nature medicine

ARTICLES

https://doi.org/10.1038/s41591-020-01183-8

Check for updates

Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals

Francesco Asnicar[1,16], Sarah E. Berry[2,16] ✉, Ana M. Valdes[3,4], Long H. Nguyen[5], Gianmarco Piccinno[1], David A. Drew[5], Emily Leeming[6], Rachel Gibson[2], Caroline Le Roy[6], Haya Al Khatib[7], Lucy Francis[7], Mohsen Mazidi[6], Olatz Mompeo[6], Mireia Valles-Colomer[1], Adrian Tett[1], Francesco Beghini[1], Léonard Dubois[1], Davide Bazzani[1], Andrew Maltez Thomas[1], Chloe Mirzayi[8], Asya Khleborodova[8], Sehyun Oh[8], Rachel Hine[7], Christopher Bonnett[7], Joan Capdevila[7], Serge Danzanvilliers[7], Francesca Giordano[7], Ludwig Geistlinger[8], Levi Waldron[8], Richard Davies[7], George Hadjigeorgiou[7], Jonathan Wolf[7], José M. Ordovás[9,10], Christopher Gardner[11], Paul W. Franks[12,13], Andrew T. Chan[5,13,14,17], Curtis Huttenhower[13,14,17], Tim D. Spector[6,17] and Nicola Segata[1,15,17] ✉

The gut microbiome is shaped by diet and influences host metabolism; however, these links are complex and can be unique to each individual. We performed deep metagenomic sequencing of 1,203 gut microbiomes from 1,098 individuals enrolled in the Personalised Responses to Dietary Composition Trial (PREDICT 1) study, whose detailed long-term diet information, as well as hundreds of fasting and same-meal postprandial cardiometabolic blood marker measurements were available. We found many significant associations between microbes and specific nutrients, foods, food groups and general dietary indices, which were driven especially by the presence and diversity of healthy and plant-based foods. Microbial biomarkers of obesity were reproducible across external publicly available cohorts and in agreement with circulating blood metabolites that are indicators of cardiovascular disease risk. While some microbes, such as *Prevotella copri* and *Blastocystis* spp., were indicators of favorable postprandial glucose metabolism, overall microbiome composition was predictive for a large panel of cardiometabolic blood markers including fasting and postprandial glycemic, lipemic and inflammatory indices. The panel of intestinal species associated with healthy dietary habits overlapped with those associated with favorable cardiometabolic and postprandial markers, indicating that our large-scale resource can potentially stratify the gut microbiome into generalizable health levels in individuals without clinically manifest disease.

Dietary contributions to health and chronic conditions, such as obesity, metabolic syndrome, cancer and cardiovascular disease, are of universal importance. Obesity and associated mortality/morbidity have risen dramatically over the past decades[1], with the gut microbiome implicated as one of several potentially causal human-environment interactions[2,3]. Surprisingly, the details of the microbiome's role in obesity and cardiometabolic health have proven difficult to define reproducibly in large human populations[4], probably due to the complexity of habitual diets, the difficulty of measuring them at scale and disentangling them from other lifestyle variables[5,6]and the personalized nature of the microbiome[7].

To overcome these challenges, we launched the PREDICT 1 trial of diet–microbiome interactions in metabolic health[8]. PREDICT 1 included >1,000 participants profiled pre- and post-standardized dietary challenges using intensive in-clinic biometric and blood measures, habitual dietary data collection, continuous glucose monitoring and stool metagenomics. The study was inspired by previous large-scale diet–microbiome interaction profiles, which identified gut microbiome configurations and microbial taxa associated with postprandial glucose responses[9,10], obesity-associated biometrics such as body mass index (BMI) and adiposity[11–13] and blood lipids and inflammatory markers[14,15].

[1]Department of Cellular, Computational and Integrative Biology, University of Trento, Trento, Italy. [2]Department of Nutritional Sciences, King's College London, London, UK. [3]School of Medicine, University of Nottingham, Nottingham, UK. [4]Nottingham National Institute for Health Research Biomedical Research Centre, Nottingham, UK. [5]Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [6]Department of Twin Research, King's College London, London, UK. [7]Zoe Global Ltd, London, UK. [8]City University of New York, New York, NY, USA. [9]Jean Mayer-United States Department of Agriculture-Human Nutrition Research Center on Aging, Tufts University, Boston, MA, USA. [10]Institutos Madrileño de Estudios Avanzados Food Institute, Campus of International Excellence Universidad Autónoma de Madrid & Consejo Superior de Investigaciones Científicas, Madrid, Spain. [11]Stanford University, Stanford, CA, USA. [12]Department of Clinical Sciences, Lund University, Malmö, Sweden. [13]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [14]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [15]European Institute of Oncology Scientific Institute for Research, Hospitalization and Healthcare, Milan, Italy. [16]These authors contributed equally: Francesco Asnicar, Sarah E. Berry. [17]These authors jointly supervised this work: Andrew T. Chan, Curtis Huttenhower, Tim D. Spector, Nicola Segata. ✉e-mail: sarah.e.berry@kcl.ac.uk; nicola.segata@unitn.it

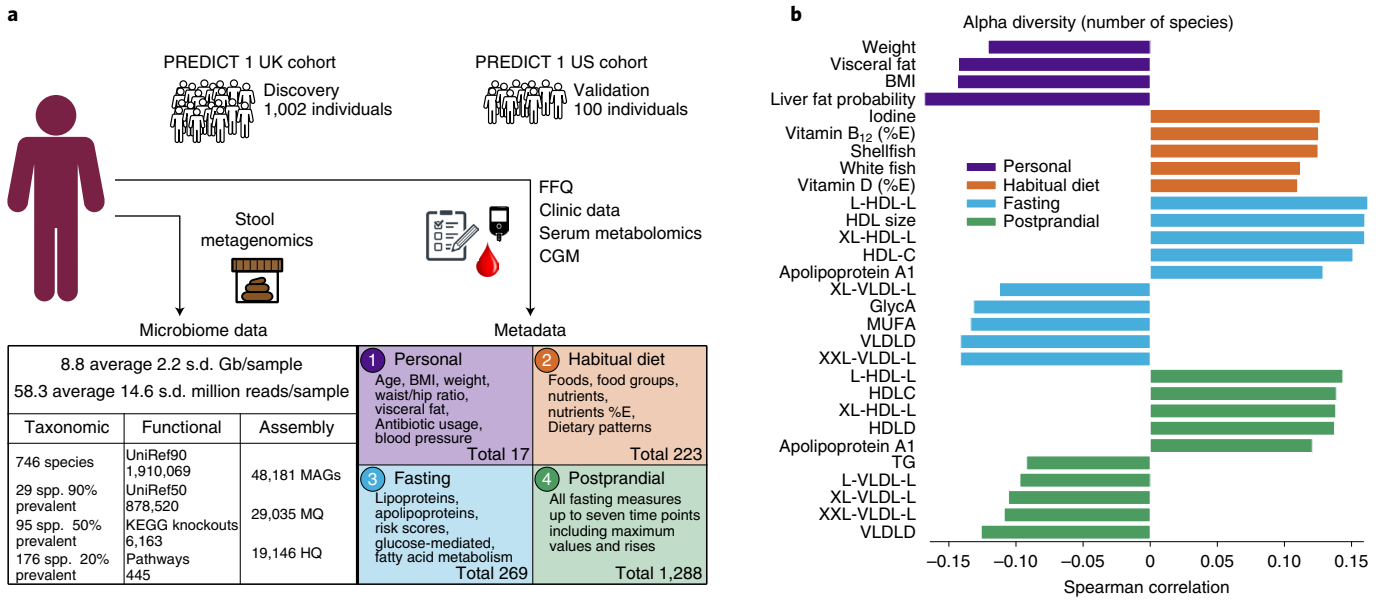


Fig. 1 | The PREDICT 1 study associates gut microbiome structure with habitual diet and blood cardiometabolic markers. a, The PREDICT 1 study assessed the gut microbiome of 1,098 volunteers from the UK and USA via metagenomic sequencing of stool samples. Phenotypic data obtained through in-person assessment, blood/biospecimen collection and the return of validated study questionnaires queried a range of relevant host/environmental factors including: (1) personal characteristics, such as age, BMI and estimated visceral fat; (2) habitual dietary intake using semiquantitative FFQs; (3) fasting; and (4) postprandial cardiometabolic blood and inflammatory markers, total lipid and lipoprotein concentrations, lipoprotein particle sizes, apolipoproteins, derived metabolic risk scores, glycemic-mediated metabolites and metabolites related to fatty acid metabolism. **b**, Overall microbiome alpha diversity, estimated as the total number of confidently identified microbial species in a given sample (richness), was correlated with HDLD (positive) and estimated hepatic steatosis (negative). The five strongest positive and negative Spearman correlations with $q < 0.05$ are reported for each of the four categories. The top species based on Shannon diversity are reported in Extended Data Fig. 1a; all correlations, *P* and *q* values are reported in Supplementary Table 1. The '%E' label represents foods and nutrients normalized by the estimated daily energy intake in kcal.

Results

Large metagenomically profiled cohorts with rich clinical, cardiometabolic and dietary information. PREDICT 1 (refs. [8,16]) is an intervention study of diet–microbiome–cardiometabolic interactions (Methods), including a discovery cohort in the UK ($n = 1{,}002$) and a validation population in the USA ($n = 100$). We collected demographic information, habitual diet data, cardiometabolic blood biomarkers and postprandial responses to standardized test meals in the clinic and in free-living settings[8,16] (Fig. 1a). At-home stool collection yielded 1,098 baseline and 105 follow-up microbiome samples (+14 d), which were all shotgun sequenced and then taxonomically and functionally profiled (Fig. 1a and Methods).

Microbial diversity and composition are linked with diet and fasting and postprandial biomarkers. We first leveraged a unique subpopulation of 480 monozygotic and dizygotic twins and confirmed that host genetics influences microbiome composition only to a limited extent[17]. Indeed, twin pair microbiome similarity was substantially lower than intrasubject longitudinal similarity (day 0 versus day 14, $P < 1 \times 10^{-12}$; Extended Data Fig. 1b), a testament to the personalized nature of the gut microbiome attributable to non-genetic factors (Extended Data Fig. 1c,d).

We then investigated overall intrasample (alpha) microbiome diversity as a broad summary statistic of microbiome structure and found that it was significantly associated ($q < 0.05$) in 56 of the 295 tested correlations with personal characteristics, habitual diet and metabolic indices (Fig. 1b and Supplementary Table 1a). BMI, visceral fat measurements and probability of fatty liver (using a validated prediction model[18]) were inversely associated with species richness. Among clinical circulating measures, high-density lipoprotein cholesterol (HDLC) was positively correlated with species richness. Emerging cardiometabolic biomarkers[19] that are not routinely used clinically, including lipoprotein particle size (diameter, '-D') and glycoprotein acetyl (GlycA) (inflammatory biomarker), were also associated (positively or negatively) with microbiome richness. These results associating simple indicators such as microbiome richness to cardiometabolic health indicators and diet, motivated our more detailed investigations of specific gut microbiome components.

Diversity of healthy plant-based foods in habitual diet shapes gut microbiome composition. We assessed the links between habitual diet and the microbiome using random forest models, each trained on quantitative microbiome features to predict each dietary variable from food frequency questionnaires (FFQs) (Methods). The performance of the models was quantified with receiver operating characteristic (ROC) area under the curve (AUC) for classification and correlation for regression (Methods). Several foods and food groups exceeded the 0.15 median Spearman correlation over bootstrap folds (denoted as ρ) between predicted and FFQ-estimated values (14.5%) and AUC > 0.65 (10.8%; Fig. 2a). The strongest association was for coffee (instant or ground) ($\rho = 0.43$, AUC = 0.8), with dose-dependent effects and validated in the US cohort (Fig. 2d). Tighter microbiome links were found for energy-adjusted nutrients (Fig. 2a), with almost one-third (Supplementary Table 2) showing correlations above 0.3.

We then summarized constituent foods into dietary indices (Supplementary Table 2), including the Healthy Food Diversity (HFD) index (incorporating dietary diversity and food quality)[20], the Healthy (hPDI)/Unhealthy Plant-based Dietary Indices (uPDI) (considering quality and quantity of plant-based foods), Healthy Eating Index (HEI) (extent of alignment with dietary guidelines)[21]

and the alternate Mediterranean diet (aMED) score[22], all of which are associated with reduced risk of chronic disease[22–27]. We demonstrated tight correlations between microbial composition and the HFD, hPDI/uPDI and HEI in the UK (ρ between 0.31 and 0.37; Fig. 2a); the results were consistent in the US validation cohort, with ρ reaching 0.42 for HFD and 0.31 for aMED (Fig. 2e,f and Extended Data Fig. 3), highlighting the relationship between the microbiome and health-associated dietary patterns.

Microbial species segregate into groups associated with more and less healthy plant- and animal- based foods. We proceeded to identify the specific microbial taxa most responsible for these diet-based community associations (Fig. 2b). After adjusting for age and BMI, we found 42 species (24% of those at >20% prevalence) significantly correlated with at least 5 dietary exposures ($q < 0.2$; Supplementary Table 5). This included expected associations (Extended Data Fig. 2), such as enrichment of the probiotic taxa *Bifidobacterium animalis*[28] and *Streptococcus thermophilus* with greater full-fat yogurt consumption ($\rho = 0.22$ for both). The strongest food-microbe association was between the recently characterized butyrate-producing *Lawsonibacter asaccharolyticus*[29] and coffee consumption (Fig. 2b). However, due to the low resolution of FFQ data, the complexity of dietary patterns, nutrient–nutrient interactions and clustering of healthy/less healthy food items, it is challenging to disentangle the independent associations of single foods with microbial species.

At a broader level, we found clear segregation of species (Fig. 2b) into two distinct clusters with either more healthy plant-based foods (for example, spinach, seeds, tomatoes, broccoli) or less healthy plant-based (for example, juices, sweetened beverages, refined grains) and animal-based foods, as defined by the PDI[30] (Supplementary Table 4). Taxa linked to healthy plant-based foods (Fig. 2b,c and Extended Data Fig. 2) mostly included butyrate producers, such as *Roseburia hominis*, *Agathobaculum butyriciproducens*, *Faecalibacterium prausnitzii* and *Anaerostipes hadrus*, as well as uncultivated species, predicted to have this metabolic capability (*Roseburia* bacterium CAG:182 and *Firmicutes* bacterium CAG:95). Clades correlating with several less healthy plant-based and animal-based foods included several *Clostridium* species (*Clostridium innocuum*, *Clostridium symbiosum*, *Clostridium spiroforme*, *Clostridium leptum*, *Clostridium saccharolyticum*). The segregation of species according to animal-based healthy foods (for example, eggs, white and oily fish) or animal-based less healthy foods (for example, meat pies, bacon, dairy desserts) using a new categorization (Methods), was also distinct and overlapping with taxa signatures for healthy and less healthy plant foods (Fig. 2c and Extended Data Fig. 2). The few foods not fitting into the healthy cluster despite being classified as healthy plant foods, were (ultra)-processed foods[31] (for example, sauces, baked beans; Extended Data Fig. 2). This emphasizes the importance of food quality (for example, highly processed versus unprocessed), food source (for example, plant versus animal) and food type (that is, not all plant foods are healthy) both in overall health and microbiome ecology.

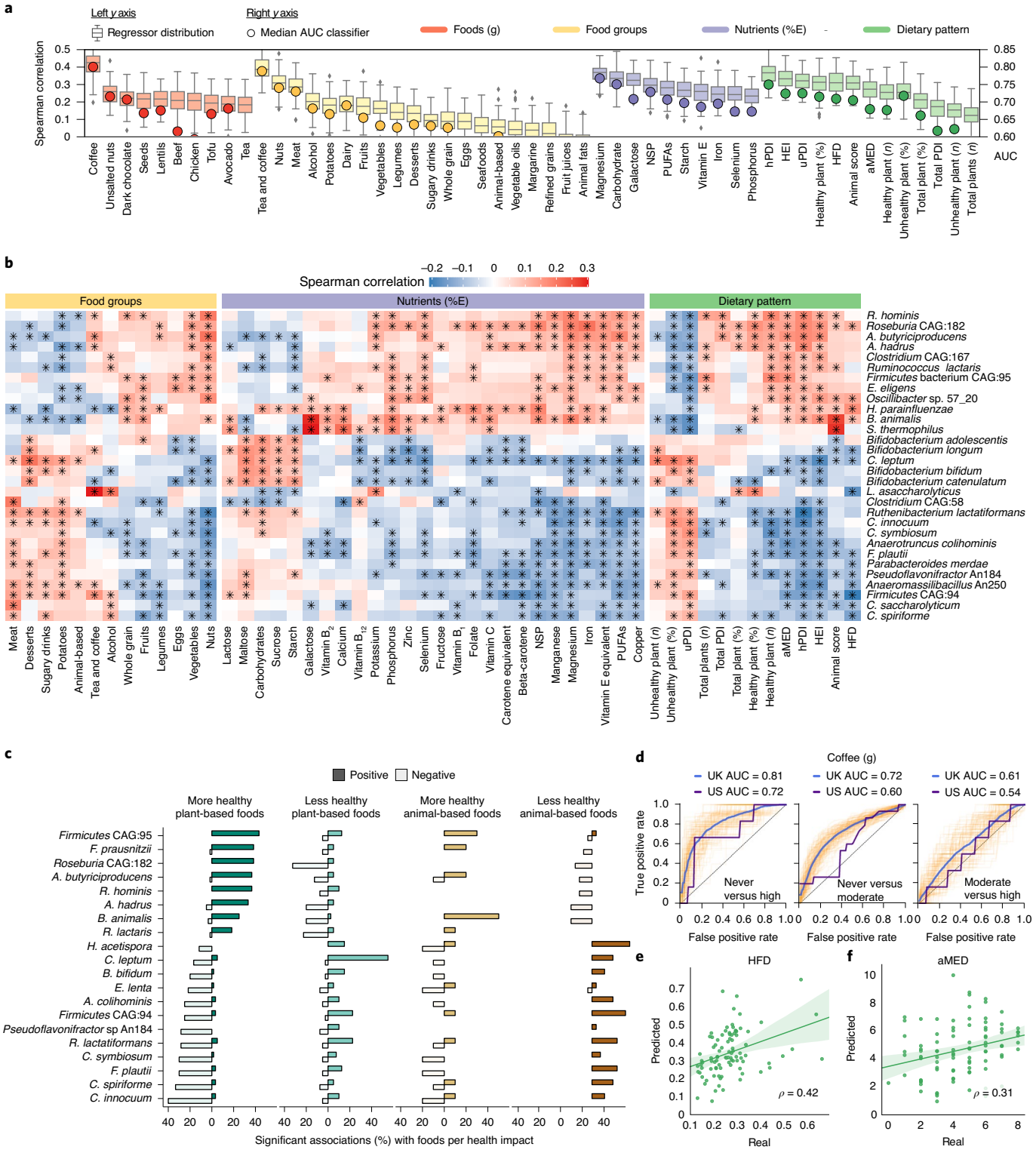
The strongest microbiome habitual diet associations are driven by poorly characterized microbes. Many of the strongest microbial associations with diet occurred with only recently isolated or still uncultured taxa including five species defined using coabundance gene groups (CAGs) from metagenomics[32]. Among indices, the hPDI significantly correlated with 60 of the 176 prevalent species, highlighting together with the HFD (Fig. 2e) the impact of dietary diversity and quality on gut microbial responsiveness. Among other dietary indices and nutrients, we observed general concordance with the two sets of microbes associated with healthy and less healthy foods. A greater animal-based food score (definition in Supplementary Table 4) was associated with the healthy cluster (Fig. 2c and Extended Data Fig. 2), suggesting that a diet rich in healthier animal-based foods is associated with the more favorable diet–microbiome signature, although this may also reflect an overall healthier dietary pattern. The healthy and unhealthy PDIs, which differentially affect disease risk[25,30], also had distinct clusters, again emphasizing the oversimplification of conventional plant and animal-based food groupings. The taxa with the highest correlations in the two clusters are *Firmicutes* bacterium CAG:95 and *C. leptum* for healthy and unhealthy diet, respectively. The lack or paucity of cultivated representatives for these two taxa may explain why these links were previously overlooked[9,12]. The US validation cohort generally confirmed these associations despite its smaller sample size: among the subset of derived pattern/index scores shared between the UK and US cohorts, of the 54 associations that were significant both in the UK cohort (false discovery rate (FDR) $q < 0.2$) and in the US cohort ($P < 0.05$), 70.4% were concordant.

Microbial indicators of obesity are reproducible across varied populations. Microbiome links to obesity have attracted much interest, although results have varied in human populations[3,4]. Our machine learning approach (Methods) found visceral fat to be more strongly linked to gut microbial composition than BMI[33], a finding again validated in US participants (Fig. 3a). Some obesity-associated taxa were also indicators of poor dietary patterns after controlling for BMI (for example, *Clostridium* CAG:58, *Flavonifractor plautii*), whereas markers of lower visceral fat mass (for example,

Fig. 2 | Food quality, regardless of source, is linked to overall and feature-level composition of the gut microbiome. a, Specific components of habitual diet comprising foods, nutrients and dietary indices are linked to the composition of the gut microbiome with variable strengths as estimated by machine learning regression and classification models. Box plots report the correlation between the real value of each component and the value predicted by regression models across 100 training/testing folds (Methods). The circles denote the median AUC values across 100 folds for a corresponding binary classifier between the highest and lowest quartiles (Methods). NSP, non-starch polysaccharide. **b**, Single Spearman correlations adjusted for BMI and age between microbial species and components of habitual diet with the asterisks denoting significant associations (FDR $q < 0.2$). The 30 microbial species with the highest number of significant associations across habitual diet categories are reported. All indices of dietary patterns are reported, whereas only food groups and nutrients (energy-adjusted) with at least 7 associations among the top 30 microbial species are reported. Rows and columns are hierarchically clustered (complete linkage, Euclidean distance). Full heatmaps of foods and unadjusted nutrients are reported in Extended Data Fig. 2; the full set of correlations, *P* and *q* values are available in Supplementary Tables 5 and 6 for UK and US, respectively. **c**, Number of significant positive and negative associations (Spearman correlation, $P < 0.2$) between foods and taxa categorized by more and less healthy plant-based foods and more and less healthy animal-based foods according to the PDI. The taxa shown are the 20 species with the highest total number of significant associations regardless of category. **d**, The association between the gut microbiome and coffee consumption in UK participants is dose-dependent, that is, stronger when assessing heavy (for example, >4 cups per day) versus never drinkers, and was validated in the US cohort when applying the UK model. The reported ROC curves represent the performance of the classifier at varying classification thresholds with regard to the true positive (that is, recall) and false positive rates (that is, precision). **e**,**f**, Among general dietary patterns and indices, the HFD (**e**) and aMED (**f**) were validated in the US cohort, thus showing consistency between the two populations on these two important dietary indices. Other validations of the UK model applied to the US cohort are reported in Extended Data Fig. 3. The box plots show the first and third quartiles (boxes) and the median (middle line); the whiskers extend up to 1.5× the IQR.

F. prausnitzii) were more strongly linked to healthier foods and patterns of intake, illustrating that diet and obesity microbiome signatures overlap but are not identical (Fig. 3b).

Of the 17 species surpassing $q < 0.05$, 3 had an (absolute) $\rho > 0.1$ in the US cohort and 2 of these were concordant with those in the UK cohort (Fig. 3c). Across harmonized independent datasets, all but two median associations were consistent with the PREDICT 1 UK signatures and 12 of the 14 were concordant despite different sample collection and DNA extraction methods. Microbiome models to predict BMI in the UK cohort were further validated in six independent datasets available in curatedMetagenomicData[34] (Methods). Despite interpopulation differences[11,35], the UK model improved cohort-specific cross-validation accuracy in most cases, on par with the leave-one-dataset-out (LODO) approach (Fig. 3d).



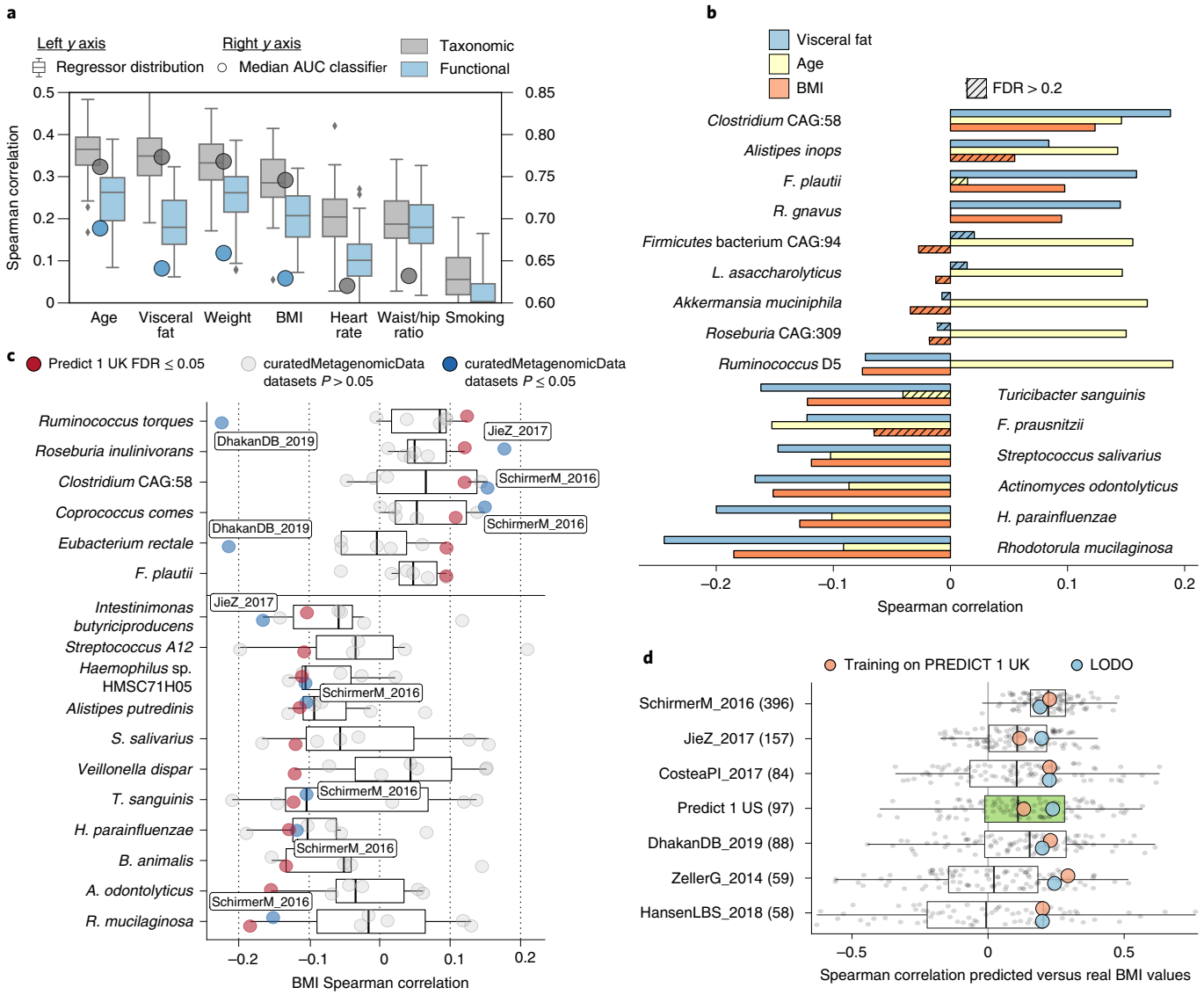


Fig. 3 | Random forest machine learning models trained on microbial or functional profiles can predict obesity phenotypic markers, even on independent cohorts. a, Whole-microbiome machine learning models can assess personal factors with random forest regression (box plots and left-side *y* axis) using only taxonomic or functional (that is, pathway) microbiome features. Classification models (circles and right-side *y* axis) exceeded an AUC of 0.65 except for waist-to-hip ratio and smoking. **b**, We observed the highest correlations between the relative abundance of microbial species and age, BMI and visceral fat. The link between microbial features and visceral fat was of greater effect and more often significant than with traditional BMI. **c**, Using several independent datasets[34], we confirmed the correlations between single microbial species and BMI, with the blue points denoting significant associations at $P < 0.05$. The statistical test used was a two-sided *z*-test (Methods). **d**, The machine learning model for BMI trained on PREDICT 1 data was reproducible in several external datasets (Extended Data Fig. 5), achieving correlations with true values exceeding those obtained in the cross-validation of a single given dataset in five of seven cases. When the PREDICT 1 microbiome model was expanded to include other datasets (excluding those used for testing, that is, LODO approach), performance remained comparable, confirming the generalizability of the PREDICT 1 model on obesity-related indicators. The box plots show the first and third quartiles (boxes) and the median (middle line); the whiskers extend up to 1.5× the IQR.

Fasting cardiometabolic markers associated with specific microbiome structures. To explore the connections between the gut microbiome and cardiometabolic health, we developed and evaluated microbiome-based machine learning models for each selected clinical and emerging cardiometabolic biomarker. We found modest concordance between microbiome models and several traditional clinical fasting cardiometabolic biomarkers (Fig. 4a) including blood pressure, lipids (triglycerides (TGs), total cholesterol, HDLC, low-density lipoprotein cholesterol (LDLC)), fasting glucose and glycosylated hemoglobin (percentage HbA_{1c}) as well as a clinical prediction score estimating the latent 10-year risk of heart disease (atherosclerotic cardiovascular disease score)[36].

For other blood biomarkers (Fig. 1a), we found stronger correlations between the microbiome and an inflammatory surrogate (GlycA; Fig. 4a), circulating polyunsaturated fatty acids (PUFAs) (both omega-6 (fatty aid ω6/fatty acid) and total PUFA (PUFA/FA) to total fatty acid ratios, $\rho = 0.3$ and 0.32, respectively), as well as emerging lipid measures linked to host health, including HDL and very-low-density lipoprotein (VLDL) particle size (-D, $\rho = 0.29$ for both) and the lipid content of lipoprotein subfractions (including total lipids in very large HDL and total lipids in large HDL, $\rho = 0.3$ and 0.28, respectively). GlycA and VLDLD are associated with increased risk for metabolic syndrome, CVD and type 2 diabetes, whereas HDLD and its lipid constituents, omega-6 and total

PUFA, have inverse associations[37,38]. Similarly, most glycemic indicators such as insulin and C-peptide were also coupled to human gut microbiome composition ($\rho = 0.17$ and 0.22, respectively) as well as derived predictors of insulin sensitivity (quantitative insulin-sensitivity check index (QUICKI), $\rho = 0.22$)[39] and hepatic steatosis (liver fat probability, $\rho = 0.2$).

Species-based predictors proved more accurate than pathway abundance profiles (Extended Data Fig. 4a), which is consistent with other reports[40]. Our primary findings were generally replicated in the US cohort (Fig. 4a), corroborating the existence of a strong, previously overlooked link between the gut microbiome and surrogates of cardiometabolic health.

The gut microbiome is a better predictor of postprandial TGs and insulin concentrations than of glucose levels. Fasting blood assays are standard for research and clinical investigations; however, individuals consume multiple mixed-nutrient meals throughout the day and spend most of their waking hours in the postprandial state, resulting in repeated elevations in circulating TG, glucose and related metabolites[8]. While postprandial glucose responses may, in part, be predicted by the gut microbiome[9], real-life variations in both postprandial lipid and glucose-mediated metabolites have not been explored. We assessed them by considering the overall magnitude of the response by incremental AUC (iAUC), peak concentrations and change from fasting (that is, rise).

First, we measured postprandial TG, glucose, C-peptide, insulin and circulating metabolite concentrations at regular intervals (0–6 h) in the clinic after 2 sequential test meals (890 kcal, 50 g fat and 85 g carbohydrates at 0 h (breakfast) and 500 kcal, 22 g fat and 71 g carbohydrates at 4 h (lunch); Fig. 4b,c). Notably, we found that postprandial TG (0–6 h iAUC), insulin and C-peptide (both 0–2 h iAUC) responses were more strongly associated with the gut microbiome ($\rho = 0.15$, 0.2, 0.24, respectively; AUC > 0.65) compared with postprandial glucose (0–2 h iAUC) responses ($\rho = 0.13$ and AUC = 0.6; Fig. 4b), findings that were replicated in our US cohort (Fig. 4b–g). We also measured glucose concentrations during the 13-d at-home period[16] after isocaloric standardized meals with different macronutrient compositions (Supplementary Table 3). However, contrary to our clinic meal responses (Fig. 4b) and previous work[9], the glucose 0–2 h iAUCs after these meals did not achieve high correlations with the microbiome (all $\rho < 0.07$ and AUC < 0.59; Fig. 4c). While this may be dependent on meal composition and the effect of multiple meals consumed after stool collection, these results suggest that the microbiome is a stronger predictor of postprandial lipemia than glycemia.

Postprandial rises in lipid- and glucose-mediated measures are differentially predicted by the microbiome compared with fasting levels. Postprandial measures depend both on the corresponding fasting levels and meal-induced rise. Therefore, we compared the differential prediction accuracy of the gut microbiome for fasting levels, postprandial (peak) total levels and postprandial rises (Fig. 4h). For lipid- and glucose-mediated (clinic day) measures, despite a similar strength of association between peak (6 h), magnitude (iAUC) and fasting TG concentrations, the rise (6–0 h) was not similarly correlated (Fig. 4a–e,f). In contrast, the microbiome associations with glycemic measures were comparable between fasting, peak and rise (Fig. 4a–d).

Of particular interest were lipoprotein subfraction concentrations, composition and size (Extended Data Fig. 4b,c), which are remodeled postprandially into potentially atherogenic lipoproteins (for example, large VLDL particles, TG-enriched LDL and HDL particles)[41]. These particles were predicted at comparable accuracy for both fasting and postprandial peak 6-h concentrations (Fig. 4a–e,f–h); notably, HDLD and VLDLD achieved modestly stronger correlations ($\rho = 0.32$ for both) postprandially (Fig. 4f). However, as with TG, we found that the microbiome was substantially less predictive for the postprandial rise in all lipid metabolite measures compared with fasting and postprandial peak concentration (Fig. 4a–e,f–h). For example, HDLD is closely associated with gut microbial composition at fasting and 6 h postprandially ($\rho = 0.29$ and 0.32; AUC = 0.71 and 0.72, respectively; Fig. 4a–e,f–h), but not with the rise (Fig. 4f). These differential associations suggest that the microbiome may influence postprandial lipid-mediated measures via effects on fasting measures.

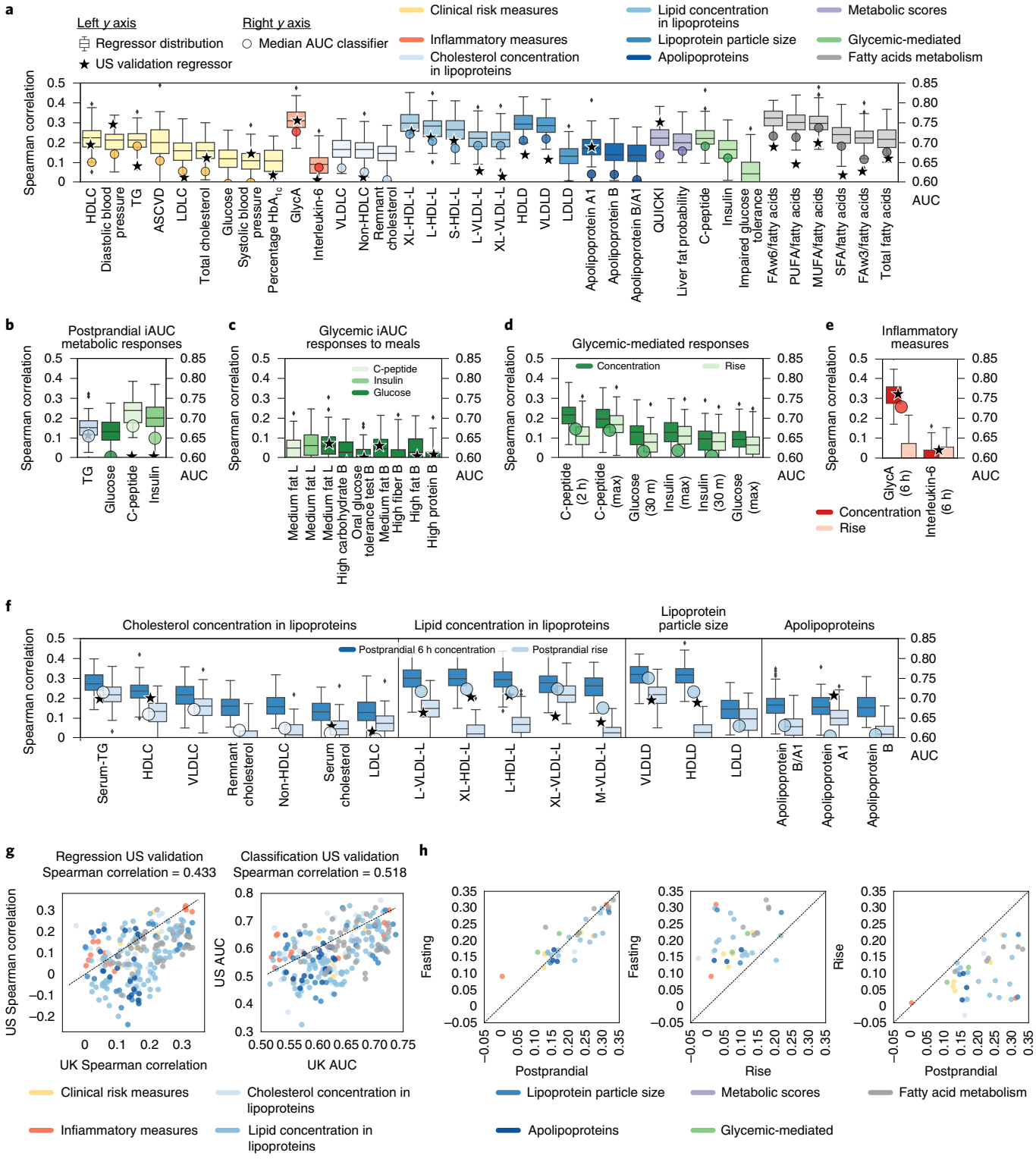
Distinct microbial signatures discriminate between positive and negative metabolic health indices under fasting conditions. Motivated by the observed potential of the gut microbiome to predict the fasting and postprandial levels of circulating metabolic markers, we next assessed the microbiome features driving these associations. Among three general risk indices of cardiovascular health (atherosclerotic cardiovascular disease, liver fat probability and insulin sensitivity or QUICKI; Fig. 4a), we found six species significantly and concordantly correlated with all three (negatively or positively, $P < 0.05$), hinting at a global underlying microbial signature of improved metabolic health. These taxa included *Clostridium* CAG:58 (higher cardiometabolic risk) and *Haemophilus parainfluenzae* (lower risk) that we had previously linked with healthy and less healthy dietary patterns (Fig. 2b).

We found similarly distinct separations between two opposing, clearly defined clusters of species either positively or negatively correlated with fasting cardiometabolic measures (Fig. 5a,b). Species correlated with positive markers included some prevalent taxa generally regarded as healthy (*F. prausnitzii*) but also eight uncultivated and undercharacterized bacteria. The positive cluster included many distinct genera, pointing at a rich functional diversity. In contrast, the cluster negatively correlated with

Fig. 4 | Fasting and postprandial cardiometabolic responses to standardized test meals associated with the microbiome. a, The strongest observed links according to the correlation of predicted versus collected measures between the gut microbiome and fasting metabolic blood markers. For measures of lipid concentration in lipoproteins, we report the five strongest correlations only. Indices are grouped in nine distinct categories and the box plots report the correlation between the prediction of random forest regression models trained on microbial taxa or pathway abundances across 100 training/testing folds; the stars report the regressor performance when trained on the UK cohort and evaluated on the independent US validation cohort (left-side *y* axis). The circles denote the AUC values for the random forest classification (right-side *y* axis). **b**–**f**, Performance of our microbiome-based machine learning model in estimating postprandial absolute levels and postprandial increases in cardiometabolic markers. The stars denote the regression model results in our US validation cohort for postprandial measurements (not rises; Extended Data Fig. 4b,c). **b**, Random forest regression and classification performance in predicting postprandial metabolic responses for clinic meal 1 (breakfast) measured as iAUC at 6 h for TGs and iAUC at 2 h for glucose, C-peptide and insulin. **c**, Glycemic-mediated postprandial iAUCs at 2 h for the other meals (Supplementary Table 7). **d**, Glycemic-mediated markers of absolute levels versus rise. **e**, Postprandial inflammatory measures (concentration and rise). **f**, Postprandial lipoprotein measures (6 h concentration and rise). **g**, Overall agreement between random forest regression and classification tasks for the UK models applied to the independent US cohort. **h**, Random forest microbiome-based model performance with postprandial changes (concentrations and rise) in lipoprotein concentration, composition and size. Fasting and postprandial performance indices (correlation of the regressors' outputs) were more tightly linked to gut community structure than were their corresponding postprandial rises. The box plots show the first and third quartiles (boxes) and median (middle line); the whiskers extend up to 1.5× the IQR.

positive markers included eight *Clostridium* species and the recurrent negatively connotated *Ruminococcus gnavus* and *F. plautii*. Large HDL particles (and their lipid compositions; Extended Data Figs. 6 and 7), which have strong inverse associations with cardiometabolic outcomes[38], were associated with the healthy cluster. Conversely, lipoproteins associated with an increased risk of CVD and type 2 diabetes (VLDL of all sizes and lipid composition) and atherogenicity[42] (small LDL, medium HDL and small HDL TG), were associated with the less healthy cluster (Extended Data Figs. 6 and 7).

Circulating omega-6 and total PUFA were associated with the healthy cluster (Fig. 5a and Supplementary Table 5). Due to the lack of endogenous production of PUFA, circulating levels closely reflect dietary intake[43] and are linked to a reduced risk of chronic disease[38]. In contrast, circulating monounsaturated fatty acids (MUFAs), which do not closely reflect dietary intake and unlike dietary MUFA have been linked to increased risk of chronic disease[38], were associated with the unhealthy cluster, with an undercharacterized *Firmicutes* species (CAG:170) and *Clostridium bolteae* responsible for the strongest negative and positive associations, respectively.



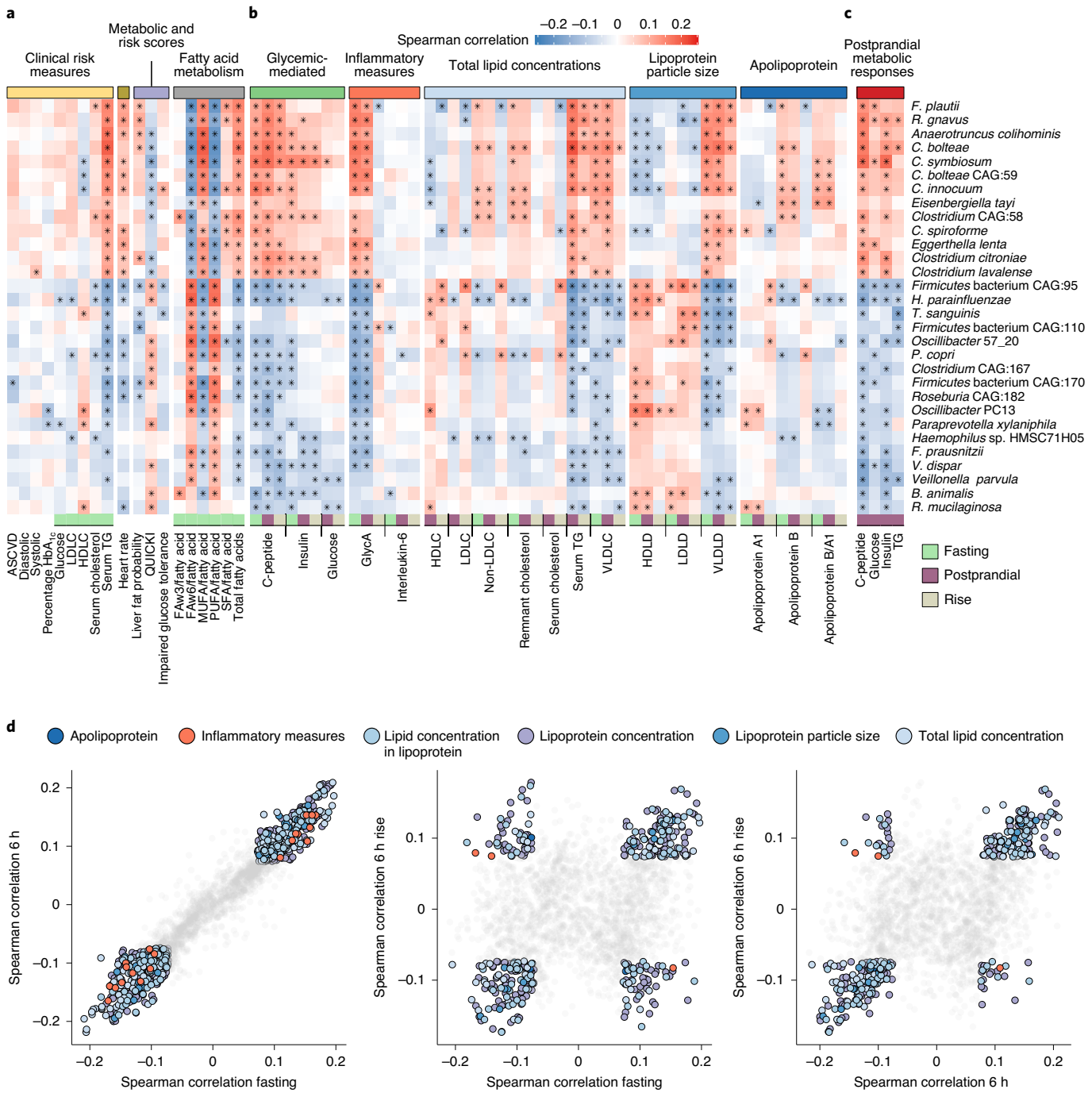


Fig. 5 | Species-level segregation into healthy and unhealthy microbial signatures of fasting and postprandial cardiometabolic markers. a,**b**, Associations (Spearman correlation, $q < 0.2$ marked with stars) between single microbial species and fasting clinical risk measures (**a**) and glycemic, inflammatory and lipemic indices (**b**). **c**, Correlation between microbial species and the iAUC for glucose and C-peptide estimations based on clinical measurements before and after standardized meals. The 30 species with the highest number of significant correlations with distinct fasting and postprandial indices are shown. Rows are hierarchically clustered (complete linkage, Euclidean distance). **d**, Microbe–metabolite correlations are very consistent when evaluated for fasting versus postprandial (6 h) conditions (left). Associations with postprandial variations (rise) conversely often show opposing relationships, with several species positively correlated with fasting measures being negatively correlated with postprandial variation of the same metabolite (or vice versa; center). This was mitigated somewhat when comparing absolute postprandial responses with rise (right). The complete set of correlations, *P* and *q* values are available in Supplementary Table 5 and 6 for UK and US, respectively.

Both favorable and unfavorable microbial signatures of metabolic health are maintained under postprandial conditions. Links between postprandial levels of cardiometabolic and inflammatory measures corresponded with the segregation of healthful versus detrimental taxa observed under fasting conditions (Fig. 5b,c and Extended Data Figs. 6 and 7). Notably, fasting and postprandial GlycA were strongly linked with the microbiome (47 species significantly correlated at 6 h and 64 at fasting), substantially

exceeding interleukin-6 levels (5 and 16 significant associations; Fig. 5b,c). *C. boltae* and *R. gnavus* correlated the most with increased fasting and postprandial inflammation, whereas *H. parainfluenzae* and *Firmicutes* bacterium CAG:95 were the strongest associations with reduced GlycA levels. VLDL lipoprotein subfractions (markers of adverse cardiometabolic effects) were also consistently associated with the less healthy cluster both at fasting and postprandially.

Postprandial rises, rather than absolute postprandial levels, were in some cases uncoupled from the microbial associations with fasting markers (Fig. 5d). For example, change in GlycA (Fig. 5b) was differentially associated with clusters compared to fasting and postprandial levels (especially for *F. plautii*, *Firmicutes* bacterium CAG:95 and *Firmicutes* bacterium CAG:110), probably due to the small reduction in GlycA postprandially. Other immunological markers and some lipid and cholesterol levels paralleled this behavior (Extended Data Fig. 6), possibly reflecting postprandial lipoprotein remodeling[44].

We observed the same favorable versus unfavorable clustering of microbiome features when analyzing microbial pathways and gene families (Extended Data Fig. 8) supporting taxa segregation by their underlying biochemical activities. The strengths of microbe–blood marker associations were confirmed by random forest feature relevance analysis (Extended Data Fig. 9); importantly, they were confirmed in the US cohort. For the 209 microbe–index correlations that were significant both in the UK ($q<0.2$) and US cohorts ($P<0.05$), the concordance in the sign of the correlation reached 88.7% for the associations in fasting conditions and 96.1% postprandially.

***P. copri* diversity and *Blastocystis* presence are markers of improved postprandial glucose responses.** Some ecologically unusual microbes hypothesized to have population-scale health effects solely based on their presence or absence appeared in our microbial signatures[45]. Among them, *P. copri*[45,46] had conflicting previous accounts for its role in glucose homeostasis[47,48] possibly due to subspecies diversity[49,50]. Our data found *P. copri* to be associated with beneficial cardiometabolic markers, being negatively correlated with estimated visceral fat ($\rho=-0.11$, $P=0.0006$), fasting VLDL-D ($\rho=-0.08$, $P=0.011$) and fasting GlycA ($\rho=-0.14$, $P<0.0001$) among others (Supplementary Table 5). While almost no diet indices were associated with *P. copri*, postprandial rises in glucose ($\rho=-0.11$, $P<0.001$) and polyunsaturated/omega-6 fatty acids ($\rho=0.15$ and 0.14, respectively, and $P<0.001$) were top-scoring correlations for this bacterium and were stronger than corresponding fasting and postprandial levels in contrast with what we observed for the overall microbiome (Fig. 4a,b). *P. copri* was present in at least one of its subtypes[49] in 29.8% of the PREDICT 1 individuals and *P. copri* carriers had lower C-peptide (−9.2%, $P=0.002$), insulin (−14%, $P=0.006$) and TG levels (−3.2%, $P=0.003$) compared to *P. copri*-negative individuals (Extended Data Fig. 10 and Supplementary Table 8). Similarly, postprandial glucose after breakfast was significantly less pronounced in individuals with *P. copri* (−20.4% glucose iAUC at 2 h, $P=0.002$; Extended Data Fig. 10c) and visceral fat was significantly lower (−12.5%, $P=0.005$; Extended Data Fig. 10a). This set of diverse associations supports that the presence of *P. copri* in the gut microbiome could be beneficial in glucose homeostasis and host metabolism.

Blastocystis is a unicellular eukaryotic parasite increasingly regarded as a commensal member of the gut microbiome[51–53]. It shares with *P. copri* a limited prevalence in Western-lifestyle populations[53] and a high abundance when present. We found evidence that *Blastocystis*-positive individuals (25.7% in our cohort) also had favorable glucose homeostasis and lower estimated visceral fat (−15.7% glucose iAUC, −22.1% visceral fat, $P<0.002$; Extended Data Fig. 10). The latter confirms that *Blastocystis* is less prevalent in individuals with high BMI, as suggested previously[53]. Interestingly, the effect of the simultaneous presence of *P. copri* and *Blastocystis* (12% of individuals) appeared to further promote healthier metabolic function. Visceral fat was 17.3% lower on average ($P<0.005$; Supplementary Table 8) for individuals positive for both *P. copri* and *Blastocystis* compared to individuals with only one or the other and 23.3% lower ($P=8.9\times10^{-6}$) compared with individuals lacking both.

A clear microbial signature of cardiometabolic health levels consistent across diet, obesity indicators and cardiometabolic risks. We observed above a consistent set of microbial species that were strongly linked to (1) food indices reflecting different levels of healthy diets, (2) indicators of obesity and cardiometabolic health, (3) fasting circulating metabolites connected with cardiometabolic risk and (4) postprandial responses. To test the consistency of such a signature, we selected representative cardiometabolic health indicators from each category and ranked microbial species based on their correlation coefficient. We found remarkable agreement among microbes associated with different positive or negative indicators of cardiometabolic health (Fig. 6 and Supplementary Table 9).

In particular, *Firmicutes* bacterium CAG:95 was the uncultivated species with the most beneficial score. Of the health-associated microbial species, only *F. prausnitzii* and, partially, *P. copri* were already convincingly linked with health in previous investigations[54]. The beneficial signature also included *Eubacterium eligens* and *H. parainfluenzae*, without previous clear roles in health, and additional species without cultivated representatives such as *Roseburia* bacterium CAG:182, *Oscillibacter* sp. 57_20, *Firmicutes* bacterium CAG:170, *Oscillibacter* sp. PC13 and *Clostridium* sp. CAG:167. Species conversely consistent with indicators of poor overall health (Fig. 6) included the already discussed set of *Clostridia* (*C. spiroforme*, *C. bolteae* CAG:59, *C. bolteae*, *Clostridium* sp. CAG:58, *C. symbiosum*, *C. innocuum* and *C. leptum*) and the mucolytic microbes *R. gnavus* and *F. plautii*, again previously found to be associated with disease[55,56]. Overall, this set of 30 species serves as a marker of overall good or poor cardiometabolic health and dietary patterns in nondiseased human hosts.

Discussion

PREDICT 1 represents the first diet–microbiome study to identify both individual components of the microbiome and an overall gut microbial signature associated with multiple measures of dietary intake and cardiometabolic health. These signatures were reproduced across UK and US populations, across multiple previously published study populations and for multiple dietary and health indicators. Notably, microbiome signatures grouped both microbiome and dietary components into health-associated and anti-associated clusters, the latter in agreement with dietary quality and diversity scores[20,57]. The diversity and quality of a healthy diet (HFD and PDI) was particularly predictable by the microbiome, surpassing other indices including the Mediterranean diet previously linked with microbiome composition[58]. The segregation of favorable and unfavorable microbial clusters according to the heterogeneity of the food source (healthy or unhealthy animal or plant), quality (processed versus unprocessed) and dietary patterns highlights the importance of looking beyond nutrients and single foods in diet–microbiome research. The substantially greater detail and consistency in our results relative to previous diet–microbiome work[9,11–13,15] may be due to the quality in dietary recording, metagenomic profiling and the large sample size. However, given the limitations of FFQ dietary data, future diet–microbiome studies would benefit further from higher resolution dietary assessment methodologies, such as weighed food record data.

Several aspects of the consistent gut microbiome signatures across diet, obesity and cardiometabolic health measures are

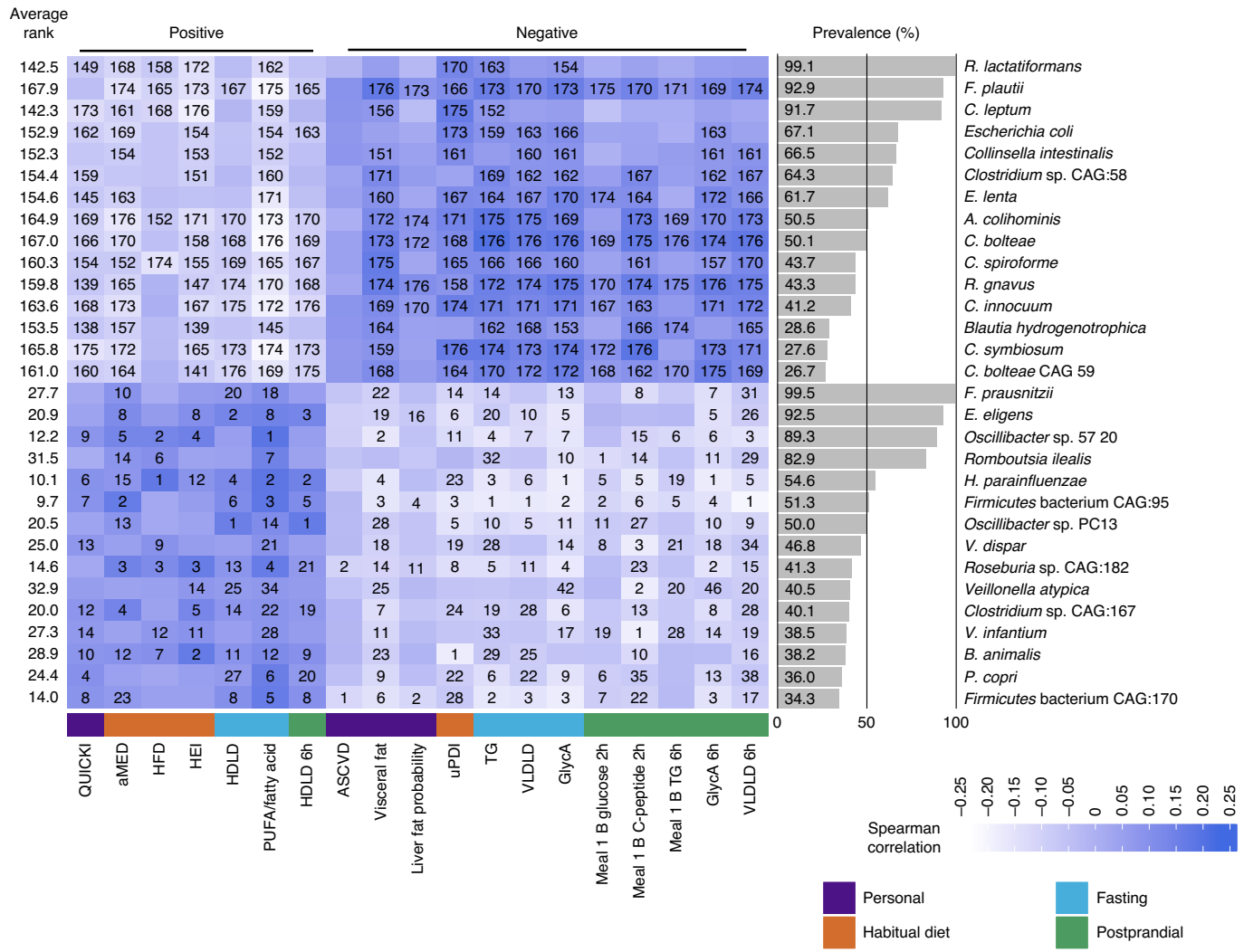


Fig. 6 | Panel of the 30 species showing the strongest overall correlations with a selection of markers of nutritional and cardiometabolic health. The 30 species with the highest and lowest average ranks with diverse positive and negative cardiometabolic health and healthy diet indicators, respectively, are shown. The rank of each microbe's correlation with individual indicators is written within cells when significant ($P < 0.05$). For each of the main categories of indices, we selected up to five representative markers (for 'personal' we considered only four since the remaining were highly correlated with visceral fat or not relevant in this context). Indices can be considered positive and negative depending on whether higher or lower values are a proxy for more or less healthy conditions. Partial correlations were computed using the pcor.test (two-sided) with params 'method=spearman' (Methods). Correlations and ranks are available in Supplementary Table 9. *P* values and FDR-adjusted *P* values are available in Supplementary Table 2.

striking for their potential new epidemiology and microbial biochemistry. A surprising proportion of diet- or health-associated taxa in these results are largely uncharacterized or represented solely by metagenomic assemblies[5]. Other microbes found in this study to have dietary or cardiometabolic associations, such as *Prevotella* or *Blastocystis* spp., have been characterized in greater biochemical detail but their population structure in the human microbiome has only recently begun to be appreciated[49,53]. The latter in particular may be only one of many examples of nonbacterial microbiome members not amenable to most current high-throughput approaches but with unexpected and potentially key positive roles in humans.

Likewise, these new contributions of the gut microbiome to human dietary responses may help to explain some of the heterogeneity seen among previous population studies[4,9,59]. First, diet–microbiome–blood marker associations were overall strongest with regard to circulating lipid levels relative to glycemic indices. It is possible that the relative contribution of gut microbes is higher for circulating lipid levels than carbohydrate derivatives, through either direct processes or indirectly through gastrointestinal or systemic bile acid signaling[60]. Alternatively, host metabolism may play a greater role in circulating glucose and insulin levels relative to microbial bioactivity. The lipoprotein features most closely associated with the microbiome (such as total lipids in large HDL) are also more strongly associated with cardiovascular risk compared with typically measured lipids (for example, total cholesterol, HDLC, LDLC), suggesting that their utility as clinical biomarkers or as targets for beneficial gut microbiome manipulation warrants further investigation.

Overall, this is the first study to identify a shared diet–metabolic health microbial signature, segregating favorable and unfavorable taxa with multiple measures of both dietary intake and cardiometabolic health. As a resource, these results will aid both in the utilization of the gut microbiome as a biomarker for cardiometabolic risk and in strategies for reshaping the microbiome to improve personalized dietary health.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-01183-8.

Received: 10 March 2020; Accepted: 16 November 2020;
Published online: 11 January 2021

References

1. Ng, M. et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **384**, 766–781 (2014).
2. Brown, J. M. & Hazen, S. L. Microbial modulation of cardiovascular disease. *Nat. Rev. Microbiol.* **16**, 171–181 (2018).
3. Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
4. Sze, M. A. & Schloss, P. D. Looking for a signal in the noise: revisiting obesity and the microbiome. *mBio* **7**, e01018-16 (2016).
5. 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).
6. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
7. Gilbert, J. A. et al. Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
8. Berry, S. E. et al. Human postprandial responses to food and potential for precision nutrition. *Nat. Med.* **26**, 964–973 (2020).
9. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
10. Mendes-Soares, H. et al. Model of personalized postprandial glycemic response to food developed for an Israeli cohort predicts responses in Midwestern American individuals. *Am. J. Clin. Nutr.* **110**, 63–75 (2019).
11. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
12. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
13. Thingholm, L. B. et al. Obese individuals with and without type 2 diabetes show different gut microbial functional capacity and composition. *Cell Host Microbe* **26**, 252–264.e10 (2019).
14. Schirmer, M. et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1897 (2016).
15. Fu, J. et al. The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circ. Res.* **117**, 817–824 (2015).
16. Berry, S. et al. Personalised REsponses to DIetary Composition Trial (PREDICT): an intervention study to determine inter-individual differences in postprandial response to foods. Preprint at https://protocolexchange.researchsquare.com/article/pex-802/v1 (2020).
17. Xie, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584.e3 (2016).
18. Atabaki-Pasdar, N. et al. Predicting and elucidating the etiology of fatty liver disease: a machine learning modeling and validation study in the IMI DIRECT cohorts. *PLoS Med.* **17**, e1003149 (2020).
19. Vojinovic, D. et al. Relationship between gut microbiota and circulating metabolites in population-based cohorts. *Nat. Commun.* **10**, 5813 (2019).
20. Vadiveloo, M., Dixon, L. B., Mijanovich, T., Elbel, B. & Parekh, N. Development and evaluation of the US Healthy Food Diversity index. *Br. J. Nutr.* **112**, 1562–1574 (2014).
21. Guenther, P. M. et al. Update of the healthy eating index: HEI-2010. *J. Acad. Nutr. Diet.* **113**, 569–580 (2013).
22. Fung, T. T. et al. Diet-quality scores and plasma concentrations of markers of inflammation and endothelial dysfunction. *Am. J. Clin. Nutr.* **82**, 163–173 (2005).
23. Reedy, J. et al. Higher diet quality is associated with decreased risk of all-cause, cardiovascular disease, and cancer mortality among older adults. *J. Nutr.* **144**, 881–889 (2014).
24. Mitrou, P. N. et al. Mediterranean dietary pattern and prediction of all-cause mortality in a US population: results from the NIH-AARP Diet and Health Study. *Arch. Intern. Med.* **167**, 2461–2468 (2007).
25. Satija, A. et al. Plant-based dietary patterns and incidence of type 2 diabetes in US men and women: results from three prospective cohort studies. *PLoS Med.* **13**, e1002039 (2016).
26. Vadiveloo, M., Parekh, N. & Mattei, J. Greater healthful food variety as measured by the US Healthy Food Diversity index is associated with lower odds of metabolic syndrome and its components in US adults. *J. Nutr.* **145**, 564–571 (2015).
27. Onvani, S., Haghighatdoost, F., Surkan, P. J., Larijani, B. & Azadbakht, L. Adherence to the Healthy Eating Index and Alternative Healthy Eating Index dietary patterns and mortality from all causes, cardiovascular disease and cancer: a meta-analysis of observational studies. *J. Hum. Nutr. Diet.* **30**, 216–226 (2017).
28. Redondo-Useros, N. et al. Associations of probiotic fermented milk (PFM) and yogurt consumption with *Bifidobacterium* and *Lactobacillus* components of the gut microbiota in healthy adults. *Nutrients* **11**, 651 (2019).
29. Sakamoto, M., Iino, T., Yuki, M. & Ohkuma, M. *Lawsonibacter asaccharolyticus* gen. nov., sp. nov., a butyrate-producing bacterium isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **68**, 2074–2081 (2018).
30. Satija, A. et al. Healthful and unhealthful plant-based diets and the risk of coronary heart disease in U.S. adults. *J. Am. Coll. Cardiol.* **70**, 411–422 (2017).
31. Monteiro, C. A. et al. The UN Decade of Nutrition, the NOVA food classification and the trouble with ultra-processing. *Public Health Nutr.* **21**, 5–17 (2018).
32. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
33. Beaumont, M. et al. Heritable components of the human fecal microbiome are associated with visceral fat. *Genome Biol.* **17**, 189 (2016).
34. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
35. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
36. D'Agostino, R. B. Sr et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
37. Kettunen, J. et al. Biomarker glycoprotein acetyls is associated with the risk of a wide spectrum of incident diseases and stratifies mortality risk in angiography patients. *Circ. Genom. Precis. Med.* **11**, e002234 (2018).
38. Würtz, P. et al. Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* **131**, 774–785 (2015).
39. Hrebícek, J., Janout, V., Malincíková, J., Horáková, D. & Cízek, L. Detection of insulin resistance by simple quantitative insulin sensitivity check index QUICKI for epidemiological assessment and prevention. *J. Clin. Endocrinol. Metab.* **87**, 144–147 (2002).
40. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
41. Wojczynski, M. K. et al. High-fat meal effect on LDL, HDL, and VLDL particle size and number in the Genetics of Lipid-Lowering Drugs and Diet Network (GOLDN): an interventional study. *Lipids Health Dis.* **10**, 181 (2011).
42. Skeggs, J. W. & Morton, R. E. LDL and HDL enriched in triglyceride promote abnormal cholesterol transport. *J. Lipid Res.* **43**, 1264–1274 (2002).
43. Hodson, L., Skeaff, C. M. & Fielding, B. A. Fatty acid composition of adipose tissue and blood in humans and its use as a biomarker of dietary intake. *Prog. Lipid Res.* **47**, 348–380 (2008).
44. Cohn, J. S. Postprandial lipemia: emerging evidence for atherogenicity of remnant lipoproteins. *Can. J. Cardiol.* **14**, 18B–27B (1998).
45. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
46. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
47. Kovatcheva-Datchary, P. et al. Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of *Prevotella*. *Cell Metab.* **22**, 971–982 (2015).
48. Pedersen, H. K. et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016).
49. Tett, A. et al. The *Prevotella copri* complex comprises four distinct clades underrepresented in westernized populations. *Cell Host Microbe* **26**, 666–679.e7 (2019).
50. De Filippis, F. et al. Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. *Cell Host Microbe* **25**, 444–453.e3 (2019).
51. Clark, C. G., van der Giezen, M., Alfellani, M. A. & Stensvold, C. R. Recent developments in *Blastocystis* research. *Adv. Parasitol.* **82**, 1–32 (2013).
52. Lukeš, J., Stensvold, C. R., Jirků-Pomajbíková, K. & Wegener Parfrey, L. Are human intestinal eukaryotes beneficial or commensals? *PLoS Pathog.* **11**, e1005039 (2015).
53. Beghini, F. et al. Large-scale comparative metagenomics of *Blastocystis*, a common member of the human gut microbiome. *ISME J.* **11**, 2848–2863 (2017).
54. Sokol, H. et al. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl Acad. Sci. USA* **105**, 16731–16736 (2008).

55. Hall, A. B. et al. A novel *Ruminococcus gnavus* clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).
56. Valles-Colomer, M. et al. The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4**, 623–632 (2019).
57. Kim, H., Caulfield, L. E. & Rebholz, C. M. Healthy plant-based diets are associated with lower risk of all-cause mortality in US adults. *J. Nutr.* **148**, 624–631 (2018).
58. Meslier, V. et al. Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. *Gut* **69**, 1258–1268 (2020).
59. Kurilshikov, A. et al. Gut microbial associations to plasma metabolites linked to cardiovascular phenotypes and risk. *Circ. Res.* **124**, 1808–1820 (2019).
60. Ko, C.-W., Qu, J., Black, D. D. & Tso, P. Regulation of intestinal lipid metabolism: current concepts and relevance to disease. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 169–183 (2020).

Methods

The PREDICT 1 study. The PREDICT 1 clinical trial (NCT03479866) aimed to quantify and predict individual variations in metabolic responses to standardized meals. We integrated data from a cohort of twins and unrelated adults from the UK to explore genetic, metabolic, microbiome composition, meal composition and meal context data to distinguish predictors of individual responses to meals. We then validated these predictions in an independent cohort of adults from the USA. The trial was a single-arm, single-blinded intervention study that commenced in June 2018 and was completed in May 2019. Ethical approval for the study was obtained in the UK from the Research Ethics Committee and Integrated Research Application System (IRAS 236407) and in the USA from the Institutional Review Board (Partners Healthcare IRB 2018P002078). The trial was run in accordance with the Declaration of Helsinki (2013) and good clinical practice. Study procedures were only carried out after having received written informed consent from each participant.

For the full protocol, see Berry et al.[16]. Briefly, 1,002 generally healthy adults from the UK (non-twins and identical (monozygotic) and nonidentical (dizygotic) twins) and 100 healthy adults from the USA (non-twins; validation cohort) were enrolled in the study (see Berry et al.[8] for the eligibility criteria) and completed the baseline clinic measurements. The study consisted of a 1-d clinical visit at baseline followed by a 13-d at-home period. At baseline (day 1), participants arrived fasted and were given a standardized metabolic challenge meal for breakfast (0 h; 86 g carbohydrate, 53 g fat) and lunch (4 h; 71 g carbohydrate, 22 g fat). Fasting and postprandial (9 time points; 0–6 h) venous blood was collected to determine the serum concentrations of glucose, TG, insulin, C-peptide (as a surrogate for insulin) and metabolomics (nuclear magnetic resonance). Stool samples, anthropometry and a questionnaire querying habitual diet, lifestyle and medical health were obtained at baseline. During the home phase (days 2–14), participants consumed standardized test meals in duplicate varying in sequence and in macronutrient composition while wearing digital devices to continuously monitor their blood glucose (continuous glucose monitor; CGM), physical activity and sleep. Capillary blood was collected using dried blood spot cards during the clinic visit and at home to analyze fasting and postprandial concentrations of TG and C-peptide. Participants were supported throughout the study with reminders and communication from study staff delivered through the Zoe study app. A second stool sample was collected at home by participants after completion of the study; all devices and samples were mailed back to study staff. To monitor compliance, all test meals consumed by participants were logged in the Zoe app (with an accompanying picture) and reviewed in real time by the study nutritionists. Only test meals that were consumed according to the standardized meal protocol (outlined in Berry et al.[8]) were included in the analysis.

The recruitment criteria, meal intervention challenges, outcome variables and sample collection and analysis procedures relevant to this paper are described elsewhere[8,16]. The core characteristics of study participants at baseline were not significantly different between UK and US cohorts[8].

Overview of microbiome sequencing and profiling. We performed deep shotgun metagenomic sequencing (mean 8.8 ± 2.2 gigabase pairs per sample) in stool samples from a total of 1,098 PREDICT 1 participants (UK, $n = 1{,}001$; USA, $n = 97$). From a random subset of these participants ($n = 105$), we additionally sequenced fecal metagenomes from a second stool sample collected 14 d after the first collection (Fig. 1a) for a total of 1,168 metagenomes. Computational analysis was performed using the bioBakery suite of tools[61] to obtain species-level microbial abundances for the 769 taxa identified using the newly updated MetaPhlAn v.3.0 tool[62], functional potential profiling of >1.91 M microbial gene families, 445 Kyoto Encyclopedia of Genes and Genomes pathways with HUMAnN v.2.0[63] and reconstruction of 48,181 metagenome-assembled genomes (MAGs) of medium or high quality using our validated pipeline[5], which includes assembly with MEGAHIT[64], binning with MetaBAT 2[65] and quality control with CheckM[66].

Microbiome sample collection. Participants were mailed a pre-visit study pack with a stool collection kit and relevant questionnaires and asked to collect an at-home stool sample at two time points (one before their in-person clinical visit on day 0 and the next at the conclusion of their home phase on day 14). Those who did not collect a sample before their in-person, baseline visit completed the collection as soon as possible during the home phase. Baseline samples in the UK were collected using the EasySampler Stool Collection Kit (ALPCO), whereas post-study samples, as well as the entirety of the US collection, was conducted using the FECOTAINER stool sample kit (Excretas Medical BV). For baseline samples, one fresh unfixed sample was deposited into a sterile universal collection container (catalog no. L0263-10; Sarstedt Australia) and one into a tube containing DNA/RNA Shield buffer (catalog no. R1101; Zymo Research). Samples were stored at ambient temperature until returned to the study staff. Follow-up samples were collected similarly but only sampled into a DNA/RNA Shield buffer tube and sent by standard mail to study staff. On receipt in the laboratory, samples were homogenized, aliquoted and stored at −80 °C in QIAGEN PowerBeads 1.5-ml tubes. This sample collection procedure was tested and validated internally comparing different storage conditions (fresh, frozen, buffer), different DNA extraction kits (PowerSoil Pro, FastDNA, Protocol Q, Zymo) and different sequencing technologies (16S ribosomal RNA, shotgun metagenomics and arrays) (data not shown).

DNA extraction and sequencing. DNA was isolated by QIAGEN Genomic Services using DNeasy 96 PowerSoil Pro from all day 0 (baseline) DNA/RNA Shield-fixed microbiome samples. A random subset of day 14 (end of at-home phase) samples ($n = 105$) were also extracted. Optical density measurement was done using spectrophotometer quantification (Tecan Infinite 200). Before library preparation and sequencing, the quality and quantity of the samples were assessed using the Fragment Analyzer system (Agilent Technologies) according to manufacturer's guidelines. Samples with a high-quality DNA profile were further processed. The NEBNext Ultra II FS DNA Module (catalog no. E7810S/L; New England Biolabs) was used for DNA fragmentation, end repair and A-tailing. For adapter ligation, the NEBNext Ultra II Ligation Module (catalog no. E7595S/L; New England Biolabs) was used. The quality and yield after sample preparation were measured with the Fragment Analyzer system. The size of the resulting product was consistent with the expected size of approximately 500–700 bp. Libraries were sequenced for 300-bp paired-end reads using the Illumina NovaSeq 6000 platform according to the manufacturer's protocols. The 1.1-nM library was used for flow cell loading. The NovaSeq control software NCS v.1.5 was used. Image analysis, base calling and quality checking were performed with the Illumina data analysis pipeline RTA3.3.5 and bcl2fastq v.2.20.

Metagenome quality control and preprocessing. All sequenced metagenomes were quality control edited using the preprocessing pipeline as implemented in https://github.com/SegataLab/preprocessing. Preprocessing consisted of three main steps: (1) read-level quality control; (2) screening of contaminants, that is, host sequences; and (3) split and sorting of cleaned reads. Initial quality control involves the removal of low-quality reads (quality score <Q20), fragmented short reads (<75 bp) and reads with >2 ambiguous nucleotides. Contaminant DNA was identified using Bowtie 2 (ref. [67]) using the -sensitive-local parameter, allowing confident removal of the phi X 174 Illumina spike-in and human-associated reads (hg19). Sorting and splitting allowed for the creation of standard forward, reverse and unpaired reads output files for each metagenome.

Microbiome taxonomic and functional potential profiling. The metagenomic analysis was performed following the general guidelines[68] and relying on the bioBakery computational environment[61]. The taxonomic profiling and quantification of organisms' relative abundances of all metagenomic samples were quantified using MetaPhlAn v.3.0 (ref. [62]). The updated species-specific database of markers was built using 99,237 reference genomes representing 16,797 species retrieved from GenBank (January 2019). From this set of reference genomes, we extracted a total of 1,132,166 markers used to profile 13,393 species. This set of species also included 83 species defined by the CAG group approach[32] that were very genetically distinct from species represented by isolate genomes and for which the use of unique marker genes limited the potential issues of using metagenomic assemblies reconstructed over multiple samples. Compared to the previous version of the MetaPhlAn2 database (mpa_v20_m200), the updated database profiled 7,116 more species. Metagenomes were mapped internally in MetaPhlAn v.3.0 against the marker gene database with Bowtie 2 v.2.3.4.3 with the parameter 'very-sensitive'. The resulting alignments were filtered to remove reads aligned with an MAPQ value <5, representing an estimated probability of the likelihood of the alignments.

To estimate the microbiome species richness of an individual from the taxonomic profiles of PREDICT 1 participants, we computed two alpha diversity measures: the number of species found in the microbiome ('observed richness'); and the Shannon entropy estimation. We did not perform rarefaction before the alpha diversity calculations because of the low s.d. in sequencing depths and the verified missing correlation between the metadata of interest and sequencing depth. Microbiome dissimilarity between participants (beta diversity) was computed using the Bray–Curtis dissimilarity on microbiome taxonomic profiles.

Functional potential analysis of the metagenomic samples was performed using HUMAnN2 (v.0.11.2 and UniRef database release 2014-07) (ref. [63]), which computed the pathway profiles and gene family abundances.

Metagenomic assembly. Metagenomic samples were processed to obtain MAGs following the procedure we used elsewhere[5]. In brief, we used MEGAHIT v1.2.9 (ref. [64]) with the parameter -k-max 127 for assembly; assembled contigs ≥1.5 kilobases (kb) were considered for the binning step performed using MetaBAT2 v.2.14 (ref. [65]) with the parameters -m 1500 -unbinned. Quality control of the obtained MAGs was performed using CheckM v.1.0.18 (ref. [66]) using default parameters. High- and medium-quality microbial genomes were integrated into the existing database of >150,000 human MAGs.

Collection and processing of habitual diet information. Habitual diet information was collected using FFQs. For the UK, the European Prospective Investigation into Cancer and Nutrition (EPIC) FFQ was used; in the USA, the Harvard semiquantitative FFQ was used.

For the UK, we used an adapted 131-item EPIC FFQ that was developed and validated against pre-established nutrient biomarkers for the EPIC Norfolk[69]. The questionnaire captured average intakes in the past year. UK nutrient intakes were determined using the FETA software (v. 2.53) to calculate macro- and micronutrient data[70]. Sixteen additional foods in the modified FFQ were reviewed by two dietitians and one nutritionist who manually matched food items to corresponding foods within *McCance and Widdowson's The Composition of Food*[71], with portions allocated according to the *Food Portion Sizes*[72]. US participants completed the Harvard 2007 Grid 131-item FFQ previously validated against 2-week dietary records[73]. Nutrient intakes were estimated using the Harvard nutrient database (version SFFQ 043019; https://regepi.bwh.harvard.edu/health/nutrition/index.html). Submitted FFQs were excluded if more than 10 food items were left unanswered or if the total energy intake estimate derived from the FFQ as a ratio of the participant's estimated basal metabolic rate (determined by the Harris–Benedict equation[74]) was more than 2 s.d. outside the mean of this ratio (<0.52 or >2.58).

The following dietary indices were calculated as described below and according to categorization listed in Supplementary Tables 2 and 4.

HFD index. The HFD index considers the number, distribution and health value of consumed foods. To obtain this index, FFQ foods were first aggregated into 15 food groups according to the HFD[20]. Health values were then derived from the German Nutrition Society dietary guidelines (https://www.dge.de/en/) and the weight of each food group was multiplied by its corresponding health value. Scores were divided by the maximum (health value = 0.26) to bind values between 0 and 1 before multiplication with the Berry index. The original HFD was used instead of the US-HFD for the following reasons: the original HFD gives greater emphasis to plant-based foods and less to meat than the US-HFD, which would more closely align with hypothesized microbiome-plant food/fiber interactions; converting UK g per serving to US volume measures (as required for the US-HFD) would introduce additional error to the FFQ estimates.

HEI 2010. The HEI 2010 (ref. [21]) assesses to which extent an individual's food intake aligns with the *Dietary Guidelines for Americans* 2010 (ref. [75]) developed by the US Department of Agriculture. These guidelines cover a total of 12 food groups and nutrients. The HEI has 9 adequacy (encouraged) and 3 moderation (discouraged) components; first, a density approach is used to set per 1,000 kcal calories; and second, least restrictive standards are employed, that is, those that are easiest to achieve among recommendations that vary by energy level, sex and/or age. Total fruits, whole fruits, total vegetables, greens and beans, whole grains, dairy (lean portion only), total protein foods (lean portion of meat only), seafood and plant proteins and fatty acids (PUFAs + MUFAs/SFAs) are considered adequate, whereas refined grains, sodium and empty calories (considered added sugars, solid fats and alcohol above 13 g per 1,000 kcal) are considered detrimental and should be consumed in moderation. The index ranges from 0 (not in agreement with the guidelines) to 100 (completely in agreement with the guidelines).

PDI. Three versions of the PDI[30] were considered: the original PDI; the healthy hPDI; and the uPDI. Eighteen food groups (amalgamated from the FFQ food groups; Supplementary Table 2) were assigned either positive or reverse scores after segregation into quintiles, as outlined in Supplementary Table 4 (ref. [30]). Participants with an intake above the highest quintile for the positive score received a score of 5. Those below the lowest quintile intake received a score of 1. A reverse value was applied for the reverse scores. The scores for each participant were summed to create the final score. For the PDI, a positive score was applied to the healthy and less healthy/unhealthy plant foods and a reverse score was applied to the animal-based foods. For the hPDI, positive scores were applied to the healthy plant foods and a reverse score to the less healthy/ unhealthy plant foods and animal-based foods. For the uPDI, a positive score was applied to the less healthy/ unhealthy plant foods and a reverse score was applied to the healthy plant foods and animal-based foods.

Animal score. The animal-based score categorized animal foods into healthy and less healthy/unhealthy categories according to previous epidemiological studies[76–84]. A similar approach to the PDI scoring was applied to the animal-based food groups, with either a positive (healthy) or reverse (less healthy/unhealthy) quintile scoring (Supplementary Tables 2 and 4).

The aMED score. Adherence to the aMED diet was calculated by following the method outlined by Fung et al.[22]. Nine food/nutrient categories were included (Supplementary Table 4) and the score ranged from 0 to 9 (least to most Mediterranean). To form groups, weekly intake frequencies were first multiplied for assigned foods by the amount in g per serving and then divided by seven to determine g per day. Next, food gram amounts were summed to make the final category total. For all food categories and the fatty acid intake ratio, the median intake of each category was calculated. A score of 0 (no aMED) or 1 (aMED) was given for each category depending on whether the participant was above or below the median intake. For alcohol intake, a range was used for score assignment: females: 5–25 g d^{-1}; males: 10–50 g d^{-1} were assigned a score of 1, while those above or below this range were assigned a score of 0. Finally, the aMED was then generated by the summation of each category score.

Food groups. For individual analyses of food groups-microbe interaction, food groups were formed by aggregation of FFQ foods into the 18 PDI food groups plus margarine and alcohol (Supplementary Table 4).

Percentage of plants within the diet. The percentage of plants within the diet was calculated as the weight (g) of plant foods within the total weight (g) of the diet after adjustment of FFQ foods into quantities (g) per week.

Number of plant foods. For the number of plant foods, each plant food item within the FFQ above the value of 0 g was allocated a score of 1 and summed for each participant. For the total number of plants and the number of healthy and unhealthy plants, FFQ food items were allocated into groups according to the PDI food groupings.

Collection and processing of fasting and postprandial markers. Venous blood samples were collected as outlined in the accompanying protocol paper[16]. Briefly, participants were cannulated and venous blood was collected at fasting (before a test breakfast) and at 9 time points postprandially (15, 30, 60, 120, 180, 240, 270, 300 and 360 min). Plasma glucose and serum C-peptide and insulin were measured at all time points. Serum TG was measured at hourly intervals and serum metabolomics (nuclear magnetic resonance by Nightingale Health using the 2020 platform) at 0, 4 and 6 h. Fasting samples were analyzed for lipid profile, thyroid-stimulating hormone, alanine aminotransferase, liver function panel and complete blood count analysis.

Continuous glucose monitoring on days 2–14 was measured every 15 min using Freestyle Libre Pro continuous glucose monitors (Abbott) fitted on the upper, nondominant arm at participants' baseline clinical visits. Given the CGM device requires time to calibrate once fitted to a participant, CGM data collected 12 h and onwards after activating the device were used for analysis.

Dry blood spot analysis of TG and C-peptide was completed by participants on the first 4 d of the home phase while consuming test meals. The time points were dependent on the test meal as described elsewhere[8,16]. Test cards were stored in aluminum sachets with desiccant once completed and placed in the refrigerator at the end of the study day or until participants mailed them back to the study site. Dry blood spot cards were frozen at −80 °C on receipt in the laboratory until being shipped to Vitas for analysis (Vitas Analytical Services).

Specific time points and increments for TG, glucose, insulin and C-peptide were selected for the current analysis to reflect the different pathophysiological processes for each measure as described in our protocol[16]. The incremental area under the postprandial TG (0–6 h), glucose (0–2 h) and insulin (0–2 h) curves (iAUCs) were computed using the trapezium rule[85].

Detailed descriptions of sample collection, processing and analysis have been reported elsewhere[8,16].

Machine learning. The machine learning framework employed was based on the scikit-learn Python package[86]. The machine learning algorithms used for the prediction and classification of personal, habitual diet, fasting and postprandial metadata are based on random forest regression and classification. We selected random forest-based methods a priori since it has been repeatedly shown to be particularly suitable and robust to the statistical challenges inherent to microbiome abundance data[40,87]. For both the regression and classification tasks, a cross-validation approach was implemented, which was based on 100 bootstrap iterations and an 80/20 random split of training and testing folds. To specifically avoid overfitting as a result of our twin population and their shared factors, we removed any twin from the training fold if their twin was present in the test fold.

For the regression task, we trained a random forest regressor to learn the feature to predict and simple linear regression to calibrate the output for the test folds on the range of values in the training folds. From the scikit-learn package, we used the RandomForestRegressor with the n_estimators=1000, criterion = mse and max_features = sqrt parameters and LinearRegression with default parameters. For the classification task, we divided the continuous features into two classes: the top and bottom quartiles. From the scikit-learn package we used the RandomForestClassifier function with the n_estimators=1000, max_features = sqrt parameters.

We used random forest classification and regression on both species-level taxonomic relative abundance and functional potential profiles. For taxonomic abundances, we used the species-level relative abundances as estimated by MetaPhlAn v.3.0 (see above normalized using the arcsin-sqrt transformation for compositional data). For functional profiles, we considered both raw relative abundance estimates of single microbial gene families and pathway-level relative abundance as provided by HUMAnN2.

As an additional control, we verified that when randomly swapping the target labels or values (classification and regression, respectively), the performances reflected a random prediction, hence an AUC very close to 0.5 and a nonsignificant correlation between the real and predicted values approaching 0.

Statistical analysis. Spearman correlations (reported with ρ in the text) were computed using the cor.test from the stats R package (version 3.5.1) and pcor.test from the ppcor R package (version 1.1), respectively. Correlations and *P* values were computed for each couple of metadata and species; *P* values were corrected using the FDR through the Benjamini–Hochberg procedure, which are reported in the text as *q* values. We considered significant correlations with a $q < 0.2$. Significant species were selected by ranking them according to their number of significant associations for the panel of metadata considered; then, the top 30 unique species were considered for each panel of metadata. In the heatmaps for partial correlations, the asterisk indicates that the correlation index for the corresponding species metadata pair is significant at an FDR ≤ 0.2.

The contribution of metadata variables to microbiota community variation was determined by distance-based redundancy analysis (dbRDA) on species-level Bray–Curtis dissimilarity and Aitchison distance with the capscale function in the vegan R package (version 2.5.6)[88]. Correction for multiple testing (Benjamini–Hochberg, FDR) was applied and significance was defined at an FDR < 0.1. The cumulative contribution of metadata variables or metadata categories was determined by forward model selection on dbRDA (stepwise dbRDA) with the ordiR2step function in vegan, with variables that showed a significant contribution to microbiota community variation in the previous step. Because of the high consistency between the two distance functions, we performed the cumulative distribution analysis using the Bray–Curtis dissimilarity. Only metadata variables with <15% missing data and without high collinearity with other variables (Spearman $\rho < 0.8$) were used as input in the stepwise model.

Data validation on the US cohort and on the curatedMetagenomicData datasets. As independent validation, we considered the publicly available datasets collected in the curatedMetagenomicData v.1.16.0 R package[34]. Of the 57 datasets available, we selected those that had samples with the following characteristics: (1) gut samples collected from healthy adult individuals at first collection (days_from_first_collection=0 or not applicable); (2) samples with age and BMI data available and BMI interquartile range (IQR) of these samples between 3.5 and 7.5 (±2 regarding the PREDICT 1 UK IQR of 5.5; Extended Data Fig. 5). For each dataset with samples meeting the above criteria, only datasets with at least 50 samples were considered: CosteaPI_2017 (ref. [89]) (84 samples out of 279); DhakanDB_2019 (ref. [90]) (88 samples out of 110); HansenLBS_2018 (ref. [91]) (58 samples out of 208); JieZ_2017 (ref. [92]) (157 samples out of 385); SchirmerM_2016 (ref. [14]) (396 samples out of 471); and ZellerG_2014 (ref. [93]) (59 samples out of 199).

We used the previously selected validation datasets from curatedMetagenomicData in two analyses: one based on machine learning to verify the reproducibility of the machine learning model we trained using the PREDICT 1 UK samples; and the second to verify the species-level correlations found in the PREDICT 1 UK cohort. For the first task, we applied a regression algorithm to predict BMI and age. Three different cross-validation approaches were used. First, using each dataset independently in 100 bootstrap iterations and an 80/20 random split of training and testing folds. Second, one more iteration was performed using the PREDICT 1 UK dataset as the training fold and each dataset as the testing fold. Third, a final prediction was made using LODO, meaning that all datasets (PREDICT 1 UK, PREDICT 1 USA and the curatedMetagenomicData datasets) were considered together and each validation dataset was successively used as the test fold while all others were used for training. An additional validation performed using the curatedMetagenomicData datasets was done by applying a pairwise Spearman correlation for each species in each curatedMetagenomicData dataset against BMI and age. For each correlation, we selected the top associated species in PREDICT 1 UK (FDR, $q <= 0.05$) and reported their correlation in curatedMetagenomicData. For those species found also in the PREDICT 1 USA dataset, we also reported their correlation.

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

Data availability

The metagenomes are deposited in European Bioinformatics Institute European Nucleotide Archive under accession no. PRJEB39223. The non-metagenomic data used for analysis in this study are held by the Department of Twin Research at King's College London. The data can be released to bona fide researchers using our normal procedures overseen by the Wellcome Trust and its guidelines as part of our core funding. We receive around 100 requests per year for our datasets and have three meetings per month with independent members to assess proposals. The application can be found at https://twinsuk.ac.uk/resources-for-researchers/access-our-data/. This means that data need to be anonymized and conform to GDPR standards.

Code availability

Computational analyses were performed using the bioBakery suite of tools; species-level microbial abundances were computed using MetaPhlAn v.3.0 (https://github.com/biobakery/MetaPhlAn). Functional potential profiling was carried out with HUMAnN v.2.0 (https://github.com/biobakery/humann; Methods).

References

61. McIver, L. J. et al. bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**, 1235–1237 (2018).
62. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
63. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
64. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
65. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
66. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Meth.* **9**, 357–359 (2012).
68. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
69. Bingham, S. A. et al. Nutritional methods in the European Prospective Investigation of Cancer in Norfolk. *Public Health Nutr.* **4**, 847–858 (2001).
70. Mulligan, A. A. et al. A new tool for converting food frequency questionnaire data into nutrient and food group values: FETA research methods and availability. *BMJ Open* **4**, e004503 (2014).
71. *McCance and Widdowson's The Composition of Foods* 7th edn (Public Health England, 2014).
72. *Food Portion Sizes* 3rd edn (Food Standards Agency, 2002).
73. Rimm, E. B. et al. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. *Am. J. Epidemiol.* **135**, 1114–1126 (1992).
74. Frankenfield, D. C., Muth, E. R. & Rowe, W. A. The Harris–Benedict studies of human basal metabolism: history and limitations. *J. Am. Diet. Assoc.* **98**, 439–445 (1998).
75. McGuire, S. U.S. Department of Agriculture and U.S. Department of Health and Human Services, Dietary Guidelines for Americans, 2010. 7th Edition, Washington, DC: U.S. Government Printing Office, January 2011. *Adv. Nutr.* **2**, 293–294 (2011).
76. World Health Organization & Brouwer, I. A. Effect of *trans*-fatty acid intake on blood lipids and lipoproteins: a systematic review and meta-regression analysis. *World Health Organization* https://apps.who.int/iris/handle/10665/246109 (2016).
77. Zhong, V. W. et al. Associations of dietary cholesterol or egg consumption with incident cardiovascular disease and mortality. *JAMA* **321**, 1081–1095 (2019).
78. de Souza, R. J. et al. Intake of saturated and trans unsaturated fatty acids and risk of all cause mortality, cardiovascular disease, and type 2 diabetes: systematic review and meta-analysis of observational studies. *BMJ* **351**, h3978 (2015).
79. Michaëlsson, K. et al. Milk intake and risk of mortality and fractures in women and men: cohort studies. *BMJ* **349**, g6015 (2014).
80. Mazidi, M. et al. Consumption of dairy product and its association with total and cause specific mortality: a population-based cohort study and meta-analysis. *Clin. Nutr.* **38**, 2833–2845 (2019).
81. Petsini, F., Fragopoulou, E. & Antonopoulou, S. Fish consumption and cardiovascular disease related biomarkers: a review of clinical trials. *Crit. Rev. Food Sci. Nutr.* **59**, 2061–2071 (2019).
82. Rimm, E. B. et al. Seafood long-chain n-3 polyunsaturated fatty acids and cardiovascular disease: a science advisory from the American Heart Association. *Circulation* **138**, e35–e47 (2018).
83. Kim, K. et al. Role of total, red, processed, and white meat consumption in stroke incidence and mortality: a systematic review and meta-analysis of prospective cohort studies. *J. Am. Heart Assoc.* **6**, e005983 (2017).
84. Dairy and alternatives in your diet. *NHS* https://www.nhs.uk/live-well/eat-well/milk-and-dairy-nutrition/ (2018).
85. Matthews, J. N., Altman, D. G., Campbell, M. J. & Royston, P. Analysis of serial measurements in medical research. *BMJ* **300**, 230–235 (1990).
86. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
87. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* **12**, e1004977 (2016).
88. Oksanen, J. et al. Vegan: Community ecology package. R package v.1.17-4 https://cran.r-project.org/web/packages/vegan/index.html (2010).
89. Costea, P. I. et al. Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).
90. Dhakan, D. B. et al. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience* **8**, giz004 (2019).

91. Hansen, L. B. S. et al. A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. *Nat. Commun.* **9**, 4630 (2018).
92. Jie, Z. et al. The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun.* **8**, 845 (2017).
93. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).

Acknowledgements

We thank the participants of the PREDICT 1 study. We thank N. Atabaki-Pasdar for generating the liver fat score. We thank the staff of Zoe Global, the Department of Twin Research and the Massachusetts General Hospital and all the members of the Segata, Berry and Spector laboratories for their tireless work in contributing to the running of the study, data collection and data processing. We thank Nightingale Health and Affinity Biomarker Laboratories for their support and analytical work. This work was supported by Zoe Global and received support from grants from the Wellcome Trust (no. 212904/Z/18/Z) and Medical Research Council/British Heart Foundation Ancestry and Biological Informative Markers for Stratification of Hypertension (no. MR/M016560/1). The work was also supported by the European Research Council (ERC-STG project MetaPG-716575 to N.S.), MIUR 'Futuro in Ricerca' (grant no. RBFR13EWWI_001 to N.S.), the European H2020 program (ONCOBIOME-825410 and MASTER-818368 projects to N.S.), the National Cancer Institute of the National Institutes of Health (grant no. 1U01CA230551 to N.S.) and the Premio Internazionale Lombardia e Ricerca 2019 to N.S. S.E.B. was supported in part by a grant funded by the Biotechnology and Biological Sciences Research Council (grant no. BB/NO12739/1). P.W.F. was supported in part by grants from the European Research Council (grant no. CoG-2015_681742_NASCENT), Swedish Research Council (grant no. IRC15-0067) and Novo Nordisk Foundation. A.T.C. was supported in part as a Stuart and Suzanne Steele MGH Research Scholar. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, Chronic Disease Research Foundation, Zoe Global and the National Institute for Health Research-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

Author contributions

J.W., G.H. and T.D.S. obtained the funding. S.E.B., A.M.V., J.W., G.H., H.A.K., R.D., A.T.C., N.S., P.W.F. and T.D.S. designed the study and developed the concept. S.E.B., N.S., F.A., H.A.K., A.T.C., D.A.D. and T.D.S. collected the data. F.A., S.E.B., N.S., L.F., E.L., R.G., M.M., O.M., G.P., C.L.R., M.V.-C., S.O., F.G., A.T., F.B., C.M., A.K., L.D., D.B., A.M.T., C.B., L.W., L.G., J.C.P., S.D. and R.H. analyzed the data. S.E.B., H.A.K., D.A.D., G.H., J.W. and N.S. coordinated the study. F.A., S.E.B., A.M.V., L.H.N., D.A.D., E.L., R.G., J.W., C.G., J.M.O., C.H., P.W.F., T.D.S. and N.S. wrote the manuscript. All authors reviewed and revised the final manuscript.

Competing interests

T.D.S., S.E.B., A.M.V., F.A., P.W.F., C.H. and N.S. are consultants to Zoe Global. J.W., G.H., R.D., J.C.P., C.B., R.H., L.F., F.G. and S.D. are or have been employees of Zoe Global. The other authors declare no competing interests.

Additional information

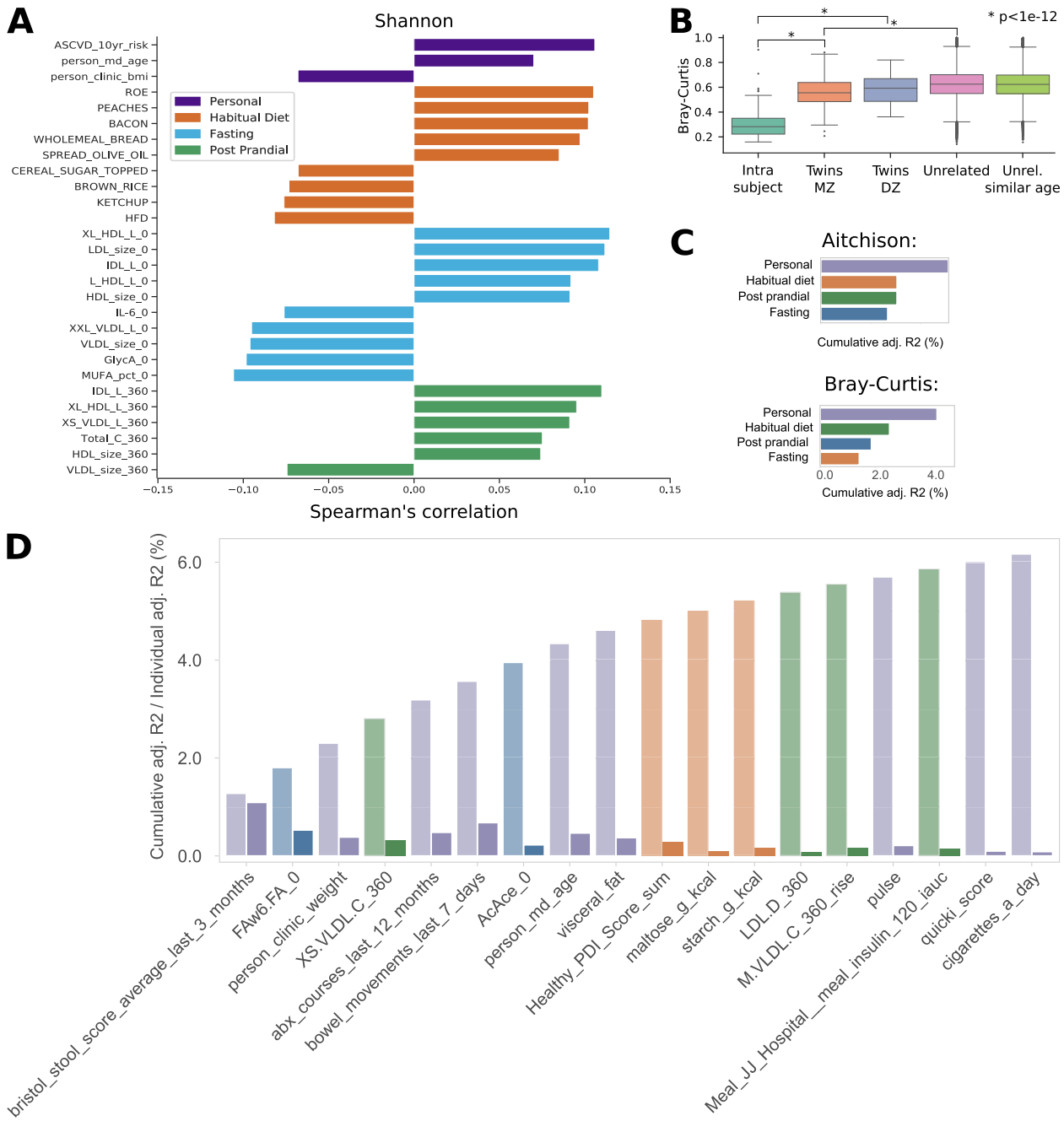
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-01183-8.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-01183-8.

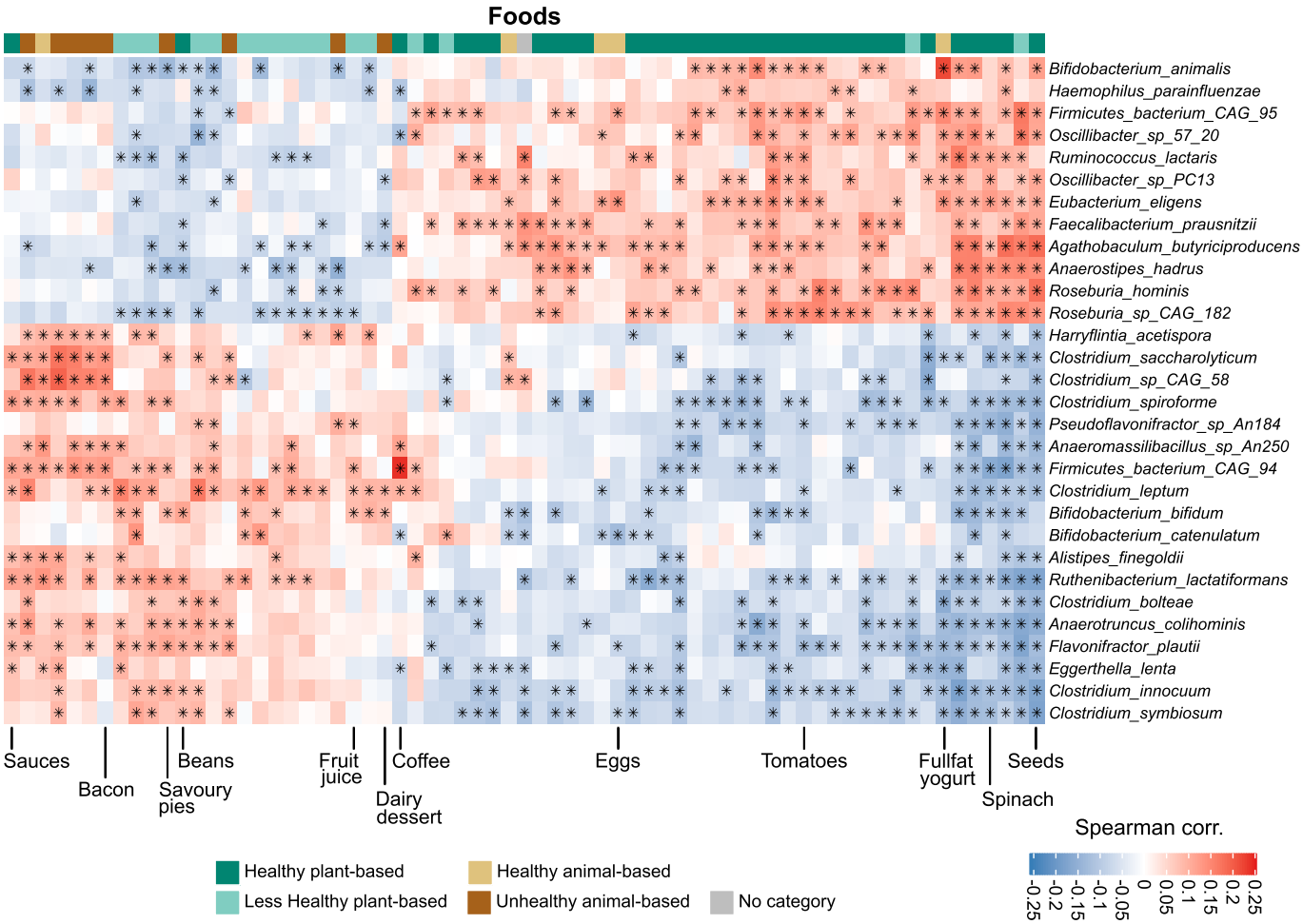
Correspondence and requests for materials should be addressed to S.E.B. or N.S.

Peer review information Jennifer Sargent was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

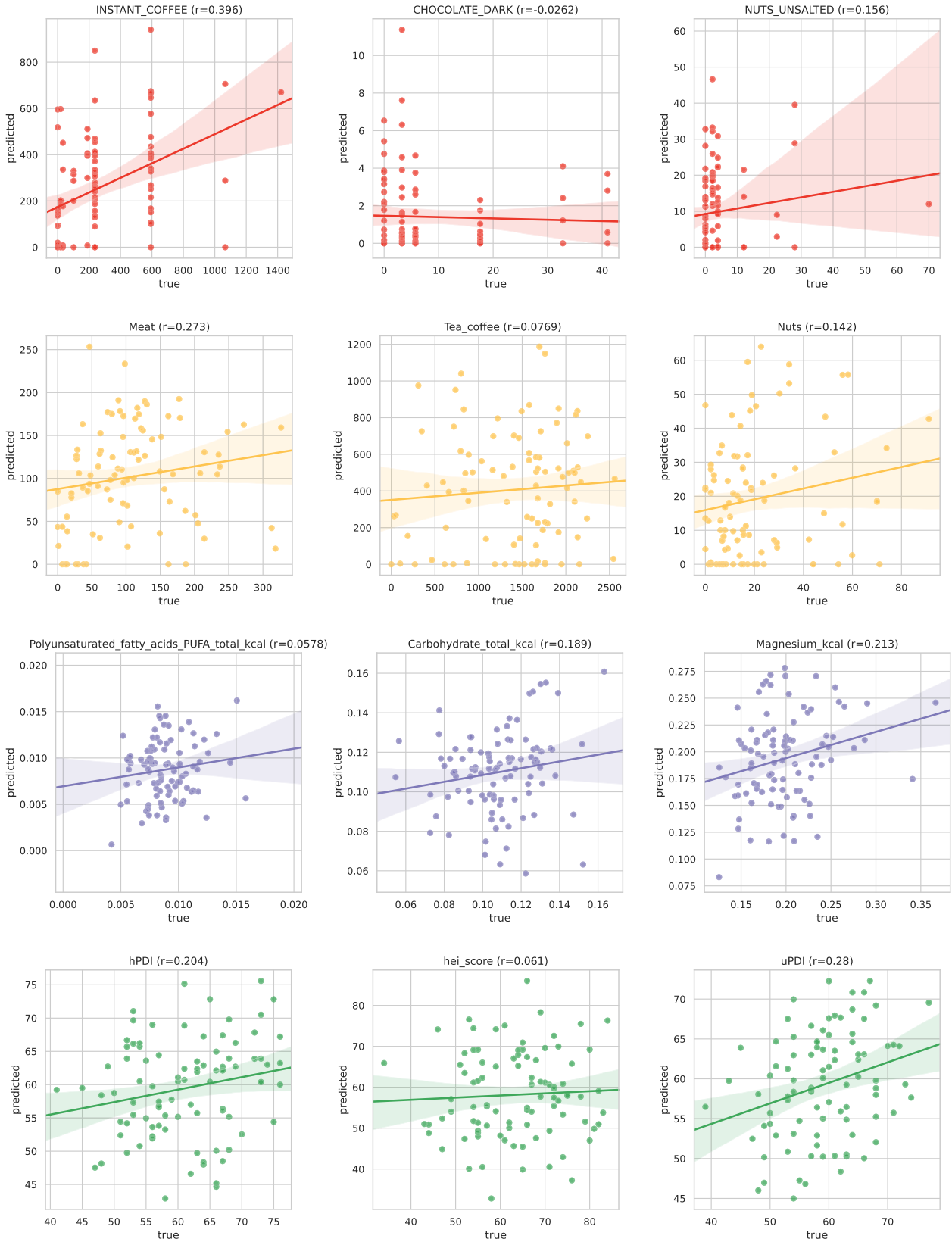
Reprints and permissions information is available at www.nature.com/reprints.



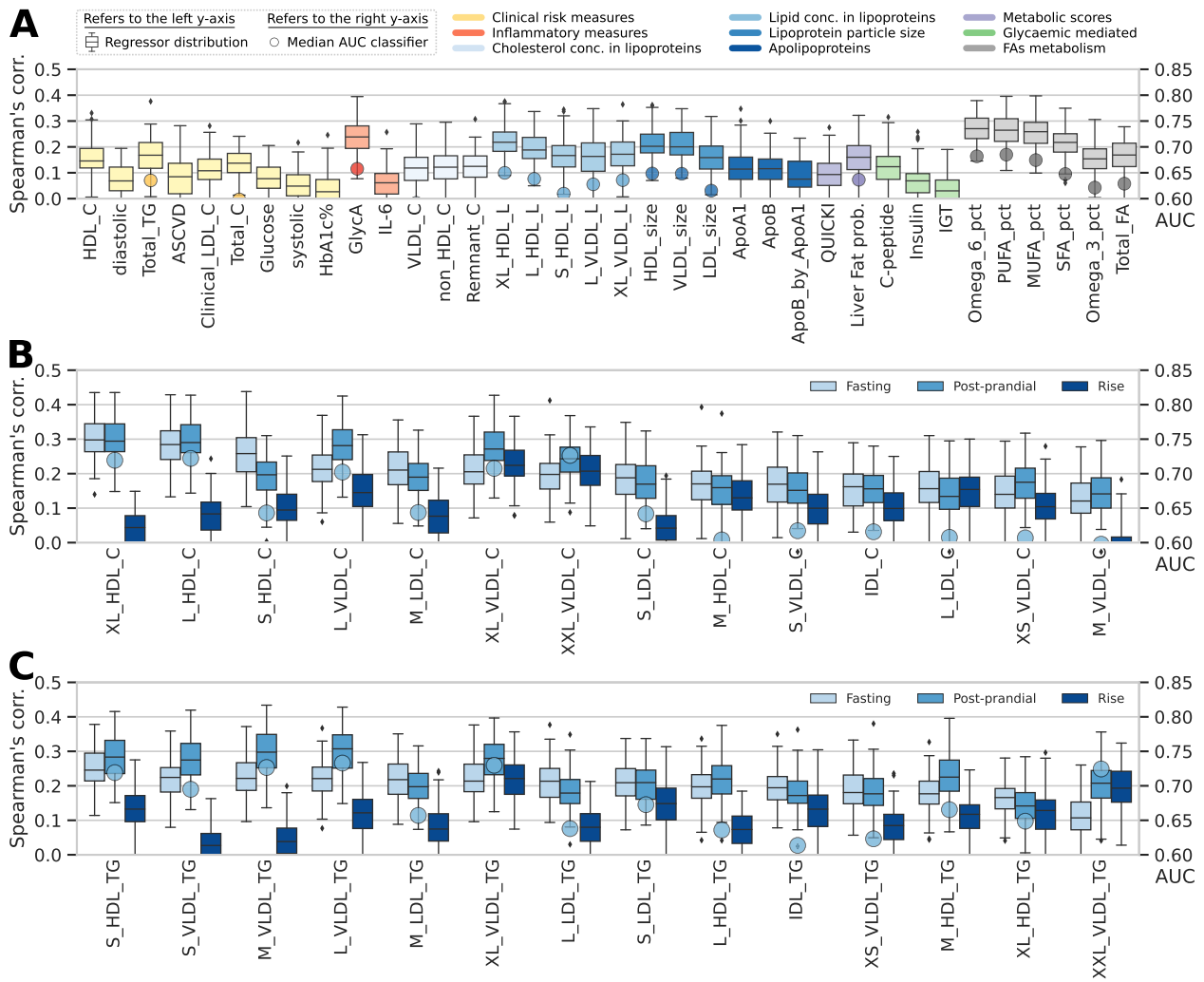
Extended Data Fig. 1 | Alpha diversity linked with personal factors, habitual diet, fasting, and postprandial markers. a, Microbiome alpha diversity computed using the Shannon index correlated markers from the four categories: personal, habitual diet, fasting, and post-prandial. Reported are the five strongest positive and negative Spearman correlations for each category with $p < 0.05$. All correlations and p-values available in the Supplementary Table 1. **b**, Inter-sample microbiome distances (beta-diversity) were substantially lower, that is closer, among samples from the same individuals (two weeks apart) compared to those amongst different individuals. Gut microbial communities in monozygotic twins were slightly more similar than in dizygotic twins (Mann–Whitney U test two-sided $p = 0.06$), which, in turn, were more similar than unrelated individuals ($p < 1e\text{-}12$), even after adjusting for age ($p < 1e\text{-}12$). **c**, After excluding twin status (that is non-twin, vs. mono vs. dizygotic twins) from the model, personal factors still accounted for the greatest proportion of variance explained in overall microbial diversity, followed by dietary habits, fasting and postprandial cardiometabolic blood markers (by cumulative stepwise dbRDA). **d**, Cumulative (left bars) contributions and individual (right bars) contributions for each metadata variable based on Bray-Curtis dissimilarity. Box plots show first and third quartiles (boxes) and the median (middle line), whiskers extends up-to 1.5× the interquartile range.



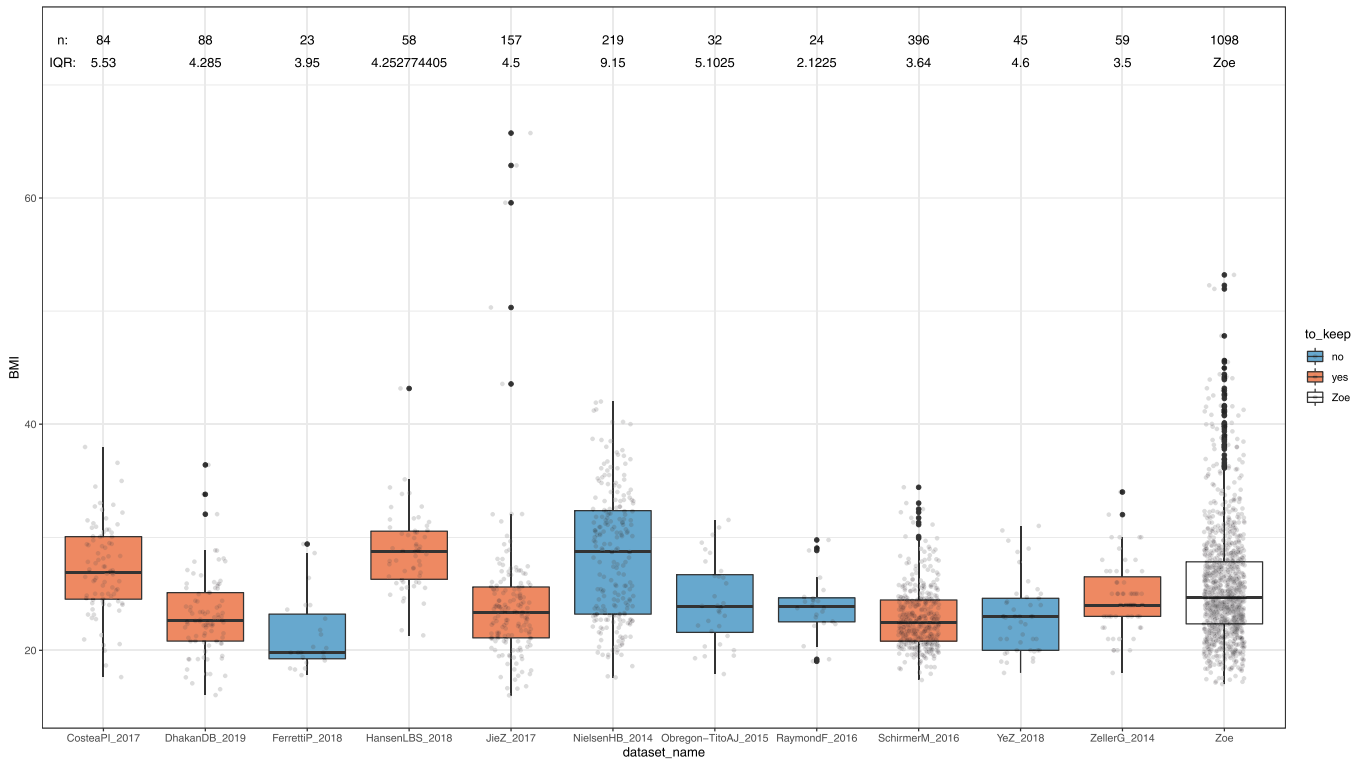
Extended Data Fig. 2 | Species-level correlation with single foods. The figure shows the species-level correlations (Spearman) with single food quantities as estimated from the food frequency questionnaires. Only foods with at least 5 significant associations (q-value≤0.2) are displayed. Species are sorted by the number of significant associations, and the top 30 are reported in the figure.



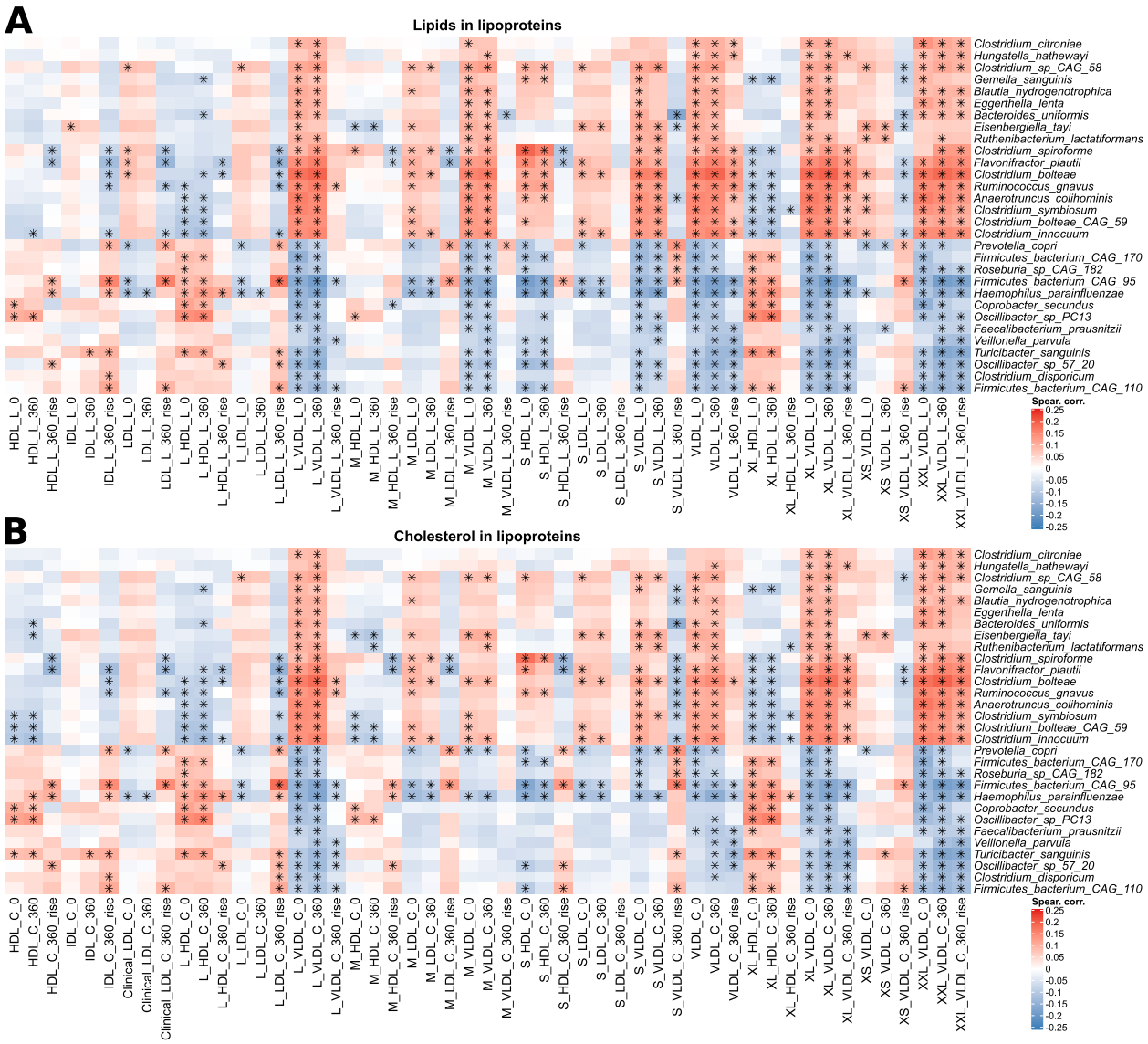
Extended Data Fig. 3 | Top foods, food groups, nutrients, and dietary patterns validated in the PREDICT 1 US cohort. The application of the RF regression model trained on the PREDICT 1 UK cohort on the PREDICT 1 US participants, validating the associations with food-related variables found in the PREDICT 1 UK.



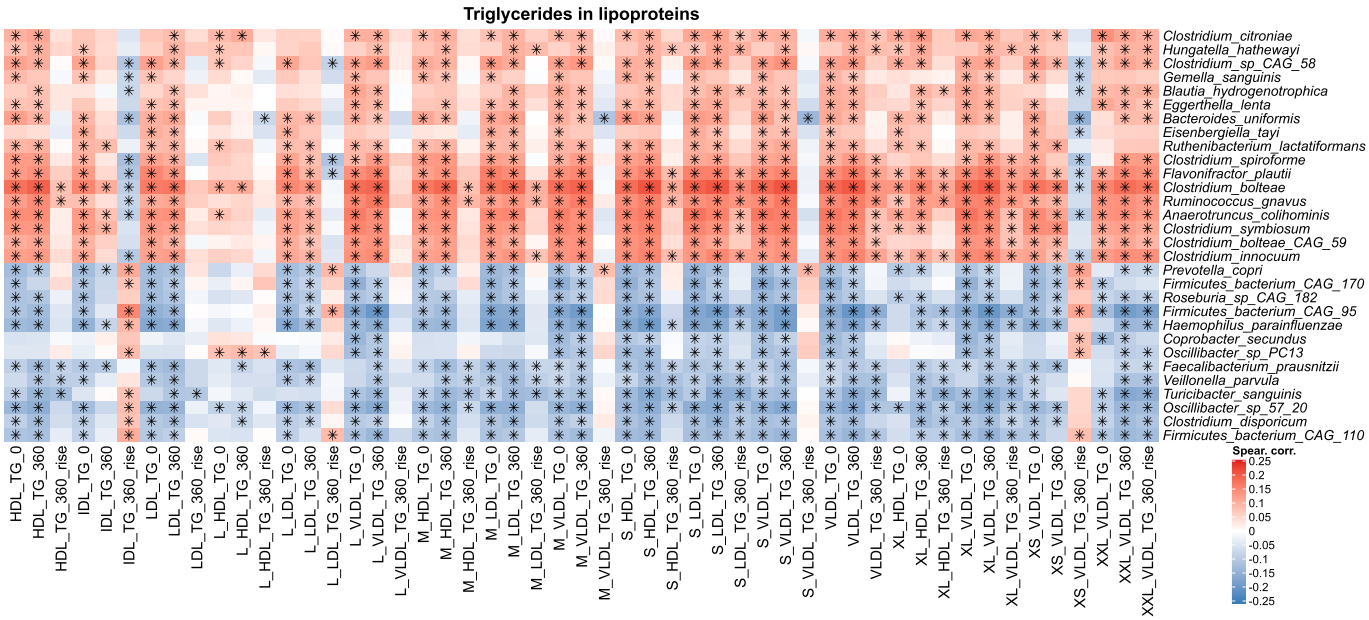
Extended Data Fig. 4 | Performance for random Forest regression and classification on microbiome functional potential in predicting fasting measurements, total cholesterol and triglycerides in different lipoproteins. The figure shows the performance of both RF regression and classification tasks trained on microbiome gene families profiles in predicting (**a**) the fasting measurements presented in Fig. 4a, sorted as in Fig. 4a. **b**, Predicting performances of the total cholesterol and (**c**) of triglycerides in different sizes of lipoproteins. For each lipoprotein, we considered its concentration values at both fasting and postprandial (6 h), and also the difference (rise) between the post-prandial concentration and the fasting one. Box plots show the distribution of the Spearman correlations (left axis) between real and predicted values using RF regression. Box plots show first and third quartiles (boxes) and the median (middle line), whiskers extends up-to 1.5× the interquartile range. Circles show the median AUC (right axis) of RF classification in predicting the bottom quartile of the distribution vs. the top quartile.



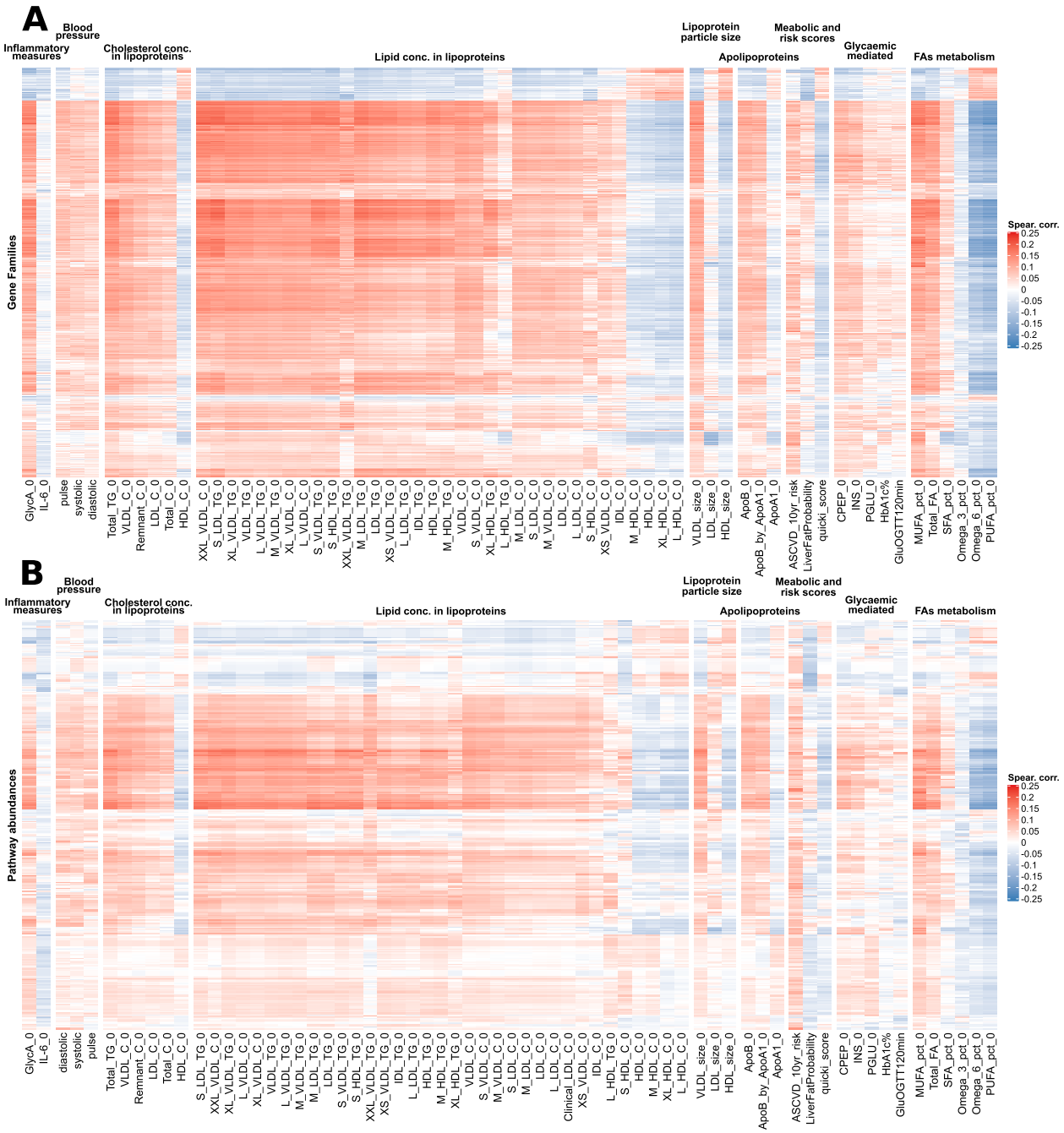
Extended Data Fig. 5 | Distributions of BMI in each curatedMetagenomicData dataset. The figure shows the distributions of BMI values for the datasets available in curatedMetagenomicData. This was used to further select those datasets with a comparable range of values (interquartile range between 3.5 and 7.5) as the one in the PREDICT 1 UK dataset (IQR of 5.5), to be used as validation datasets for the associations found. Box plots show first and third quartiles (boxes) and the median (middle line), whiskers extends up-to 1.5× the interquartile range.



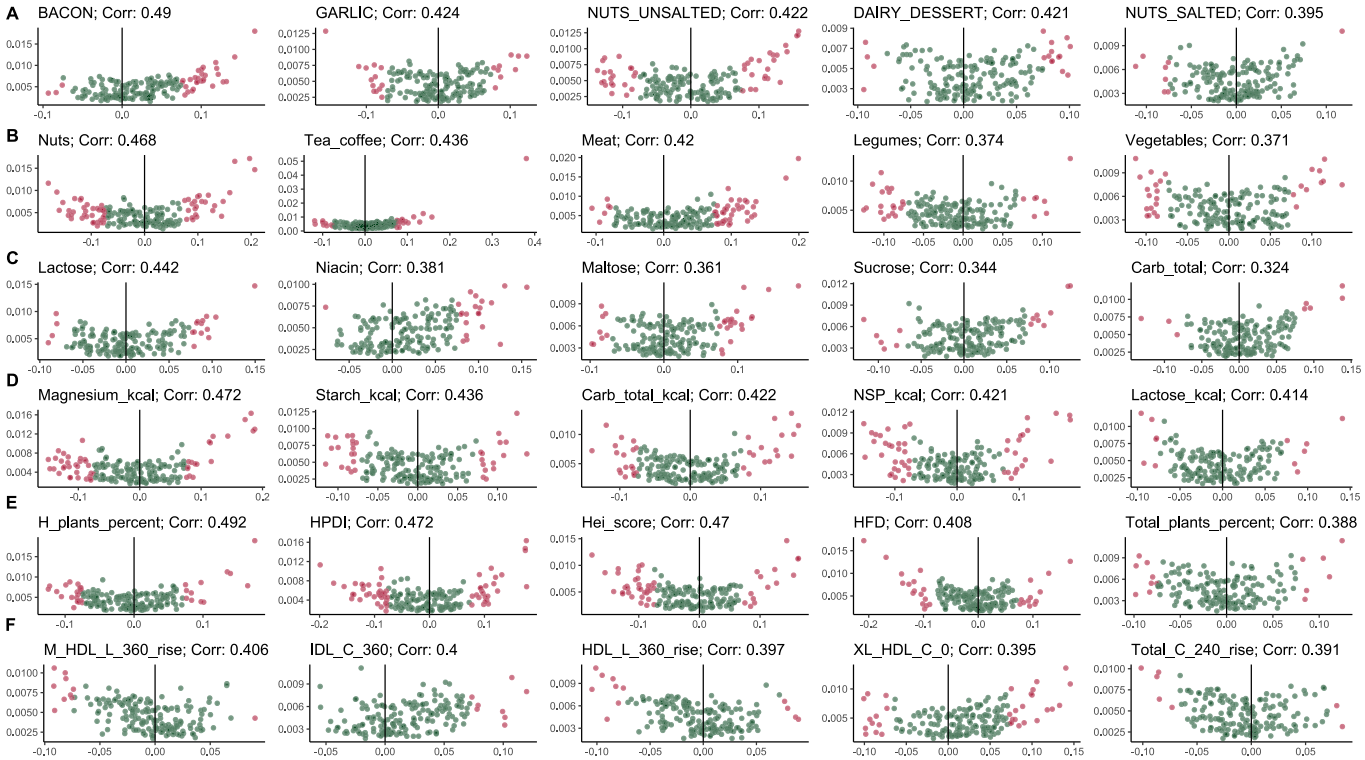
Extended Data Fig. 6 | Pairwise partial Spearman correlations between bacterial species and total lipids and cholesterol in lipoproteins. a, The heatmap shows the species-level correlations with total lipids in lipoprotein variables at fasting, post-prandial (6 h), and the difference (rise) between the postprandial and fasting concentrations. The 30 species with the highest number of significant associations (FDR ≤ 0.2) are shown. The asterisk indicates a significant correlation between species and metadata variable using a t-test two-sided, corrected with FDR with $q < 0.2$. **b**, The heatmap shows the species-level correlations with total cholesterol in lipoprotein variables at fasting, post-prandial (6 h), and the difference (rise) between the postprandial and fasting concentrations. The 30 species with the highest number of significant associations (FDR ≤ 0.2) are shown. The asterisk indicates a significant correlation between species and metadata variable using a t-test two-sided, corrected with FDR with $q < 0.2$. All correlations, p-values, and q-values are available in the Supplementary Table 6.



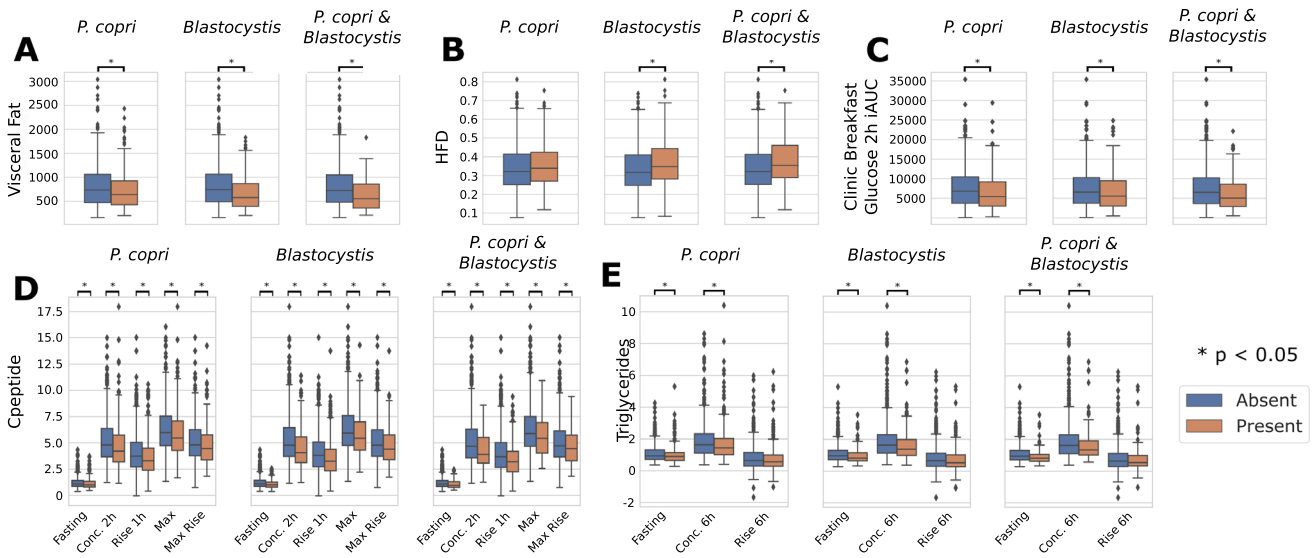
Extended Data Fig. 7 | Species-level correlations with triglycerides in lipoproteins. The heatmap shows the species-level correlations with triglycerides in lipoprotein variables at fasting, post-prandial (6 h), and the difference (rise) between the postprandial and fasting concentrations. The 30 species with the highest number of significant associations (FDR ≤ 0.2) are shown. The asterisk indicates a significant correlation between species and metadata variable using a t-test two-sided, corrected with FDR with $q < 0.2$. All correlations, p-values, and q-values are available in the Supplementary Table 6.



Extended Data Fig. 8 | Pairwise partial Spearman correlations between bacterial gene families and pathway abundances with clinical and metabolic risk scores, glycaemic and inflammatory measures, and lipoproteins. a, The heatmap shows gene families correlations with the set of metadata presented in Fig. 5a–c reporting the top 2,000 genes selected among those with at least 20% prevalence on their number of significant correlations ($q < 0.2$). Gene families' correlations are showing the same clusters as the species-level correlations in Fig. 5a–c. **b**, The heatmap shows pathway abundances correlations with the set of metadata presented in Fig. 5a–c reporting all the pathways at 20% prevalence (349 in total). Pathway abundances correlations are showing the same cluster structure as the species-level correlations in Fig. 5a–c.



Extended Data Fig. 9 | Concordance of Random Forest scores with species-level partial correlations. Volcano plots of the scores assigned to each species by Random Forest and their partial correlation, showing an overall concordance between the two independent approaches. We considered the top 5 metadata variables for the six metadata categories: **a**, Foods, bacon (g) (corr. 0.49), garlic (g) (corr. 0.424), unsalted nuts (g) (0.422), dairy dessert (g) (corr. 0.421), salted nuts (g) (corr. 0.395). **b**, Food groups, nuts (corr. 0.468), tea and coffee (corr. 0.436), meat (corr. 0.42), legumes (corr. 0.374), vegetables (corr. 0.371). **c**, Nutrients, lactose (corr. 0.442), niacin (corr. 0.381), maltose (corr. 0.361), sucrose (corr. 0.344), total carbohydrates (corr. 0.324). **d**, Nutrients normalized by daily energy intake, magnesium (corr. 0.472), starch (corr. 0.436), total carbohydrates (corr. 0.422), non-starch polysaccharides (NSP) (corr. 0.421), lactose (corr. 0.414). **e**, Dietary patterns, healthy plant percentage (corr. 0.492), healthy PDI (corr. 0.472), hei score (corr. 0.47), HFD (corr. 0.408), total plants percentage (0.388). **f**, Lipoproteins, M-HDL-L 6 h rise (corr. 0.406), IDL-C 6 h (corr. 0.4), HDL-L 6 h rise (corr. 0.397), XL-HDL-C 0 h (corr. 0.395), Total Cholesterol 4 h rise (corr. 0.391).



Extended Data Fig. 10 | *Prevotella copri* and/or *Blastocystis* presence are indicators of a more favourable postprandial glucose response to meals. **a–c**, Differential analysis of visceral fat, HFD and glucose iAUC 2 h after standardised breakfast according to presence-absence of one and both of *P. copri* and *Blastocystis*. The analysis reveals that both these species are indicators of reduced visceral fat, good cholesterol and meal-driven increase of glucose. **d**,**e**, Differential analysis of C-peptide and triglycerides at different time points according to presence-absence of one and both of *P. copri* and *Blastocystis*. The distributions of the concentrations for C-peptide and triglycerides were typically lower when one or both are absent. An asterisk between two box plots represents a significant p-value ($p < 0.05$) according to the Mann-Whitney U test (two-sided, Supplementary Table 8). Box plots show first and third quartiles (boxes) and the median (middle line), whiskers extends up-to 1.5× the interquartile range. P-values are available in Supplementary Table 8.

nature research

Corresponding author(s): Nicola Segata

Last updated by author(s): Nov 4, 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	Data from questionnaires, clinical visits and laboratory data was entered using comma delimited files, excel spreadsheets and microsoft access. CGM and accelerometer data was imported from text files into the analysis pipeline. Microbiome data was analyzed using the open source software described in the Methods.
Data analysis	Analyses were carried out using version 3.4.2 R Core Team, or with Python 3.0 using the open source libraries mentioned in the Methods. Dietary analysis was undertaken using FETA software http://www.srl.cam.ac.uk/epic/epicffq/. Microsoft Excel 365 (version 1904) was used to process spreadsheets. All open source software used in the analysis is reported in the Methods.

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

Metagenomes are deposited in EBI ENA under accession number PRJEB39223. The non-metagenomic data used for analysis in this study are held by the department of Twin Research at King's College London. The data can be released to bona fide researchers using our normal procedures overseen by the Wellcome Trust and its guidelines as part of our core funding. We receive around 100 requests per year for our datasets and have a meeting 3 times per month with independent members to assess proposals. The application is at https://twinsuk.ac.uk/resources-for-researchers/access-our-data/. This means that the data need to be anonymized and conform to GDPR standards.

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	1002 individuals were recruited for the UK cohort (allowing for 80% power to detect correlations r=0.15 with p<0.0001 to allow adjustment for 500 test), 100 for the US cohort to enable replication of the larger effects (80% for r=0.28 with p<0.05) , the final sample size excludes individuals who dropped out (didn't finish the clinic visit)
Data exclusions	Pre-established exclusion criteria were T2D. Exclusion and inclusion criteria available in "Berry, S.E., Valdes, A.M., Drew, D.A. et al. Human postprandial responses to food and potential for precision nutrition. Nat Med 26, 964–973 (2020)."
Replication	the US cohort has been used as replication for findings in the larger UK cohort
Randomization	There were 3 different test meal protocols. Meal order in each test meal protocol was randomised using Microsoft Access for each participant, using a 2-block randomisation and 1 non-randomised block. For additional details on the test meal protocols please see: https://protocolexchange.researchsquare.com/article/pex-802/v1
Blinding	Participants were blinded to the nutrient composition of test meals. Test meals were labelled with a barcode and randomization code. Laboratory analysis was performed on blinded coded samples.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Study participants were healthy individuals aged between 18-65 years and able to provide written informed consent. In total, 1002 individuals were recruited for the UK cohort (279 males and 723 females, average age 45.58 years and std 11.88 years) and 100 for the US validation cohort (32 males and 68 females, average age 41.33 years and std 12.82 years).
Recruitment	Participants enrolled already in the TwinsUK cohort were recruited as part of studies which are included in the cohort's annual newsletter and also mentioned on our website: http://www.twinsuk.ac.uk/ Non-twins were recruited via independent recruitment agencies, project specific website and online advertising including the use of social media platforms and social media. The study specifically aimed to recruit from the Twins UK cohort to enable genetic contributions to be estimated and ~30% of participants were recruited from outside of this cohort. The Twins UK cohort characteristics is representative of the UK population and we do not report any selection bias.
Ethics oversight	Ethical approval for the study was obtained in the UK from the Research Ethics Committee (REC approval18/LO/0663) and Integrated Research Application System (IRAS 236407) and in the US from the Institutional Review Board (Partners Healthcare IRB 2018P002078). The trial was registered on ClinicalTrials.gov (registration number: NCT03479866) and was run in accordance with the Declaration of Helsinki and Good Clinical Practice. Study procedures were only carried out after having received written informed consent from each participant.

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	NCT03479866
Study protocol	https://clinicaltrials.gov/ct2/show/record/NCT03479866
Data collection	Data was collected at St Thomas' Hospital, London and Massachusetts General Hospital, Boston between June 2018 and May 2019
Outcomes	Gut microbiome profile, triglyceride blood concentration, glucose blood concentration, record of sleep pattern using a wearable device (i.e. fitness watch), record of physical activity using a wearable device (i.e. fitness watch)