Microbiome connections with host metabolism and habitual diet from the PREDICT 1 metagenomic study

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

1. Department CIBIO, University of Trento, Italy
2. Dept of Nutritional Sciences, King's College London, London, UK
3. School of Medicine, University of Nottingham, Nottingham, UK
4. Nottingham NIHR Biomedical Research Centre, Nottingham UK
5. Dept of Twin Research, King's College London, London UK
6. Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
7. City University of New York, New York, NY, US
8. Zoe Global Ltd, London, UK
9. JM-USDA-HNRCA, Tufts University, Boston, MA, USA
10. Stanford University, Stanford, CA, USA
11. University of Lund, Lund, Sweden
12. Harvard T.H. Chan School of Public Health, Boston, MA, USA
13. The Broad Institute of MIT and Harvard, Cambridge, MA, USA

[^] These authors contributed equally
[*] These authors jointly supervised this work
[#] Corresponding authors: Sarah Berry (sarah.e.berry@kcl.ac.uk), Nicola Segata (nicola.segata@unitn.it)

Abstract

The gut microbiome is shaped by diet and influences host metabolism, but these links are complex and can be unique to each individual. We performed deep metagenomic sequencing of >1,100 gut microbiomes from individuals with detailed long-term diet information, as well as hundreds of fasting and same-meal postprandial cardiometabolic blood marker measurements. We found strong associations between microbes and specific nutrients, foods, food groups, and general dietary indices, driven especially by the presence and diversity of healthy and plant-based foods. Microbial biomarkers of obesity were reproducible across cohorts, and blood markers of cardiovascular disease and impaired glucose tolerance were more strongly associated with microbiome structure. While some microbes such as *Prevotella copri* and *Blastocystis* spp., were indicators of reduced postprandial glucose metabolism, several species were more directly predictive for postprandial triglycerides and C-peptide. The panel of intestinal species associated with healthy dietary habits overlapped with those associated with favourable cardiometabolic and postprandial markers, indicating our large-scale resource can potentially stratify the gut microbiome into generalizable health levels among individuals without clinically manifest disease.

Introduction

Dietary contributions to health, and particularly to long-term chronic conditions such as obesity, metabolic syndrome, and cardiac events, are of universal importance. This is especially true as obesity and associated mortality and morbidity have risen dramatically over the past decades, and continue to do so worldwide [1]. Despite intense study, the reasons for this relatively rapid change have remained unclear, with the gut microbiome implicated as one of several potentially causal human-environmental interactions [2–5]. Surprisingly, the details of the microbiome's role in obesity and cardiometabolic health have proven difficult to define reproducibly in large, diverse human populations[6] - contrary to their behaviour in mice - likely due to the complexity of habitual diets, the difficulty of measuring them at scale, and the highly personalized nature of the microbiome[7].

To overcome these challenges, we launched the Personalised Responses to Dietary Composition Trial (PREDICT 1) observational and interventional study of diet-microbiome interactions in metabolic health. PREDICT 1 included over 1,000 participants in the United Kingdom (UK) and the United States (US) who were profiled pre- and post- standardised dietary challenges using a combination of intensive in-clinic biometric and blood measures, nutritionist-administered free-living dietary recall and logging, habitual dietary data collection, continuous glucose monitoring, and stool shotgun metagenomic sequencing. The study was inspired by and generally concordant with previous large-scale diet-microbiome interaction profiles, identifying both overall gut microbiome configurations and specific microbial taxa and functions associated with postprandial glucose responses [8,9], obesity-associated biometrics such as body mass index (BMI) and adiposity [10–12], and blood lipids and inflammatory markers [13–15]. By combining PREDICT's extensive dietary and blood biomarker measures with high-precision microbiome analysis, we were also able to extend these findings to specific beneficial (e.g. *Faecalibacterium prausnitzii*) and detrimental (e.g. *Ruminococcus gnavus*) organisms, as well as to a highly-reproducible gut microbial signature of overall health that reproduced across multiple blood and dietary measures within PREDICT and in several previously published cohorts [16].

Results

Large metagenomically-profiled cohorts with rich clinical, cardiometabolic, and dietary information

We performed a multi-national, single-arm (pre-post) intervention study of diet-microbiome-cardiometabolic interactions, including a discovery cohort based in the United Kingdom (UK) and a validation population in the United States (US). The UK cohort recruited 1,002 generally healthy adults (non-twins, identical [monozygotic; MZ], and non-identical [dizygotic; DZ] twins) with detailed demographic information, quantitative habitual diet data, cardiometabolic blood biomarkers, and postprandial responses to both standardized test meals in the clinic and in free-living settings [17] (**Fig. 1A**). At-home collection of stool by our validated protocol (**Methods**) yielded 1,001 baseline samples for gut microbiome analysis. The US population employed the same enrollment and biospecimen collection protocols for 100 healthy, unrelated individuals (97 stool samples received). The data from the US cohort was analysed separately to the UK data to test the machine learning models trained in the UK cohort and independently validate microbiome-feature correlations. From a randomly-selected subset of UK participants (n=70), we additionally sequenced faecal metagenomes from a second stool sample 14 days after the first collection. All metagenomes were shotgun sequenced, taxonomically and functionally profiled, and assembled to provide metagenome-assembled genomes (MAGs, **Fig. 1A**, **Methods**). Collectively, these UK and US-based results comprise the PREDICT 1 study.

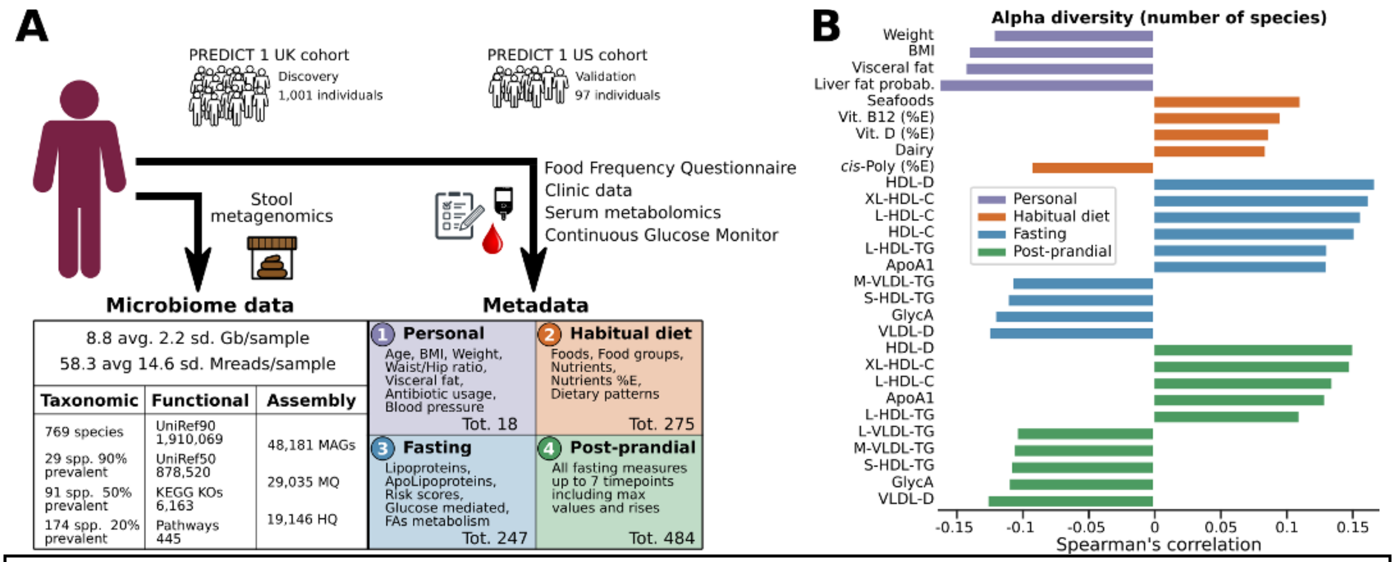

Fig. 1: The PREDICT 1 study associates gut microbiome structure with habitual diet and blood cardiometabolic markers. (A) The PREDICT 1 study assessed the gut microbiome of 1,098 volunteers from the UK and US via metagenomic sequencing of stool samples. Phenotypic data obtained through in-person assessment, 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 semi-quantitative food frequency questionnaires (FFQs); (3) fasting; and (4) postprandial cardiometabolic blood and inflammatory markers, total lipid and lipoprotein concentrations, lipoprotein particle sizes, apolipoproteins, derived metabolic risk scores, glycaemic-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 HDL-D (positive) and estimated hepatic steatosis (negative). Up to ten strongest absolute Spearman correlations are reported for each category with q<0.05. Top species based on Shannon diversity are reported in **Supplementary Fig. 1A** and all correlations are in **Supplementary Table 1**.

Microbial diversity and composition are linked with diet and fasting and postprandial biomarkers

We first leveraged a unique subpopulation of our study comprised of 480 twins to disentangle the confounding effects of shared genetics from other factors on microbiome composition. Our data confirmed that host genetics influences microbiome composition only to a small extent [18], as intra-twin pair microbiome similarities were significantly greater than those among unrelated individuals (p<1e-12, **Supplementary Fig. 1B**), and monozygotic twins showed slightly more similar microbiomes than dizygotic twins (p=0.06). Intra twin-pair microbiome similarity, regardless of zygosity, remained substantially lower than intra-subject longitudinal sampling (day 0 *vs.* day 14, p<1e-12, **Supplementary Fig. 1B**), a testament to the highly personalized nature of the gut microbiome attributable to a variable extent to non-genetic factors (**Supplementary Fig. 1C,D**).

We initially investigated overall intra-sample (alpha) diversity of the gut microbiome as a broad summary statistic of microbiome structure [19]. In our cohort of healthy individuals, we found links between alpha diversity (specifically species richness) and personal characteristics (e.g. age and anthropometry), habitual diet, and metabolic indices (**Fig. 1B**) with 109 significant associations (p<0.05) among the total 295 Spearman's correlation tests, and 56 after FDR-correction (q<0.05, **Supplementary Table 1A**). Participant BMI, absorptiometry-based visceral fat measurements, and probability of fatty liver (using a validated prediction model [20]) were inversely associated with species richness. Consistent with previous findings for BMI [5,21], our findings suggest that the link between the microbiome and body habitus may be mediated in part by hepatic insulin resistance, particularly given the gut microbiome's strong association with liver disease and activity observed in this cohort and previously [22]. With respect to habitual dietary factors, we found 18 of 126 total nominally significant (p<0.05) correlations (5 at q<0.05, **Fig. 1B**).

Among clinical circulating measures, HDL cholesterol (HDL-C) was positively correlated with species richness. However, emerging cardiometabolic biomarkers with strong associations with cardiometabolic diseases [23–26] that are not routinely used clinically, including lipoprotein particle size (diameter, "-D"), lipoprotein composition (cholesterol "-C" and TG "-TG"), apo-lipoproteins and GlycA (inflammatory biomarker; glycoprotein acetyls), were even more strongly associated with richness than the remaining traditional clinical measures (TG, Total-C, LDL-C and fasting glucose). These emerging biomarkers of reduced risk of chronic disease were positively associated with microbial diversity (e.g. extra large and large HDL-C, HDL-D, Apolipoprotein-A1) both at fasting and postprandially, whilst those associated with increased risk of chronic disease were inversely correlated with microbial diversity (e.g. GlycA, VLDL-D small-HDL-TG). These results for species richness provide initial evidence that the microbiome is modestly, but significantly, associated with some key classical and emerging cardiometabolic health indicators and diet, motivating more detailed investigations of the links between cardiometabolic health, diet, and specific gut microbiome components.

Diversity of healthy plant-based foods in habitual diet shapes gut microbiome composition

We assessed links between habitual diet (over the past year) and the microbiome in PREDICT 1 using detailed, validated semi-quantitative food frequency questionnaires (FFQs). These links were quantified using random forest (RF) regression and classification models, each trained on the whole set of quantitative microbiome features to predict one habitual diet feature (with training/testing via repeated bootstrapping, **Methods**). The performance of the models was evaluated with ROC AUCs for classification and with correlation between predicted and collected values for regression, thus quantifying the degree to which each dietary feature could be estimated based on microbiome composition.

Dietary features assessed in this manner included individual food items, food groups, nutrients (energy adjusted and non-adjusted), and dietary patterns (**Fig. 2**). We assessed individual foods and food groups, the latter after collapsing items into bins according to Plant-based Diet Index (PDI) [27] groupings (**Supplementary Table 2**). Several foods and food groups exceeded 0.15 median Spearman's correlation over bootstrap folds (denoted as "ρ") between predicted and FFQ-estimated values (20/165 or 12.1%) and AUC>0.65 (14/165, 8.5%; **Fig. 2A**). The strongest association among food items was coffee (ρ=0.45), which appeared to be dose-dependent (**Fig. 2D**) and validated in the US cohort when the model trained in the UK cohort was applied in the US (**Fig. 2E**). We found particularly tight coupling between energy-adjusted derived nutrients and the taxonomic composition of the microbiome, especially compared to foods and food groups (**Fig. 2A**). Almost one-third of the energy-normalized nutrients (**Supplementary Table 2**) had correlations above 0.3 (14/47) with the highest correlations achieved for saturated fatty acids (SFAs, ρ=0.46, AUC 0.82), zinc (ρ=0.39, AUC 0.76), and starch (ρ=0.39, AUC 0.75).

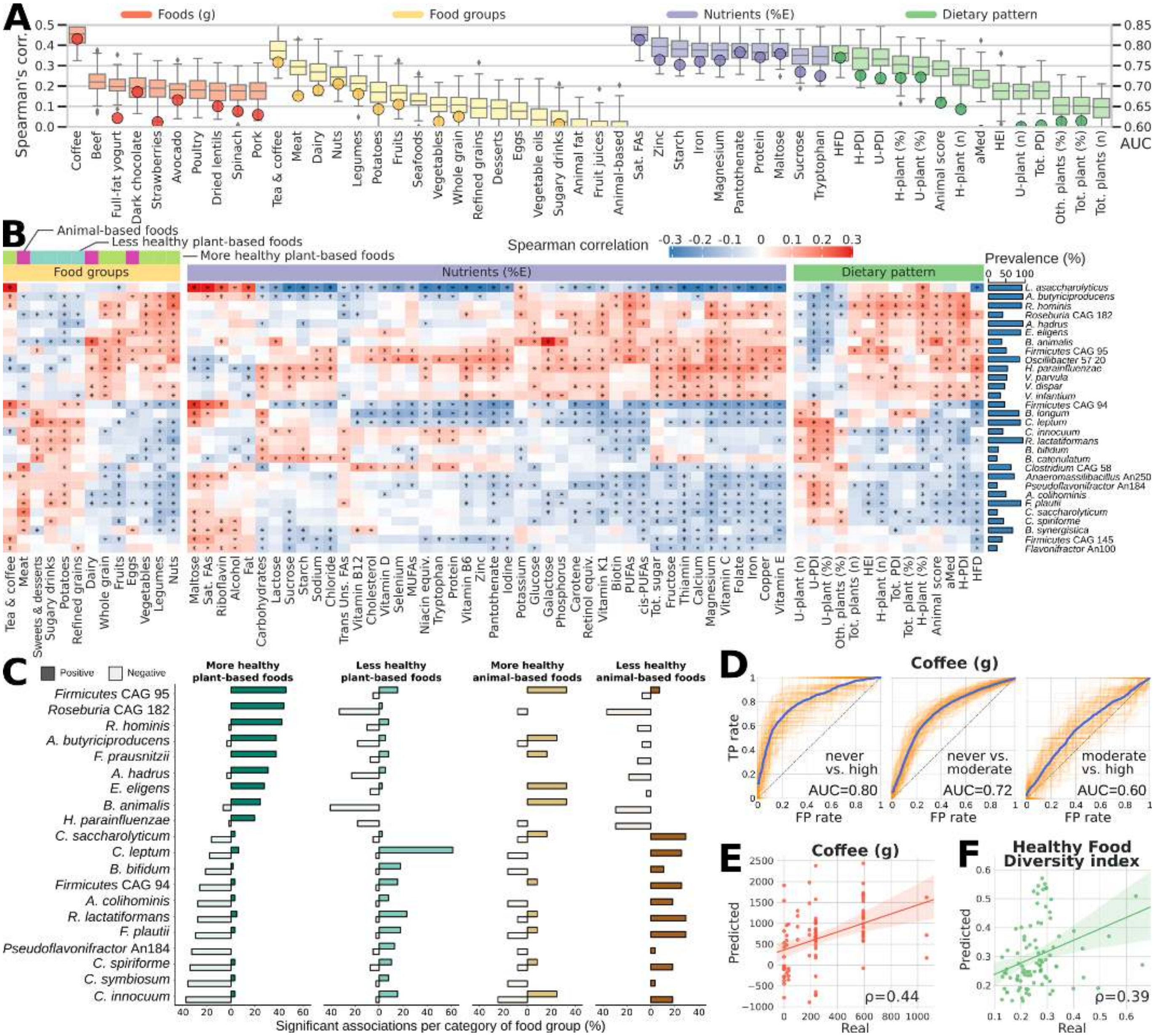

Fig. 2: Food quality, regardless of source, is linked to overall and feature-level composition of the gut microbiome. (A) Specific components of habitual diet comprising foods, nutrients, and dietary indices are linked to the composition of the gut microbiome with variable strengths as estimated by machine learning regression and classification models. Boxplots report the correlation between the real value of each component and the value predicted by regression models across 100 training/testing folds (**Methods**). Circles denote median area-under-the-curve (AUC) values across 100 folds for a corresponding binary classifier between the highest and lowest quartiles (**Methods**). **(B)** Single Spearman correlations adjusted for BMI and age between microbial species and components of habitual diet with 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. Full heatmaps of foods and unadjusted nutrients are reported in **Supplementary Fig. 2**, and the full set of correlations is available in **Supplementary Table 4**. **(C)** Number of significant positive and negative associations (Spearman's 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. 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, i.e. stronger when assessing heavy (e.g. >4 cups/d) vs. never drinkers, **(E)** and was validated in the US cohort when applying the UK model across cohorts. **(F)** Among general dietary patterns and indices, the Healthy Food Diversity index (HFD) was validated in the US cohort when using the UK model, thus showing consistency between the two populations on this important dietary index. The other indices are also generally validated as reported in **Supplementary Fig. 3**.

Because of the complex and interacting nature of dietary intake, as well as to offer practical recommendations, we then summarized constituent foods and food groups into several established dietary indices (**Supplementary Table 2**), including the Healthy Food Diversity index (HFD) [28], the Healthy and Unhealthy Plant-based Dietary Indices (H-PDI and U-PDI), and the Alternate Mediterranean Diet score (aMED) [29]. The HFD, unlike the other food scores, incorporates a measure of dietary diversity (greater is considered better) and food quality according to dietary guidelines, whereas the PDI characterizes a given diet on the basis of type and quantity of the plant-based foods categorized as 'more-healthy'/'healthy' or 'less-healthy'/'unhealthy' based on epidemiological evidence [27]**.** These scores have been associated with lower cardiovascular disease risk [29], T2D risk [27], metabolic syndrome [30], and all-cause mortality [31]. The aMED dietary score is based on dietary patterns in Mediterranean countries and has been associated with reduced risk of chronic disease and mortality [32,33]. We demonstrated tight correlations between values predicted from gut microbial composition and all the indices (HFD, H-PDI, U-PDI, and aMED) in the UK (ρ=0.36, 0.34, 0.33, and 0.23, respectively) and in the US validation cohort (ρ=0.39, 0.23, 0.31, and 0.38, respectively; **Fig. 2A** and **Supplementary Fig. 3**), highlighting the relationship between the microbiome and healthy dietary patterns. Additionally, these results indicate that diet-microbiome associations are consistent and generalizable from UK to US populations, adding confidence to the suggested biological targets explored below and alleviating concerns of overfitting.

Microbial species segregate into groups associated with more healthy and less healthy plant- and animal-based foods

We proceeded to undertake feature-level testing to identify the specific microbial taxa most responsible for these diet-based community associations (**Fig. 2B**). By focusing on prevalent species (i.e., those detected in >20% of samples) and adjusting for age and BMI, we found that 30 species (17%) were significantly correlated with at least five defined dietary exposures at False Discovery Rate (FDR) q<0.2 (**Supplementary Table 4**). This included a confirmation of expected associations (**Supplementary Fig. 2**), such as the relative enrichment of the probiotic taxa *Bifidobacterium animalis* [34] and *Streptococcus thermophilus* with greater full-fat yogurt consumption (ρ=0.22 and 0.20 respectively). The strongest food/microbe association was between the recently characterized butyrate-producing *Lawsonibacter asaccharolyticus* [35] and coffee consumption (**Fig. 2B**). However, due to the low precision of dietary data collected by FFQ, the complexity of dietary patterns, nutrient-nutrient interactions, and clustering of 'healthy'/'less-healthy' food items within diets, it is challenging to disentangle the independent associations of single nutrients and single foods with microbial species. Indeed, considering the top 30 species most strongly associated with various dietary determinants (based on number of significant correlations; **Fig. 2B**), we found a clear segregation of species into two distinct clusters with either more healthy plant-based foods (e.g. spinach, seeds, tomatoes, broccoli) or with less healthy plant-based (e.g. juices, sweetened beverages, and refined grains) and animal-based foods, as defined by the PDI [36] (**Supplementary Table 4**).

Taxa linked to diets rich in more healthy plant-based foods (**Fig. 2B,C** and **Supplementary Fig. 2**) mostly included butyrate producers, such as *Roseburia hominis*, *Agathobaculum butyriciproducens*, *Faecalibacterium prausnitzii*, and *Anaerostipes hadrus*, as well as other uncultivated species from clades typically capable of butyrate production (*Roseburia* CAG 182) or predicted to have this metabolic capability (*Firmicutes* CAG 95, with 92% of its 166 MAGs encoding for butyrate kinases). Clades correlating with several 'less-healthy' plant-based and animal-based foods included several *Clostridium* species (*Clostridium innocuum*, *C. symbiosum*, *C. spiroforme*, *C. leptum*, *C. saccharolyticum*). The relationship between *C. leptum* and the intake of unhealthy foods is particularly worth noting, as prior experimental evidence has demonstrated their counts can be modulated by diet in mice [37]. The segregation of species according to animal-based 'healthy' foods (e.g. eggs, white and oily fish) or animal-based 'less-healthy' foods (e.g. meat pies, bacon and dairy desserts) using a novel categorisation developed for this analysis based on epidemiological evidence outlined in **Methods**, was also distinct and was similar to taxa linked to patterns for 'healthy' and 'less-healthy' plant foods (**Fig. 2C** and **Supplementary Fig. 2**). The few food items that did not fit into the 'healthy' cluster despite being categorised as 'healthy plant' foods, were (ultra) processed foods according to the NOVA classification [38] (e.g. sauces, tomato ketchup, and baked beans; Group 4 and 3, respectively; **Supplementary Fig. 2**). This emphasises the importance of food quality (e.g. highly processed vs. unprocessed), food source (e.g. plant vs. animal), and food heterogeneity (i.e. not all plant foods are healthy and animal foods unhealthy, nor vice versa) both in overall health and in microbiome ecology.

Poorly characterized microbes drive the strongest microbiome-habitual diet associations

Many of the strongest microbial associations with food items, food groups, and dietary indices occurred with only recently-isolated organisms or still uncultured taxa including, for example, five species defined using co-abundance gene groups (CAGs) from metagenomics [39]. Among indices, the HFD, which prioritizes diversity of all food items while considering dietary guidelines, was most tightly coupled to feature-level abundances (**Fig. 2A**), significantly correlated with 41 of the 174 prevalent species (i.e. those found in >20% samples), highlighting the synergistic impact of dietary diversity, dietary quality, and gut microbial responsiveness. Among species whose abundance was highly correlated to the HFD (**Fig. 2B**) were taxa also associated with 'healthy' or 'less-healthy' foods, such as *Firmicutes* CAG 94 (ρ=-0.25) and *Roseburia* CAG 182 (ρ=0.13). The highest correlation was observed for *Lawsonibacter asaccharolyticus* (ρ=-0.29), the aforementioned and recently characterized [35] and sequenced species [40]. This microbe has two additional known genomes with the conflicting species name of *Clostridium phoceensis* [41], and we predicted that it encodes butyrate-producing enzymes from metagenome-assembled genomes enzymes [42] (49 of the 53 MAGs in the *L. asaccharolyticus* SGB15154 encode for butyrate kinase EC 2.7.2.7). The link between the HFD and *L. asaccharolyticus* is particularly noteworthy and not likely a consequence of our previously observed association with coffee, as the HFD index does not include non-caloric beverages, including coffee, mineral water, and tea, as well as alcoholic beverages. This may suggest alternative and complementary strategies to modulate this microbe through both coffee intake and adherence to a diverse diet.

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, which is derived based on the relative amount of 'healthy' (positive score) and 'less-healthy' (inverse score) animal foods consumed (**Supplementary Table 4**), was associated with the 'healthy' cluster, suggesting that a diet rich in healthier animal-based foods is associated with the more favourable diet-microbiome signature, although this likely also reflects an overall healthier dietary pattern by healthy animal-based food consumers. The healthy and unhealthy PDI, which have been shown to differentially affect disease risk [27,36] also had distinct clusters, again emphasising the oversimplification of conventional plant and animal-based food groupings. The strongest representatives for the two clusters (i.e. taxa with the highest correlations) are *Firmicutes* CAG 95 and *Firmicutes* CAG 94 for healthy and unhealthy diet, respectively, and the lack of cultivated representatives for these two candidate species may explain why these links were previously overlooked even in large analyses [8,11]. The PREDICT 1 validation cohort in the US generally confirmed these associations despite its comparatively smaller sample size: among the subset of derived pattern/index scores shared between the UK and US cohorts, of the 52 associations that were significant both in the UK cohort (FDR q<0.2) and in the US cohort (p<0.05), 78.8% were concordant for the direction of the correlation.

Microbial indicators of obesity are reproducible across varied populations

Microbiome links to obesity have attracted much interest although results have varied in human populations [5,6]. We thus explored them in the PREDICT 1 populations with RF regression and classification (as above, **Methods**) using either taxonomic or functional features. We found visceral fat measured by DEXA scan to be more strongly linked to gut microbial composition than BMI [43], a finding we validated in our US participants when applying UK-trained models (**Fig. 3A**). Some obesity-associated taxa—assessed either by BMI or visceral fat—were also associated with poor dietary patterns after controlling for BMI (e.g. *Clostridium* CAG 58, *Flavonifractor plautii*), whereas markers of healthier low visceral fat mass (e.g. *Faecalibacterium prausnitzii*) were more strongly linked to healthier foods and patterns of intake, illustrating that diet and obesity signatures overlap but are not identical (**Fig. 3B**).

Microbiome models to predict BMI developed and trained on the UK-based cohort were validated not only in the PREDICT US cohort, but also in six additional independent datasets [13,44–48] that have been uniformly pre-processed and harmonized using *curatedMetagenomicData* [16] (cMD), lending credence and generalizability to our findings (**Methods**). Despite substantial differences [49,50] in the microbiomes among people from different populations, the PREDICT 1 UK model improved cohort-specific cross-validation accuracy in the majority of cases, on par with the leave-one-out approach that notably also includes the UK cohort (**Fig. 3D**). Interestingly, BMI was not predictable at all for two included datasets when using just their own samples. However, predictions and classification improved when using the PREDICT 1 UK model. Of the 17 species surpassing our FDR threshold of $q<0.05$, three had an (absolute) $\rho>0.1$ in the smaller US cohort and two of these three were concordant with those in the UK cohort (*I. butyriciproducens* negatively and *R. torques* positively correlated with BMI; **Fig. 3C**). Across our harmonized independent cMD datasets, all but two median association estimates were consistent with the PREDICT 1 UK signatures, and 12 of the 14 were concordant despite different sample collection and DNA extraction methods.

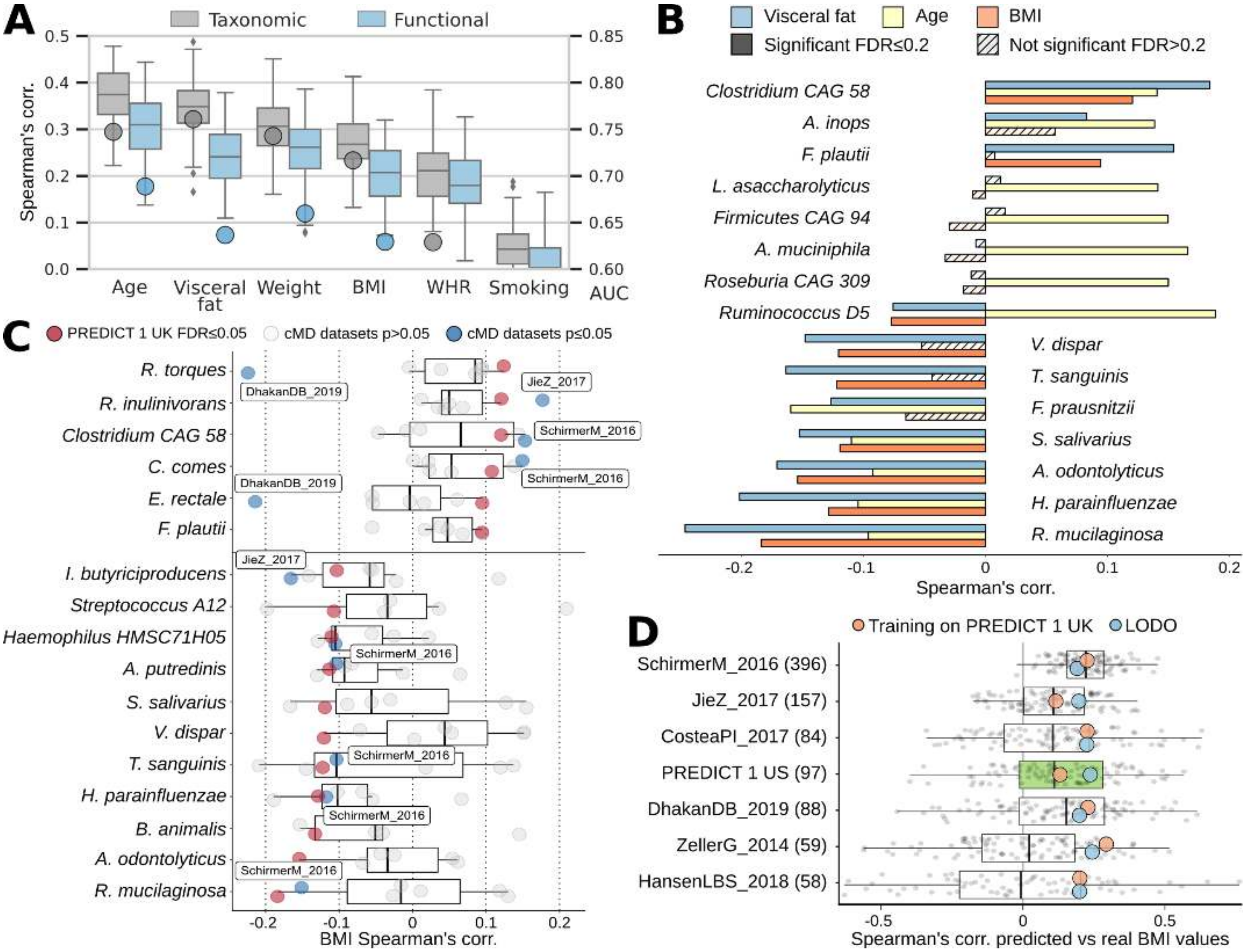

Fig. 3: Random forest machine learning models based on microbial or functional profiles are capable of predicting obesity phenotypic markers, even when tested against separate, independent cohorts. (A) Whole-microbiome machine learning models can assess personal factors with RF regression (boxplots and left-side vertical axis) using only taxonomic or functional (i.e. pathway) microbiome features. Classification models (circles and right-side vertical axis) exceed AUC 0.65 except for waist-to-hip ratio (WHR) 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 [16] we confirmed correlations between single microbial species and BMI with blue points denoting significant associations at p<0.05. **(D)** The machine learning model for BMI trained on PREDICT 1 data is reproducible in several external datasets (**Supplementary Fig. 5**), achieving correlations with true values exceeding those obtained in cross-validation of a single given dataset in five of seven cases. When the PREDICT 1 microbiome model is expanded to include other datasets (excluding those ones used for testing, i.e. leave-one-dataset-out/LODO approach) the performance remains comparable, affirming the generalizability of the PREDICT 1 model on obesity-related indicators.

Fasting cardiometabolic markers associated with specific microbiome structures

To explore the connections between the gut microbiome and markers of cardiometabolic health, we performed fine-scale evaluations of microbial community membership and their biochemical functions against established clinical and emerging cardiometabolic biomarkers. We developed ML prediction models for each of these outcomes built using both species-level taxonomic abundances and functional potential profiles and tested how accurately they were able to estimate host biomarkers.

We found modest concordance between microbiome classifiers and several traditional clinical fasting cardiometabolic biomarkers (**Fig. 4A**). These include near-term metrics, such as systolic and diastolic blood pressure, heart rate, lipids (TG, TC, HDL-C, LDL-C) and fasting glucose, as well as glycosylated haemoglobin (HbA1c), a widely-used clinical test reflecting mean glucose levels over weeks-to-months. Notably, the difference between total and high-density lipoprotein (HDL) cholesterol (e.g. non-HDL), recently considered a clinically useful aggregate count of atherogenic cholesterol fractions [51], was also linked to gut microbial features (ρ=0.17; AUC 0.61). These associations were largely recapitulated in a clinical prediction model incorporating most of these factors to estimate latent 10-year risk of heart disease or stroke using the AtheroSclerotic CardioVascular Disease (ASCVD) algorithm [52].

From our remaining compendium of blood biomarkers (**Fig. 1A**), we found stronger correlations between the microbiome and an inflammatory surrogate (glycoprotein acetyls, GlycA, **Fig. 4A**), as well as various emerging lipid measures linked to host health, such as HDL and VLDL particle size (HDL-D and VLDL-D, ρ=0.3 and 0.28 respectively), the lipid content of lipoprotein subfractions (including XL-HDL-L and L-HDL-L, ρ=0.39 and 0.37 respectively), and circulating polyunsaturated fatty acids (PUFA) fatty acid (omega-6 [FAω6/FA] and PUFA [PUFA/FA] to total fatty acid ratios, ρ=0.31 for both). GlycA [26] and VLDL-D have been strongly associated with increased risk for the metabolic syndrome, CVD, and T2D, whereas HDL-D and its lipid constituents, omega-6, and PUFA have strong inverse associations [23,24,53]. The strongest association for all circulating markers was observed for large HDL particle lipid concentrations (XL-HDL-L and L-HDL-L, with ρ=0.41 and 0.38, and AUC=0.70 and 0.69, respectively), which also have the strongest inverse association with CVD and T2D of all the lipid measures [23,24,53]. Similarly, the majority of glycaemic indicators such as insulin, C-peptide (a surrogate of insulin secretion), and to a much lesser extent, impaired glucose tolerance (IGT) were also coupled to human gut microbiome composition (**Fig. 4A,B**). Derived predictors of insulin sensitivity (QUantitative Insulin sensitivity Check Index or QUICKI) [54] and hepatic steatosis (Liver Fat Probability) were also reasonably captured using microbiome-based ML classifiers (ρ=0.22 and 0.18; AUC 0.66 and 0.64 respectively).

Species-based predictors proved more accurate for RF-based learning tasks than pathway abundance profiles (**Supplementary Fig. 4**), consistent with other microbiome-wide training exercises [55]. Despite a smaller study population and a more restricted panel of fasting circulating metabolites, our primary findings were generally replicated in the US validation cohort (**Fig. 4A**), corroborating the existence of a strong, previously overlooked link between the gut microbiome and surrogate markers of cardiometabolic health.

The gut microbiome is a better predictor of postprandial triglycerides and insulin concentrations than of glucose levels

Fasting blood assays are the standard for most research and clinical investigations; however, in free-living conditions, individuals consume multiple meals throughout the day and therefore spend most of their waking hours in the postprandial state. Mixed nutrient meals (carbohydrate, fat and protein) result in person-specific food-induced elevations in triglycerides (TG), glucose, insulin, and other related metabolites, impacting personalized cardiometabolic responses and downstream health outcomes [56]. Whilst prior efforts have demonstrated that postprandial glucose responses may, in part, be predicted by the gut microbiome [8], the relationship between the microbiome and ‘real-life’ variations in both postprandial lipid and glucose-mediated metabolites has not been explored. We therefore assessed postprandial metabolic responses to foods of varying nutrient composition in the clinic and free-living settings by considering the overall magnitude of the response by iAUC, as well as its peak concentrations, and its change from fasting (i.e. rise).

Firstly, we measured postprandial TGs, glucose, C-peptide, insulin, and circulating metabolite concentrations at regular intervals (0-6h) in the clinic after the administration of two formulated, sequential test meals (890 kcal, 50g fat and 85g carb at 0h [breakfast] and 500 kcal, 22g fat and 71g carb at 4h [lunch]; **Fig. 4D,E**). Notably, we found that the magnitude of postprandial TG (0-6h iAUC), insulin, and C-peptide (both 0-2h iAUC) responses were more strongly associated with the gut microbiome (ρ=0.15, 0.19, and 0.21, respectively; AUC >0.63 for each) compared with postprandial glucose (0-2h iAUC) responses (ρ=0.12 and AUC 0.59, **Fig. 4D**), findings replicated in our US validation cohort (**Fig. 4D**).

Following the in-person clinic day, we also measured glucose concentrations via continuous glucose monitoring over the subsequent 13-day at-home period [17] that included responses to isocaloric standardized meals, in duplicate, with different macronutrient compositions (fat, carbohydrate, protein and fibre; **Supplementary Table 3**). However, contrary to our clinic meal responses (**Fig. 4D**) and previous work [8], the glucose 0-2h iAUCs following these meals did not achieve high correlations with the microbiome regardless of their macronutrient composition (all ρ<0.11 and AUC<0.58, **Fig. 4E**). Whilst this may be due to the lower energy, fat, and carbohydrate dose in at-home isocaloric meals (500 kcal) compared to our successive clinic meals (total 1,390 kcal for breakfast and lunch), reducing discrimination between interindividual responses, Zeevi et al [8] found associations using meals of <500kcal. However, the stool sample in our study was collected within 24h of the metabolic clinic meal(s), whereas the standardised at-home meals were consumed (in random order) between days 2-13 post-home stool collection, introducing additional variability due to short-term fluctuations in microbiome composition [57]. Taken together, these results suggest that the microbiome is a stronger predictor of postprandial lipaemia (TG) than glycaemia, with the strength of association for glycaemic responses influenced by overall metabolic load and short-term variations in microbial composition rather than differences in macronutrient composition.

Postprandial rises in lipid- and glucose-mediated measures are differentially predicted by the microbiome compared with fasting levels

Postprandial measures (iAUC and peak) depend both on the corresponding fasting measure and the 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. 4C**). When looking at lipid and glucose-mediated metabolites from the clinic day measures, despite a similar strength of association between peak (6h), magnitude (iAUC) and fasting TG concentrations, the rise (6-0h) was not similarly correlated (**Fig. 4A,G,H**). In contrast, the microbiome associations with glycaemic measures were comparable between fasting, peak, and rise (**Fig. 4A,F**).

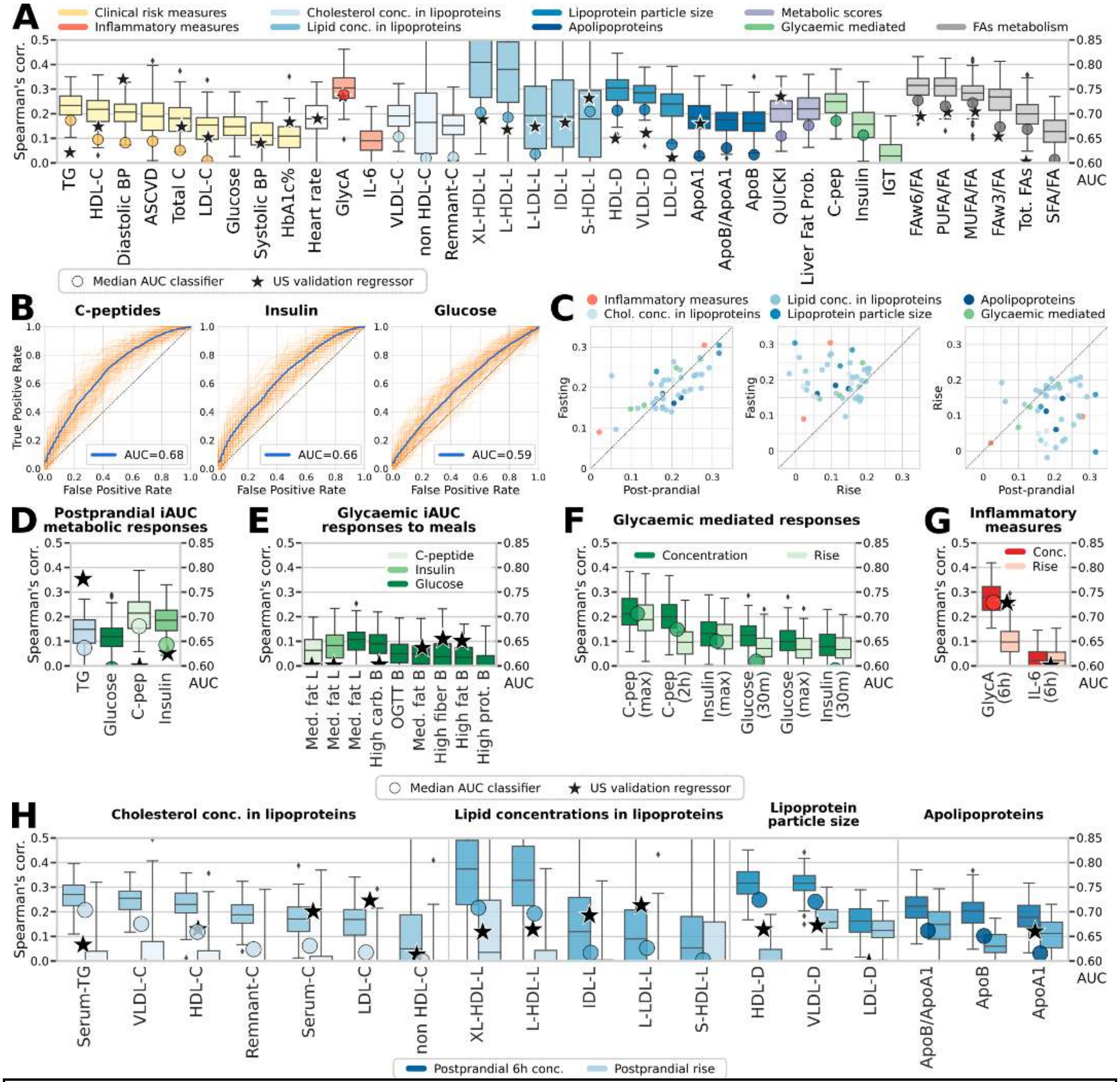

Fig. 4: Fasting and postprandial cardiometabolic responses to standardized test meals associated with the microbiome. (A) The strongest observed links according to correlation of the predicted versus collected measures between the gut microbiome and fasting metabolic blood markers. For measures of lipid concentration in lipoproteins, we report the five strongest correlations only. Indices are grouped in nine distinct categories, and boxplots report the correlation between the prediction of RF regression models trained on microbial taxa or pathway abundances across 100 training/testing folds. Circles denote AUC values for RF classification, while stars report regressor performance when trained on the UK cohort and evaluated on the independent US validation cohort. **(B)** Fasting levels of C-peptide, insulin, and glucose were strongly linked to the gut microbiome, with AUCs higher than what we observed for HbA1c and IGT (Panel A). **(C)** Fasting and postprandial performance indices (correlation of the regressors' outputs) were more tightly linked to gut community structure than were their corresponding postprandial rises. **(D-H)** Performance of our microbiome-based ML-model in estimating postprandial absolute levels and postprandial increases in cardiometabolic markers. Stars denote regression model results in our US validation cohort for postprandial measurements (not rises; **Supplementary Fig. 6** and **7**). **(D)** RF regression and classification performance in predicting postprandial metabolic responses for clinic Meal 1 (breakfast) measured as iAUC at 6h for triglycerides and iAUC at 2h for glucose, C-peptide, and insulin. **(E)** Glycaemic-mediated postprandial iAUCs at 2h for the other meals (**Supplementary Table 7**), and **(F)** glycaemic-mediated markers absolute levels vs. rise. **(G)** Postprandial inflammatory measures (concentration and rise). **(H)** RF microbiome-based model performance with postprandial changes (concentrations and rise) in lipoprotein concentration, composition, and size.

Of particular interest were the lipoprotein subfraction concentrations, composition, and size (**Supplementary Fig. 6** and **7**), which are remodelled postprandially, resulting in the generation of atherogenic lipoproteins (e.g. Large VLDL particles and TG-enriched LDL, and HDL particles) [58,59]. We found that these atherogenic particles were predicted at comparable accuracy for both fasting and postprandial peak 6h concentrations (**Fig. 4A,C,H**), and notably, HDL and VLDL size ("-D", key lipoproteins associated with cardiometabolic risk) achieve modestly stronger correlations (ρ=0.32 and 0.31, respectively) postprandially (**Fig. 4H**). However, as with TG, we found that the microbiome was substantially less predictive for the postprandial rise (6h - fasting) in all lipid metabolite measures compared with fasting and postprandial 6h peak concentration (**Fig. 4A,C,H**). For example, HDL-D is closely associated with gut microbial composition at fasting and 6h postprandially (ρ=0.30 and 0.32; AUC 0.71 and 0.72 respectively; **Fig. 4A,C,H**), but not with the rise (**Fig. 4H**). These differential associations suggest that the microbiome may influence postprandial lipid-mediated measures via effects on fasting measures, but may impact the postprandial glucose rise more independently of fasting levels.

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 sought to identify the specific taxa and functions driving these associations. Among three general risk indices of cardiovascular health (ASCVD, liver fat probability, and insulin sensitivity or QUICKI) which demonstrated significant although rather modest correlation of predictions (~0.2) using our microbiome-wide RF model (**Fig. 4**), we found eight species that were significantly correlated with all three (negatively or positively, $p<0.05$). Seven of these eight were concordantly correlated in the direction of a more healthful metabolic profile (i.e. correlated for greater QUICKI values and lower ASCVD and fatty liver risk), hinting at a global underlying microbial signature of improved metabolic health. These taxa included *Flavonifractor plautii* and *Clostridium innocuum* (higher cardiometabolic risk, **Fig. 5A-C**) and *Oscillibacter sp 57_20*, *Haemophilus parainfluenzae*, and *Eubacterium eligens* (lower risk, **Fig. 5A-C**) that we had previously linked with healthy and less-healthy dietary habits (**Fig. 2**).

We found similarly distinct separations between two opposing and clearly defined clusters of species either positively or negatively correlated with fasting cardiometabolic measures (**Fig. 5A**), including blood pressure, inflammatory markers, lipid concentrations, lipoprotein sizes and fractions, and apolipoproteins (**Fig. 5A,B**). As per the association with diet, species correlated with positive markers included some taxa generally regarded as healthy (e.g. *F. prausnitzii*) but also many uncultivated and under-characterized bacteria (7 from the cluster of 18). With the notable exception of three species of *Prevotella* (*P. copri*, *P. clara*, and *P. xylaniphila*) the positive cluster included many distinct genera, pointing at a large functional richness and diversity. In contrast, the cluster of species negatively correlated with positive markers again included many *Clostridium* species (5 of the 12 in the cluster) and the recurrent negatively connotated *R. gnavus* and *F. plautii*. Large HDL particles (and their lipid compositions, **Supplementary Fig. 8-10**), which have strong inverse associations with cardiometabolic outcomes [23,24] as well as with the microbiome (**Fig. 4A**), were associated with the healthy cluster. Conversely, lipoproteins associated with increased risk of CVD and T2D (VLDL of all sizes; XXL, XL, L, M, S and lipid composition) and atherogenicity [60] (S-LDL, M-HDL and S-HDL TG), were associated with the less-healthy cluster (**Supplementary Fig. 8-10**).

Circulating omega-6 and total polyunsaturated fatty acids (PUFA), which reflect dietary intake due to the lack of endogenous production of these fatty acids [61], were associated with the healthy cluster for which *Firmicutes* bacterium CAG95 was the most correlated representative, and *F. plautii* the strongest negative correlation (**Fig. 5A** and **Supplementary Table 5**). Both omega-6 and PUFA have been linked to reduced risk of chronic disease, whether measured from dietary inventories [62] or directly assayed from the circulation [23,24,63]. In contrast, circulating monounsaturated fatty acids (MUFA) in blood were associated with the unhealthy cluster, with an under-characterized *Oscillibacter* species (sp. 57_20) and *Clostridium bolteae* responsible for the strongest negative and positive associations respectively. Measures of circulating MUFA [24,23] but not dietary intake of MUFA [64,65] have been associated with increased risk of CVD and T2D. Differences in circulating vs. estimated dietary intakes of MUFA may be a function of endogenous MUFA production, as well as the divergent animal and plant dietary sources of MUFA [66,67], complicating their relationship with chronic health outcomes [61]. Taken together with our findings (**Fig. 2**), these results suggest that food sources of MUFA play an important role in the relationship between MUFA and health.

Both favourable and unfavourable microbial signatures of metabolic health were maintained under postprandial conditions

Links between postprandial levels of cardiometabolic and inflammatory measures corresponded with the segregation of healthful vs. detrimental taxa observed under fasting conditions (**Fig. 5B** and **Supplementary Fig. 8-10**). Notably, fasting and postprandial GlycA, which we found highly correlated with postprandial TG concentrations [68], were strongly linked with the microbiome (62 species significantly correlated at 6 hours and 67 at fasting), substantially exceeding IL-6 (5 and 26 significant postprandial and fasting associations, **Fig. 5B**). *F. plautii* and *R. gnavus* were the two species most correlated with increased inflammation both in fasting and postprandial conditions, whereas *H. parainfluenzae* and *Firmicutes bacterium CAG95* 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 frequently uncoupled from the microbial associations with fasting markers; several positive correlations between microbial species and fasting and peak metabolites measures became negative when correlating the same species with the rise from fasting (and vice versa, **Fig. 5D**). For example, the rise in total LDL cholesterol and size (-D, **Fig. 5B**) was differentially associated with clusters compared to fasting levels (especially for *T. sanguinis*, *B. animalis*, and *R. mucilaginosa*). S- and XL- HDL total lipid (-L) and cholesterol (-C) levels also paralleled this behaviour (**Supplementary Fig. 8,9**), possibly reflecting postprandial lipoprotein remodelling and reciprocal exchange of TG and cholesterol, between these particles and TG-rich lipoproteins (chylomicrons and VLDL) [69]. In contrast, the associations of the microbial species with absolute fasting and postprandial peak levels were fully consistent (**Fig. 5D**), again reflecting the close relationship between fasting levels and postprandial responses.

We observed the same “favourable” vs. “unfavourable” clustering of microbiome features when analyzing microbial pathways and gene families (**Supplementary Fig. 11,12**). This supports the segregation of many taxa, even at the species level (and likely more so among strains), by their underlying biochemical activities in the microbiome. The strengths of microbe-blood marker associations measured using Spearman’s correlation were consistent with the estimated microbe relevance by the random forest model (**Supplementary Fig. 13**). Importantly, these associations were confirmed in the PREDICT 1 US validation cohort; we had a total of 62,366 microbe-index correlations for indices present in both cohorts, and for the 292 that were significant both in the UK cohort ($q<0.2$) and in the US cohort ($p<0.05$) the concordance in the sign of the correlation reached 90.8% for the associations in fasting conditions and 91.2% postprandially.

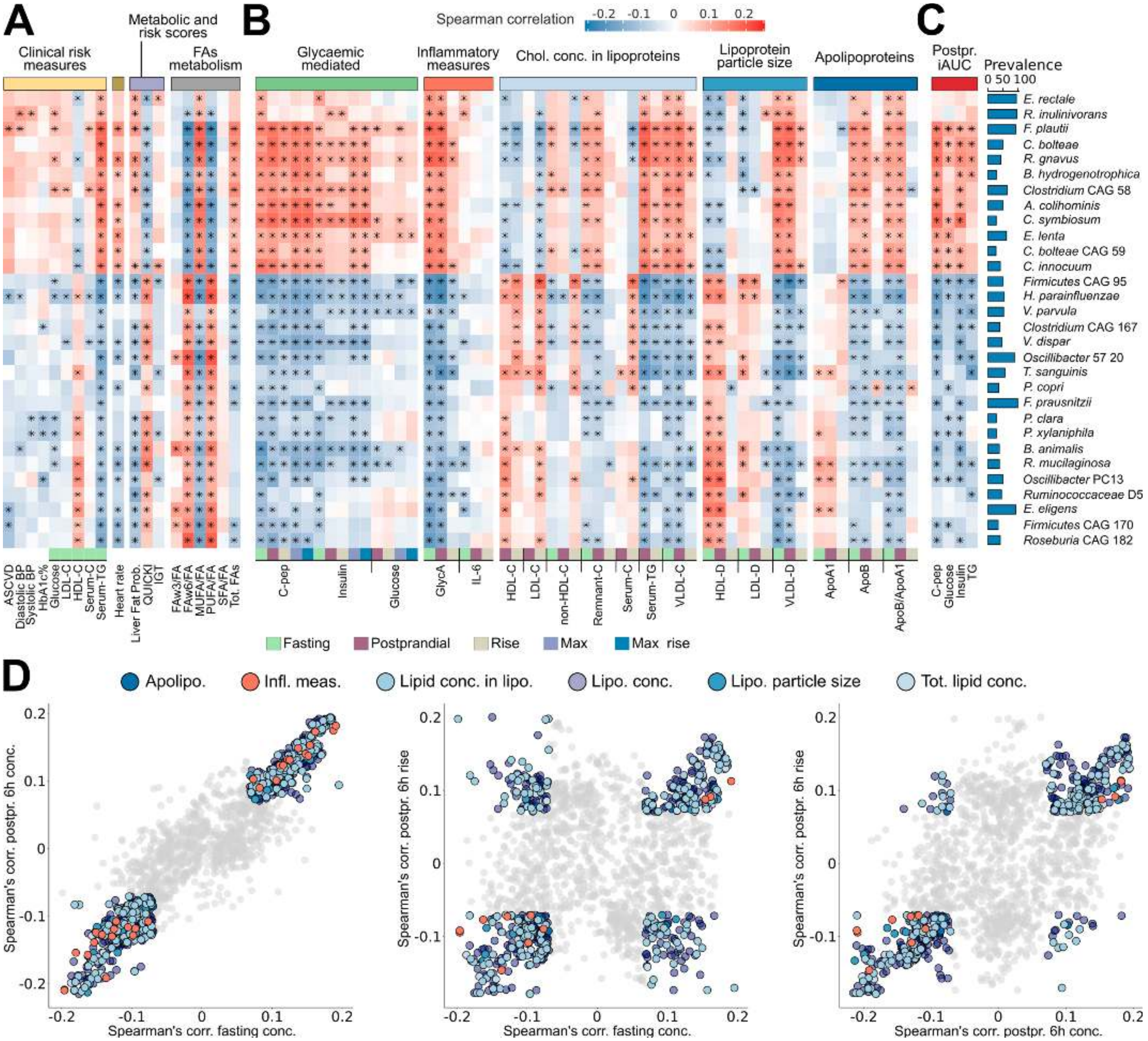

Fig. 5: Species-level segregation into healthy and unhealthy microbial signatures of fasting and postprandial cardiometabolic markers. (A) Associations (Spearman correlation, q<0.2 marked with stars) between single microbial species and fasting clinical risk measures and **(B)** glycaemic, inflammatory, and lipaemic indices. **(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. (**D**) Microbe-metabolite correlations are very consistent when evaluated for fasting versus postprandial (6h) conditions (left panel). 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, central panel). This was mitigated somewhat when comparing absolute postprandial responses with rise (right panel). (**E**) Significant species-metabolite Spearman's correlations are generally in agreement with the relevance score assigned by the RF classifier to each species in the learning models (**Supplementary Fig. 13**).

Prevotella 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 among our microbial signatures. Among them, *Prevotella copri* is a frequent and highly abundant inhabitant of the gut [70,71], but its beneficial or detrimental role in human health remains controversial [72,73]. Previous reports have yielded conflicting accounts of *P. copri* in glucose homeostasis, with some studies suggesting health benefits [74,75] and others suggesting deleterious effects [76] possibly due to subspecies diversity [77,78]. Our data largely find *P. copri* to be associated with beneficial cardiometabolic markers, being weakly negatively correlated with estimated visceral fat (ρ=-0.09, p=0.009, q=0.098), fasting VLDL-D (ρ=-0.07, p=0.06, q=0.21), and fasting GlycA (ρ=-0.12, p=0.0001, q=0.005) among others (**Supplementary Table 4**). While almost no habitual diet foods, nutrients, or scores were associated with *P. copri*, this bacterium showed a very strong correlation with postprandial increases of several circulating metabolic markers when compared with corresponding absolute fasting or postprandial levels. Postprandial rises in glucose (ρ=-0.12, p<0.0002) and polyunsaturated and omega-6 fatty acids (ρ=0.11 and 0.10, respectively, and p<0.001) were among the top-scoring correlations and were more strongly connected with the microbiome than were corresponding fasting and postprandial levels, in sharp contrast with what we observed for the overall microbiome (**Fig. 4A,D**), suggesting a potentially unique role for *P. copri* in host metabolism.

As *P. copri* has a relatively low prevalence in Western-lifestyle populations but is highly abundant when present [77], we tested whether the presence of one or more of the subtypes of this species [77] is associated with markers of improved glucose metabolism. *P. copri* is present in the form of at least one of its subtypes in 29.8% of the PREDICT 1 individuals, and we identified significant differences in *P. copri* carriers consisting of lower C-peptide (-9.2%, p=0.002), insulin (-14%, p=0.006), and lower TG levels (-3.2%, p=0.003) compared to individuals without this species (**Supplementary Fig. 14**). Similarly, postprandial blood glucose spikes after breakfast were significantly less pronounced in individuals with *P. copri* (-20.4% glucose iAUC at 2h, p=0.002, **Supplementary Fig. 14C**), and visceral fat was significantly lower (-12.5%, p=3E-7, **Supplementary Fig. 14A**). Although these observations are only associative, and the direct effect of *P. copri* on these markers of glucose metabolism is unknown, this positive association further supports that the presence of *P. copri* in the gut microbiome could be beneficial in glucose homeostasis.

Blastocystis is a unicellular eukaryotic parasite increasingly regarded as a commensal member of the gut microbiome rather than a potential pathogen [79–81]. It shares with *P. copri* a limited prevalence in Western-lifestyle populations [82] coupled with high relative abundance when present, unique among eukaryotic organisms in the gut to date. By assessing microbiome characteristics in presence or absence of *Blastocystis*, we found evidence that *Blastocystis*-positive individuals (28.1% in our cohort) also have a favourable glucose homeostasis and lower estimated visceral fat (-14.9% glucose iAUC, -21.7% visceral fat, p<0.01, **Supplementary Fig. 14**). The latter confirms that *Blastocystis* is less prevalent in overweight and obese individuals compared to individuals with BMI in the normal range, as previously shown [82] in multiple cohorts [5,39,83,84]. Interestingly, the effect of the simultaneous presence of *P. copri* and *Blastocystis* (12.8% of the individuals) appears to further promote healthier metabolic function. Visceral fat is 9.4% lower on average (p=0.028, **Supplementary Table 8**) for individuals positive for both *P. copri* and *Blastocystis* compared to individuals with only one or the other and 22.6% lower (p=3.3E-7) compared with individuals lacking both. Triglycerides and C-peptide were also consistently lower (although not individually significant, **Supplementary Table 8**) when both microbes were present.

A clear microbial signature of health levels consistent across diet, obesity indicators, and cardiometabolic risks

In the preceding analyses, we observed a consistent set of microbial species that were strongly linked to (1) foods and food indices reflecting different levels of a “healthy” diet, (2) indicators of obesity and of general health, (3) fasting circulating metabolites connected with cardiometabolic risks, and (4) postprandial responses to food. To test the consistency of such a signature, we selected a representative set of “health” indicators from each of the four categories (diet, personal characteristics, fasting and postprandial biomarkers) and ranked each microbial species based on their correlation coefficient. By averaging the ranks of the association (or inverted ranks for “unhealthy” indicators), we found remarkable agreement among microbes associated with different positive or negative indicators of health (**Fig. 6, Supplementary Table 9**).

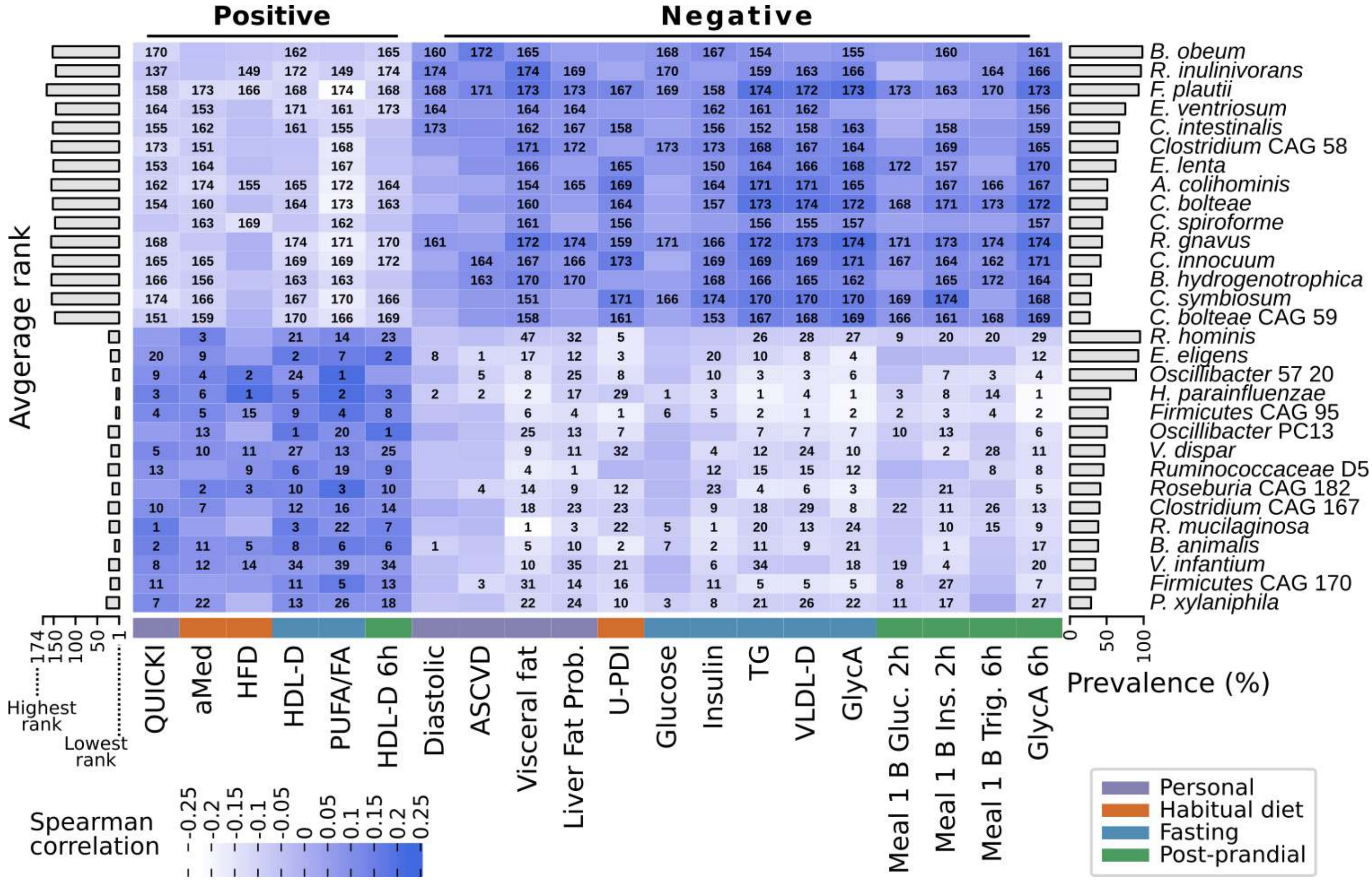

Fig. 6: The panel of 30 species showing the strongest overall correlations with a selection of markers of nutritional and cardiometabolic health. The 30 species with the highest and lowest average ranks with diverse positive and negative health indicators, respectively, are shown here. The rank of each microbe’s correlation with individual health indicators is written within cells when significant (p<0.05). For each of the main categories of indices, we selected up to five representative quantitative markers (for “Personal” we considered only four as 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.

In particular, *Firmicutes* CAG 95 is the uncultivated species with the most beneficial score (average rank 7.14) and ranked within the top 5 correlated species for 13 of the 20 indicators. Of the “health”-associated microbial species only *R. hominis* (23.76) was already convincingly linked with health in case/control disease investigations [85], even though others such as *F. prausnitzii* [86] and *P. copri* were highly ranked (average ranks 31.7 and 37.2 respectively, 18th and 21st best ranks) but not in the top 15. The beneficial signature also included several known species such as *E. eligens* (16.6) and *H. parainfluenzae* (6.4) without clear roles in health, and additional species without cultivated representatives such as *Roseburia* CAG 182 (15.5), *Oscillibacter* sp 57_20 (13.6), *Firmicutes bacterium* CAG 170 (20.1), *Oscillibacter* sp PC13 (24.5), *Clostridium* sp CAG 167 (24.8), and *Ruminococcaceae bacterium* D5 (24.8). Species that were conversely consistent with indicators of poor overall health (**Fig. 6**) included the already discussed set of Clostridia (*C. spiroforme* - 149.7, *C. bolteae* CAG 59 - 149.9, *C. bolteae* - 154.8, *Clostridium* CAG 58 - 157.5, *C. symbiosum* - 157.4, *C. innocuum* - 155.1). The two strongest microbial indicators of poor cardiometabolic and diet-related health were the mucolytic microbe *R. gnavus* (158.8) and *F. plautii* (169.1), again previously found to be associated with disease conditions [87–92]. Overall, this set of 30 species serves as a marker of overall good or poor general health and dietary patterns in non-diseased human hosts.

Discussion

PREDICT 1 represents the first diet-microbiome clinical intervention 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 reproduced across UK and US populations, across multiple previously-published study populations, and for multiple dietary, biometric, and blood markers of health and cardiometabolic risk, including individual food items, nutrients, dietary patterns, adiposity, BMI, circulating lipids, inflammatory markers, blood glucose, and interactions between baseline and postprandial response levels. Notably, microbiome signatures robustly grouped both microbiome and dietary components into health-associated and anti-associated clusters, the latter in agreement with dietary quality and diversity scores (such as the Plant-based Diet Index [PDI] and Healthy Food Diversity [HFD] index) known to be health-associated [28,93] and often unlinked from macronutrient source (e.g. more vs. less healthy plant- and animal-based foods). The diversity of a healthy diet (measured by the HFD and PDI) was particularly predictable by the microbiome, surpassing other indices such as the Mediterranean diet index that has been independently linked with microbiome composition [94]. The segregation of favourable and unfavourable microbial clusters according to the heterogeneity of the food source (healthy or unhealthy animal or plant), quality (processed vs 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 prior diet-microbiome work [8,10–12,14,95] may be due to the quality in the metagenomic profiling and the large sample size. However, given the limitations of FFQ dietary data (which can be highly scalable but noise-prone [96]), future diet-microbiome studies would benefit further from more detailed weighed food record data complemented with nutritionist/dietitian support.

Several aspects of the gut microbiome associations and matched signatures across diet, obesity, and metabolic health measures are striking with respect to their potential novel epidemiology and microbial biochemistry. A surprising proportion of diet- or health-associated taxa in these results are represented solely by existing or newly-generated metagenomic assemblies [42], in addition to very recently isolated organisms with limited cultured strains. This was true for *Lawsonibacter asaccharolyticus*, the taxon most strongly associated with individual food items (particularly coffee) and nutrient intake, for which only two recent publications with limited and conflicting microbial physiology and taxonomy exist [35,41]. Both of the taxa most abundant in diets rich in healthy plant-based foods were represented only by previous metagenomic assemblies [39] (*Firmicutes* CAG 95 and *Roseburia* CAG 182), as was the strongest microbial association with adiposity (*Clostridium* CAG 58) and several of the most reproducible microbes associated with (un)healthy blood markers (*C. bolteae* CAG 59, *Clostridium* CAG 167). Other microbes found here to have dietary or cardiometabolic associations, such as *Prevotella* spp. or *Blastocystis* spp., have been characterized in greater biochemical detail, but their prevalence and population structure in the human microbiome have only recently begun to be appreciated [77,82]. The latter in particular may be only one of many examples of eukaryotic, fungal, or viral members of the gut microbiome not amenable to most current high-throughput experimental or analytical approaches, but with unexpected and potentially key positive roles in dietary metabolism or cardiometabolic health.

Likewise, these new, highly specific contributions of the gut microbiome to human dietary responses may help to explain some of the heterogeneity and apparent contradictions seen among previous population studies [6,8,95,97]. First, diet-microbiome-blood marker associations were overall strongest with respect to circulating lipid levels (triglycerides, lipoproteins, etc.) relative to glycemic indices (e.g. blood glucose, insulin sensitivity). This may have both biochemical and clinical implications. It is possible that gut microbial metabolism contributes relatively more to circulating lipid levels than to carbohydrate derivatives, either directly or via mediating processes such as gastrointestinal or systemic bile acid signalling [97,98]. 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 L-HDL-L) are also more strongly associated with cardiovascular risk compared with typically measured lipids (e.g. TC, HDL-C, LDL-C), suggesting a closer look may be warranted at their utility as clinical biomarkers or as targets for beneficial gut microbiome manipulation.

Finally, an important conclusion of these results with respect to overall microbiome epidemiology is the limitation and coarseness of phenotypic associations achievable by using simple diversity or microbiome summary statistics. Even when we identified a variety of significant species-specific dietary and molecular associations in the gut, their effect sizes were often limited, likely reflecting both strain-specific functionality not assessed in these profiles [42,50,99–101] and ecological signals among multiple interacting microbes as captured by our richer machine learning models [102]. Similarly, with respect to host physiology, many postprandial responses relative to individual-specific fasting values (e.g. triglyceride levels, lipoproteins, insulin concentrations) were moderately more associated with the gut microbiome than the pre-existing fasting values themselves. This may speak to the interaction of both host metabolism and microbial metabolism impacting digestive and metabolic pathways, shaping long- and short-term diet-host effects on health and disease [103]. Overall, this is the first study to identify a shared diet-metabolic-health microbial signature, segregating favourable and unfavourable taxa with multiple measures of both dietary intake and cardiometabolic health. We hope that these initial PREDICT 1 results, targeted clinical and microbial follow-up based on them, and future iterations of the PREDICT study will aid as a resource both in utilization of the gut microbiome as a biomarker for cardiometabolic risk and in strategies for reshaping the microbiome to improve personalized dietary health.

Methods

The PREDICT 1 study

The PREDICT 1 clinical trial (NCT03479866) aimed to quantify and predict individual variations in metabolic responses to standardised 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 US. The trial was a single-arm, single-blinded intervention study that commenced in June 2018 and completed in May 2019.

For full protocol, see Berry et al, 2020 [17]. In brief; 1,002 generally healthy adults from the United Kingdom (UK; non-twins, and identical [monozygotic; MZ] and non-identical [dizygotic; DZ] twins) and 100 healthy adults from the United States (US; non-twins; validation cohort) were enrolled into the study (see Berry *et al* [56] for eligibility criteria) and completed baseline clinic measurements. The study consisted of a 1-day clinical visit at baseline followed by a 13-day at-home period. At baseline (Day 1), participants arrived fasted and were given a standardised metabolic challenge meal for breakfast (0h; 86g carbohydrate, 53g fat) and lunch (4h; 71g carbohydrate, 22g fat). Fasting and postprandial (9 timepoints; 0-6h) venous blood was collected to determine serum concentrations of glucose, triglycerides (TG), insulin, C-peptide (as a surrogate for insulin), and metabolomics (NMR). 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 standardised 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 following completion of the study, and 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 standardised meal protocol (outlined in Berry et al 2020[56]) 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 [56,104]. The trial was approved in the UK by the Research Ethics Committee and Integrated Research Application System (IRAS 236407) and in the US by the Partners Healthcare Institutional Review Board (IRB 2018P002078). The core characteristics of study participants at baseline were not significantly different between UK and US cohorts [56].

Overview of microbiome sequencing and profiling

We performed deep shotgun metagenomic sequencing (mean 8.8±2.2 gigabases/sample) in stool samples from a total of 1,098 PREDICT 1 participants (UK n=1,001; US n=97). From a random subset of these participants (n=70), we additionally sequenced faecal metagenomes from a second stool sample collected 14 days after the first collection (**Fig. 1A**) for a total of 1,168 metagenomes. Computational analysis was performed using the bioBakery suite of tools [105] to obtain species-level microbial abundances for the 769 taxa identified using the newly updated MetaPhlAn 2.96 tool [106], functional potential profiling of >1.91 M microbial gene families, 445 KEGG pathways with HUMAnN 2.0 [107], and reconstruction of 48,181 metagenome-assembled genomes (MAGs) of medium or high-quality using our validated pipeline [42], which includes assembly with MegaHIT [108], binning with MetaBAT2 [109], and quality-control with CheckM [110].

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 timepoints (one prior to their in-person clinical visit on day 0 and the next at the conclusion of their home-phase, day 14). Those who did not collect a sample prior to 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 collection kit (ALPCO, NH, US), whereas post-study samples, as well as the entirety of the US collection was conducted using the Fecotainer collection kit (Excretas Medical BV, Enschede, the Netherlands). For baseline samples, one fresh unfixed sample was deposited into a sterile universal collection container (Sarstedt, Australia, Cat #L0263-10) and one into a tube containing DNA/RNA Shield buffer (Zymo Research, CA, US, Cat #R1101). Samples were stored at ambient temperature until return 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. Upon receipt in the laboratory, samples were homogenized, aliquoted, and stored at -80°C in Qiagen PowerBeads 1.5 mL tubes (Qiagen, Germany). This sample collection procedure was tested and validated internally comparing different storage conditions (fresh, frozen, buffer), different DNA extraction kits (PowerSoilPro, FastDNA, ProtocolQ, Zymo), and different sequencing technologies (16S rRNA, 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=70) 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 (Agilent Technologies, Inc.) according to manufacturer's guidelines. Samples with a high-quality DNA profile were further processed. The NEBNext® Ultra II FS DNA module (cat# NEB #E7810S/L) was used for DNA fragmentation, end-repair, and A-tailing. For adapter ligation, the NEBNext® Ultra II Ligation module (cat# NEB #E7595S/L) was used. The quality and yield after sample preparation were measured with the Fragment Analyzer. 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 NovaSeq6000 platform according to manufacturer's protocols. 1.1 nM library was used for flow cell loading. NovaSeq control software NCS v1.5 was used. Image analysis, base calling, and the quality check was performed with the Illumina data analysis pipeline RTA3.3.5 and Bcl2fastq v2.20.

Metagenome quality control and pre-processing

All sequenced metagenomes were QCed using the pre-processing pipeline as implemented in https://github.com/SegataLab/preprocessing. Pre-processing consists of three main steps: (1) read-level quality control; (2) screening of contaminant, i.e. 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 (Langmead and Salzberg 2012) using the --sensitive-local parameter, allowing confident removal of the phiX174 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 [100] and relying on the bioBakery computational environment [105]. The taxonomic profiling and quantification of organisms relative abundances of all metagenomic samples have been quantified using MetaPhlAn2 (version 2.9.21 and marker database release 2.9.4) [106]. 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 1,077,785 markers able to profile 10,586 species. Compared to the previous version of the MetaPhlAn2 database (mpa_v20_m200), the updated database is able to profile 8,102 more species. Metagenomes were mapped internally in MetaPhlAn2 against the marker genes database with BowTie2 version 2.3.4.3 with the parameter "very-sensitive". The resulting alignments were filtered to remove reads aligned with a MAPQ value <5, representing an estimated probability of the likelihood of the alignments.

For estimating 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. Microbiome dissimilarity between participants (beta diversity) was computed using the Bray-Curtis dissimilarity and the Aitchison distance on microbiome taxonomic profiles.

Functional potential analysis of the metagenomic samples was performed using HUMAnN2 (version 0.11.2 and UniRef database release 2014-07) [107] that computed pathway profiles and gene-family abundances.

Metagenomic assembly

Metagenomic samples were processed to obtain metagenome-assembled genomes (MAGs) following the procedure we used elsewhere [42]. In brief, we used MEGAHIT (version 1.2.9) [108] with parameters "--k-max 127" for assembly and assembled contigs ≥1.5kb were considered for the binning step performed using MetaBAT2 (version 2.14) [109] with parameters: "-m 1500 --unbinned". Quality control of the obtained MAGs was performed using CheckM (version 1.0.18) [110] using default parameters. High-quality 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 food frequency questionnaires (FFQ). For the UK, the European Prospective Investigation into Cancer and Nutrition (EPIC) FFQ was used and in the US, the Harvard semi-quantitative FFQ was used.

For the UK, we used the 131-item EPIC FFQ that was developed and validated against pre-established nutrient biomarkers for the EPIC Norfolk [111]. The questionnaire captured average intakes in the past year. Nutrient intakes were determined via consultation with McCance and Widdowson's 6th edition, an established nutrient database [112]. US participants completed the Harvard 2007 Grid 131-item FFQ previously validated against two-week dietary records [113]. Nutrient intakes were estimated using the Harvard Nutrient Database. Submitted FFQs were excluded if greater than 10 food items were left unanswered, or if the total energy intake estimate derived from FFQ as a ratio of the subject's estimated basal metabolic rate (determined by the Harris-Benedict equation [114]) was more than two standard deviations outside the mean of this ratio (<0.52 or >2.58).

The following dietary indices were calculated as described below and according to categorisation listed in **Supplementary Table 2,4**.

Healthy Food Diversity Index. The Healthy Food Diversity (HFD) index considers the number, distribution, and health value of consumed foods. To obtain this index, food frequency questionnaire foods were first aggregated into 15 food groups according to the HFD [28]. Health values were then derived from the German Nutrition Society (DGE) dietary guidelines (https://www.dge.de/en/) and the weight of each food group was multiplied by its corresponding health value (hv). Scores were divided by the maximum (hv=0.26) to bind values between 0-1 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 hypothesised microbiome-plant food/fibre interactions, and converting UK g/serving to US volume measures (as required for the US-HFD) would introduce additional error to the FFQ estimates.

The plant-based diet index. Three versions of the plant-based diet index [36] were considered: the original plant-based diet index (PDI), the healthy plant-based index (h-PDI), and the unhealthy plant-based index (u-PDI). 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** [36]. 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 applied to the animal-based foods. For the h-PDI, positive scores were applied to the "healthy" plant foods, and a reverse score to the "less-healthy"/"unhealthy" plant foods and the animal-based foods. For the u-PDI, a positive score was applied to the "less-healthy"/"unhealthy" plant foods and a reverse score applied to the "healthy" plant foods and the animal-based foods.

Animal score. The animal-based score categorised animal foods into "healthy" and "less-healthy"/"unhealthy" categories according to previous epidemiological studies [115–124]. 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 Table 2** and **4**).

The aMED score. Adherence to the aMED diet was calculated by following the method outlined by Fung *et al.* [29]. 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 grams per serving and then divided by 7 to determine grams per day. Next, food gram amounts were summed to make the final category total. For all food categories as well as 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; males:10-50 g/d 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 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 diet: The percentage of plants within diet was calculated as weight in grams of plant foods within total weight (g) of 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 0g 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 [104]. In brief, participants were cannulated, and venous blood was collected at fasting (prior to a test breakfast) and at 9 timepoints postprandially (15, 30, 60, 120, 180, 240, 270, 300, and 360 minutes). Plasma glucose and serum C-peptide and insulin were measured at all timepoints. Serum TG was measured at hourly intervals, and serum metabolomics (NMR by Nightingale Health, Helsinki, Finland) at 0, 4 and 6h. Fasting samples were analysed for lipid profile, thyroid-stimulating hormone, alanine aminotransferase, liver function panel, and complete blood count (CBC) analysis.

Continuous glucose monitoring on day 2-14 was measured every 15 minutes using Freestyle Libre Pro continuous glucose monitors (Abbott, Abbott Park, IL, US), fitted on the upper, non-dominant arm at participants’ baseline clinical visit. Given the CGM device requires time to calibrate once fitted to a participant, CGM data collected 12 hours and onwards after activating the device was used for analysis.

Dry blood spot analysis of TG and C-peptide was completed by participants on the first 4 days of the home-phase while consuming test meals. The timepoints were dependent on the test meal as described elsewhere [56,104]. 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. DBS cards were frozen at -80 °C upon receipt in the laboratory until being shipped to Vitas for analysis (Vitas Analytical Services, Oslo, Norway).

Specific timepoints 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 [104]. The incremental area under the postprandial TG (0-6h), glucose (0-2h), and insulin (0-2h) curves (iAUC) were computed using the trapezium rule [125].

Detailed descriptions of sample collection, processing and analysis have been reported in [17,56].

Machine learning

The machine learning (ML) framework employed is based on the scikit-learn Python package [126]. The ML algorithms used for the prediction and classification of personal, habitual diet, fasting, and postprandial metadata are based on Random Forest (RF) regressor and classification. We selected RF-based methods *a priori* as it has been repeatedly shown to be particularly suitable and robust to the statistical challenges inherent to microbiome abundance data [55,102] For both the regression and classification tasks, a cross-validation approach was implemented, 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 an RF 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 “n_estimators=1000, criterion='mse'” 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 “n_estimators=1000” parameter.

We used RF classification and regression on both species-level taxonomic relative abundance and functional potential profiles. For taxonomic abundances, we used the relative abundances of MetaPhlAn2 (see above) with all the abundances of all microbial clades from phylum to species normalized using the arcsin-sqrt transformation for compositional data. For functional profiles, we considered both raw relative abundance estimates of single microbial gene families, as well as pathway-level relative abundance as provided by HUMAnN2.

As an additional control, we verified that when random swapping the target labels or values (classification and regression, respectively), the performances were reflecting a random prediction, hence an AUC very close to 0.5 and a non-significant correlation between the predicted with values approaching 0.

Statistical analysis

Spearman's correlations (reported with "ρ" in the text), have been computed using the *cor.test* from the *stats* R package and a modified version of the *pcor.test* from the *ppcor* package (available at http://www.yilab.gatech.edu/pcor.R) that permits to control for a set of covariates rather than single ones, respectively. Correlations and the p-values were computed for each couple of metadata and species and p-values were corrected using 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 have been selected by ranking them according to their number of significant associations for the panel of metadata considered, and then the top thirty unique species are 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 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 [127]. Correction for multiple testing (Benjamini–Hochberg, FDR) was applied and significance was defined at 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. Only metadata variables with <15% missing data and without high collinearity with other variables (Spearman's rho <0.8) were used as input in the stepwise model.

Data validation on the US cohort and on the cMD datasets

As independent validation, we considered the publically available datasets collected in the *curatedMetagenomicData* version 1.16.0 R package (cMD) [16]. Of the 57 datasets available we selected those that have samples with the following characteristics: (1) gut samples collected from healthy adult individuals at first collection ("days_from_first_collection"=0 or NA), (2) samples with age and BMI data available and BMI interquartile range (IQR) of these samples between 3.5 and 7.5 (± 2 with respect to the PREDICT 1 UK IQR of 5.5, **Supplementary Fig. 5**). For each dataset with samples meeting the above criteria, only datasets with at least 50 samples were considered: CosteaPI_2017 (84 samples out of 279), DhakanDB_2019 (88 samples out of 110), HanenLBS_2018 (58 samples out of 208), JieZ_2017 (157 samples 385), SchirmerM_2016 (396 samples out of 471), and ZellerG_2014 (59 samples out of 199).

We used the previously selected validation datasets from cMD in two analyses: one based on machine learning to verify the reproducibility of the ML 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 training fold and each dataset as testing fold. Third, a final prediction was made using Leave-One-Dataset-Out cross-validation (LODO), meaning that all datasets (PREDICT 1 UK, PREDICT 1 UK, and the cMD 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 cMD datasets was done by applying a pairwise Spearman correlation for each species in each cMD 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 cMD. For those species found also in the PREDICT 1 US, we reported their correlation as well.

Data availability

Metagenomes are being deposited in EBI ENA and will be made publicly available upon acceptance of the paper.

Conflict of interest statement

TD Spector, SE Berry, AM Valdes, F Asnicar, PW Franks, C Huttenhower, N Segata, are consultants to Zoe Global Ltd (“Zoe”). J Wolf, G Hadjigeorgiou, R Davies, J Capdevila, C Bonnett, R Hine, L Francis and S Danzanvilliers are or have been employees of Zoe. Other authors have no conflict of interest to declare.

Funding

This work was supported by Zoe Global Ltd and also received support from grants from the Wellcome Trust (212904/Z/18/Z) and the Medical Research Council (MRC)/British Heart Foundation Ancestry and Biological Informative Markers for Stratification of Hypertension (AIMHY; MR/M016560/1). The work was also supported by the European Research Council (ERC-STG project MetaPG) to NS; by MIUR ‘Futuro in Ricerca’ (grant No. RBFR13EWWI_001) to NS; and by the European H2020 program (ONCOBIOME-825410 project and MASTER-818368 project) to NS. PWF was supported in part by grants from the European Research Council. DAD is supported by NIH NIDDK 1K01DK120742. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, Chronic Disease Research Foundation (CDRF), Zoe Global Ltd and the National Institute for Health Research (NIHR)-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.

Authors contribution

Obtained funding: JW, GH, TDS. Study design and developed concept: SEB, AMV, JW, GH, RD, ATC, NS, PWF, TDS. Data collection: SEB, NS, FA, ATC, DAD, TDS. Data analysis: FA, SEB, NS, LF, EL, RG, MM, OM, GP, CLR, MVC, SO, AT, FB, CM, AK, LD, DB, AMT, CB, LW, LG, JCP, SD, RH. Study coordination: SEB, DAD, GH, JW, NS. Writing the manuscript FA, SEB, AMV, LHN, DAD, EL, RG, JW, CDG, JMO, CH, PWF, TDS, NS. All authors reviewed and revised the final manuscript.

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. Mozaffarian, D. Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity: A Comprehensive Review. *Circulation* **133**, 187–225 (2016).
4. Musso, G., Gambino, R. & Cassader, M. Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. *Annu. Rev. Med.* **62**, 361–380 (2011).
5. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
6. Sze, M. A. & Schloss, P. D. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *MBio* **7**, (2016).
7. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
8. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
9. 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).
10. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
11. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
12. 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).
13. Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1897 (2016).
14. Fu, J. *et al.* The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids. *Circ. Res.* **117**, 817–824 (2015).
15. Org, E. *et al.* Relationships between gut microbiota, plasma metabolites, and metabolic syndrome traits in the METSIM cohort. *Genome Biol.* **18**, 70 (2017).
16. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
17. Berry, S. *et al.* Personalised REsponses to DIetary Composition Trial (PREDICT): an intervention study to determine inter-individual differences in postprandial response to foods. *Protocol Exchange* (2020).
18. 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).
19. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4680–4687 (2011).
20. Atabaki-Pasdar, N. *et al.* Predicting and elucidating the etiology of fatty liver disease using a machine learning-based approach: an IMI DIRECT study. *Genetic and Genomic Medicine* (2020) doi:10.1101/2020.02.10.20021147.
21. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
22. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
23. Würtz, P. *et al.* Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* **131**, 774–785 (2015).
24. Ahola-Olli, A. V. *et al.* Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *Diabetologia* **62**, 2298–2309 (2019).
25. Vojinovic, D. *et al.* Relationship between gut microbiota and circulating metabolites in population-based cohorts. *Nat. Commun.* **10**, 5813 (2019).
26. Duprez, D. A. *et al.* Comparison of the Predictive Value of GlycA and Other Biomarkers of Inflammation for Total Death, Incident Cardiovascular Events, Noncardiovascular and Noncancer

Inflammatory-Related Events, and Total Cancer Events. *Clin. Chem.* **62**, 1020–1031 (2016).

27. 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).
28. 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).
29. 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).
30. 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).
31. Kim Hyunju *et al.* Plant-Based Diets Are Associated With a Lower Risk of Incident Cardiovascular Disease, Cardiovascular Disease Mortality, and All-Cause Mortality in a General Population of Middle-Aged Adults. *J. Am. Heart Assoc.* **8**, e012865 (2019).
32. 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).
33. 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).
34. 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**, (2019).
35. 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).
36. 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).
37. Eslinger, A. J., Eller, L. K. & Reimer, R. A. Yellow pea fiber improves glycemia and reduces Clostridium leptum in diet-induced obese rats. *Nutr. Res.* **34**, 714–722 (2014).
38. 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).
39. 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).
40. Sakamoto, M., Ikeyama, N., Yuki, M. & Ohkuma, M. Draft Genome Sequence of Lawsonibacter asaccharolyticus JCM 32166T, a Butyrate-Producing Bacterium, Isolated from Human Feces. *Genome Announc.* **6**, (2018).
41. Hosny, M. *et al.* Description of Clostridium phoceensis sp. nov., a new species within the genus Clostridium. *New Microbes New Infect* **14**, 85–92 (2016).
42. 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).
43. Beaumont, M. *et al.* Heritable components of the human fecal microbiome are associated with visceral fat. *Genome Biol.* **17**, 189 (2016).
44. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, (2014).
45. 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).
46. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).
47. Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun.* **8**, 845 (2017).
48. 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**, (2019).
49. Falony, G. *et al.* Population-level analysis of gut microbiome variation. Science (80-)[Internet]. 2016 Apr 29; 352 (6285): 560-4.
50. 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).

51. Cui, Y. *et al.* Non-high-density lipoprotein cholesterol level as a predictor of cardiovascular disease mortality. *Arch. Intern. Med.* **161**, 1413–1419 (2001).
52. 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).
53. 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).
54. 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).
55. 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).
56. Berry, S. *et al.* Decoding Human Postprandial Responses to Food and their Potential for Precision Nutrition: the PREDICT 1 Study. *submitted* (2020).
57. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
58. 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).
59. Cohn, J. S. Postprandial lipemia and remnant lipoproteins. *Clin. Lab. Med.* **26**, 773–786 (2006).
60. Skeggs, J. W. & Morton, R. E. LDL and HDL enriched in triglyceride promote abnormal cholesterol transport. *J. Lipid Res.* **43**, 1264–1274 (2002).
61. 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).
62. Li, J., Guasch-Ferré, M., Li, Y. & Hu, F. B. Dietary intake and biomarkers of linoleic acid and mortality: systematic review and meta-analysis of prospective cohort studies. *Am. J. Clin. Nutr.* (2020) doi:10.1093/ajcn/nqz349.
63. Marklund, M. *et al.* Biomarkers of Dietary Omega-6 Fatty Acids and Incident Cardiovascular Disease and Mortality. *Circulation* **139**, 2422–2436 (2019).
64. Chowdhury, R. *et al.* Association of dietary, circulating, and supplement fatty acids with coronary risk: a systematic review and meta-analysis. *Ann. Intern. Med.* **160**, 398–406 (2014).
65. Zong, G. *et al.* Intake of individual saturated fatty acids and risk of coronary heart disease in US men and women: two prospective longitudinal cohort studies. *BMJ* **355**, i5796 (2016).
66. Wu, J. H. Y., Micha, R. & Mozaffarian, D. Dietary fats and cardiometabolic disease: mechanisms and effects on risk factors and outcomes. *Nat. Rev. Cardiol.* **16**, 581–601 (2019).
67. Zong, G. *et al.* Monounsaturated fats from plant and animal sources in relation to risk of coronary heart disease among US men and women. *Am. J. Clin. Nutr.* **107**, 445–453 (2018).
68. Berry *et al.* Postprandial lipemia and CVD; does the magnitude, peak concentration or duration impact intermediary cardiometabolic risk factors differentially? PREDICT I Study. *Proc. Nutr. Soc.* **in review**,.
69. Cohn, J. S. Postprandial lipemia: emerging evidence for atherogenicity of remnant lipoproteins. *Can. J. Cardiol.* **14 Suppl B**, 18B–27B (1998).
70. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
71. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
72. Cani, P. D. Human gut microbiome: hopes, threats and promises. *Gut* **67**, 1716–1725 (2018).
73. Ley, R. E. Gut microbiota in 2015: Prevotella in the gut: choose carefully. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 69–70 (2016).
74. 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).
75. De Vadder, F. *et al.* Microbiota-Produced Succinate Improves Glucose Homeostasis via Intestinal Gluconeogenesis. *Cell Metab.* **24**, 151–157 (2016).
76. Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity.

Nature **535**, 376–381 (2016).

77. Tett, A. *et al.* The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* (2019) doi:10.1016/j.chom.2019.08.018.
78. 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).
79. Clark, C. G., van der Giezen, M., Alfellani, M. A. & Stensvold, C. R. Recent developments in Blastocystis research. *Adv. Parasitol.* **82**, 1–32 (2013).
80. Alfellani, M. A. *et al.* Variable geographic distribution of Blastocystis subtypes and its potential implications. *Acta Trop.* **126**, 11–18 (2013).
81. Lukeš, J., Stensvold, C. R., Jirků-Pomajbíková, K. & Wegener Parfrey, L. Are Human Intestinal Eukaryotes Beneficial or Commensals? *PLoS Pathog.* **11**, e1005039 (2015).
82. Beghini, F. *et al.* Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. *ISME J.* **11**, 2848–2863 (2017).
83. Andersen, L. O., Bonde, I., Nielsen, H. B. & Stensvold, C. R. A retrospective metagenomics approach to studying Blastocystis. *FEMS Microbiol. Ecol.* **91**, (2015).
84. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
85. Machiels, K. *et al.* A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut* **63**, 1275–1283 (2014).
86. 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. U. S. A.* **105**, 16731–16736 (2008).
87. Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).
88. Azzouz, D. *et al.* Lupus nephritis is linked to disease-activity associated expansions and immunity to a gut commensal. *Ann. Rheum. Dis.* **78**, 947–956 (2019).
89. Ni, Y. H., Chua, H.-H., Chou, H.-C. C., Chiang, B.-L. & Liu, H.-H. Gut Dysbiosis Featured by Abundant Ruminococcus Gnavus Heralds the Manifestation of Allergic Diseases in Infants. *Gastroenterology* **152**, S214 (2017).
90. 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).
91. Gupta, A. *et al.* Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. *mSystems* **4**, (2019).
92. Jiang, H. *et al.* Altered fecal microbiota composition in patients with major depressive disorder. *Brain Behav. Immun.* **48**, 186–194 (2015).
93. 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).
94. 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* (2020) doi:10.1136/gutjnl-2019-320438.
95. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, (2018).
96. Cade, J. E., Burley, V. J., Warm, D. L., Thompson, R. L. & Margetts, B. M. Food-frequency questionnaires: a review of their design, validation and utilisation. *Nutr. Res. Rev.* **17**, 5–22 (2004).
97. Kurilshikov, A. *et al.* Gut Microbial Associations to Plasma Metabolites Linked to Cardiovascular Phenotypes and Risk. *Circ. Res.* **124**, 1808–1820 (2019).
98. 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.* (2020) doi:10.1038/s41575-019-0250-7.
99. Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016).
100. 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).
101. Yan, Y., Nguyen, L., Franzosa, E. & Huttenhower, C. Strain-level epidemiology of microbial

communities and the human microbiome. *in review* (2020).

102. 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).
103. Rowland, I. *et al.* Gut microbiota functions: metabolism of nutrients and other food components. *Eur. J. Nutr.* **57**, 1–24 (2018).
104. Berry, S. *et al.* Personalised REsponses to DIetary Composition Trial (PREDICT): an intervention study to determine inter-individual differences in postprandial response to foods. *Protocol Exchange* (2020).
105. McIver, L. J. *et al.* bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**, 1235–1237 (2018).
106. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
107. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
108. 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).
109. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
110. 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).
111. Bingham, S. A. *et al.* Nutritional methods in the European Prospective Investigation of Cancer in Norfolk. *Public Health Nutr.* **4**, 847–858 (2001).
112. Holland, B. *et al. McCance and Widdowson's the composition of foods*. (Royal Society of Chemistry, 1991).
113. 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**, 1114–26; discussion 1127–36 (1992).
114. 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).
115. WHO | Effect of *trans*-fatty acid intake on blood lipids and lipoproteins: a systematic review and meta-regression analysis. (2016).
116. Zhong, V. W. *et al.* Associations of Dietary Cholesterol or Egg Consumption With Incident Cardiovascular Disease and Mortality. *JAMA* **321**, 1081–1095 (2019).
117. 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).
118. Michaëlsson, K. *et al.* Milk intake and risk of mortality and fractures in women and men: cohort studies. *BMJ* **349**, g6015 (2014).
119. 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).
120. 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).
121. 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).
122. 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**, (2017).
123. WHO | Effect of *trans*-fatty acid intake on blood lipids and lipoproteins: a systematic review and meta-regression analysis. (2016).
124. Website, N. H. S. Dairy and alternatives in your diet. *nhs.uk* https://www.nhs.uk/live-well/eat-

well/milk-and-dairy-nutrition/.

125. Matthews, J. N., Altman, D. G., Campbell, M. J. & Royston, P. Analysis of serial measurements in medical research. *BMJ* **300**, 230–235 (1990).

126. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

127. Oksanen, J. *et al.* Vegan: community ecology package. R package version 1.17-4. *URL http://CRAN. R-project. org/package= vegan* (2010).