

Microbiome connections with host metabolism and habitual diet from 1098 deeply phenotyped individuals

Francesco Asnicar^{1,^}, Sarah E. Berry^{2,^,#}, Ana M. Valdes^{3,4}, Long H. Nguyen⁵, Gianmarco Piccinno¹, David A. Drew⁵, Emily Leeming⁶, Rachel Gibson², Caroline Le Roy⁶, Haya Al Khatib⁷, Lucy Francis⁷, Mohsen Mazidi⁶, Olatz Mompeo⁶, Mireia Valles-Colomer¹, Adrian Tett¹, Francesco Beghini¹, Léonard Dubois¹, Davide Bazzani¹, Andrew Maltez Thomas¹, Chloe Mirzayi⁸, Asya Khleborodova⁸, Sehyun Oh⁸, Rachel Hine⁷, Christopher Bonnett⁷, Joan Capdevila Pujol⁷, Serge Danzanvilliers⁷, Francesca Giordano⁷, Ludwig Geistlinger⁸, Levi Waldron⁸, Richard Davies⁷, George Hadjigeorgiou⁷, Jonathan Wolf⁷, José M. Ordovás^{9,10}, Christopher Gardner¹¹, Paul W. Franks^{12,13}, Andrew T. Chan^{5,13,14,*}, Curtis Huttenhower^{13,14,*}, Tim D. Spector^{6,*}, Nicola Segata^{1,15,*,#}

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. Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
6. Dept of Twin Research, King's College London, London UK
7. Zoe Global Ltd, London, UK
8. City University of New York, New York, NY, US
9. JM-USDA-HNRCA, Tufts University, Boston, MA, USA
10. IMDEA Food Institute, CEI UAM+CSIC, Madrid, Spain
11. Stanford University, Stanford, CA, USA
12. Department of Clinical Sciences, Lund University, Malmö, Sweden
13. Harvard T.H. Chan School of Public Health, Boston, MA, USA
14. The Broad Institute of MIT and Harvard, Cambridge, MA, USA
15. IEO, European Institute of Oncology IRCCS, Milan, Italy

[^] 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 1,203 gut microbiomes from 1,098 individuals enrolled in the PREDICT 1 study, whose detailed long-term diet information, as well as hundreds of fasting and same-meal postprandial cardiometabolic blood marker measurements were available. We found many significant associations between microbes and specific nutrients, foods, food groups, and general dietary indices, driven especially by the presence and diversity of healthy and plant-based foods. Microbial biomarkers of obesity were reproducible across external publicly available cohorts, and in agreement with circulating blood metabolites that are indicators of cardiovascular disease risk. While some microbes such as *Prevotella copri* and *Blastocystis* spp., were indicators of favourable postprandial glucose metabolism, the overall microbiome composition was predictive for a large panel of cardiometabolic blood markers including fasting and postprandial glycaemic, lipaemic, and inflammatory indices. 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.

54 Introduction

55 Dietary contributions to health and chronic conditions, such as obesity, metabolic syndrome, cancer and
56 cardiovascular disease, are of universal importance. Obesity and associated mortality/morbidity have
57 risen dramatically over the past decades¹, with the gut microbiome implicated as one of several
58 potentially causal human-environment interactions^{2,3}. Surprisingly, the details of the microbiome's role in
59 obesity and cardiometabolic health have proven difficult to define reproducibly in large human
60 populations⁴, likely due to the complexity of habitual diets, the difficulty of measuring them at scale and
61 disentangling them from other lifestyle variables^{5,6}, and the personalized nature of the microbiome⁷.

62 To overcome these challenges, we launched the Personalised Responses to Dietary Composition Trial
63 (PREDICT 1) of diet-microbiome interactions in metabolic health⁸. PREDICT 1 included >1,000
64 participants profiled pre- and post- standardised dietary challenges using intensive in-clinic biometric and
65 blood measures, habitual dietary data collection, continuous glucose monitoring, and stool
66 metagenomics. The study was inspired by previous large-scale diet-microbiome interaction profiles,
67 identifying gut microbiome configurations and microbial taxa associated with postprandial glucose
68 responses^{9,10}, obesity-associated biometrics such as body mass index (BMI) and adiposity¹¹⁻¹³, and
69 blood lipids and inflammatory markers^{14,15}.

70 Results

71 *Large metagenomically-profiled cohorts with rich clinical, cardiometabolic, and dietary* 72 *information*

73 PREDICT 1^{8,16} is an intervention study of diet-microbiome-cardiometabolic interactions (see **Methods**),
74 including a discovery cohort in the UK (n=1,002) and a validation population in the US (n=100). We
75 collected demographic information, habitual diet data, cardiometabolic blood biomarkers, and
76 postprandial responses to standardized test meals in the clinic and in free-living settings^{8,16} (**Fig. 1A**).
77 At-home stool collection yielded 1,098 baseline and 105 follow-up microbiome samples (+14 days) that
78 were all shotgun sequenced, and then taxonomically and functionally profiled (**Fig. 1A, Methods**).

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

80 We first leveraged a unique subpopulation of 480 monozygotic and dizygotic twins and confirmed that
81 host genetics influences microbiome composition only to a limited extent¹⁷. Indeed, twin-pair microbiome
82 similarity remained substantially lower than intra-subject longitudinal similarity (day 0 vs. day 14, p<1e-
83 12, **Extended Data Fig. 1B**), a testament to the personalized nature of the gut microbiome attributable to
84 non-genetic factors (**Extended Data Fig. 1C-D**).

85 We then investigated overall intra-sample (alpha) microbiome diversity as a broad summary statistic of
86 microbiome structure and found it was significantly associated (q<0.05) in 56 of the 295 tested
87 correlations with personal characteristics, habitual diet, and metabolic indices (**Fig. 1B, Supplementary**
88 **Table 1A**). BMI, visceral fat measurements, and probability of fatty liver (using a validated prediction
89 model¹⁸) were inversely associated with species richness. Among clinical circulating measures, High-
90 Density Lipoprotein cholesterol (HDL-C) was positively correlated with species richness. Emerging
91 cardiometabolic biomarkers¹⁹ that are not routinely used clinically, including lipoprotein particle size
92 (diameter, “-D”), and GlycA (inflammatory biomarker; glycoprotein acetyls), were also associated
93 (positively or negatively) with microbiome richness. These results associating simple indicators such as
94 microbiome richness to cardiometabolic health indicators and diet, motivated our more detailed
95 investigations of specific gut microbiome components.

97 We assessed links between habitual diet and the microbiome using random forest (RF) models, each
98 trained on quantitative microbiome features to predict each dietary variable from food frequency
99 questionnaires (FFQs) (**Methods**). The performance of the models was quantified with ROC AUCs for
100 classification and with correlation for regression (**Methods**). Several foods and food groups exceeded
101 0.15 median Spearman's correlation over bootstrap folds (denoted as " ρ ") between predicted and FFQ-
102 estimated values (14.5%) and AUC>0.65 (10.8%; **Fig. 2A**). The strongest association was for coffee
103 (instant or ground) (ρ =0.43, AUC=0.8), with dose-dependent effects and validated in the US cohort (**Fig.**
104 **2D**). Tighter microbiome links were found for energy-adjusted nutrients (**Fig. 2A**) with almost one-third
105 (**Supplementary Table 2**) showing correlations above 0.3.

106 We then summarized constituent foods into dietary indices (**Supplementary Table 2**), including the
107 Healthy Food Diversity index (incorporating dietary diversity and food quality, HFD) ²⁰, the
108 Healthy/Unhealthy Plant-based Dietary Indices (considering quality and quantity of plant-based foods,
109 hPDI/uPDI), Healthy Eating Index (extent of alignment with dietary guidelines, HEI) ²¹ and the Alternate
110 Mediterranean Diet score (aMED) ²², all of which are associated with reduced risk of chronic disease <sup>22-
111 27</sup>. We demonstrated tight correlations between microbial composition and the HFD, hPDI/uPDI and HEI
112 in the UK (ρ between 0.31 and 0.37, **Fig. 2A**) and the results were consistent in the US validation cohort,
113 with ρ reaching 0.42 for HFD and 0.31 for aMED (**Fig. 2E-F** and **Extended Data Fig. 3**), highlighting the
114 relationship between the microbiome and health-associated dietary patterns.

115 **Microbial species segregate into groups associated with more healthy and less healthy plant- and** 116 **animal- based foods**

117 We proceeded to identify the specific microbial taxa most responsible for these diet-based community
118 associations (**Fig. 2B**). After adjusting for age and BMI, we found 42 species (24% of those at >20%
119 prevalence) significantly correlated with at least five dietary exposures (q <0.2, **Supplementary Table 5**).
120 This included expected associations (**Extended Data Fig. 2**), such as enrichment of the probiotic taxa
121 *Bifidobacterium animalis* ²⁸ and *Streptococcus thermophilus* with greater full-fat yogurt consumption
122 (ρ =0.22 for both). The strongest food/microbe association was between the recently characterized
123 butyrate-producing *Lawsonibacter asaccharolyticus* ²⁹ and coffee consumption (**Fig. 2B**). However, due
124 to the low resolution of FFQ data, the complexity of dietary patterns, nutrient-nutrient interactions, and
125 clustering of 'healthy'/'less-healthy' food items, it is challenging to disentangle the independent
126 associations of single foods with microbial species.

127 At a broader level, we found clear segregation of species (**Fig. 2B**) into two distinct clusters with either
128 more healthy plant-based foods (e.g. spinach, seeds, tomatoes, broccoli) or with less healthy plant-
129 based (e.g. juices, sweetened beverages, refined grains) and animal-based foods, as defined by the PDI
130 ³⁰ (**Supplementary Table 4**). Taxa linked to healthy plant-based foods (**Fig. 2B-C**, **Extended Data Fig.**
131 **2**) mostly included butyrate producers, such as *Roseburia hominis*, *Agathobaculum butyriciproducens*,
132 *Faecalibacterium prausnitzii*, and *Anaerostipes hadrus*, as well as uncultivated species, predicted to
133 have this metabolic capability (*Roseburia* CAG182 and *Firmicutes* CAG95). Clades correlating with
134 several 'less-healthy' plant-based and animal-based foods included several *Clostridium* species
135 (*Clostridium innocuum*, *C. symbiosum*, *C. spiroforme*, *C. leptum*, *C. saccharolyticum*). The segregation of
136 species according to animal-based 'healthy' foods (e.g. eggs, white and oily fish) or animal-based 'less-
137 healthy' foods (e.g. meat pies, bacon, dairy desserts) using a novel categorisation (**Methods**), was also
138 distinct and overlapping with taxa signatures for 'healthy' and 'less-healthy' plant foods (**Fig. 2C** and

139 **Extended Data Fig. 2).** The few foods not fitting into the 'healthy' cluster despite being classified as
140 'healthy plant' foods, were (ultra) processed foods³¹ (e.g. sauces, baked beans; **Extended Data Fig. 2**).
141 This emphasises the importance of food quality (e.g. highly processed vs. unprocessed), food source
142 (e.g. plant vs. animal), and food type (i.e. not all plant foods are healthy) both in overall health and in
143 microbiome ecology.

144 ***The strongest microbiome-habitual diet associations are driven by poorly characterized microbes***

145 Many of the strongest microbial associations with diet occurred with only recently-isolated or still
146 uncultured taxa including five species defined using co-abundance gene groups (CAGs) from
147 metagenomics³². Among indices, the hPDI significantly correlated with 60 of the 176 prevalent species,
148 highlighting together with HFD (**Fig. 2E**) the impact of dietary diversity and quality on gut microbial
149 responsiveness. Among other dietary indices and nutrients, we observed general concordance with the
150 two sets of microbes associated with healthy and less-healthy foods. A greater animal-based food score
151 (definition in **Supplementary Table 4**), was associated with the 'healthy' cluster (**Fig. 2C** and **Extended**
152 **Data Fig. 2**), suggesting that a diet rich in healthier animal-based foods is associated with the more
153 favourable diet-microbiome signature, although this may also reflect an overall healthier dietary pattern.
154 The healthy and unhealthy PDI, which differentially affect disease risk^{25,30} also had distinct clusters,
155 again emphasising the oversimplification of conventional plant and animal-based food groupings. The
156 taxa with the highest correlations in the two clusters are *Firmicutes* CAG95 and *Clostridium leptum* for
157 healthy and unhealthy diet, respectively. The lack or paucity of cultivated representatives for these two
158 taxa may explain why these links were previously overlooked^{9,12}. The US validation cohort generally
159 confirmed these associations despite its smaller sample size: among the subset of derived pattern/index
160 scores shared between the UK and US cohorts, of the 54 associations that were significant both in the
161 UK cohort (FDR $q < 0.2$) and in the US cohort ($p < 0.05$), 70.4% were concordant.

162 ***Microbial indicators of obesity are reproducible across varied populations***

163 Microbiome links to obesity have attracted much interest although results have varied in human
164 populations^{3,4}. Our machine learning (ML) approach (**Methods**) found visceral fat to be more strongly
165 linked to gut microbial composition than BMI³³, a finding again validated in US participants (**Fig. 3A**).
166 Some obesity-associated taxa were also indicators of poor dietary patterns after controlling for BMI (e.g.
167 *Clostridium* CAG58, *Flavonifractor plautii*), whereas markers of lower visceral fat mass (e.g.
168 *Faecalibacterium prausnitzii*) were more strongly linked to healthier foods and patterns of intake,
169 illustrating that diet and obesity microbiome signatures overlap but are not identical (**Fig. 3B**).

170 Microbiome models to predict BMI in the UK cohort were further validated in six independent datasets
171 available in *curatedMetagenomicData*³⁴ (**Methods**). Despite inter-population differences^{11,35}, the UK
172 model improved cohort-specific cross-validation accuracy in the majority of cases, on par with the leave-
173 one-dataset-out approach (**Fig. 3D**). Of the 17 species surpassing $q < 0.05$, three had an (absolute) $\rho > 0.1$
174 in the US cohort and two of these were concordant with those in the UK cohort (**Fig. 3C**). Across
175 harmonized independent datasets, all but two median associations were consistent with the PREDICT 1
176 UK signatures, and 12 of the 14 were concordant despite different sample collection and DNA extraction
177 methods.

179 ***Fasting cardiometabolic markers associated with specific microbiome structures***

180 To explore the connections between the gut microbiome and cardiometabolic health, we developed and
181 evaluated microbiome-based ML models for each selected clinical and emerging cardiometabolic
182 biomarker. We found modest concordance between microbiome models and several traditional clinical

fasting cardiometabolic biomarkers (**Fig. 4A**) including blood pressure, lipids (TG, TC, HDL-C, LDL-C), fasting glucose, and glycosylated haemoglobin (HbA1c%) as well as a clinical prediction score estimating latent 10-year risk of heart disease (ASCVD score)³⁶.

For other blood biomarkers (**Fig. 1A**), we found stronger correlations between the microbiome and an inflammatory surrogate (GlycA, **Fig. 4A**), circulating polyunsaturated fatty acids (PUFA) (both omega-6 [FA_{ω6}/FA] and total PUFA [PUFA/FA] to total fatty acid ratios, $\rho=0.3$ and 0.32 , respectively), as well as emerging lipid measures linked to host health, including HDL and VLDL particle size (-D, $\rho=0.29$ for both) and the lipid content of lipoprotein subfractions (including XL-HDL-L and L-HDL-L, $\rho=0.3$ and 0.28 respectively). GlycA and VLDL-D are associated with increased risk for metabolic syndrome, CVD, and T2D, whereas HDL-D and its lipid constituents, omega-6, and total PUFA have inverse associations^{37,38}. Similarly, the majority of glycaemic indicators such as insulin and C-peptide were also coupled to human gut microbiome composition ($\rho=0.17$ and 0.22 , respectively), as well as derived predictors of insulin sensitivity (QUICKI, $\rho=0.22$)³⁹ and hepatic steatosis (Liver Fat Probability, $\rho=0.2$).

Species-based predictors proved more accurate than pathway abundance profiles (**Extended Data Fig. 4A**), consistent with other reports⁴⁰. Our primary findings were generally replicated in the US cohort (**Fig. 4A**), corroborating the existence of a strong, previously overlooked link between the gut microbiome and surrogates of cardiometabolic health.

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

Fasting blood assays are standard for research and clinical investigations; however, individuals consume multiple mixed-nutrient meals throughout the day and spend most of their waking hours in the postprandial state, resulting in repeated elevations in circulating TG, glucose and related metabolites⁸. Whilst postprandial glucose responses may, in part, be predicted by the gut microbiome⁹, ‘real-life’ variations in both postprandial lipid and glucose-mediated metabolites have not been explored. We assessed them by considering the overall magnitude of the response by iAUC, peak concentrations, and change from fasting (i.e. rise).

Firstly, we measured postprandial TG, glucose, C-peptide, insulin, and circulating metabolite concentrations at regular intervals (0-6h) in the clinic after two 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. 4B-C**). Notably, we found that postprandial TG (0-6h iAUC), insulin, and C-peptide (both 0-2h iAUC) responses were more strongly associated with the gut microbiome ($\rho=0.15$, 0.2 , 0.24 , respectively; AUCs >0.65) compared with postprandial glucose (0-2h iAUC) responses ($\rho=0.13$ and AUC=0.6, **Fig. 4B**), findings replicated in our US cohort (**Fig. 4B-G**). We also measured glucose concentrations during the 13-day at-home period¹⁶ following isocaloric standardized meals, with different macronutrient compositions (**Supplementary Table 3**). However, contrary to our clinic meal responses (**Fig. 4B**) and previous work⁹, the glucose 0-2h iAUCs following these meals did not achieve high correlations with the microbiome (all $\rho<0.07$ and AUC <0.59 , **Fig. 4C**). Whilst this may be dependent on meal composition and the effect of multiple meals consumed after stool collection, these results suggest that the microbiome is a stronger predictor of postprandial lipaemia than glycaemia.

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

224 Postprandial measures depend both on the corresponding fasting levels and the meal-induced rise.
225 Therefore, we compared the differential prediction accuracy of the gut microbiome for fasting levels,
226 postprandial (peak) total levels, and postprandial rises (**Fig. 4H**). For lipid and glucose-mediated (clinic
227 day) measures, despite a similar strength of association between peak (6h), magnitude (iAUC) and
228 fasting TG concentrations, the rise (6-0h) was not similarly correlated (**Fig. 4A-E-F**). In contrast, the
229 microbiome associations with glycaemic measures were comparable between fasting, peak, and rise
230 (**Fig. 4A-D**).

231 Of particular interest were the lipoprotein subfraction concentrations, composition, and size (**Extended**
232 **Data Fig. 4B-C**), which are remodelled postprandially into potentially atherogenic lipoproteins (e.g. Large
233 VLDL particles and TG-enriched LDL, and HDL particles)⁴¹. These particles were predicted at
234 comparable accuracy for both fasting and postprandial peak 6h concentrations (**Fig. 4A-F-H**), and
235 notably, HDL-D and VLDL-D achieve modestly stronger correlations ($\rho=0.32$ for both) postprandially
236 (**Fig. 4F**). However, as with TG, we found that the microbiome was substantially less predictive for the
237 postprandial rise in all lipid metabolite measures compared with fasting and postprandial peak
238 concentration (**Fig. 4A-F-H**). For example, HDL-D is closely associated with gut microbial composition at
239 fasting and 6h postprandially ($\rho=0.29$ and 0.32 ; AUC 0.71 and 0.72 respectively; **Fig. 4A-F-H**), but not
240 with the rise (**Fig. 4F**). These differential associations suggest that the microbiome may influence
241 postprandial lipid-mediated measures via effects on fasting measures.

242 **Distinct microbial signatures discriminate between positive and negative metabolic health**
243 **indices under fasting conditions**

244 Motivated by the observed potential of the gut microbiome to predict the fasting and postprandial levels
245 of circulating metabolic markers, we next assessed the microbiome features driving these associations.
246 Among three general risk indices of cardiovascular health (ASCVD, liver fat probability, and insulin
247 sensitivity or QUICKI, **Fig. 4A**), we found six species significantly and concordantly correlated with all
248 three (negatively or positively, $p<0.05$), hinting at a global underlying microbial signature of improved
249 metabolic health. These taxa included *Clostridium* CAG58 (higher cardiometabolic risk) and
250 *Haemophilus parainfluenzae* (lower risk) that we had previously linked with healthy and less-healthy
251 dietary patterns (**Fig. 2B**).

252 We found similarly distinct separations between two opposing clearly defined clusters of species either
253 positively or negatively correlated with fasting cardiometabolic measures (**Fig. 5A-B**). Species correlated
254 with positive markers included some prevalent taxa generally regarded as healthy (*F. prausnitzii*) but also
255 eight uncultivated and under-characterized bacteria. The positive cluster included many distinct genera,
256 pointing at a rich functional diversity. In contrast, the cluster negatively correlated with positive markers
257 included eight *Clostridium* species and the recurrent negatively connotated *R. gnavus* and *F. plautii*.
258 Large HDL particles (and their lipid compositions, **Extended Data Figs. 6 and 7**), which have strong
259 inverse associations with cardiometabolic outcomes³⁸, were associated with the healthy cluster.
260 Conversely, lipoproteins associated with increased risk of CVD and T2D (VLDL of all sizes and lipid
261 composition) and atherogenicity⁴² (S-LDL, M-HDL and S-HDL TG), were associated with the less-
262 healthy cluster (**Extended Data Figs. 6 and 7**).

263 Circulating omega-6 and total polyunsaturated fatty acids (PUFA), were associated with the healthy
264 cluster (**Fig. 5A** and **Supplementary Table 5**). Due to the lack of endogenous production of PUFA,
265 circulating levels closely reflect dietary intake⁴³ and are linked to a reduced risk of chronic disease³⁸. In

contrast, circulating monounsaturated fatty acids (MUFA) (which do not closely reflect dietary intake and unlike dietary MUFA have been linked to increased risk of chronic disease³⁸), were associated with the unhealthy cluster, with an under-characterized *Firmicutes* species (CAG 170) and *Clostridium bolteae* responsible for the strongest negative and positive associations respectively.

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 **Extended Data Figs. 6** and **7**). Notably, fasting and postprandial GlycA were strongly linked with the microbiome (47 species significantly correlated at 6 hours and 64 at fasting), substantially exceeding IL-6 (5 and 16 significant associations, **Fig. 5B**). *C. boltae* and *R. gnavus* were the most correlated with increased fasting and postprandial inflammation, 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 in some cases uncoupled from the microbial associations with fasting markers (**Fig. 5D**). For example, the change in GlycA (**Fig. 5B**) was differentially associated with clusters compared to fasting and postprandial levels (especially for *F. plautii*, *Firmicutes* CAG95, and *Firmicutes* CAG110); likely due to the small reduction in GlycA postprandially. Other immunological markers and some lipid and cholesterol levels paralleled this behaviour (**Extended Data Fig. 6**), possibly reflecting postprandial lipoprotein remodelling⁴⁴.

We observed the same “favourable” vs. “unfavourable” clustering of microbiome features when analyzing microbial pathways and gene families (**Extended Data Fig. 8**) supporting taxa segregation by their underlying biochemical activities. The strengths of microbe-blood marker associations were confirmed by RF feature relevance analysis (**Extended Data Fig. 9**) and, importantly, confirmed in the US cohort. For the 209 microbe-index correlations 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 88.7% for the associations in fasting conditions and 96.1% postprandially.

***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 in our microbial signatures⁴⁵. Among them, *Prevotella copri*^{45,46} had conflicting previous accounts for its role in glucose homeostasis^{47,48} possibly due to subspecies diversity^{49,50}. Our data find *P. copri* to be associated with beneficial cardiometabolic markers, being negatively correlated with estimated visceral fat ($\rho = -0.11$, $p = 0.0006$), fasting VLDL-D ($\rho = -0.08$, $p = 0.011$), and fasting GlycA ($\rho = -0.14$, $p < 0.0001$) among others (**Supplementary Table 5**). While almost no diet indices were associated with *P. copri*, postprandial rises in glucose ($\rho = -0.11$, $p < 0.001$) and polyunsaturated/omega-6 fatty acids ($\rho = 0.15$ and 0.14 , respectively, and $p < 0.001$) were top-scoring correlations for this bacterium and were stronger than corresponding fasting and postprandial levels, in contrast with what we observed for the overall microbiome (**Fig. 4A-B**). *P. copri* is present in at least one of its subtypes⁴⁹ in 29.8% of the PREDICT 1 individuals, and *P. copri*-carriers had lower C-peptide (-9.2%, $p = 0.002$), insulin (-14%, $p = 0.006$), and TG levels (-3.2%, $p = 0.003$) compared to *P. copri*-negative individuals (**Extended Data Fig. 10** and **Supplementary Table 8**). Similarly, postprandial glucose after

breakfast was significantly less pronounced in individuals with *P. copri* (-20.4% glucose iAUC at 2h, $p=0.002$, **Extended Data Fig. 10C**), and visceral fat was significantly lower (-12.5%, $p=0.005$, **Extended Data Fig. 10A**). This set of diverse associations supports that the presence of *P. copri* in the gut microbiome could be beneficial in glucose homeostasis and host metabolism.

Blastocystis is a unicellular eukaryotic parasite increasingly regarded as a commensal member of the gut microbiome^{51–53}. It shares with *P. copri* a limited prevalence in Western-lifestyle populations⁵³ and a high abundance when present. We found evidence that *Blastocystis*-positive individuals (25.7% in our cohort) also have favourable glucose homeostasis and lower estimated visceral fat (-15.7% glucose iAUC, -22.1% visceral fat, $p<0.002$, **Extended Data Fig. 10**). The latter confirms that *Blastocystis* is less prevalent in individuals with high BMI, as previously suggested⁵³. Interestingly, the effect of the simultaneous presence of *P. copri* and *Blastocystis* (12% of the individuals) appears to further promote healthier metabolic function. Visceral fat is 17.3% lower on average ($p<0.005$, **Supplementary Table 8**) for individuals positive for both *P. copri* and *Blastocystis* compared to individuals with only one or the other and 23.3% lower ($p=8.9E-6$) compared with individuals lacking both.

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

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

In particular, *Firmicutes* CAG95 is the uncultivated species with the most beneficial score. Of the “health”-associated microbial species only *F. prausnitzii*, and - partially - *P. copri* were already convincingly linked with health in previous investigations⁵⁴. The beneficial signature also included *E. eligens* and *H. parainfluenzae* without previous clear roles in health, and additional species without cultivated representatives such as *Roseburia* CAG182, *Oscillibacter* sp 57_20, *Firmicutes bacterium* CAG170, *Oscillibacter* sp PC13, and *Clostridium* sp CAG167. Species conversely consistent with indicators of poor overall health (**Fig. 6**) included the already discussed set of Clostridia (*C. spiroforme*, *C. bolteae* CAG59, *C. bolteae*, *Clostridium* sp CAG58, *C. symbiosum*, *C. innocuum*, *C. leptum*) and the mucolytic microbe *R. gnavus* and *F. plautii*, again previously found to be associated with disease conditions^{55,56}. Overall, this set of 30 species serves as a marker of overall good or poor cardiometabolic health and dietary patterns in non-diseased human hosts.

Discussion

PREDICT 1 represents the first diet-microbiome study to identify both individual components of the microbiome and an overall gut microbial signature associated with multiple measures of dietary intake and cardiometabolic health. These signatures reproduced across UK and US populations, across multiple previously-published study populations, and for multiple dietary and health indicators. Notably, microbiome signatures grouped both microbiome and dietary components into health-associated and anti-associated clusters, the latter in agreement with dietary quality and diversity scores^{20,57}. The diversity and quality of a healthy diet (HFD and PDI) was particularly predictable by the microbiome, surpassing other indices including the Mediterranean diet previously linked with microbiome composition⁵⁸. 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

355 patterns highlights the importance of looking beyond nutrients and single foods in diet-microbiome
356 research. The substantially greater detail and consistency in our results relative to prior diet-microbiome
357 work ^{9,11–13,15} may be due to the quality in the dietary recording, metagenomic profiling, and the large
358 sample size. However, given the limitations of FFQ dietary data, future diet-microbiome studies would
359 benefit further from higher resolution dietary assessment methodologies such as weighed food record
360 data.

361 Several aspects of the consistent gut microbiome signatures across diet, obesity, and cardiometabolic
362 health measures are striking for their potential novel epidemiology and microbial biochemistry. A
363 surprising proportion of diet- or health-associated taxa in these results are largely uncharacterized or
364 represented solely by metagenomic assemblies ⁵. Other microbes found here to have dietary or
365 cardiometabolic associations, such as *Prevotella* spp. or *Blastocystis* spp., have been characterized in
366 greater biochemical detail, but their population structure in the human microbiome have only recently
367 begun to be appreciated ^{49,53}. The latter in particular may be only one of many examples of non-bacterial
368 microbiome members not amenable to most current high-throughput approaches, but with unexpected
369 and potentially key positive roles in humans.

370 Likewise, these new contributions of the gut microbiome to human dietary responses may help to explain
371 some of the heterogeneity seen among previous population studies ^{4,9,59}. First, diet-microbiome-blood
372 marker associations were overall strongest with respect to circulating lipid levels relative to glycemic
373 indices. It is possible that the relative contribution of gut microbes is higher for circulating lipid levels than
374 carbohydrate derivatives, through either direct processes or indirectly through gastrointestinal or
375 systemic bile acid signalling ⁶⁰. Alternatively, host metabolism may play a greater role in circulating
376 glucose and insulin levels relative to microbial bioactivity. The lipoprotein features most closely
377 associated with the microbiome (such as L-HDL-L) are also more strongly associated with cardiovascular
378 risk compared with typically measured lipids (e.g. TC, HDL-C, LDL-C), suggesting that their utility as
379 clinical biomarkers or as targets for beneficial gut microbiome manipulation warrants further investigation.

380 Overall, this is the first study to identify a shared diet-metabolic-health microbial signature, segregating
381 favourable and unfavourable taxa with multiple measures of both dietary intake and cardiometabolic
382 health. These results will aid as a resource both in utilization of the gut microbiome as a biomarker for
383 cardiometabolic risk and in strategies for reshaping the microbiome to improve personalized dietary
384 health.
385

Authors contribution
Obtained funding: JW, GH, TDS. Study design and developed concept: SEB, AMV, JW, GH, HAK, RD, ATC, NS, PWF, TDS. Data collection: SEB, NS, FA, HAK, ATC, DAD, TDS. Data analysis: FA, SEB, NS, LF, EL, RG, MM, OM, GP, CLR, MV-C, SO, FG, AT, FB, CM, AK, LD, DB, AMT, CB, LW, LG, JCP, SD, RH. Study coordination: SEB, HAK, 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.

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

Acknowledgements
We express our sincere thanks to the participants of the PREDICT 1 study. We thank Naeimeh Atabaki-Pasdar for generating the liver fat score. We thank the staff of Z. Global, the Department of Twin Research and Massachusetts General Hospital, and all the members of the Segata, Berry and Spector labs for their tireless work in contributing to the running of the study, data collection, and data processing. We thank Nightingale Health and Affinity Biomarker Labs for their support and analytical work.

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; by the European H2020 program (ONCOBIOME-825410 project and MASTER-818368 project) to NS; by the National Cancer Institute of the National Institutes of Health (1U01CA230551) to NS; and by the Premio Internazionale Lombardia e Ricerca 2019 to N.S. SEB was supported in part by a grant funded by the BBSRC (BB/NO12739/1). PWF was supported in part by grants from the European Research Council (CoG-2015_681742_NASCENT), Swedish Research Council, SSF (IRC15-0067) and Novo Nordisk Foundation. ATC was supported in part as a Stuart and Suzanne Steele MGH Research Scholar. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, Chronic Disease Research Foundation (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.

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

429 **Code availability statement**

430 Computational analyses were performed using the bioBakery suite tools and species-level microbial
431 abundances were computed using MetaPhlAn 3.0 (<https://github.com/biobakery/MetaPhlAn>) and
432 functional potential profiling with HUMAnN 2.0 (<https://github.com/biobakery/humann>, see **Methods**).

433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488

References

1. Ng, M. *et al.* Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **384**, 766–781 (2014).
2. Brown, J. M. & Hazen, S. L. Microbial modulation of cardiovascular disease. *Nat. Rev. Microbiol.* **16**, 171–181 (2018).
3. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
4. Sze, M. A. & Schloss, P. D. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *MBio* **7**, (2016).
5. Pasoli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
6. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
7. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
8. Berry, S. *et al.* Human postprandial responses to food and potential for precision nutrition. *Nature Medicine* (2020).
9. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
10. Mendes-Soares, H. *et al.* Model of personalized postprandial glycemic response to food developed for an Israeli cohort predicts responses in Midwestern American individuals. *Am. J. Clin. Nutr.* **110**, 63–75 (2019).
11. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
12. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
13. Thingholm, L. B. *et al.* Obese Individuals with and without Type 2 Diabetes Show Different Gut Microbial Functional Capacity and Composition. *Cell Host Microbe* **26**, 252–264.e10 (2019).
14. Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1897 (2016).
15. Fu, J. *et al.* The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids. *Circ. Res.* **117**, 817–824 (2015).
16. Berry, S. *et al.* Personalised REsponses to Dietary Composition Trial (PREDICT): an intervention study to determine inter-individual differences in postprandial response to foods. *Protocol Exchange* (2020).
17. Xie, H. *et al.* Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Syst* **3**, 572–584.e3 (2016).
18. Atabaki-Pasdar, N. *et al.* Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts. *PLoS Med.* **17**, e1003149 (2020).
19. Vojinovic, D. *et al.* Relationship between gut microbiota and circulating metabolites in population-based cohorts. *Nat. Commun.* **10**, 5813 (2019).
20. Vadiveloo, M., Dixon, L. B., Mijanovich, T., Elbel, B. & Parekh, N. Development and evaluation of the US Healthy Food Diversity index. *Br. J. Nutr.* **112**, 1562–1574 (2014).
21. Guenther, P. M. *et al.* Update of the Healthy Eating Index: HEI-2010. *J. Acad. Nutr. Diet.* **113**, 569–580 (2013).
22. Fung, T. T. *et al.* Diet-quality scores and plasma concentrations of markers of inflammation and endothelial dysfunction--. *Am. J. Clin. Nutr.* **82**, 163–173 (2005).
23. Reedy, J. *et al.* Higher diet quality is associated with decreased risk of all-cause, cardiovascular disease, and cancer mortality among older adults. *J. Nutr.* **144**, 881–889 (2014).
24. Mitrou, P. N. *et al.* Mediterranean dietary pattern and prediction of all-cause mortality in a US population: results from the NIH-AARP Diet and Health Study. *Arch. Intern. Med.* **167**, 2461–2468 (2007).
25. Satija, A. *et al.* Plant-Based Dietary Patterns and Incidence of Type 2 Diabetes in US Men

- and Women: Results from Three Prospective Cohort Studies. *PLoS Med.* **13**, e1002039 (2016).
26. Vadiveloo, M., Parekh, N. & Mattei, J. Greater healthful food variety as measured by the US Healthy Food Diversity index is associated with lower odds of metabolic syndrome and its components in US adults. *J. Nutr.* **145**, 564–571 (2015).
27. Onvani, S., Haghighatdoost, F., Surkan, P. J., Larijani, B. & Azadbakht, L. Adherence to the Healthy Eating Index and Alternative Healthy Eating Index dietary patterns and mortality from all causes, cardiovascular disease and cancer: a meta-analysis of observational studies. *J. Hum. Nutr. Diet.* **30**, 216–226 (2017).
28. Redondo-Useros, N. *et al.* Associations of Probiotic Fermented Milk (PFM) and Yogurt Consumption with Bifidobacterium and Lactobacillus Components of the Gut Microbiota in Healthy Adults. *Nutrients* **11**, (2019).
29. Sakamoto, M., Iino, T., Yuki, M. & Ohkuma, M. *Lawsonibacter asaccharolyticus* gen. nov., sp. nov., a butyrate-producing bacterium isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **68**, 2074–2081 (2018).
30. Satija, A. *et al.* Healthful and Unhealthful Plant-Based Diets and the Risk of Coronary Heart Disease in U.S. Adults. *J. Am. Coll. Cardiol.* **70**, 411–422 (2017).
31. Monteiro, C. A. *et al.* The UN Decade of Nutrition, the NOVA food classification and the trouble with ultra-processing. *Public Health Nutr.* **21**, 5–17 (2018).
32. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
33. Beaumont, M. *et al.* Heritable components of the human fecal microbiome are associated with visceral fat. *Genome Biol.* **17**, 189 (2016).
34. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
35. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
36. D'Agostino, R. B., Sr *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
37. Kettunen, J. *et al.* Biomarker Glycoprotein Acetyls Is Associated With the Risk of a Wide Spectrum of Incident Diseases and Stratifies Mortality Risk in Angiography Patients. *Circ Genom Precis Med* **11**, e002234 (2018).
38. Würtz, P. *et al.* Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* **131**, 774–785 (2015).
39. Hřebíček, J., Janout, V., Malincíková, J., Horáková, D. & Cízek, L. Detection of insulin resistance by simple quantitative insulin sensitivity check index QUICKI for epidemiological assessment and prevention. *J. Clin. Endocrinol. Metab.* **87**, 144–147 (2002).
40. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
41. Wojczynski, M. K. *et al.* High-fat meal effect on LDL, HDL, and VLDL particle size and number in the Genetics of Lipid-Lowering Drugs and Diet Network (GOLDN): an interventional study. *Lipids Health Dis.* **10**, 181 (2011).
42. Skeggs, J. W. & Morton, R. E. LDL and HDL enriched in triglyceride promote abnormal cholesterol transport. *J. Lipid Res.* **43**, 1264–1274 (2002).
43. Hodson, L., Skeaff, C. M. & Fielding, B. A. Fatty acid composition of adipose tissue and blood in humans and its use as a biomarker of dietary intake. *Prog. Lipid Res.* **47**, 348–380 (2008).
44. Cohn, J. S. Postprandial lipemia: emerging evidence for atherogenicity of remnant lipoproteins. *Can. J. Cardiol.* **14 Suppl B**, 18B–27B (1998).
45. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
46. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
47. Kovatcheva-Datchary, P. *et al.* Dietary Fiber-Induced Improvement in Glucose Metabolism Is Associated with Increased Abundance of *Prevotella*. *Cell Metab.* **22**, 971–982 (2015).
48. Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin

- sensitivity. *Nature* **535**, 376–381 (2016).
49. Tett, A. *et al.* The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* (2019) doi:10.1016/j.chom.2019.08.018.
50. De Filippis, F. *et al.* Distinct Genetic and Functional Traits of Human Intestinal *Prevotella copri* Strains Are Associated with Different Habitual Diets. *Cell Host Microbe* **25**, 444–453.e3 (2019).
51. Clark, C. G., van der Giezen, M., Alfellani, M. A. & Stensvold, C. R. Recent developments in Blastocystis research. *Adv. Parasitol.* **82**, 1–32 (2013).
52. Lukeš, J., Stensvold, C. R., Jirků-Pomajbíková, K. & Wegener Parfrey, L. Are Human Intestinal Eukaryotes Beneficial or Commensals? *PLoS Pathog.* **11**, e1005039 (2015).
53. Beghini, F. *et al.* Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. *ISME J.* **11**, 2848–2863 (2017).
54. Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16731–16736 (2008).
55. Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).
56. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat Microbiol* **4**, 623–632 (2019).
57. Kim, H., Caulfield, L. E. & Rebholz, C. M. Healthy Plant-Based Diets Are Associated with Lower Risk of All-Cause Mortality in US Adults. *J. Nutr.* **148**, 624–631 (2018).
58. Meslier, V. *et al.* Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. *Gut* (2020) doi:10.1136/gutjnl-2019-320438.
59. Kurilshikov, A. *et al.* Gut Microbial Associations to Plasma Metabolites Linked to Cardiovascular Phenotypes and Risk. *Circ. Res.* **124**, 1808–1820 (2019).
60. Ko, C.-W., Qu, J., Black, D. D. & Tso, P. Regulation of intestinal lipid metabolism: current concepts and relevance to disease. *Nat. Rev. Gastroenterol. Hepatol.* (2020) doi:10.1038/s41575-019-0250-7.

Figure's legends

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). The five strongest positive and negative Spearman's correlations with $q < 0.05$ are reported for each of the four categories. Top species based on Shannon diversity are reported in **Extended Data Fig. 1A** and all correlations, p-values, and q-values are reported in **Supplementary Table 1**.

Fig. 2: Food quality, regardless of source, is linked to overall and feature-level composition of the gut microbiome. (A) Specific components of habitual diet comprising foods, nutrients, and dietary indices are linked to the composition of the gut microbiome with variable strengths as estimated by machine learning regression and classification models. Box plots report the correlation between the real value of each component and the value predicted by regression models across 100 training/testing folds (**Methods**). 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's 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 (NSP: non-starch polysaccharides). Rows and columns are hierarchically clustered (complete linkage, Euclidean distance). Full heatmaps of foods and unadjusted nutrients are reported in **Extended Data Fig. 2**, and the full set of correlations, p-values, and q-values are available in **Supplementary Table 5**. (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, and was validated in the US cohort when applying the UK model. The reported ROC curves represent the performance of the classifier at varying classification thresholds with respect to the True Positive Rate (i.e. recall) and the False Positive Rate (i.e. precision). (E-F) Among general dietary patterns and indices, the Healthy Food Diversity index (HFD) and the Alternate Mediterranean Diet score (aMED) were validated in the US cohort, thus showing consistency between the two populations on these two important dietary indices. Other validations of the UK model applied to the US cohort are reported in **Extended Data Fig. 3**. Box plots show first and third quartiles (boxes) and the median (middle line), whiskers extends up-to $1.5 \times$ the interquartile range.

Fig. 3: Random forest machine learning models trained on microbial or functional profiles are capable of predicting obesity phenotypic markers, even on independent cohorts. (A) Whole-microbiome machine learning models can assess personal factors with RF regression (box plots and left-side y-axis) using only taxonomic or functional (i.e. pathway) microbiome features. Classification models (circles and right-side y-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 datasets³⁴ we confirmed correlations between single microbial species and BMI with blue points denoting significant associations at $p < 0.05$. The statistical test used is a two-sided z-test (see **Methods**). (D) The machine learning model for BMI trained on PREDICT 1 data is reproducible in several external datasets (**Extended Data 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, confirming the generalizability of the PREDICT 1 model on obesity-related indicators. Box plots show first and third quartiles (boxes) and the median (middle line), whiskers extends up-to $1.5 \times$ the interquartile range.

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 box plots report the correlation between the prediction of RF regression models trained on microbial taxa or pathway abundances across 100 training/testing folds and stars report regressor performance when trained on the UK cohort and evaluated on the independent US validation cohort (left-side y-axis). Circles denote AUC values for RF classification (right-side y-axis) (B-F) 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; **Extended Data Fig. 4B-C**). (B) 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. (C) Glycaemic-mediated postprandial iAUCs at 2h for the other meals (**Supplementary Table 7**), and (D) glycaemic-mediated markers absolute levels vs. rise. (E) Postprandial inflammatory measures (concentration and rise). (F) Postprandial lipoproteins measures (6h concentration and rise). (G) Overall agreement between RF regression and classification tasks for UK models applied to the independent US

cohort. **(H)** RF microbiome-based model performance with postprandial changes (concentrations and rise) in lipoprotein concentration, composition, and size. Fasting and postprandial performance indices (correlation of the regressors' outputs) were more tightly linked to gut community structure than were their corresponding postprandial rises. Box plots show first and third quartiles (boxes) and the median (middle line), whiskers extends up-to 1.5× the interquartile range.

Fig. 5: Species-level segregation into healthy and unhealthy microbial signatures of fasting and postprandial cardiometabolic markers. **(A)** Associations (Spearman's 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. Rows are hierarchically clustered (complete linkage, Euclidean distance). **(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). The complete set of correlations, p-values, and q-values are available in **Supplementary Table 5**.

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 cardiometabolic health and healthy diet indicators, respectively, are shown here. The rank of each microbe's correlation with individual indicators is written within cells when significant ($p < 0.05$). For each of the main categories of indices, we selected up to five representative markers (for "Personal" we considered only four 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. Partial correlations were computed using `pcor.test` (two-sided) with params "method=spearman" (see **Methods**). Correlations and ranks available in **Supplementary Table 9**, p-values and FDR-adjusted p-values available in **Supplementary Table 2**.

Online 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. Ethical approval for the study was obtained in the UK from the Research Ethics Committee and Integrated Research Application System (IRAS 236407) and in the US from the Institutional Review Board (Partners Healthcare IRB 2018P002078). The trial was run in accordance with the Declaration of Helsinki and Good Clinical Practice. Study procedures were only carried out after having received written informed consent from each participant.

For full protocol, see Berry et al., 2020¹⁶. 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.⁸ 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

697 analyze fasting and postprandial concentrations of TG and C-peptide. Participants were supported
698 throughout the study with reminders and communication from study staff delivered through the ZOE
699 study app. A second stool sample was collected at home by participants following completion of the
700 study, and all devices and samples were mailed back to study staff. To monitor compliance, all test
701 meals consumed by participants were logged in the Zoe app (with an accompanying picture) and
702 reviewed in real-time by the study nutritionists. Only test meals that were consumed according to the
703 standardised meal protocol (outlined in Berry et al. 2020⁸) were included in the analysis.

704 The recruitment criteria, meal intervention challenges, outcome variables, and sample collection and
705 analysis procedures relevant to this paper are described elsewhere^{8,16}. The core characteristics of study
706 participants at baseline were not significantly different between UK and US cohorts⁸.

707 ***Overview of microbiome sequencing and profiling***

708 We performed deep shotgun metagenomic sequencing (mean 8.8±2.2 gigabases/sample) in stool
709 samples from a total of 1,098 PREDICT 1 participants (UK n=1,001; US n=97). From a random subset of
710 these participants (n=105), we additionally sequenced faecal metagenomes from a second stool sample
711 collected 14 days after the first collection (**Fig. 1A**) for a total of 1,168 metagenomes. Computational
712 analysis was performed using the bioBakery suite of tools⁶¹ to obtain species-level microbial
713 abundances for the 769 taxa identified using the newly updated MetaPhlAn 3.0 tool⁶², functional
714 potential profiling of >1.91 M microbial gene families, 445 KEGG pathways with HUMAnN 2.0⁶³, and
715 reconstruction of 48,181 metagenome-assembled genomes (MAGs) of medium or high-quality using our
716 validated pipeline⁵, which includes assembly with MegaHIT⁶⁴, binning with MetaBAT2⁶⁵, and quality-
717 control with CheckM⁶⁶.

718 ***Microbiome sample collection***

719 Participants were mailed a pre-visit study pack with a stool collection kit and relevant questionnaires and
720 asked to collect an at-home stool sample at two timepoints (one prior to their in-person clinical visit on
721 day 0 and the next at the conclusion of their home-phase, day 14). Those who did not collect a sample
722 prior to their in-person, baseline visit completed the collection as soon as possible during the home-
723 phase. Baseline samples in the UK were collected using the EasySampler collection kit (ALPCO, NH,
724 US), whereas post-study samples, as well as the entirety of the US collection was conducted using the
725 Fecotainer collection kit (Excretas Medical BV, Enschede, the Netherlands). For baseline samples, one
726 fresh unfixed sample was deposited into a sterile universal collection container (Sarstedt, Australia, Cat
727 #L0263-10) and one into a tube containing DNA/RNA Shield buffer (Zymo Research, CA, US, Cat
728 #R1101). Samples were stored at ambient temperature until returned to the study staff. Follow-up
729 samples were collected similarly, but only sampled into a DNA/RNA Shield buffer tube and sent by
730 standard mail to study staff. Upon receipt in the laboratory, samples were homogenized, aliquoted, and
731 stored at -80°C in Qiagen PowerBeads 1.5 mL tubes (Qiagen, Germany). This sample collection
732 procedure was tested and validated internally comparing different storage conditions (fresh, frozen,
733 buffer), different DNA extraction kits (PowerSoilPro, FastDNA, ProtocolQ, Zymo), and different
734 sequencing technologies (16S rRNA, shotgun metagenomics, and arrays), data not shown.

735 ***DNA extraction and sequencing***

736 DNA was isolated by QIAGEN Genomic Services using DNeasy® 96 PowerSoil® Pro from all Day 0
737 (baseline) DNA/RNA shield fixed microbiome samples. A random subset of Day 14 (end of at-home
738 phase) samples (n=105) were also extracted. Optical density measurement was done using
739 Spectrophotometer Quantification (Tecan Infinite 200). Before library preparation and sequencing, the
740 quality and quantity of the samples were assessed using the Fragment Analyzer (Agilent Technologies,
741 Inc.) according to manufacturer's guidelines. Samples with a high-quality DNA profile were further
742 processed. The NEBNext® Ultra II FS DNA module (cat# NEB #E7810S/L) was used for DNA

743 fragmentation, end-repair, and A-tailing. For adapter ligation, the NEBNext® Ultra II Ligation module
744 (cat# NEB #E7595S/L) was used. The quality and yield after sample preparation were measured with the
745 Fragment Analyzer. The size of the resulting product was consistent with the expected size of
746 approximately 500-700 bp. Libraries were sequenced for 300 bp paired-end reads using the Illumina
747 NovaSeq6000 platform according to manufacturer's protocols. 1.1 nM library was used for flow cell
748 loading. NovaSeq control software NCS v1.5 was used. Image analysis, base calling, and the quality
749 check was performed with the Illumina data analysis pipeline RTA3.3.5 and Bcl2fastq v2.20.

750 ***Metagenome quality control and pre-processing***

751 All sequenced metagenomes were QCed using the pre-processing pipeline as implemented in
752 <https://github.com/SegataLab/preprocessing>. Pre-processing consists of three main steps: (1) read-level
753 quality control; (2) screening of contaminants, i.e. host sequences; and (3) split and sorting of cleaned
754 reads. Initial quality control involves the removal of low-quality reads (quality score <Q20), fragmented
755 short reads (<75 bp), and reads with >2 ambiguous nucleotides. Contaminant DNA was identified using
756 Bowtie 2 (Langmead and Salzberg 2012) using the --sensitive-local parameter, allowing confident
757 removal of the phiX174 Illumina spike-in and human-associated reads (hg19). Sorting and splitting
758 allowed for the creation of standard forward, reverse, and unpaired reads output files for each
759 metagenome.

760 ***Microbiome taxonomic and functional potential profiling***

761 The metagenomic analysis was performed following the general guidelines⁶⁷ and relying on the
762 bioBakery computational environment⁶¹. The taxonomic profiling and quantification of organisms' relative
763 abundances of all metagenomic samples have been quantified using MetaPhlAn 3.0⁶². The updated
764 species-specific database of markers was built using 99,237 reference genomes representing 16,797
765 species retrieved from Genbank (January 2019). From this set of reference genomes, we extracted a
766 total of 1,132,166 markers used to profile 13,393 species. This set of species included also 83 species
767 defined by the co-abundant gene (CAG) groups approach³² that were very genetically distinct from
768 species represented by isolate genomes and for which the use of unique marker genes limited the
769 potential issues of using metagenomic assemblies reconstructed over multiple samples. Compared to
770 the previous version of the MetaPhlAn2 database (mpa_v20_m200), the updated database is able to
771 profile 7,116 more species. Metagenomes were mapped internally in MetaPhlAn 3.0 against the marker
772 genes database with BowTie2 version 2.3.4.3 with the parameter "very-sensitive". The resulting
773 alignments were filtered to remove reads aligned with a MAPQ value <5, representing an estimated
774 probability of the likelihood of the alignments.

775 For estimating the microbiome species richness of an individual from the taxonomic profiles of PREDICT
776 1 participants, we computed two alpha diversity measures: the number of species found in the
777 microbiome ("observed richness"), and the Shannon entropy estimation. We did not perform rarefaction
778 prior to alpha diversity calculations because of the low standard deviation in sequencing depths and the
779 verified missing correlation between metadata of interest and sequencing depth. Microbiome dissimilarity
780 between participants (beta diversity) was computed using the Bray-Curtis dissimilarity on microbiome
781 taxonomic profiles.

782 Functional potential analysis of the metagenomic samples was performed using HUMAnN2 (version
783 0.11.2 and UniRef database release 2014-07)⁶³ that computed pathway profiles and gene-family
784 abundances.

785 ***Metagenomic assembly***

786 Metagenomic samples were processed to obtain metagenome-assembled genomes (MAGs) following
787 the procedure we used elsewhere⁵. In brief, we used MEGAHIT (version 1.2.9)⁶⁴ with parameters "--k-

max 127” for assembly and assembled contigs $\geq 1.5\text{kb}$ were considered for the binning step performed using MetaBAT2 (version 2.14)⁶⁵ with parameters: “-m 1500 --unbinned”. Quality control of the obtained MAGs was performed using CheckM (version 1.0.18)⁶⁶ 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 an adapted 131-item EPIC FFQ that was developed and validated against pre-established nutrient biomarkers for the EPIC Norfolk⁶⁸. The questionnaire captured average intakes in the past year. UK nutrient intakes were determined using FETA software to calculate macro- and micro-nutrient data⁶⁹. 16 additional foods in the modified FFQ were reviewed by two dietitians and one nutritionist who manually matched food items to corresponding foods within McCance and Widdowson’s 7th Edition, with portions allocated according to the Food Standards Agency 3rd Edition ‘Food Portion Sizes’. US participants completed the Harvard 2007 Grid 131-item FFQ previously validated against two-week dietary records⁷⁰. 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 equation⁷¹) 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 Tables 2 and 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 HFD²⁰. 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.

Healthy Eating Index 2010. The Healthy Eating Index 2010 (HEI-2010)²¹ assesses to which extent an individual’s food intake aligns with the Dietary Guidelines for Americans 2010⁷², developed by the US Department of Agriculture. These guidelines cover a total of twelve food groups and nutrients. The HEI has 9 adequacy (encouraged) and 3 moderation (discouraged) components; firstly a density approach is used to set per 1000 calories; and secondly least-restrictive standards are employed, i.e., those that are easiest to achieve among recommendations that vary by energy level, sex, and/or age. Total fruits, whole fruits, total vegetables, greens and beans, whole grains, dairy (lean portion only), total protein foods (lean portion of meat only), seafood and plant proteins and fatty acids (PUFAs + MUFAs/SFAs) are considered adequate, whereas refined grains, sodium and empty calories (considered added sugars, solid fats, and alcohol above 13g per 1000kcal) are considered detrimental and should be consumed in moderation. The index ranged from 0 (not in agreement with the guidelines) to 100 (completely in agreement with guidelines).

833 *The plant-based diet index.* Three versions of the plant-based diet index ³⁰ were considered: the original
834 plant-based diet index (PDI), the healthy plant-based index (h-PDI), and the unhealthy plant-based index
835 (u-PDI). Eighteen food groups (amalgamated from the FFQ food groups; **Supplementary Table 2**) were
836 assigned either positive or reverse scores after segregation into quintiles, as outlined in **Supplementary**
837 **Table 4** ³⁰. Participants with an intake above the highest quintile for the positive score received a score of
838 5. Those below the lowest quintile intake received a score of 1. A reverse value was applied for the
839 reverse scores. The scores for each participant were summed to create the final score. For the PDI, a
840 positive score was applied to the “healthy” and “less-healthy”/“unhealthy” plant foods, and a reverse
841 score applied to the animal-based foods. For the h-PDI, positive scores were applied to the “healthy”
842 plant foods, and a reverse score to the “less-healthy”/“unhealthy” plant foods and the animal-based
843 foods. For the u-PDI, a positive score was applied to the “less-healthy”/“unhealthy” plant foods and a
844 reverse score applied to the “healthy” plant foods and the animal-based foods.

845 *Animal score.* The animal-based score categorised animal foods into “healthy” and “less-
846 healthy”/“unhealthy” categories according to previous epidemiological studies ^{73–82}. A similar approach to
847 the PDI scoring was applied to the animal-based food groups, with either a positive (“healthy”) or reverse
848 (“less-healthy”/“unhealthy”) quintile scoring (**Supplementary Tables 2 and 4**).

849 *The aMED score.* Adherence to the aMED diet was calculated by following the method outlined by Fung
850 et al. ²². Nine food/nutrient categories were included (**Supplementary Table 4**) and the score ranged
851 from 0 to 9 (“least” to “most” Mediterranean). To form groups, weekly intake frequencies were first
852 multiplied for assigned foods by the amount in grams per serving and then divided by 7 to determine
853 grams per day. Next, food gram amounts were summed to make the final category total. For all food
854 categories as well as the fatty acid intake ratio, the median intake of each category was calculated. A
855 score of 0 (no aMED) or 1 (aMED) was given for each category depending on whether the participant
856 was above or below the median intake. For alcohol intake, a range was used for score assignment:
857 females: 5-25 g/d; males: 10-50 g/d were assigned a score of 1, while those above or below this range
858 were assigned a score of 0. Finally, the aMED was then generated by summation of each category
859 score.

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

863 *Percentage of plants within diet:* The percentage of plants within diet was calculated as weight in grams
864 of plant foods within total weight (g) of diet after adjustment of FFQ foods into quantities (g) per week.

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

869 ***Collection and processing of fasting and postprandial markers***

870 Venous blood samples were collected as outlined in the accompanying protocol paper ⁸³. In brief,
871 participants were cannulated, and venous blood was collected at fasting (prior to a test breakfast) and at
872 9 timepoints postprandially (15, 30, 60, 120, 180, 240, 270, 300, and 360 minutes). Plasma glucose and
873 serum C-peptide and insulin were measured at all timepoints. Serum TG was measured at hourly
874 intervals, and serum metabolomics (NMR by Nightingale Health, Helsinki, Finland, using the 2020
875 platform) at 0, 4 and 6h. Fasting samples were analysed for lipid profile, thyroid-stimulating hormone,
876 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^{8,83}. 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 protocol⁸³. The incremental area under the postprandial TG (0-6h), glucose (0-2h), and insulin (0-2h) curves (iAUC) were computed using the trapezium rule⁸⁴.

Detailed descriptions of sample collection, processing and analysis have been reported in^{8,16}.

Machine learning

The machine learning (ML) framework employed is based on the scikit-learn Python package⁸⁵. 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^{40,86}. 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', max_features='sqrt'” parameters and *LinearRegression* with default parameters. For the classification task, we divided the continuous features into two classes: the top and bottom quartiles. From the *scikit-learn* package we used the *RandomForestClassifier* function with “n_estimators=1000, max_features='sqrt'” parameters.

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

As an additional control, we verified that when randomly swapping the target labels or values (classification and regression, respectively), the performances were reflecting a random prediction, hence an AUC very close to 0.5 and a non-significant correlation between the real and predicted 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 *pcor.test* from the *ppcor* package, respectively. Correlations and p-values were

920 computed for each couple of metadata and species, and p-values were corrected using FDR through the
921 Benjamini-Hochberg procedure, which are reported in the text as q-values. We considered significant
922 correlations with a $q < 0.2$. Significant species have been selected by ranking them according to their
923 number of significant associations for the panel of metadata considered, and then the top thirty unique
924 species are considered for each panel of metadata. In the heatmaps for partial correlations, the asterisk
925 indicates that the correlation index for the corresponding species-metadata pair is significant at $FDR \leq 0.2$.

926 The contribution of metadata variables to microbiota community variation was determined by distance-
927 based redundancy analysis (dbRDA) on species-level Bray-Curtis dissimilarity and Aitchison distance
928 with the *capscale* function in the *vegan* R package⁸⁷. Correction for multiple testing (Benjamini-
929 Hochberg, FDR) was applied and significance was defined at $FDR < 0.1$. The cumulative contribution of
930 metadata variables or metadata categories was determined by forward model selection on dbRDA
931 (stepwise dbRDA) with the *ordiR2step* function in *vegan*, with variables that showed a significant
932 contribution to microbiota community variation in the previous step. Because of the high consistency
933 between the two distance functions, we performed the cumulative distribution analysis using the Bray-
934 Curtis dissimilarity. Only metadata variables with $< 15\%$ missing data and without high collinearity with
935 other variables (Spearman's $\rho < 0.8$) were used as input in the stepwise model.

936 **Data validation on the US cohort and on the cMD datasets**

937 As independent validation, we considered the publically available datasets collected in the
938 *curatedMetagenomicData* version 1.16.0 R package (cMD)³⁴. Of the 57 datasets available we
939 selected those that have samples with the following characteristics: (1) gut samples collected from
940 healthy adult individuals at first collection ("days_from_first_collection" = 0 or NA), (2) samples with
941 age and BMI data available and BMI interquartile range (IQR) of these samples between 3.5 and 7.5
942 (± 2 with respect to the PREDICT 1 UK IQR of 5.5, **Extended Data Fig. 5**). For each dataset with
943 samples meeting the above criteria, only datasets with at least 50 samples were considered:
944 CosteaPI_2017⁸⁸ (84 samples out of 279), DhakanDB_2019⁸⁹ (88 samples out of 110),
945 HansenLBS_2018⁹⁰ (58 samples out of 208), JieZ_2017⁹¹ (157 samples 385), SchirmerM_2016¹⁴
946 (396 samples out of 471), and ZellerG_2014⁹² (59 samples out of 199).

947 We used the previously selected validation datasets from cMD in two analyses: one based on machine
948 learning to verify the reproducibility of the ML model we trained using the PREDICT 1 UK samples, and
949 the second to verify the species-level correlations found in the PREDICT 1 UK cohort. For the first task,
950 we applied a regression algorithm to predict BMI and age. Three different cross-validation approaches
951 were used. First, using each dataset independently in 100 bootstrap iterations and an 80/20 random split
952 of training and testing folds. Second, one more iteration was performed using the PREDICT 1 UK dataset
953 as training fold and each dataset as testing fold. Third, a final prediction was made using Leave-One-
954 Dataset-Out cross-validation (LODO), meaning that all datasets (PREDICT 1 UK, PREDICT 1 UK, and
955 the cMD datasets) were considered together and each validation dataset was successively used as the
956 test fold while all others were used for training. An additional validation performed using the cMD
957 datasets was done by applying a pairwise Spearman correlation for each species in each cMD
958 dataset against BMI and age. For each correlation we selected the top associated species in
959 PREDICT 1 UK ($FDR q \leq 0.05$) and reported their correlation in cMD. For those species found also
960 in the PREDICT 1 US, we reported their correlation as well.

961 **Methods' references**

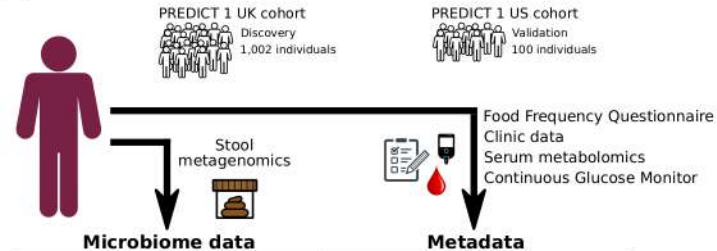
- 962 61. McIver, L. J. *et al.* bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**, 1235–1237
963 (2018).
964 62. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**,

902–903 (2015).

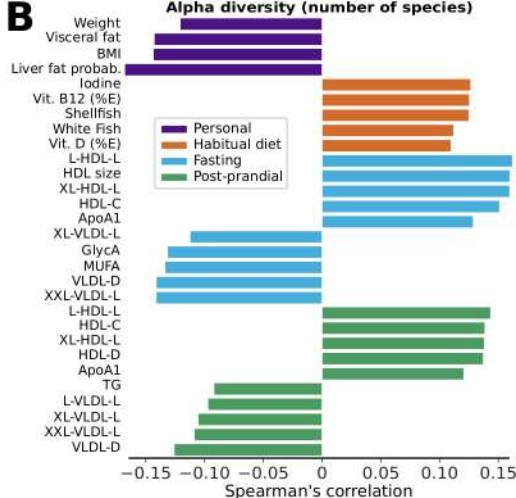
63. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
64. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
65. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
66. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
67. 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).
68. Bingham, S. A. *et al.* Nutritional methods in the European Prospective Investigation of Cancer in Norfolk. *Public Health Nutr.* **4**, 847–858 (2001).
69. Mulligan, A. A. *et al.* A new tool for converting food frequency questionnaire data into nutrient and food group values: FETA research methods and availability. *BMJ Open* **4**, e004503 (2014).
70. 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).
71. 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).
72. McGuire, S. U.S. Department of Agriculture and U.S. Department of Health and Human Services, Dietary Guidelines for Americans, 2010. 7th Edition, Washington, DC: U.S. Government Printing Office, January 2011. *Adv. Nutr.* **2**, 293–294 (2011).
73. WHO | Effect of *trans*-fatty acid intake on blood lipids and lipoproteins: a systematic review and meta-regression analysis. (2016).
74. Zhong, V. W. *et al.* Associations of Dietary Cholesterol or Egg Consumption With Incident Cardiovascular Disease and Mortality. *JAMA* **321**, 1081–1095 (2019).
75. 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).
76. Michaëlsson, K. *et al.* Milk intake and risk of mortality and fractures in women and men: cohort studies. *BMJ* **349**, g6015 (2014).
77. 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).
78. 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).
79. 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).
80. 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).
81. WHO | Effect of *trans*-fatty acid intake on blood lipids and lipoproteins: a systematic review and meta-regression analysis. (2016).
82. Website, N. H. S. Dairy and alternatives in your diet. *nhs.uk* <https://www.nhs.uk/live-well/eat-well/milk-and-dairy-nutrition/>.
83. 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).
84. Matthews, J. N., Altman, D. G., Campbell, M. J. & Royston, P. Analysis of serial measurements in medical research. *BMJ* **300**, 230–235 (1990).
85. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
86. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of

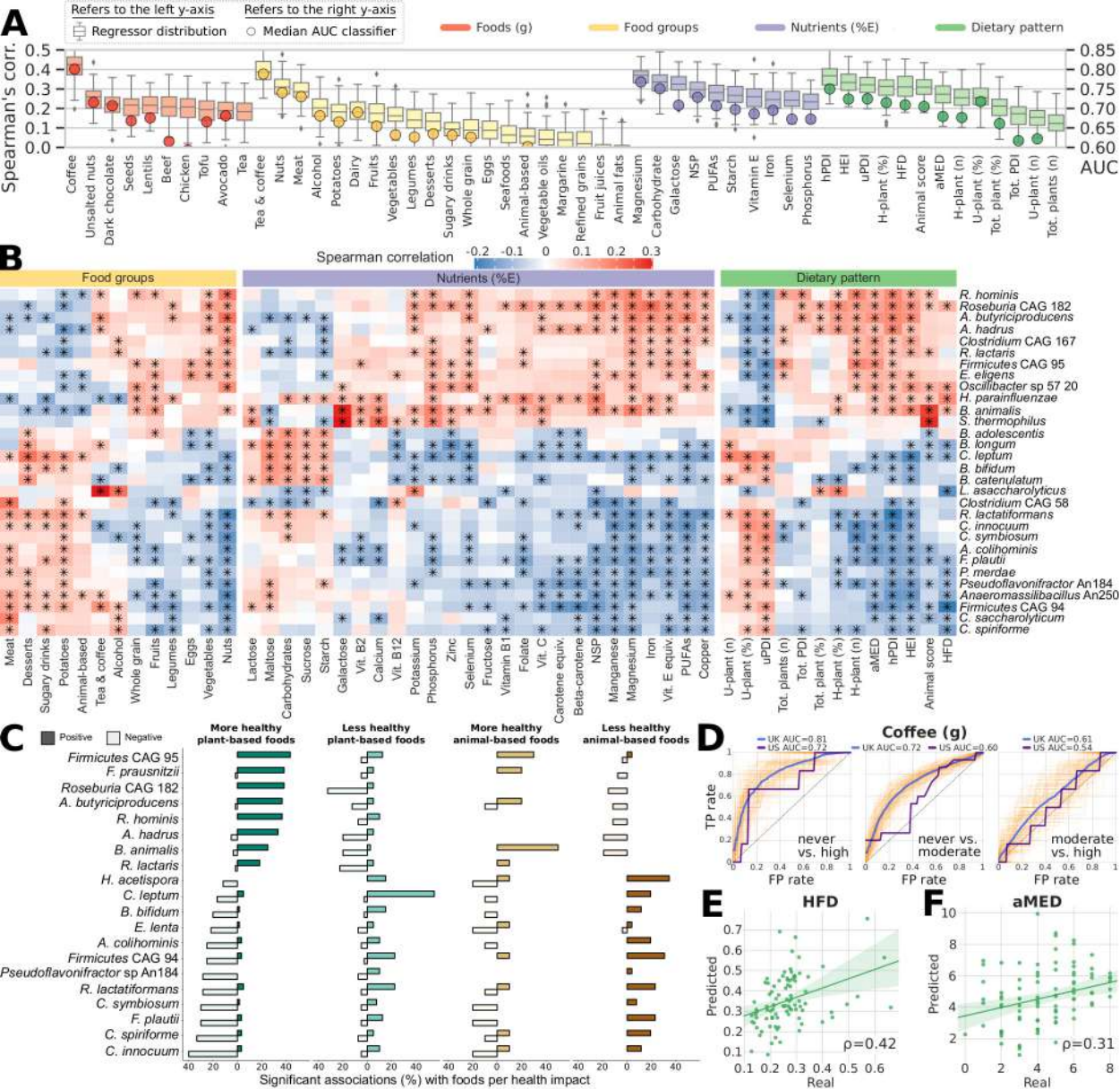
Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput. Biol.* **12**, e1004977 (2016).

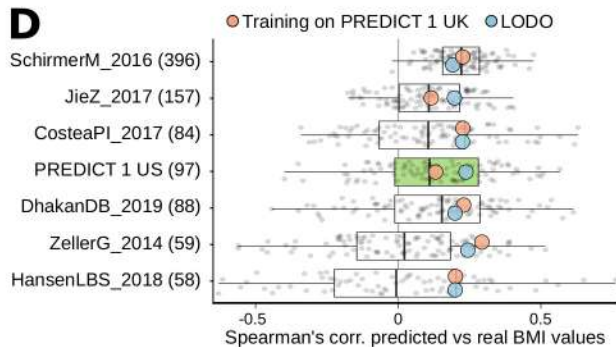
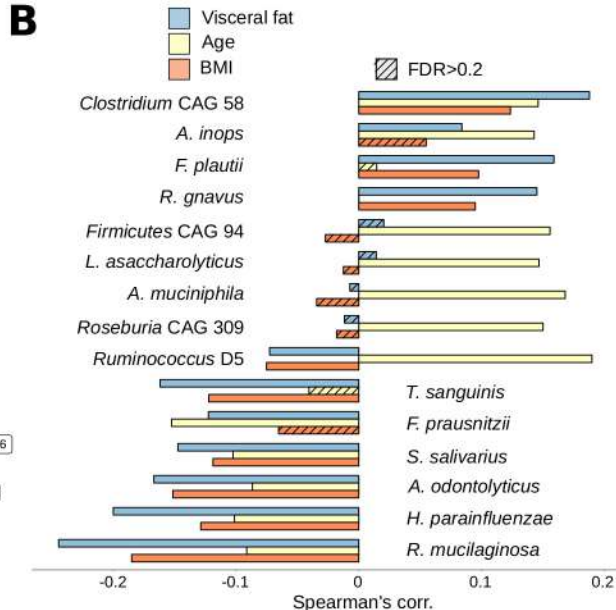
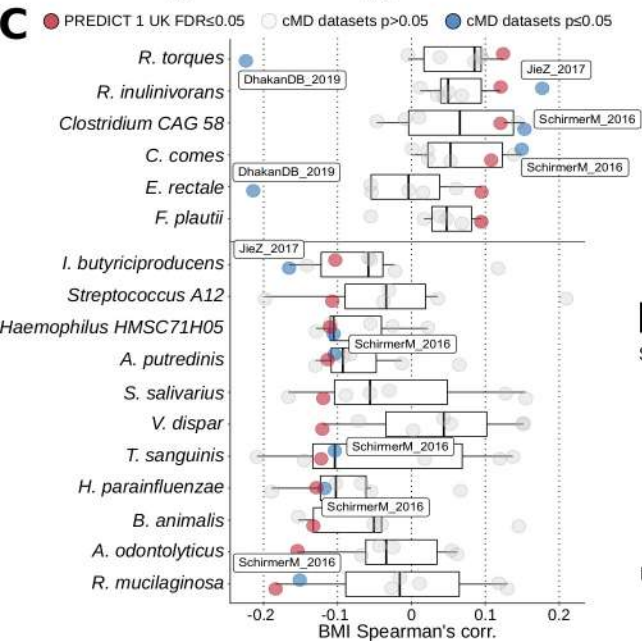
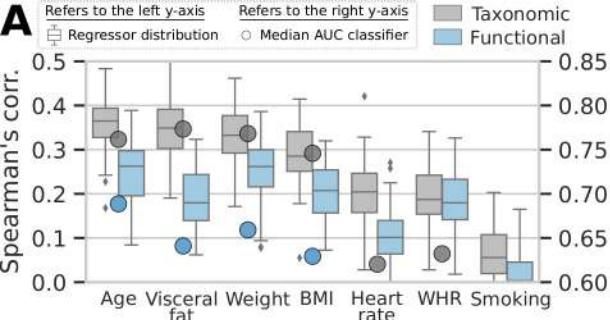
87. Oksanen, J. *et al.* Vegan: community ecology package. R package version 1.17-4. URL <http://CRAN.R-project.org/package=vegan> (2010).
88. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).
89. Dhakan, D. B. *et al.* The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience* **8**, (2019).
90. 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).
91. Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun.* **8**, 845 (2017).
92. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, (2014).

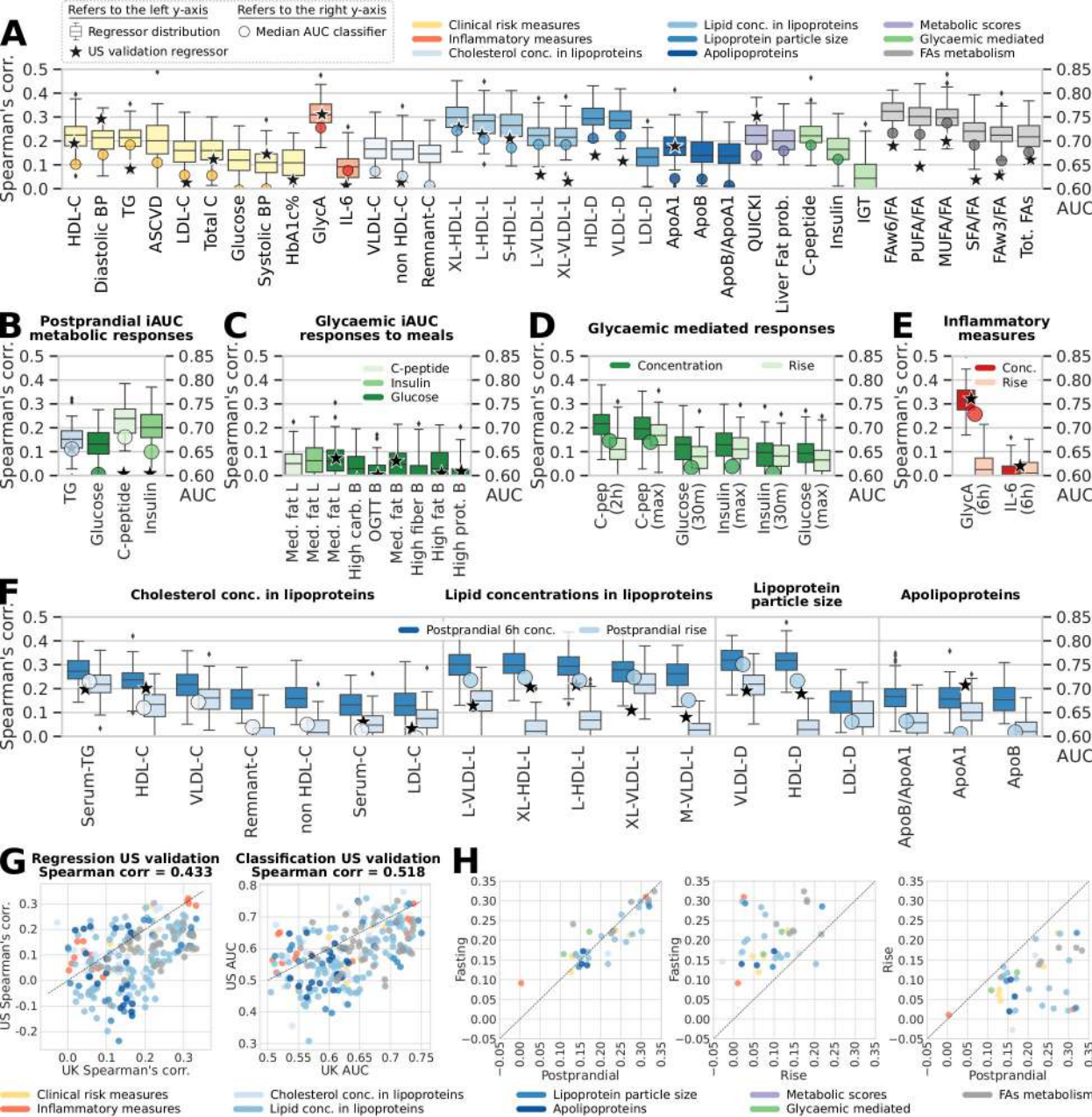
A

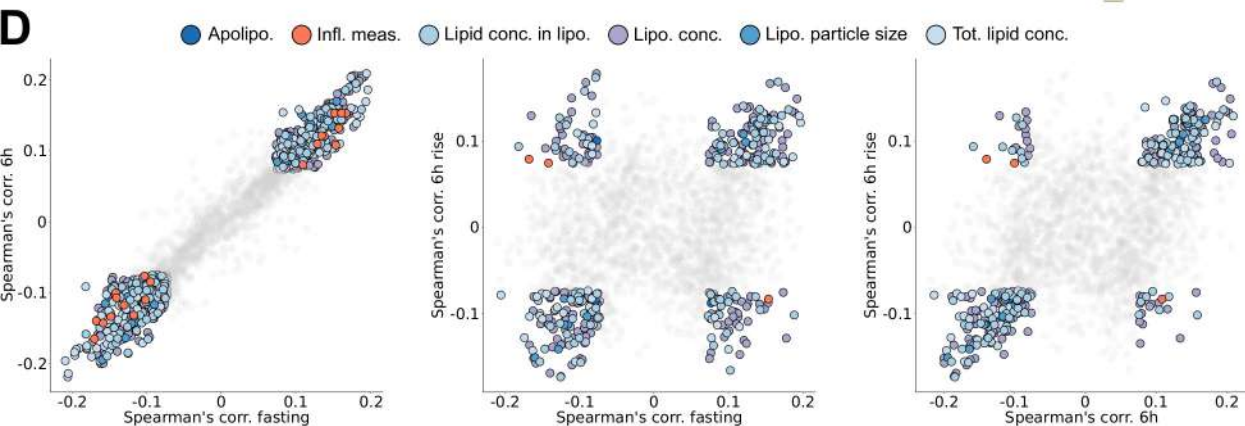
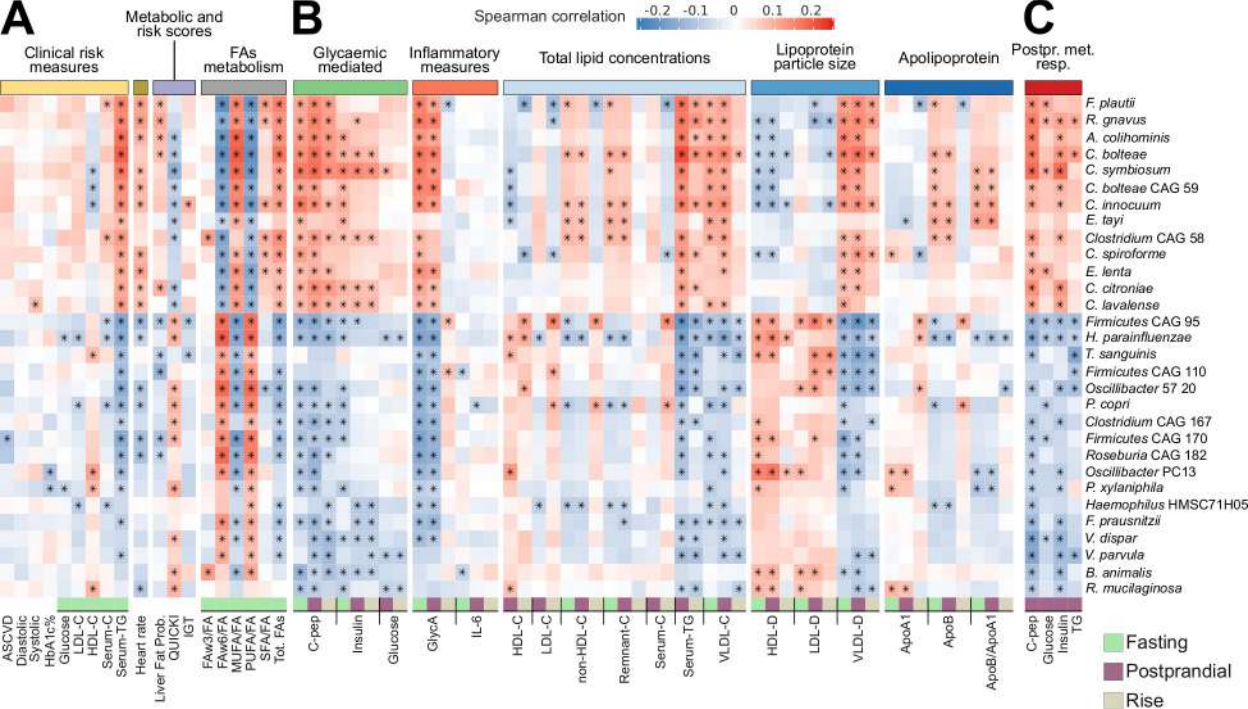
8.8 avg. 2.2 sd. Gb/sample 58.3 avg 14.6 sd. Mreads/sample			1 Personal Age, BMI, Weight, Waist/Hip ratio, Visceral fat, Antibiotic usage, Blood pressure Tot. 17		2 Habitual diet Foods, Food groups, Nutrients, Nutrients %E, Dietary patterns Tot. 223	
Taxonomic	Functional	Assembly	3 Fasting Lipoproteins, ApoLipoproteins, Risk scores, Glucose mediated, FAs metabolism Tot. 269		4 Post-prandial All fasting measures up to 7 timepoints including max values and rises Tot. 1288	
746 species	UniRef90 1,910,069	48,181 MAGs				
29 spp. 90% prevalent	UniRef50 878,520	29,035 MQ				
95 spp. 50% prevalent	KEGG KOs 6,163	19,146 HQ				
176 spp. 20% prevalent	Pathways 445					

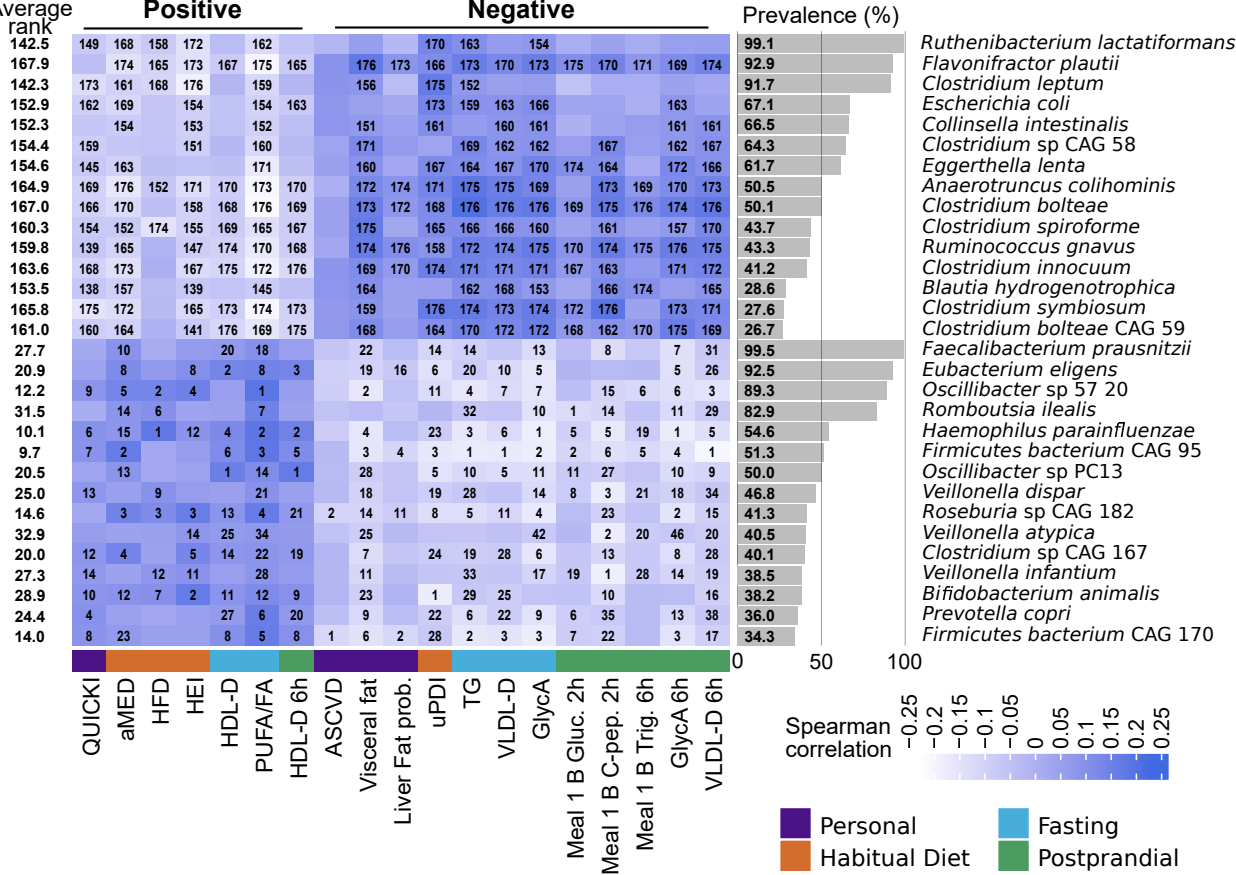
B

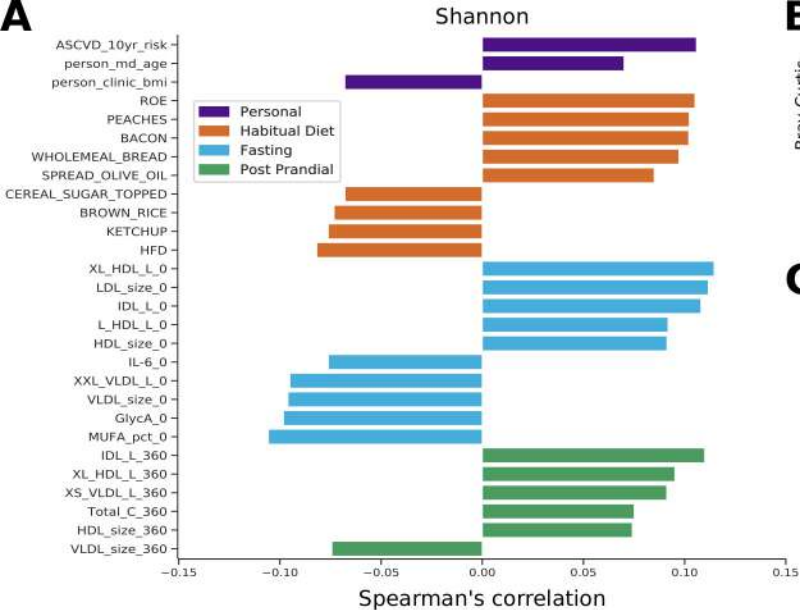
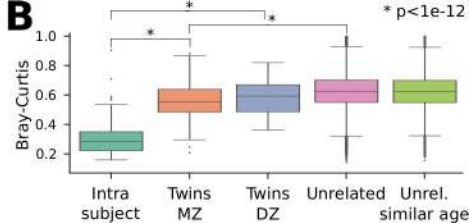
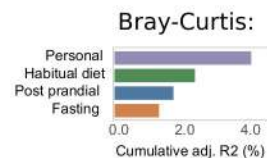
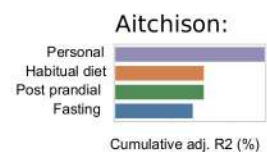
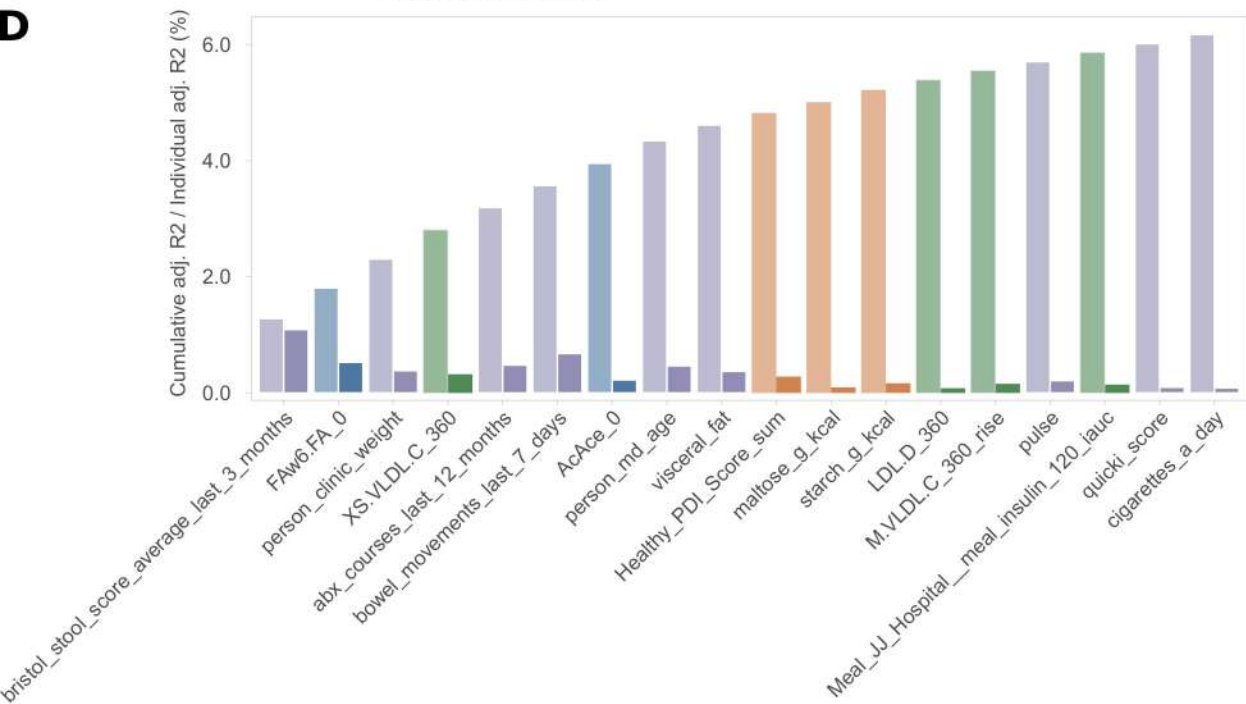




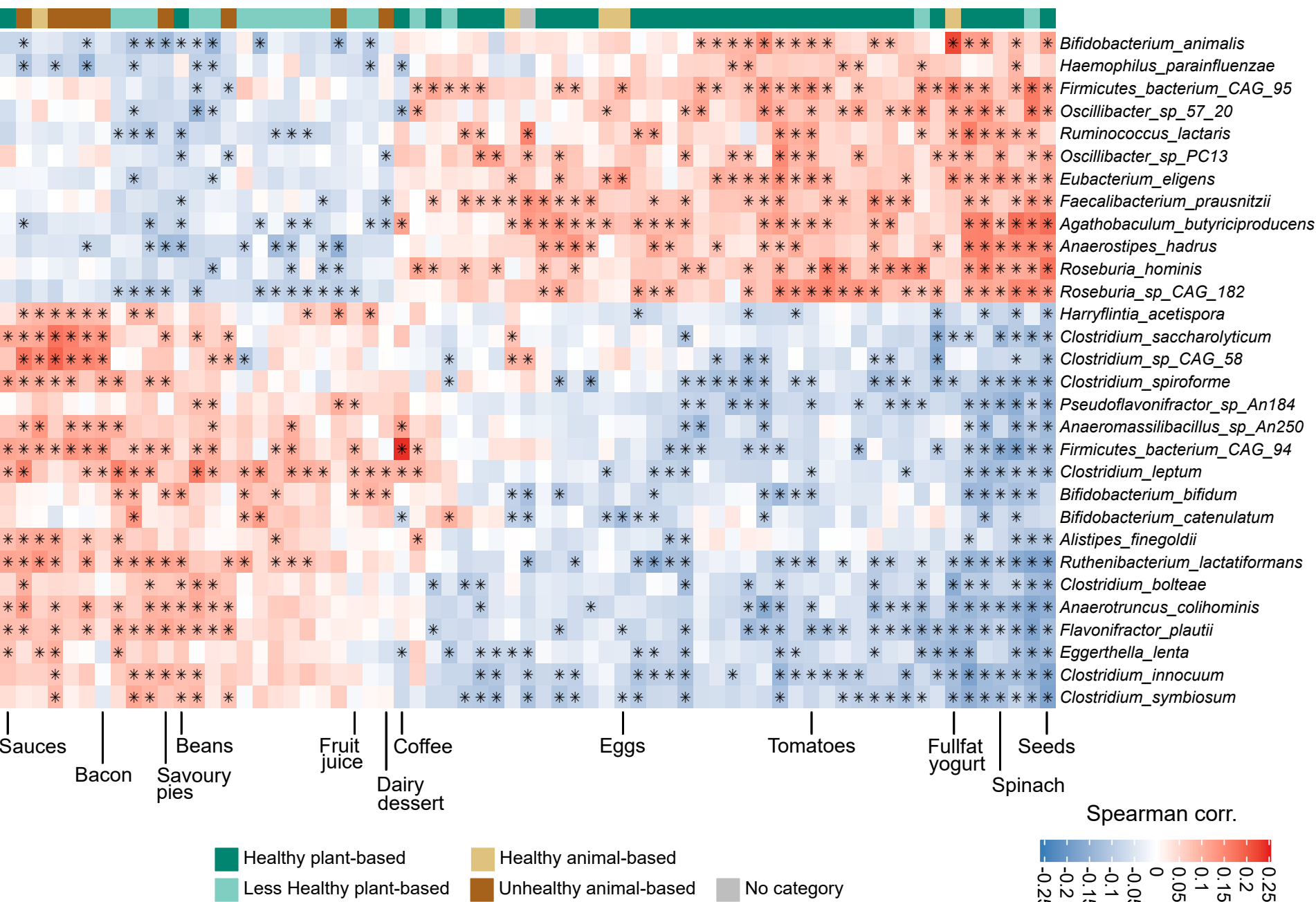


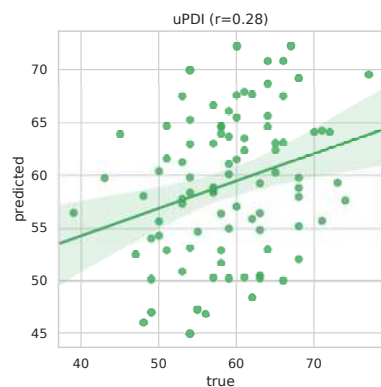
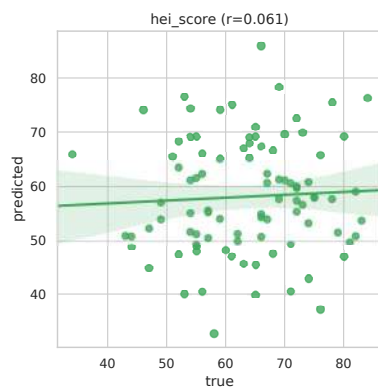
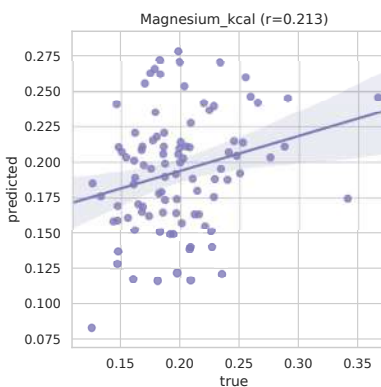
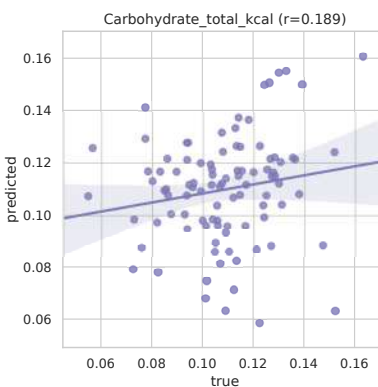
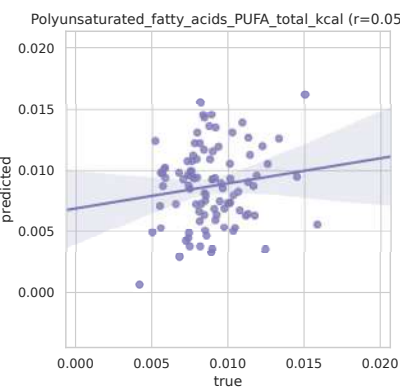
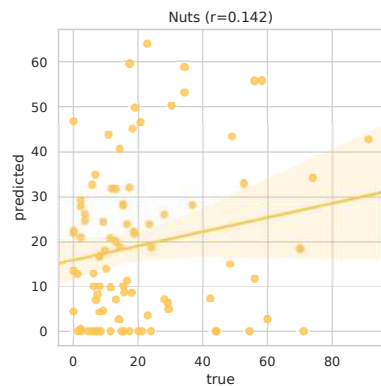
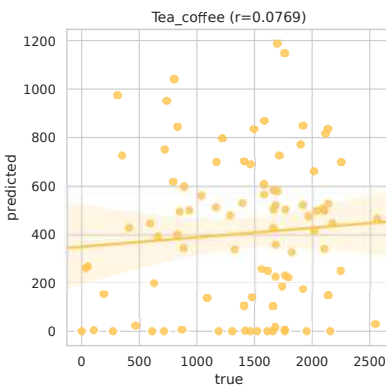
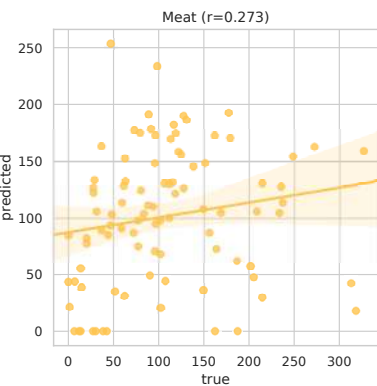
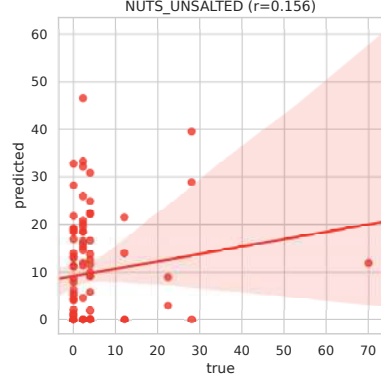
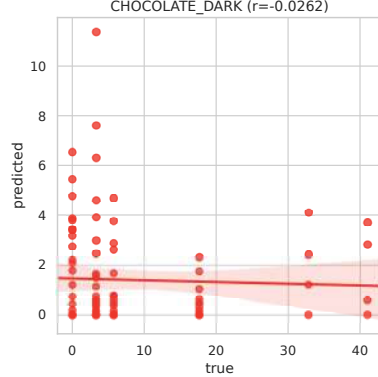
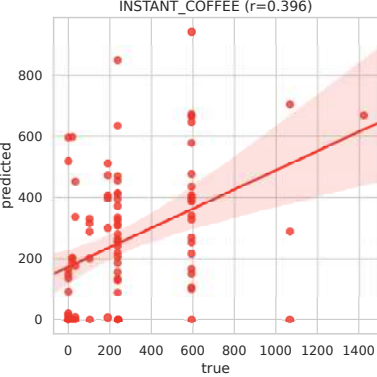


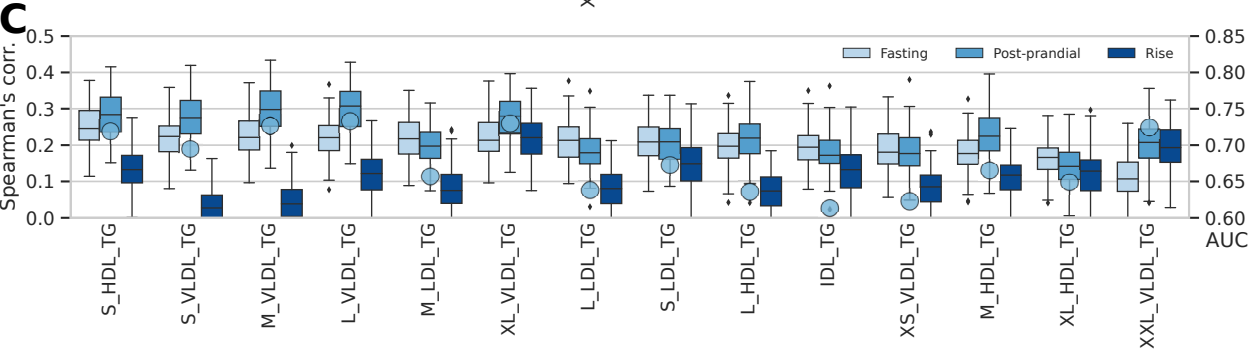
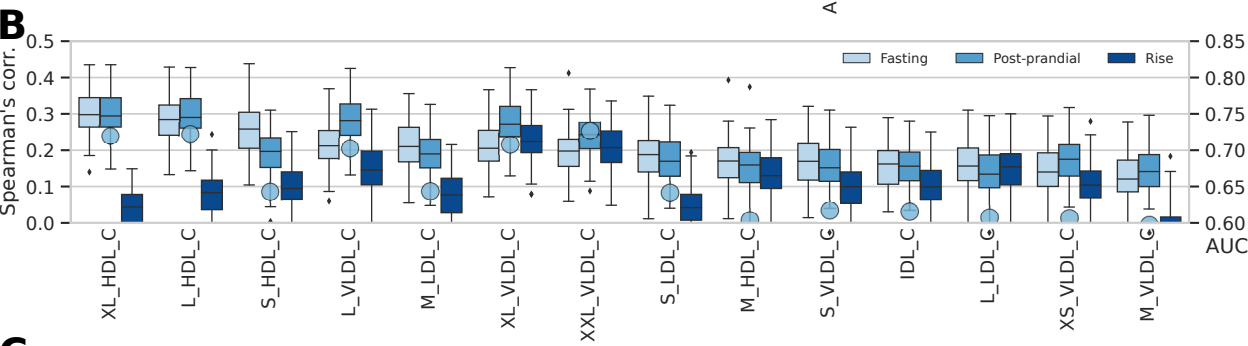
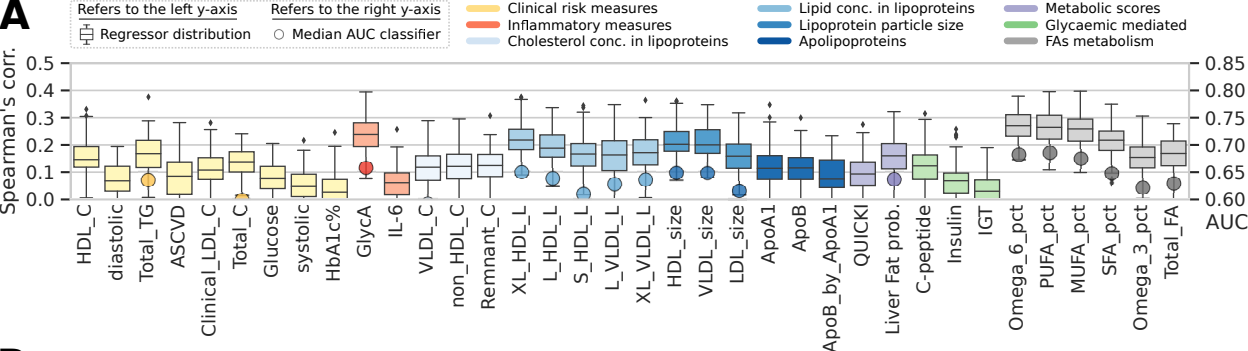


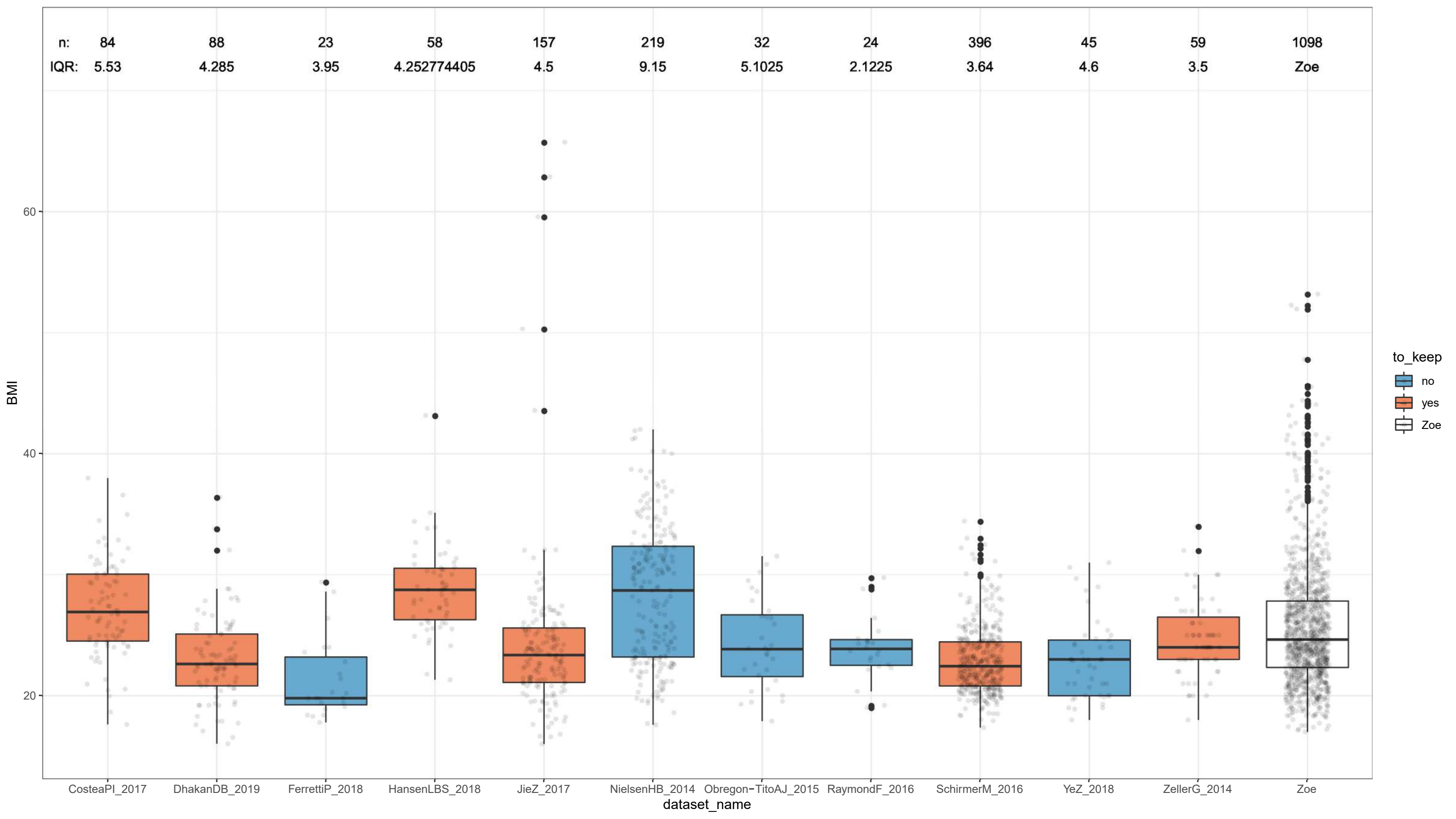
A**B****C****D**

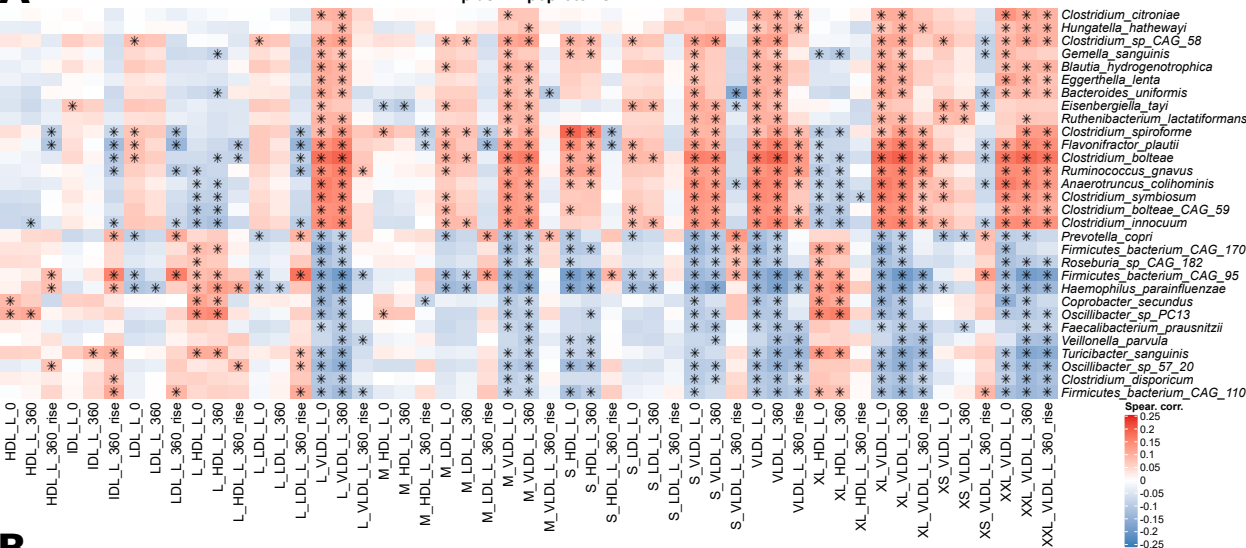
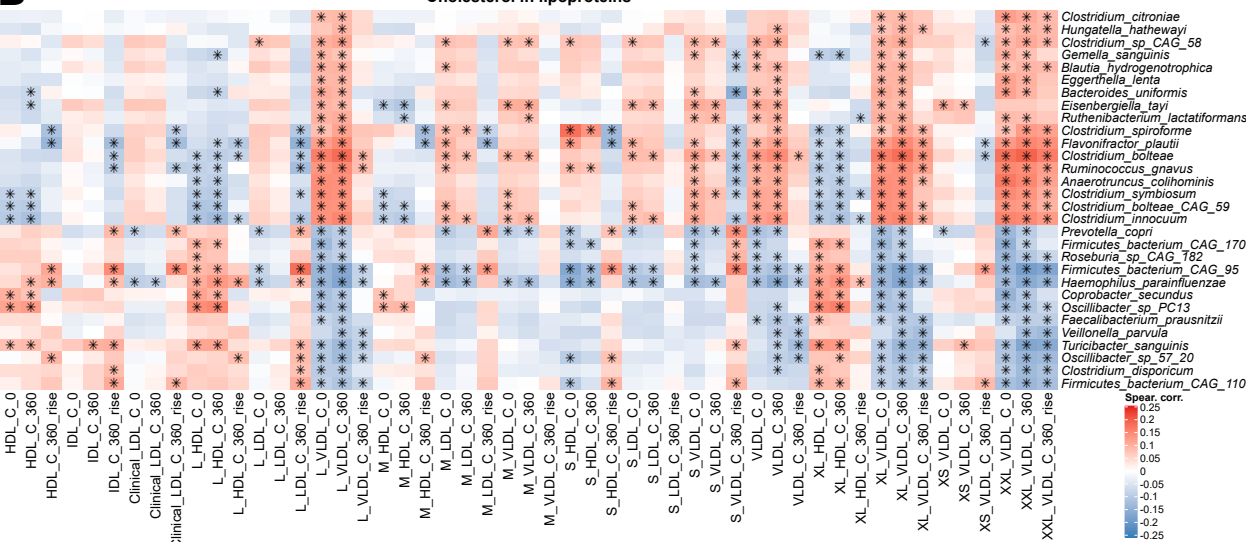
Foods









A**Lipids in lipoproteins****B****Cholesterol in lipoproteins**

Triglycerides in lipoproteins

