of 1,170 metabolites.

For every feature, we computed the mean absolute SHAP value across all instances in a specific model, reflecting the mean effect of each feature on the predictions and serving as a feature importance measure. We further used these values to compute directional mean absolute SHAP values, by multiplying them with the sign of the Spearman correlation between the population feature and the target. Here, positive values indicate that higher feature values lead, on average, to higher predicted values, whereas negative values indicate that lower feature values lead, on average, to higher predicted values.

When performing feature attribution analysis with gut microbiome data as input, we included only the relative abundance of SGB representative genomes as features, taking only features which were present in over 5% of the samples, resulting with 753 bacterial taxa. When using diet as input, we considered only features that were present in at least 5% of the samples, resulting with 398 food types from logging and items from the FFQ.

Comparing gradient boosting decision trees with a linear model

We compared the EV of every single metabolite obtained for a GBDT and a Lasso regression model. The EV of all models were calculated in fivefold CV, where in each fold we ran a hyperparameter search consistent of 10 iterations as described above. We used LightGBM as the GBDT model, and Lasso regression (sklearn, v.0.20.4) as the linear model, since its regularization scheme is better suited for a large number of features, as in the case of diet and gut microbiome composition. Because GBDT handles missing values well, we first imputed all missing values as the median of each feature to assure a fair comparison. When applying the models on the microbiome data, we used $\log_{10}$-transformed values.

Estimating relative predictive power of feature groups

To estimate the relative predictive power of different feature groups we first applied a principal component analysis over the metabolite groups data to get the first 400 PCs which constitute more than 99% of the total variance in the data (Extended Data Fig. 5b). We then used fivefold CV prediction models as described above to predict the PCs on the basis of the different feature groups independently. As baseline, we used the full model, which consists of all features combined to predict the levels of the PCs, and estimated the overall fraction of variance explained by: $\frac{\sum_{i=1}^{n\mathrm{PC}} \mathrm{EV}_i \times \mathrm{PC}_i}{\sum_{i=1}^{n\mathrm{PC}} \mathrm{PC}_i}$, where EV_i is the fraction of EV that the model recovers for PC *i*. PC_i is the fraction of variance that PC *i* explains out of the overall variation in the data. *n*PC is the number of the first PCs, those which capture the most variation. For the features we have collected, we defined this sum obtained for the full model as the total explainable variance in circulating blood metabolites. Next, for every feature group we computed a similar expression and calculated the relative predictive power by dividing this expression by that of the full model. The estimates we present are for *n*PC = 15, as the overall EV of the full model that we estimated using the first 15 PCs constitutes over 97% of the overall EV of the full model based on all 400 PCs.

Biological sub pathway prediction

We used gradient boosting decision trees from the LightGBM (v.2.1.2) package[12], in order to build a multiclass classifier to predict the biological sub pathways of metabolites as annotated by Metabolon. When developing the classifier, we considered only named metabolites from biological sub pathways which include over 10 metabolites each in our data, resulting with 28 sub pathways covering a total of 572 named molecules (sub pathway size range 11–44). The rationale behind this is that we tried to find the balance between covering as many metabolites and types of metabolites possible while keeping the number of classes reasonable.

We trained our model in a leave-one-out CV scheme, in which in every training fold we used 20% of the training samples as internal validation to perform an early stopping of 50 rounds. We then obtained a soft max of size 28 per metabolite, representing the probabilities of every metabolite being labelled as one of the 28 sub pathways. For the prediction of the unidentified molecules, we used a model trained once using all 572 metabolites. The features used for the training of the model included the normalized levels of metabolites across our main discovery cohort, the mean raw count of the metabolite and the fraction of missing values across the discovery cohort. In addition, to capture the associations between metabolites and their predictive features, we included the directional mean absolute SHAP values for every pair of metabolite–feature computed from the 'full model' as described above. The final vectors of probabilities were determined as an ensemble of three models: the first, trained only on the SHAP values; the second, trained only on metabolite levels, means and fraction of missing values; and the third, trained on all combined. Finally, the mean of these three models was computed.

When evaluating the performance of our classifier on the named labelled molecules, we concatenated all vectors of probabilities resulting from the leave-one-out procedure. For every sub pathway we computed a classification report including the classification precision (TP/(TP + FP)), recall (TP/(TP + FN)) and f1-score (2 × (precision × recall)/(precision + recall)), to account for the imbalanced class sizes. The overall accuracy was computed as the fraction of metabolites with correctly assigned labels out of all metabolites from all sub pathways which were included in the training phase. In all runs we used a fixed and predetermined set of hyperparameters (objective = multiclass, num_leaves = 25, max_depth = 4, learning_rate = 0.005, bagging_fraction = 0.8, feature_fraction = 0.8, bagging_freq = 1, bagging_seed = 2018, class_weight = balanced, n_estimators = 2000, early_stopping_rounds = 50). TP, true positive; FP, false positive; FN, false negative.

Characterization of unidentified metabolites by metabolon

Characterization of unidentified metabolites was done as previously described[21]. In brief, identification of tentative structural features for unidentified biochemicals incorporates a detailed analysis of mass spec data, that is, gathering information such as the accurate monoisotopic mass, the elution time and fragmentation pattern of the primary ion, and correlation to other molecules. The accurate monoisotopic mass is used to identify a probable structural formula for the unidentified biochemical, which is then used to search against chemical structure databases. When a candidate structure fits the accurate monoisotopic mass and fragmentation data, an authentic standard is commercially purchased or synthesized (when possible). Conformation of a proposed structure is based on a match to three primary criteria, including co-elution with the unidentified molecule of interest, and a high degree match to both the accurate monoisotopic mass and fragmentation pattern.

Interaction networks

We used a graphical layout in order to visualize the associations of features with the levels of metabolites. The nodes are either metabolites or features, and the edges are the directional mean absolute SHAP values computed from models trained only on features from the respective feature group as described above. All networks were constructed using Cytoscape[50]. The threshold for presenting SHAP values as edges was determined as 0.12, keeping the network sparse enough for convenience of visualization.

Analysis of bread intervention

In order to find the associations between metabolite levels and the consumption of both types of bread in the study cohort we computed the directional mean absolute SHAP values of the reported consumption of both white and whole-wheat bread for all metabolites. The SHAP values were computed in cross validation from models based only on the reported consumption of each type of bread. We ranked the metabolites according to their directional mean absolute SHAP value for each type of bread and used the top 5% positively and negatively driven metabolites for further analysis. The prediction models were constructed using 458 samples of distinct individuals, a subset of our cohort from which we excluded all samples of individuals which participated in the intervention study.

For each metabolite in every individual, we computed the fold change of metabolite levels between the samples taken at the end of the first week of intervention and the start of that week. Before computing fold change we imputed missing values with the minimum per metabolite and standardized their log (base 10) transformed levels. Furthermore, for each intervention group, we computed the mean fold change of every metabolite based on the 10 samples from that group. We then compared the mean fold change of the top 5% positively and negatively associated metabolites mentioned above within each intervention group by performing a rank sum test (two-sided Mann–Whitney *U*-test) over the mean fold change.

For comparing the fold change of betaine and cytosine between the two intervention groups, we used a two-sided Mann–Whitney *U*-test.

Linear mixed models-based estimates of the explained variance of metabolites using gut microbiome

For the in-sample estimation of EV for metabolites based on gut microbiome we used a linear mixed model framework that we had recently developed[33]. In brief, we used GCTA[51], a tool used in statistical genetics for the estimating of SNP-based genetic kinship. Instead of a matrix of host SNPs, as is commonly used in GCTA, we used a kinship matrix computed over the presence-absence of microbial species which were also used as features in the out-of-sample prediction models. We added the storage time as a covariate to the model. *P* values were computed using RL-SKAT[52].

Statistical analysis

For all statistical analysis and prediction models we used Python 2.7.8 with the following packages: pandas 0.23.4, numpy 1.14.2, scikit-learn 0.20.4, scipy 1.1.0, shap 0.24.0, LightGBM 2.1.2.

Reporting summary

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

Data availability

The raw metagenomic sequencing data are available from the European Nucleotide Archive under accession numbers PRJEB11532, PRJEB17643, and for TwinsUK PRJEB32731. The raw metabolomics data and the phenotypic data are available from the European Genome–phenome Archive (EGA; https://ega-archive.org/) with accession number EGAS00001004512. Known links between genetic loci and serum metabolites were taken from the GWAS Catalog[46] (https://www.ebi.ac.uk/gwas/) and the GWAS server[2,22] (http://metabolomics.helmholtz-muenchen.de/gwas/).

Code availability

Source code for analysis is available at https://github.com/noambar/SerumMetabolomePredictions.

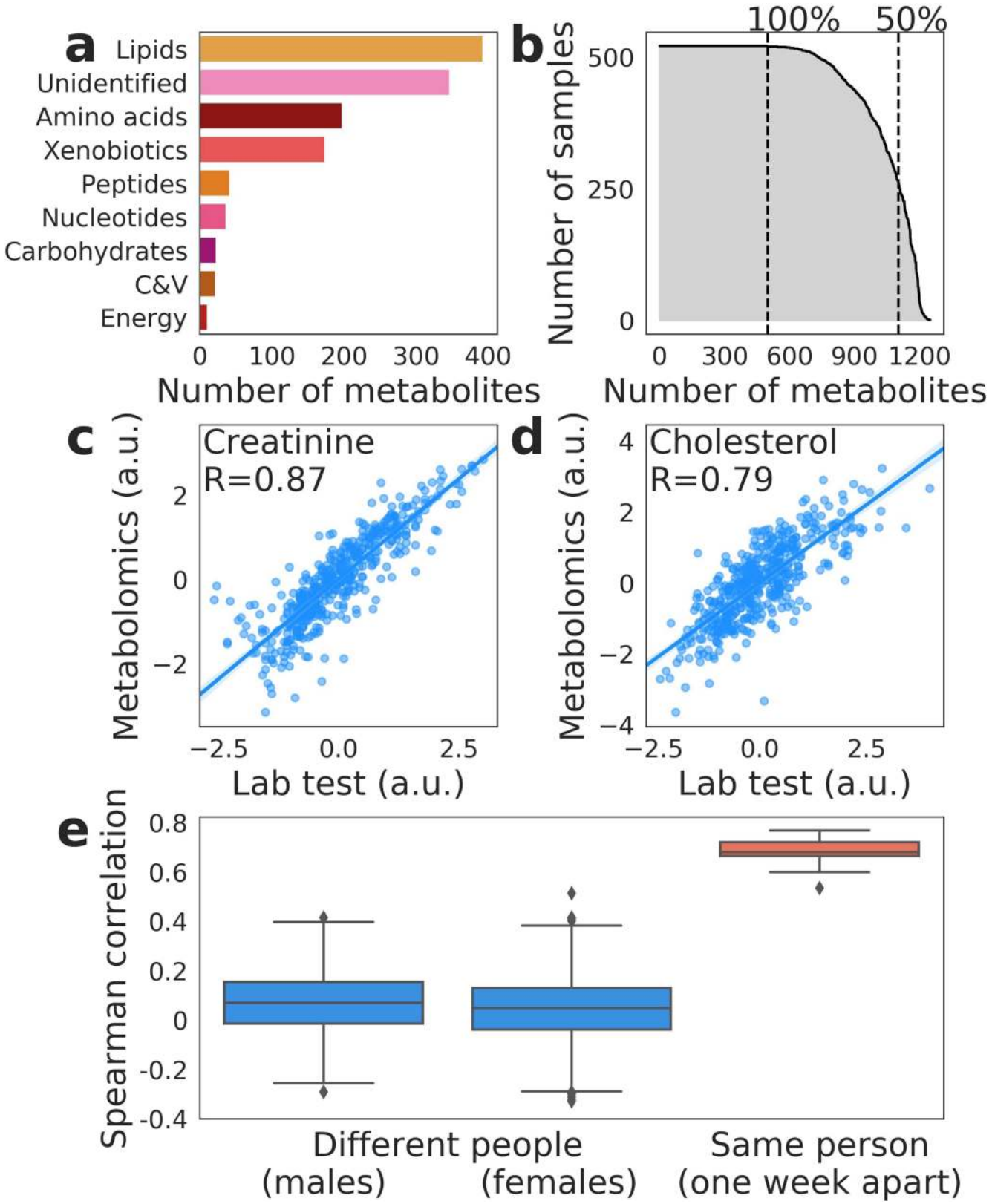
33. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
34. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
35. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
36. Zeevi, D. et al. Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).
37. Evans, A. M. et al. High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. *Metabolomics* **4**, 1000132 (2014).
38. Zierer, J. et al. The fecal metabolome as a functional readout of the gut microbiome. *Nat. Genet.* **50**, 790–795 (2018).
39. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
40. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662.e20 (2019).
41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
42. Korem, T. et al. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–1106 (2015).
43. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman and Hall, 1994).
44. Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507 (1915).
45. Wald, A. Sequential tests of statistical hypotheses. *Ann. Math. Stat.* **16**, 117–186 (1945).
46. 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 (2019).
47. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. Preprint at https://arxiv.org/abs/1705.07874v2 (2017).
48. Manor, O. & Borenstein, E. Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome. *Cell Host Microbe* **21**, 254–267 (2017).
49. SHAP https://github.com/slundberg/shap
50. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
51. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
52. Schweiger, R. et al. RL-SKAT: an exact and efficient score test for heritability and set tests. *Genetics* **207**, 1275–1283 (2017).

Acknowledgements We thank past and present members of the Segal group for discussions. N.B. received a PhD scholarship for Data Science by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center and is supported by a research grant from Madame Olga Klein Astrachan. T.K. is a CIFAR Azrieli Global Scholar in the Humans & the Microbiome Program. E.S. is supported by the Crown Human Genome Center, by D. L. Schwarz, J. N. Halpern and L. Steinberg, and by grants funded by the European Research Council and the Israel Science Foundation. The work leading to this publication has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no.115317 (DIRECT), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution from EFPIA companies. We thank A. Dutta for introducing us to the DIRECT consortium dataset.

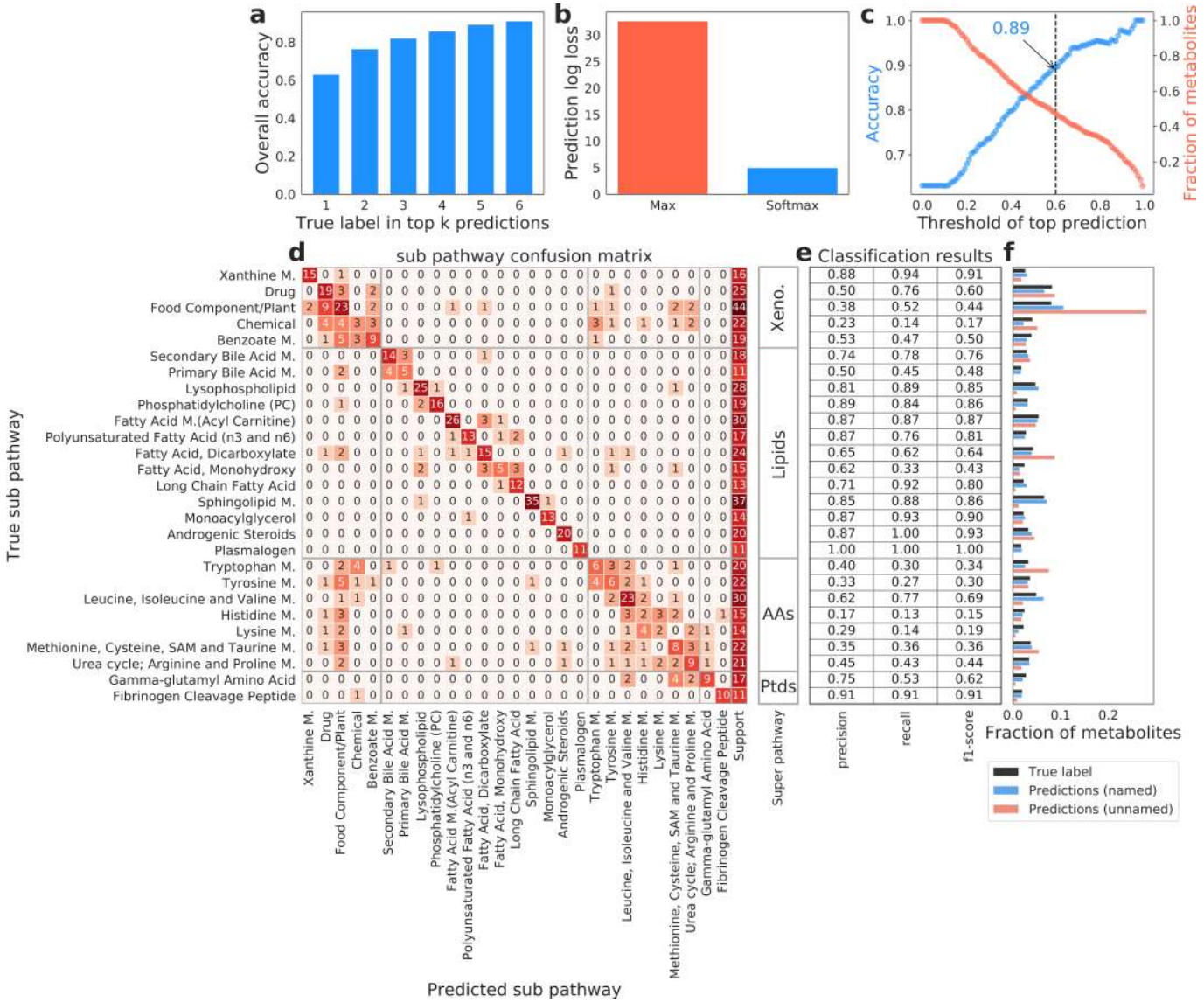
Author contributions N.B. and T.K. conceived the project, designed and conducted all analyses, interpreted the results and wrote the manuscript, and are listed in arbitrary order. O.W. and D.Z. designed statistical analyses. D.R. and S.L. conducted microbiome analysis. N.K. coordinated and designed data collection. M.L.-P. and A.W. developed protocols, performed microbiome sequencing and processed serum samples. A.W. designed the project and oversaw sample collection and processing. C.I.L.R., C.M., A.V., M.F. and T.D.S. performed the replication analysis on the TwinsUK cohort. J.A., P.W.F. and O.P. performed the replication analysis on the IMI DIRECT cohort. E.S. conceived and directed the project and analyses, designed the analyses, interpreted the results and wrote the manuscript.

Competing interests The authors declare no competing interests.

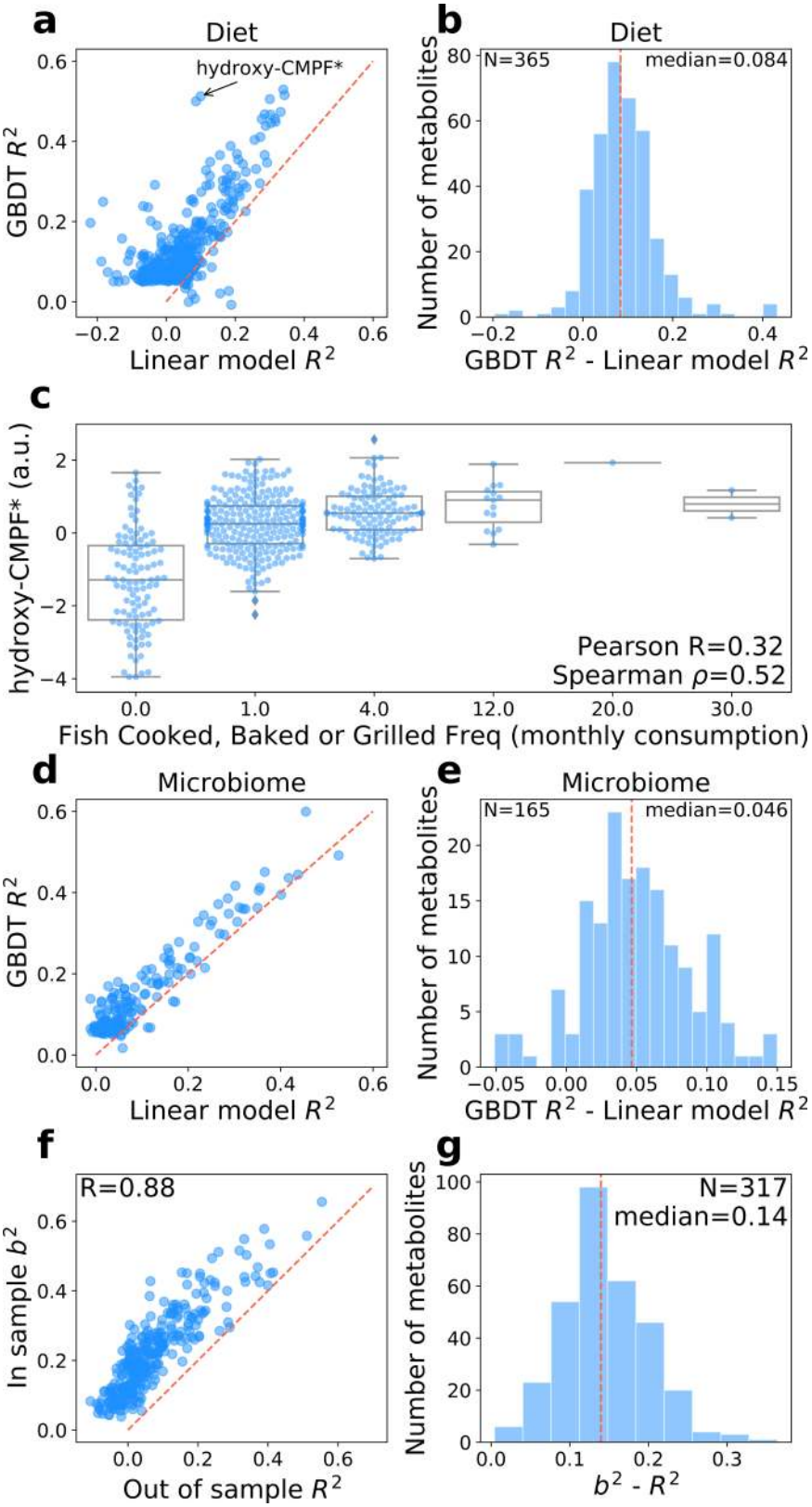
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2896-2.
Correspondence and requests for materials should be addressed to E.S.
Peer review information *Nature* thanks Katherine Pollard, Eric Topol and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Accurate and reproducible untargeted serum metabolomics. a, Breakdown of the 1,251 measured metabolites by type. **b**, Number of samples (*y* axis) in which each metabolite (*x* axis) was identified, sorted by prevalence. **c**, **d**, Mass-spectrometry measurements (*y* axis) versus standardized lab tests results (*x* axis; Methods) for creatinine (**c**; Pearson's $R = 0.87, P < 10^{-20}$) and cholesterol (**d**; $R = 0.79, P < 10^{-20}$). **e**, Spearman correlations (*y* axis: centre, median; box, IQR; whiskers, 1.5 × IQR) between standardized metabolomic profiles (Methods) of different individuals ($n = 475$; median Spearman's ρ 0.05, s.d. 0.12) stratified by sex, and between standardized metabolomic profiles of the same participant ($n = 20$; median Spearman ρ 0.68, s.d. 0.06) taken one week apart. C&V, cofactors and vitamins; a.u., arbitrary units.

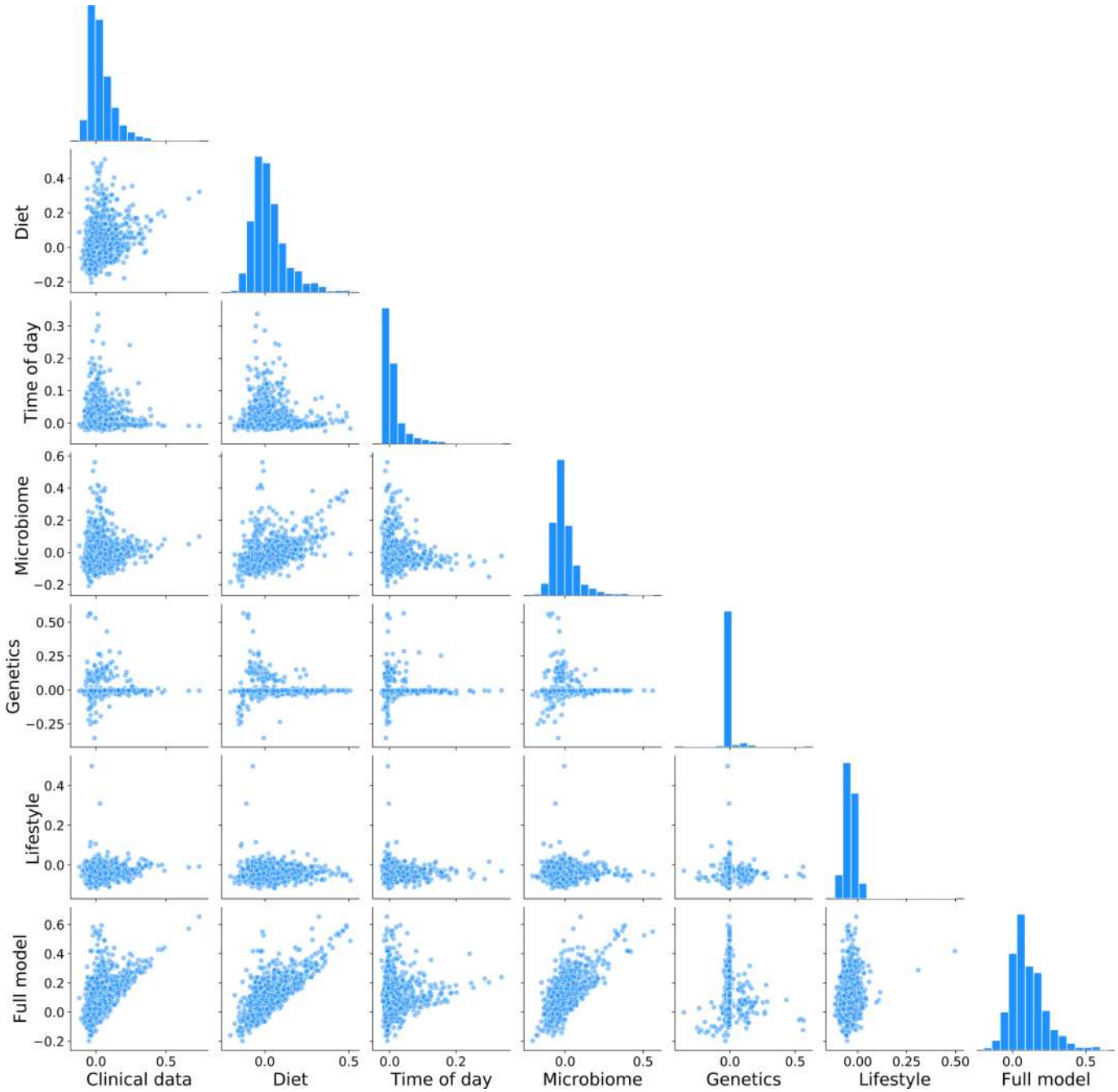


Extended Data Fig. 2 | Biological sub pathway prediction of unidentified molecules. Figure panels refer to the results of a leave-one-out cross validation prediction model of sub pathways of metabolites based on their normalized levels, raw mean, percentage of missing values, and SHAP values (Methods). Results shown are for a model trained using only sub pathways that include over 10 molecules in our data (28 sub pathways, 572 named metabolites). **a**, The overall accuracy of the sub pathway classifier (*y* axis) when a success is considered as having the true label in one of the top *k* predictions (*x* axis). **b**, The log loss of the classifier (*y* axis) computed over the resulting soft max (raw probabilities; blue) and a dichotomous matrix in which for every metabolite we only keep the top predicted sub pathway as 1 and zero-out all other predictions (red). **c**, The overall accuracy of the model (left *y* axis; blue) and the corresponding fraction of metabolites (right *y* axis; red) when considering only metabolites for which the classifier predicted a maximal probability above some threshold (*x* axis). **d**, A confusion matrix showing the predicted sub pathways (*x* axis), determined as the label with the highest probability per metabolite, versus the true annotated sub pathways (*y* axis). Each cell in the matrix counts the number of metabolites of a certain true sub pathway (*y* axis) which were assigned with some predicted sub pathway (*x* axis) by our model. The rightmost column is the sum of every row and represents the number of metabolites annotated for every sub pathway. The matrix is ordered by the higher order biological pathway (super pathway). Cell colours are log scaled. **e**, Classification results summarizing the f1-score, precision and recall per sub pathway. Rows correspond to the sub pathway annotation in **d**. **f**, For every sub pathway (*y* axis) shown are the fraction of metabolites truly annotated as such (black), predicted as such by the classifier (blue; out of the named molecules in the support of the model), and the fraction of unidentified molecules predicted as such (out of all unidentified molecules). M., metabolism; Xeno., xenobiotics; Ptds, peptides; AAs, amino acids.

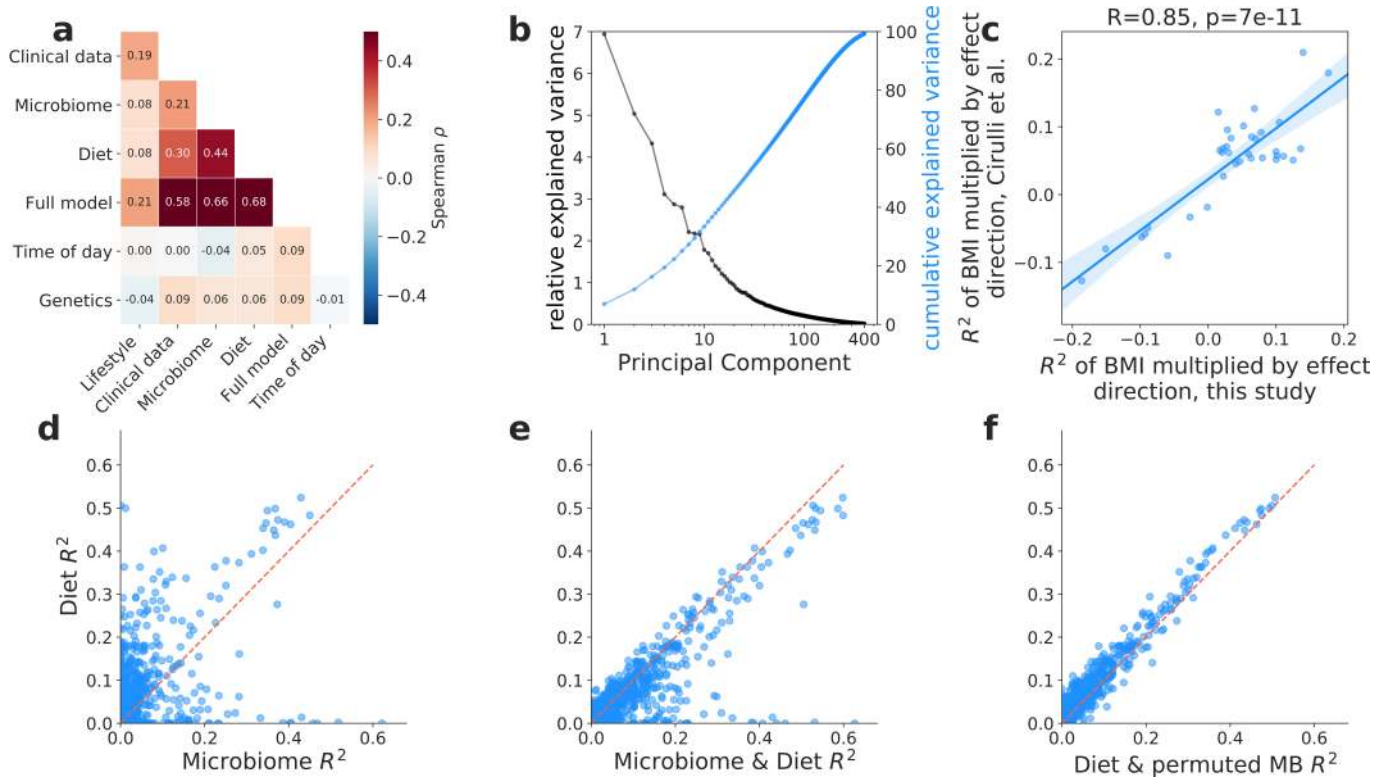


Extended Data Fig. 3 | See next page for caption.

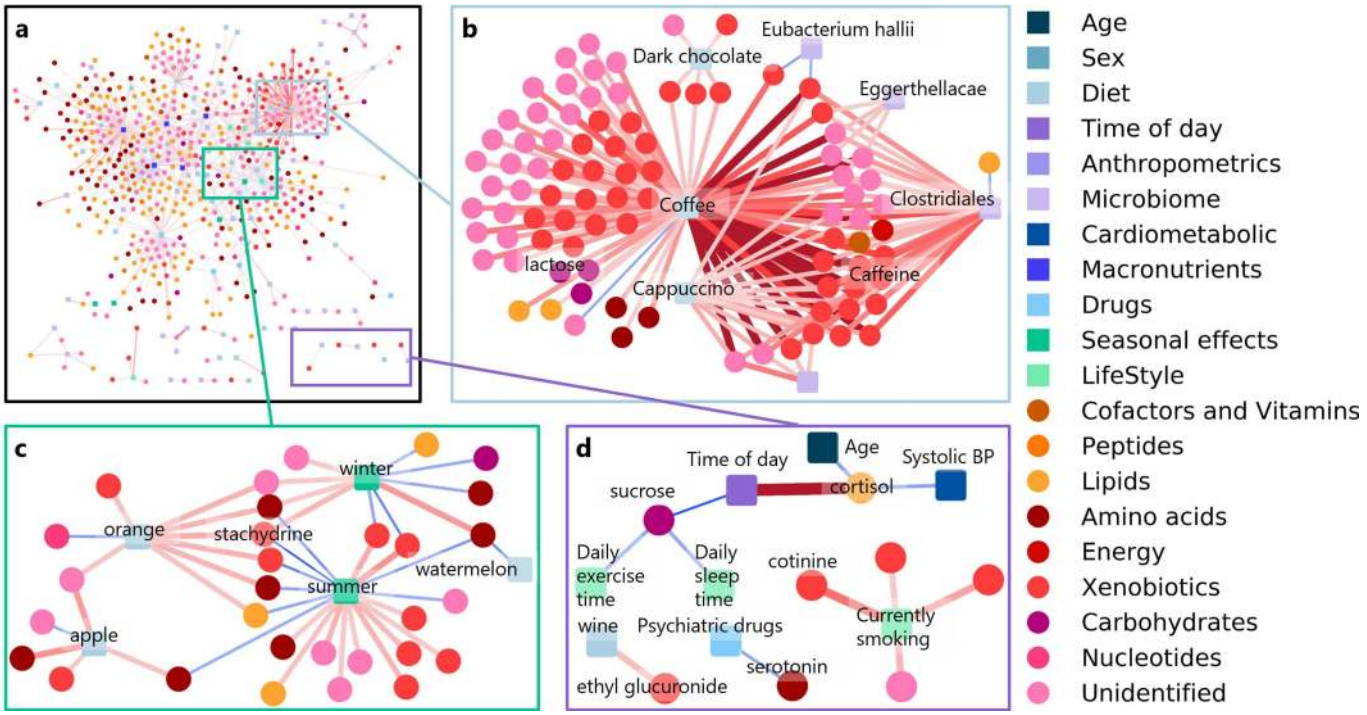
Extended Data Fig. 3 | Comparative analysis of linear versus nonlinear models and in-sample versus out-of-sample predictions. a, Metabolite prediction R^2 of GBDT versus Lasso regression models using diet data. Shown are only metabolites for which at least one model achieved significant predictions with R^2 greater than 0.05. **b**, Histogram of the differences between the R^2 of GBDT compared to Lasso regression using the diet data. **c**, The levels of the metabolite hydroxy-CMPF* (*y* axis: centre, median; box, IQR; whiskers, 1.5 × IQR) versus the monthly consumption of cooked, baked or grilled fish as reported in a food frequency questionnaire. The comparison of Spearman's and Pearson's correlation coefficients suggests that the relationship between the metabolite and the numerical values of the question are monotonic yet nonlinear, which explains why GBDT performs better in predicting the levels of hydroxy-CMPF* from diet data. The *x* axis is not to scale. **d**, **e**, Same as **a**, **b** for microbiome. **f**, Estimations of gut microbiome explainability (b^2) of metabolite levels obtained via applying a linear mixed model on the bacterial species composition as previously described (*y* axis) versus the explained variance (R^2) of metabolites from out-of-sample prediction models based on the same gut microbiome data. Shown are only metabolites with significant b^2 estimates (5% FDR). **g**, Histogram of the differences between the b^2 estimates and the R^2 of out-of-sample prediction using the gut microbiome data.



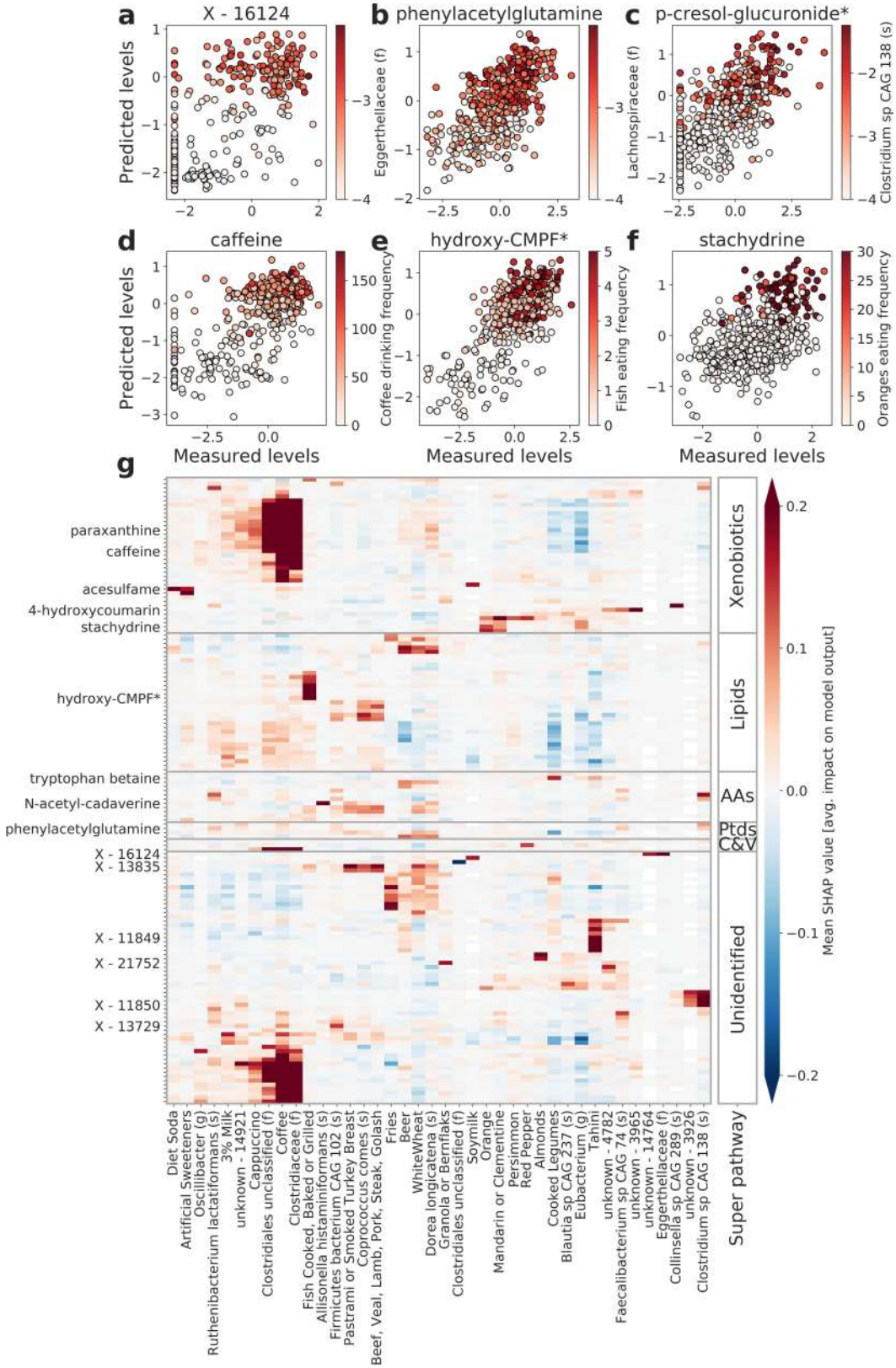
Extended Data Fig. 4 | Comparison of explained variance of metabolites for every pair of feature groups. Dot plots of the explained variance of the metabolite groups (*y* axis) from models based on every pair of feature groups (*x* axis). Panels on the diagonal show the marginal distribution of explained variance of metabolite groups for a certain feature group.



Extended Data Fig. 5 | Comparative analysis of different feature groups. **a**, Spearman correlations computed between the EV of metabolites for every pair of feature groups. **b**, The proportion of variance explained by each of the first 400 principal components (left *y* axis; black) and their cumulative EV (right *y* axis; blue). **c**, R^2 multiplied by the sign of the Pearson correlation coefficient (*x* axis) between metabolite levels and BMI in our study, versus the mean R^2 multiplied by the sign of the Pearson correlation coefficient (*y* axis) of BMI associated metabolites recently reported by a different group[13]. Shown are 36 (out of 49) BMI associated metabolites that were also measured in this cohort. *P* value for the Pearson correlation, $P = 7 \times 10^{-11}$. Line and shaded colouring represent the fitting of a linear model and the 95% confidence interval. **d**, The EV of every metabolite from prediction models based on the gut microbiome (*x* axis) versus diet (*y* axis). Dashed red line is $y = x$. **e**, Same for prediction models based on both gut microbiome and diet (*x* axis) compared to using only diet (*y* axis). **f**, Same for prediction models based on diet and permuted gut microbiome (*x* axis) compared to using only diet (*y* axis). MB, microbiome.

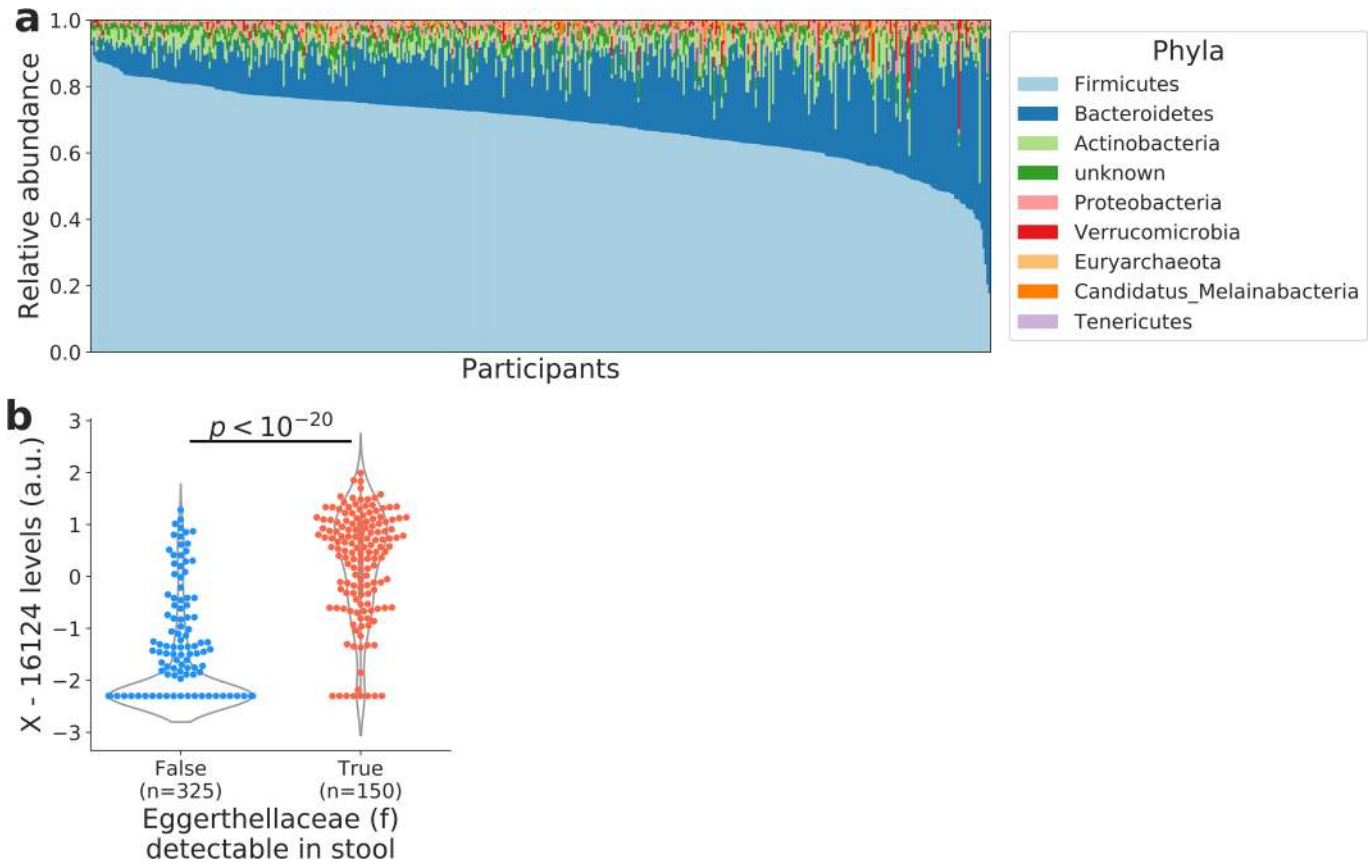


Extended Data Fig. 6 | Networks of interactions between phenotypes explain diverse metabolites. Interactions between features from different feature groups predictive of similar metabolites are presented in a graphical layout, in which nodes are either metabolites or features, and edges are the directional mean absolute SHAP values (Methods) computed from models trained only on features from the respective feature group. Circular nodes, metabolites; predictive feature nodes, squares; both coloured by relevant categories. Shown are only edges with a mean absolute SHAP value greater than 0.12. **a**, Network of associations for the following feature groups: macronutrients, diet, microbiome, lifestyle, drugs and seasonal effects. **b**, A large group of metabolites for which predictions are mainly driven by the reported consumption of coffee and the relative abundance of a bacteria from the Clostridiales order. **c**, Metabolites explained by seasonal fruit consumption. **d**, Selected examples of interactions between metabolites and features in predictive models.

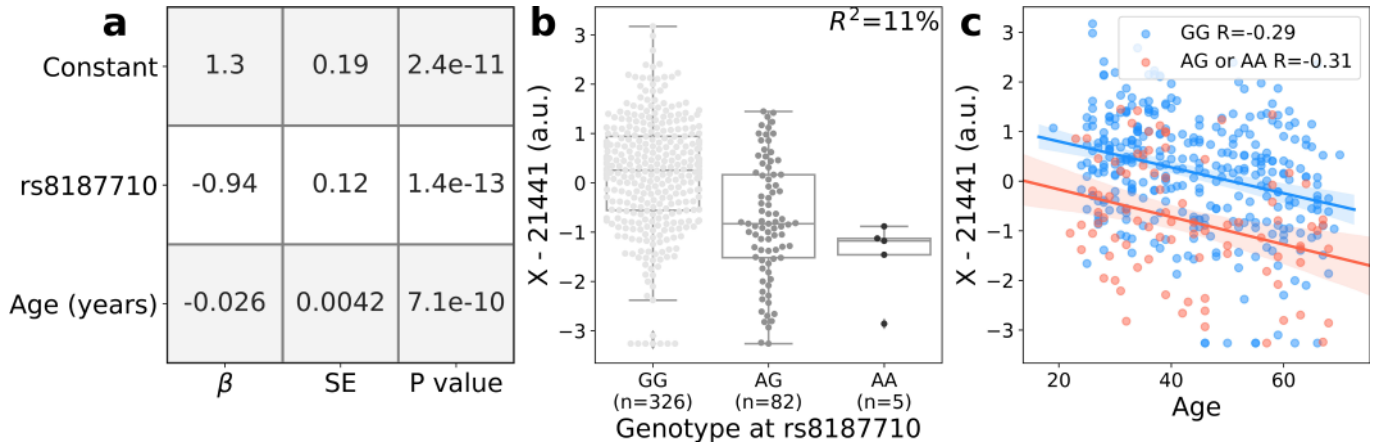


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Specific dietary features and bacterial taxa underlie the accurate prediction of circulating metabolites. a–**f**, Predicted (*y* axis) versus measured (*x* axis) levels (arbitrary units) of X-16124 (**a**; Pearson's $R = 0.77$, $P < 10^{-20}$), phenylacetylglutamine (**b**; $R = 0.63$, $P < 10^{-20}$), *p*-cresol-glucuronide (**c**; $R = 0.64$, $P < 10^{-20}$), caffeine (**d**; $R = 0.68$, $P < 10^{-20}$), hydroxy-CMPF (**e**; $R = 0.72$, $P < 10^{-20}$) and stachydrine (**f**; $R = 0.5$, $P < 10^{-20}$). Predictions of **a**–**c** are based only on microbiome data, and coloured by the relative abundance of the bacterial taxa having the highest mean absolute SHAP value for each metabolite. Predictions of **d**–**f** are based only on diet data, and coloured by the reported consumption of the dietary item having the highest mean absolute SHAP value for each metabolite. *P* values for prediction were estimated via bootstrapping. **g**, Heat map showing the directional mean absolute SHAP values (Methods) of various features (*x* axis) computed from fivefold cross validation models that predict metabolite levels (*y* axis) using two separate models, one based on diet and another on gut microbiome data. Positive (negative) SHAP values indicate that higher (lower) feature values lead, on average, to higher predicted values. Shown are the top 150 predicted metabolites using diet and gut microbiome, and the top 40 features by maximum mean absolute SHAP value across all metabolites.



Extended Data Fig. 8 | Distribution of phyla and a taxa from the Eggerthellaceae family. a, Stacked bar plots per sample (*x* axis) showing the relative abundance of bacterial phyla (*y* axis). Samples are sorted by the relative abundance of the most abundant phylum, Firmicutes. Bacteroidetes is the second most abundant phylum in our cohort. Relative abundance of a phylum is computed as the sum over relative abundances of all bacterial features belonging to that phylum. **b**, The levels of the unidentified compound X-16124 in individuals for which the bacterial taxa from the Eggerthellaceae family was detectable in stool versus individuals for which it was not ($P < 10^{-20}$, two-sided Mann–Whitney *U*-test).



Extended Data Fig. 9 | The unidentified molecule X-21441 associates with rs8187710 independent of age. a, A table showing the coefficients, standard errors and *P* values resulted from a multiple linear regression model with levels of the unidentified molecule X-21441 as the dependent variable, the allele dosage of rs8187710 (0–2) and age (years) as the independent variables: $y_{X\text{-}21441} = \text{constant} + \beta_1 \times \text{rs8187710} + \beta_2 \times \text{Age}$. **b**, The levels of X-21441 (*y* axis; centre, median; box, IQR; whiskers, 1.5 × IQR) versus the genotype of the participants (*x* axis). The number of participants with each genotype is indicated below the tick labels. The explained variance of X-21441 by rs8187710 as estimated using plink (Methods) is indicated on the upper right corner of the panel. **c**, The levels of X-21441 versus the age of the participants (*x* axis) coloured by genotype of participants. Line and shaded colouring represent the fitting of a linear model and the 95% confidence interval. SE, standard error.

Extended Data Table 1 | Basic characteristics and demographics of our main and validation cohorts

Characteristics	Main cohort (n=475)	TwinsUK (n=1004)	IMI-DIRECT (n=245)
Age (years)	44.0 +- 13.13	64.96 +- 7.77	61.81 +- 8.09
Sex, Males (%)	38.70%	3.90%	61.22%
BMI	26.3 +- 5.0	26.2 +- 4.71	30.59 +- 5.39
Smokers (%)	9.80%	42.40%	13.55%

nature research

Corresponding author(s): Eran Segal

Last updated by author(s): Aug 15, 2020

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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 software were used for data collection.
Data analysis	Python 2.7.8, with packages: pandas 0.23.4, numpy 1.14.2, scikit-learn 0.20.4, scipy 1.1.0, shap 0.24.0, LightGBM 2.1.2 MetaPhlAn2 2.1 bowtie2 2.2.5 GEM 1.376 plink v1.07

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The raw metagenomic sequencing data is available from the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena): PRJEB11532, PRJEB17643, and for the TwinsUK: PRJEB32731. The raw metabolomics data and the phenotypic data is available from the European Genome-phenome Archive (EGA; https://ega-archive.org/): EGAS00001004512. Known links between genetic loci and serum metabolites were taken from the GWAS Catalog (https://www.ebi.ac.uk/gwas/) and the GWAS server (http://metabolomics.helmholtz-muenchen.de/gwas/).

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	Repeated subsampling of subjects for full-model predictions of metabolomic PCs showed saturation of model accuracy at 300 individuals. The cohort size was set at 500 to allow for increased accuracy.
Data exclusions	No data was excluded.
Replication	Replication was performed on three fronts. First, we replicated microbiome-based predictions in 2 independent cohorts with different geography, age, and disease state. The first validation cohort included 1,004 samples of healthy participants from the TwinsUK cohort (Moayyeri et al. 2013), while the second validation cohort included 245 samples of participants of northern European ancestry (Sweden, Denmark, UK and the Netherlands) with type 2 diabetes (T2D) from the IMI-DIRECT consortium (Koivula et al. 2014). In the TwinsUK cohort, 95 out of 107 associations replicated (FDR<10%), while in the IMI-DIRECT cohort, 28 out of 50 associations replicated (FDR<10%). Data from these additional cohorts were not available to us while developing the prediction models. Second, we replicated BMI related associations which were reported by a different group [Cirulli et al. Cell Metabolism 2018]. Here, for 36 metabolites, we replicated the effect sizes in the TwinsUK cohort with a correlation of 0.85 (p=7e-11). Third, we replicated 46 genetic associations between specific SNPs and the levels of metabolites which were reported in different studies [GWAS Catalog, GWAS server].
Randomization	This study used a healthy cohort and there was no allocation to experimental groups, thus no randomization.
Blinding	There was no group allocation in this study.

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☐	☒ Clinical data

Methods

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJEguidelines for publication of clinical research and a completedCONSORT checklist must be included with all submissions.

Clinical trial registration	NCT02936362
Study protocol	Available at https://www.clinicaltrials.gov/ct2/show/NCT02936362 and https://doi.org/10.1016/j.cmet.2017.05.002
Data collection	Detailed in https://doi.org/10.1016/j.cmet.2017.05.002
Outcomes	Detailed in https://doi.org/10.1016/j.cmet.2017.05.002 - irrelevant for this study.