The East Asian Gut Microbiome is Distinct from Colocalized White Subjects and Connected to Metabolic Health

Qi Yan Ang[1,5], Diana L. Alba[2,3,5], Vaibhav Upadhyay[1,5], Jordan E. Bisanz[1], Jingwei Cai[4], Ho Lim Lee[2], Eliseo Barajas[2], Grace Wei[2], Cecilia Noecker[1], Andrew D. Patterson[4], Suneil K. Koliwad[2,3]*, and Peter J. Turnbaugh[1]*

[1]Department of Microbiology and Immunology, G.W. Hooper Research Foundation, University of California, San Francisco, CA 94143, USA

[2]Diabetes Center, University of California San Francisco, CA 94143, USA

[3]Division of Endocrinology, Diabetes, and Metabolism, Department of Medicine, University of California San Francisco, CA 94143, USA

[4]Center for Molecular Toxicology and Carcinogenesis, Department of Veterinary & Biomedical Sciences, Pennsylvania State University, PA 16802, USA

[5]These authors contributed equally

*Correspondence to: Suneil K. Koliwad, M.D., Ph.D., Suneil.Koliwad@ucsf.edu, and Peter J. Turnbaugh, Ph.D., peter.turnbaugh@ucsf.edu

ABSTRACT

East Asians experience worse metabolic health outcomes compared to other ethnic groups at lower body mass indices; however, the potential role of the gut microbiota in contributing to these health disparities remains unknown. We conducted a multi-omic study of 46 lean and obese East Asian and White participants living in the San Francisco Bay Area, revealing marked differences between ethnic groups in bacterial richness and community structure. White individuals were enriched for the mucin-degrading *Akkermansia muciniphila.* East Asian subjects had increased levels of multiple bacterial phyla, fermentative pathways detected by metagenomics, and the short-chain fatty acid end-products acetate, propionate, and isobutyrate. Differences in the gut microbiota between the East Asian and White subjects could not be explained by dietary intake, were more pronounced in lean individuals, and were associated with current geographical location. Microbiome transplantations into germ-free mice demonstrated stable diet- and host genotype-independent differences between the gut microbiotas of East Asian and White individuals that differentially impact host body composition. Taken together, our findings add to the growing body of literature describing variation between ethnicities and provide a starting point for defining the mechanisms through which the microbiome may shape disparate health outcomes in East Asians.

Keywords: human gut microbiome, ethnicity, multi-omics, metabolic syndrome, obesity, biogeography

INTRODUCTION

Culture-independent surveys have emphasized differences in gut microbial community structure between countries (Hehemann et al., 2010; Vangay et al., 2018; Yatsunenko et al., 2012); however, the factors that contribute to these differences are poorly understood. Diet is a common hypothesis for geographical variations in the gut microbiota (De Filippo et al., 2010; Devoto et al., 2019), based upon extensive data from intervention experiments in humans and mouse models (Bisanz et al., 2019; Carmody et al., 2015; David et al., 2014; Gehrig et al., 2019). However, diet is just one of the many factors that distinguishes human populations at the global scale, motivating the desire for a more holistic approach. Self-identified race/ethnicity (SIRE) provides a useful alternative, as it integrates the broader national or cultural tradition of a given social group and is closely tied to both dietary intake and genetic ancestry. Multiple studies have reported associations between the gut microbiota and ethnicity in China (Khine et al., 2019), the Netherlands (Deschasaux et al., 2018), Singapore (Xu et al., 2020), and the United States (Brooks et al., 2018; Sordillo et al., 2017). In contrast, a recent study of Asian immigrants suggested that once an individual relocates to a new country, the microbiota rapidly assumes the structure of the country of residence (Vangay et al., 2018). Thus, the degree to which microbiome signatures of ethnicity persist following immigration and their consequences for host pathophysiology remain an open question.

The links between ethnicity and metabolic disease are well-established. For example, East Asian (EA) subjects are more likely to develop health-related metabolic complications at lower body mass index (BMI) compared to their White (W) counterparts (Gu et al., 2006; Zheng et al., 2011). Moreover, Asian Americans have persistent ethnic differences in metabolic phenotypes following immigration (Jih et al., 2014), including a decoupling of BMI from total body fat percentage (Alba et al., 2018). The mechanisms contributing to these ethnic differences in fat accrual remain unknown. Human genetic polymorphisms may play a role (Wen et al., 2010; Xiang et al., 2004); however, putative alleles are often shared between members of different ethnic groups (Gravel et al., 2011). The gut microbiome might offer a possible explanation for differences in metabolic disease rates across ethnic groups (He et al., 2018), but there has been a relative scarcity of microbiome studies in this area (Gaulke and Sharpton, 2018).

These observations led us to hypothesize that ethnicity-associated differences in host metabolic phenotypes may be determined by corresponding differences in the gut microbiome. First, we sought to better understand the extent to which ethnicity is linked to the human gut microbiome in states of health and disease. We conducted a cross-sectional multi-omic analysis of the gut microbiome using paired 16S rRNA gene sequencing (16S-seq), metagenomics, and metabolomics from the Inflammation, Diabetes, Ethnicity, and Obesity (IDEO) cohort at the University of California, San Francisco. IDEO includes rich metabolic, dietary, and socioeconomic metadata (Alba et al., 2018), a restricted geographical distribution within the San Francisco Bay Area, and a balanced distribution of EA and W individuals that are both lean and obese (**Supplementary File 1A**). We report marked differences in gut microbial richness, community structure, and metabolic end-products between EA and W individuals in the IDEO cohort. We then used microbiome transplantations to assess the stability of ethnicity-associated differences in the gut microbiota in the context of genetically identical mice fed the same diet. We also explored the functional consequences of these differences for host metabolic phenotypes. Our results emphasize the importance of considering ethnicity in microbiome research and further complicate prior links between metabolic disease and the gut microbiome (Ley et al., 2006; Turnbaugh et al., 2009a; Wu et al., 2020), which may be markedly different across diverse ethnic groups.

RESULTS

Ethnicity was associated with inter-individual variations in the human gut microbiota. Principal coordinates analysis of PhILR Euclidean distances from 16S-seq data (**Supplementary File 1B,** n=22 EA, 24 W subjects) revealed a subtle but significant separation between the gut microbiotas of EA and W subjects (p=0.006, R^2=0.046, ADONIS; **Fig. 1A**). Statistical significance was robust to the distance metric used (**Supplementary File 1C**). Bacterial diversity was significantly higher in W individuals across three distinct metrics: Faith's phylogenetic diversity, ASV richness, and Shannon diversity (**Fig. 1B**). Six bacterial phyla were significantly different between ethnicities (**Fig. 1C**), of which only one phylum, *Verrucomicrobiota*, was significantly enriched in W subjects.

Phylogenetic analyses of all ASVs revealed marked variations in the direction of change across different phyla between EA and W subjects (**Figure 1-figure supplement 1A**), indicating that the phylum level trends (**Fig. 1C**) resulted from the integration of subtle shifts across multiple component members (**Fig. 1D-F**). Several significant differences were detectable at the genus level (**Fig. 1D-E**), including *Blautia*, *Bacteroides*, and *Streptococcus* which were significantly enriched in EA subjects. We also identified two ASVs that were significantly different between ethnicities: *Blautia obeum* and a *Streptococcus* species, both enriched in EA subjects **(Fig. 1F)**. There were no significant differences between ethnicities in 16S rRNA copy number (**Figure 1-figure supplement 1F**).

Next, we used a random forest classifier to define biomarkers in the gut microbiota that distinguish EA and W subjects (**Figure 1-figure supplement 1B-D**). Classifiers employing ASV data and PhILR transformed phylogenetic nodes were trained using leave-one-out cross-validation. *Blautia obeum* (ASV1) was the top contributor to the resulting classifier, followed by *Anaerostipes hadrus* (ASV45) and then *Streptococcus parasanguinis* (ASV110) (**Figure 1-figure supplement 1B)**. Both classifiers demonstrated the ability to distinguish between ethnic groups, with PhILR transformed phylogenetic nodes achieving a higher area under the curve compared to ASVs (**Figure 1-figure supplement 1C,D**). The majority (18/23) of the top ASVs identified by our classifier were also significantly different between ethnicities (**Figure 1-figure supplement 1E**).

Metagenomic sequencing provided independent confirmation of differences in the gut microbiome between ethnicities (**Supplementary File 1B**, n=21 EA, 24 W subjects). Consistent with our 16S-seq analysis, we detected a difference in the gut microbiomes between ethnicities based upon metagenomic species abundances (p=0.004, R^2=0.047, ADONIS, **Fig 2A**) and gene families (p=0.029, R^2=0.036, ADONIS). Ethnicity explained more variation in species abundances than a selection of demographic, laboratory, lifestyle, and metabolic metadata (**Fig. 2B**). Visualization of diversity and species assignments within each phylum revealed marked variation in the magnitude and direction of change between individuals of a given ethnicity (**Fig. 2C**). Genera that were found to be significantly different between ethnicities in our metagenomic data included *Akkermansia* and an unspecified *Erysipelotrichaceae* genera (**Fig. 2D**) elevated in W individuals. Four bacterial species were significantly different between ethnicities in our metagenomic data: W individuals had higher levels of *Akkermansia muciniphila*, *Bacteroidales bacterium ph8*, and *Roseburia hominis*, and lower levels of *Ruminococcus gnavus*, compared to EA individuals (**Fig. 2E**).

Next, we used NMR-based stool metabolomics to gain insight into the potential functional consequences of ethnicity-associated differences in the human gut microbiome (**Supplementary File 1B**, n=10 subjects/ethnicity). Metabolite profiles were more strongly associated with ethnicity (p=0.008, R^2=0.128, ADONIS; **Fig. 3A**) than community structure (R^2=0.029-0.055, ADONIS; **Supplementary File 1C**) or gene abundance (p=0.029, R^2=0.036, ADONIS). Feature annotations revealed elevated levels of the branched chain amino acid (BCAA) valine and the short-chain fatty acids (SCFAs) acetate and propionate in EA subjects (**Fig. 3B** and **Supplementary File 1D**). In contrast, proline, formate, alanine, xanthine, and hypoxanthine were found at higher levels in W subjects (**Fig. 3B**). To assess the statistical significance and reproducibility of these trends, we used targeted GC-MS and UPLC-MS/MS to quantify a panel of BCAAs, SCFAs, and bile acids (**Supplementary File 1E**). Confirming our NMR data, EA subjects had significantly higher levels of stool acetate (**Fig. 3C**) and propionate (**Fig. 3D**); however, we did not detect any significant differences in BCAAs or bile acids **(Figure 3-figure supplement 1**). Isobutyrate (which was not detected by NMR) was also significantly higher in EA subjects (**Fig. 3E**). In agreement with these metabolite levels, a targeted re-analysis of our metagenomic data revealed a significant enrichment in two SCFA-related pathways:

"pyruvate fermentation to butanoate" (p=0.023, fold-difference=2.216) and "superpathway of *Clostridium acetobutylicum* acidogenic fermentation" (p=0.023, fold-difference=2.182).

Consistent with prior work (Le Chatelier et al., 2013; Turnbaugh et al., 2009a), we found that gut bacterial richness in W individuals was significantly associated with both BMI (**Fig. 4A**) and body fat percentage (**Fig. 4B**). Remarkably, these associations were undetectable in EA subjects (**Figs. 4A,B**) even when other metrics of bacterial diversity were used (**Figure 4-figure supplement 1**), with the single exception of a negative correlation between Shannon diversity and BMI in EA subjects **(Figure 4-figure supplement 1C)**. Re-analysis of our data separating lean and obese individuals revealed that the previously observed differences between ethnic groups were driven by lean individuals. Compared to lean EA individuals, lean W subjects had significantly higher bacterial diversity (**Fig. 4C**) and more marked differences in gut microbial community structure (p=0.0003, R^2=0.122, ADONIS; **Fig. 4D**) and metabolite profiles (p=0.010, R^2=0.293, ADONIS; **Fig. 4E**). By contrast, obese W versus EA individuals were not different across any of these metrics (**Figs. 4C-E**), except for lower Shannon diversity in obese EA compared to W individuals (**Fig. 4C**). We also detected differences in the gut microbiotas of lean EA and W individuals at the phylum (**Fig. 5A**) and genus (**Fig. 5B**) levels that were largely consistent with our original analysis of the full dataset (**Figs. 1C,E**). More modest differences in the gut microbiota between ethnicities were observed in obese subjects (**Figs. 5A,C**).

Next, we sought to understand the potential drivers of differences in the gut microbiome between ethnic groups in lean individuals within the IDEO cohort. Consistent with prior studies (Falony et al., 2016), PERMANOVA analysis of our full 16S-seq dataset revealed that diabetes (Forslund et al., 2015), age (Ghosh et al., 2020), metformin use (Wu et al., 2017), and statin intake (Vieira-Silva et al., 2020) were significantly associated with variance in the PhILR Euclidean distances **(Figure 6-figure supplement 1)**. Metagenomic sequencing of the IDEO cohort with subsequent PERMANOVA analysis confirmed significant associations with ethnicity and statin use, while also highlighting significant associations with HOMA-IR and BMI (**Fig. 2B**), consistent with prior reports (Liu et al., 2017; Zouiouich et al., 2021). While several factors linked to body composition were different between obese EA and W subjects using a nominal p-value, only triglyceride levels were significantly different between lean EA and W subjects and this trend did not survive multiple testing correction (**Supplementary File 1A**). Although everyone in the cohort was recruited from the San Francisco Bay Area, birth location

varied widely (**Figure 6-figure supplement 2**). There was no significant difference in the proportion of subjects born in the USA between ethnicities (75% W, 54.5% EA; p=0.15, Pearson's χ^2 test). There was also no significant difference in the geographical distance between birth location and San Francisco [W median 2,318 (2.2-6,906) miles; EA median 1,986 (2.2-6,906) miles; p=0.69, Wilcoxon rank-sum test) or the amount of time spent in the San Francisco Bay Area at the time of sampling [W median 270 (8.00-741) months; EA median 282.5 (8.50-777) months; p=0.42, Wilcoxon rank-sum test).

Surprisingly, we did not detect any significant differences in either short- (**Supplementary File 1F**) or long-term (**Supplementary File 1G**) dietary intake between ethnicities. Consistent with this, Procrustes analysis did not reveal any significant associations between dietary intake and gut microbial community structure: procrustes p=0.280 (DHQIII) and p=0.080 (ASA24) relative to PhILR transformed 16S-seq ASV data. The Spearman Mantel statistic was also non-significant [r=0.0524, p=0.243 (DHQIII) and r=-0.0173, p=0.590 (ASA24)], relative to PhILR transformed 16S-seq ASV data. Despite the lack of an overall association between reported dietary intake and the gut microbiota, we were able to identify 12 ASVs and 7 metagenomic species associated with dietary intake in lean W individuals (**Figure 6-figure supplement 3A**). We also detected 20 significant species-level associations in lean EA subjects (**Figure 6-figure supplement 3B**). There were no overlapping associations between ethnicities.

Given the marked variation in the gut microbiome at the continental scale (Hehemann et al., 2010; Vangay et al., 2018; Yatsunenko et al., 2012), we hypothesized that the observed differences in lean EA and W individuals may be influenced by a participant's current address at the time of sampling. Consistent with this hypothesis, we found clear trends in ethnic group composition across ZIP codes in the IDEO cohort (**Figs. 6A,B**) that were mirrored by the 2018 US census data (Pearson r=0.52, p=0.026 for neighborhoods with greater than 50% white subjects; **Fig. 6D**). Obese individuals from both ethnicities and lean W subjects tended to live closer to the center of San Francisco relative to lean EA subjects (**Fig. 6C**). Distance between current ZIP code and the center of San Francisco and duration of residency within San Francisco were both associated with gut microbial community structure (**Figs. 6E,F**). The association between current address and the gut microbiota was robust to the central point used, as evidenced

by using the Bay Bridge as the central reference point (p=0.008, rho=0.394, Spearman correlation).

Taken together, our results support the hypothesis that there are stable ethnicity-associated signatures within the gut microbiota of lean EA vs. W individuals that are independent of diet. To experimentally test this hypothesis, we transplanted the gut microbiotas of two representative lean W and lean EA individuals into germ-free male C57BL/6J mice fed a low-fat, high-plant-polysaccharide (LFPP) diet (2 independent experiments; per group n=12 mice, 2 donors; per donor n=6 mice, 1 isolator; **Figure 7-figure supplement 1A,B**). The donors for this and the subsequent experiment were matched for their metabolic and other phenotypes to minimize potential confounding factors (**Supplementary File 1H,I**). Despite maintaining the genetically identical recipient mice on the same autoclaved LFPP diet, we detected significant differences in gut microbial community structure (**Fig. 7A**), bacterial richness (**Fig. 7C**), and taxonomic abundance (**Figs. 7D,E** and **Supplementary File 1J**) between the two ethnicity-specific recipient groups. These differences recapitulated key aspects of the gut microbiota observed in the IDEO cohort, including significantly lower bacterial richness **(Fig. 7C)** and higher abundance of *Bacteroides* **(Fig. 7D,E)** in recipient mice transplanted with microbiota from EA compared to W donors.

Next, we sought to assess the reproducibility of these findings across multiple donors and in the context of a distinctive dietary pressure. We fed 20 germ-free male mice a high-fat, high-sugar (HFHS) diet for 4 weeks prior to colonization with a gut microbiota from one of 5 W and 5 EA donors. Mice were maintained on the HFHS diet following colonization (per group n=10 mice, 5 donors; per donor n=2 mice, 1 cage; **Figure 7-figure supplement 1C**). This experiment replicated our original findings on the LFPP diet, including significantly altered gut microbial community structure between ethnicities (**Fig. 7F**), significantly increased richness in mice receiving W donor microbiota (**Fig. 7H**), and a trend towards higher levels of *Bacteroides* in mice receiving the gut microbiotas of EA donors (**Figs. 7I,J**). Of note, the variance explained by ethnicity was lower in mice fed the HFHS diet (R^2=0.126) than the LFPP diet (R^2=0.384), potentially suggesting that in the context of human obesity, excessive fat and sugar consumption may serve to diminish the signal otherwise associated with ethnicity. As expected (Nayak et al., 2021; Turnbaugh et al., 2009b; Walter et al., 2020), the input donor microbiota was distinct from that of the recipient mice (**Figs. 7B,G**); however, there was no difference between ethnic groups

in the efficiency of engraftment (**Figure 7-figure supplement 2**). In a pooled analysis of all gnotobiotic experiments accounting for one donor for multiple recipient mice, ethnicity and diet were both significantly associated with variations in the gut microbiota (**Figure 7-figure supplement 3**), consistent with the extensive published data demonstrating the rapid and reproducible impact of a HFHS diet on the mouse and human gut microbiota (Bisanz et al., 2019).

Finally, mice transplanted with gut microbiomes of EA and W individuals displayed differences in body composition. LFPP fed mice that received W donor microbiota had significantly increased adiposity in conjunction with decreased lean mass, relative to LFPP fed mice that received the EA donor microbiota (**Figs. 8A-C**). Although these trends were mirrored in recipient mice that fed the HFHS diet **(Figs. 8E-G)**, they did not reach statistical significance. There were no significant differences in glucose tolerance in either experiment (**Figs. 8D**,**H**). Together, these results suggest that dietary input may mask the metabolic consequences of ethnicity-associated differences in the gut microbiota.

DISCUSSION

Despite the potential for immigration to erase some of the geographically specific aspects of gut microbiome structure (Vangay et al., 2018), our study suggests that even in a given geographic location, there remain stable long-lasting microbial signatures of ethnicity, as revealed here for W and EA residents of the San Francisco Bay Area. The mechanisms responsible remain to be elucidated. In lean individuals within the IDEO cohort, these differences appear to be independent of immigration status, host phenotype, or dietary intake. Our experiments using inbred germ-free mice support the stability of ethnicity-associated differences in the gut microbiota on both the LFPP and HFHS diets, while also demonstrating that variations in host genetics are not necessary to maintain these signatures, at least over short timescales. Even though we conducted multiple experiments and recipient mice from the same donor generally mapped together, differences between the human donor and recipient mouse microbiotas inherent to gnotobiotic transplantation warrant further investigation, as do differences in the stability of the gut microbiotas of male versus female donors.

Our data also supports a potential role for geographic location of residence in reinforcing differences in the gut microbiota between ethnic groups. The specific reasons why current location would matter to the gut microbiota remain unclear. Current location may reflect subtle differences in dietary intake (*e.g.,* ethnic foods, food sources, or phytochemical contents) that are hard to capture using the validated nutritional surveys employed here (Garduño-Diaz et al., 2014). Alternative hypotheses include biogeographical patterns in microbial dispersion (Martiny et al., 2006) or a role for socioeconomic factors, which are correlated with neighborhood (Kakar et al., 2018).

Surprisingly, our findings demonstrate that ethnicity-associated differences in the gut microbiota are stronger in lean individuals. Obese individuals did not exhibit as clear a difference in the gut microbiota between ethnic groups, either suggesting that established obesity or its associated dietary patterns can overwrite long-lasting microbial signatures. Alternatively, there could be a shared ethnicity-independent microbiome type that predisposes individuals to obesity. Studies in other disease areas (*e.g.,* inflammatory bowel disease and cancer) with similar multi-ethnic cohorts are essential to test the generalizability of these findings and to generate hypotheses as to their mechanistic underpinnings.

Our results in humans and mouse models support the broad potential for downstream consequences of ethnicity-associated differences in the gut microbiome for metabolic syndrome and potentially other disease areas. However, the causal relationships and how they can be understood in the context of the broader differences in host phenotype between ethnicities require further study. While these data are consistent with our general hypothesis that ethnicity-associated differences in the gut microbiome are a source of differences in host metabolic disease risk, we were surprised by both the nature of the microbiome shifts and their directionality. Based upon observations in the IDEO (Alba et al., 2018) and other cohorts (Gu et al., 2006; Zheng et al., 2011), we anticipated that the gut microbiomes of lean EA individuals would promote obesity or other features of metabolic syndrome. In humans, we did find multiple signals that have been previously linked to obesity and its associated metabolic diseases in EA individuals, including increased Firmicutes (Basolo et al., 2020; Bisanz et al., 2019), decreased *A. muciniphila* (Depommier et al., 2019; Plovier et al., 2017), decreased diversity (Turnbaugh et al., 2009a), and increased acetate (Perry et al., 2016; Turnbaugh et al., 2006). Yet EA subjects also had higher levels of *Bacteroidota* and *Bacteroides*, which have been linked to improved metabolic health (Johnson et al., 2017). More importantly, our microbiome transplantations demonstrated that the recipients of the lean EA gut microbiome had less body fat despite consuming the same diet. These seemingly contradictory findings may suggest that the recipient mice lost some of the microbial features of ethnicity relevant to host metabolic disease or alternatively that the microbiome acts in a beneficial manner to counteract other ethnicity-associated factors driving disease.

EA subjects also had elevated levels of the short-chain fatty acids propionate and isobutyrate. The consequences of elevated intestinal propionate levels are unclear given the seemingly conflicting evidence in the literature that propionate may either exacerbate (Tirosh et al., 2019) or protect from (Lu et al., 2016) aspects of metabolic syndrome. Clinical data suggests that circulating propionate may be more relevant for disease than fecal levels (Müller et al., 2019), emphasizing the importance of considering both the specific microbial metabolites produced, their intestinal absorption, and their distribution throughout the body. Isobutyrate is even less well-characterized, with prior links to dietary intake (Berding and Donovan, 2018) but no association with obesity (Kim et al., 2019). Unlike SCFAs, we did not identify consistent differences in BCAAs, potentially due to differences in both extraction and standardization

techniques inherent to GC-MS and NMR analysis (Cai et al., 2016; Lynch and Adams, 2014; Qin et al., 2012).

There are multiple limitations of this study. Due to the investment of resources into ensuring a high level of phenotypic information on each cohort member coupled to the restricted geographical catchment area, the IDEO cohort was relatively small at the time of this analysis (n=46 individuals). The current study only focused on two of the major ethnicities in the San Francisco Bay Area. As IDEO continues to expand and diversify its membership, we hope to study participants from other ethnic groups. Stool samples were collected at a single time point and analyzed in a cross-sectional manner. While we used validated tools from the field of nutrition to monitor dietary intake, we cannot fully exclude subtle dietary differences between ethnicities (Johnson et al., 2019), which could be interrogated through controlled feeding studies (Basolo et al., 2020). Our mouse experiments were all performed in wild-type adult males. The use of a microbiome-dependent transgenic mouse model of diabetes (Brown et al., 2016) would be useful to test the effects of inter-ethnic differences in the microbiome on insulin and glucose tolerance. Additional experiments are warranted using the same donor inocula to colonize germ-free mice prior to concomitant feeding of multiple diets, allowing a more explicit test of the hypothesis that diet can disrupt ethnicity-associated microbial signatures. These studies, coupled to controlled experimentation with individual strains or more complex synthetic communities, would help to elucidate the mechanisms responsible for ethnicity-associated changes in host physiology and their relevance to disease.

CONCLUSIONS

Our results support the utility of considering ethnicity as a covariate in microbiome studies, due to the ability to detect signals that are difficult to capture by more specific metadata such as individual dietary intake values. On the other hand, these findings raise the importance of dissecting the sociological and biological components of ethnicity with the goal of identifying factors that shape the gut microbiota, either alone or in combination. This emerging area of microbiome research is just one component in the broader efforts to explore the boundaries and mechanistic underpinning of ethnicity with respect to multiple ethnic groups. The IDEO cohort provides a valuable research tool to conduct prospective longitudinal and intervention studies examining diabetes in diverse participants. More broadly, IDEO provides a framework to approach other disease states where self-identified race or ethnicity are thought to contribute to health outcomes related to the microbiome, including the use of gnotobiotic mouse models to examine the specific role of microbial communities in contributing to phenotypes linked to ethnicity. By understanding the biologic features that drive differences between ethnic groups, we may be able to achieve similar health outcomes and to support more precise therapies informed by a broader appreciation of both microbial and human diversity.

FIGURE LEGENDS

Figure 1. The gut microbiota is distinct between East Asian and White subjects living in the Bay Area. (A-C) Each point represents a single individual's gut microbiota based upon 16S-seq. **(A)** Principal coordinate analysis of PhILR Euclidean distances reveals significant separation between ethnic groups (ADONIS test values shown). Additional distance calculations for complementary distance matrix calculations are shown in **Supplementary File 1C**. **(B)** Calculations of alpha diversity between EA and W subjects. *p*-values determined using Wilcoxon rank-sum tests. **(C)** CLR abundances of all bacterial phyla between EA and W subjects. *p*-values determined using Wilcoxon rank-sum tests. **(D)** Stacked bar plots showing the average percent relative abundances at the genus level for EA and W subjects respectively. The most abundant taxa are shown as differently colored bars, with lower abundance taxa grouped as a single bar ("Remainder"). **(E,F)** Volcano plot of ALDEx2 differential abundance testing on **(E)** genera and **(F)** ASVs detected by 16S-seq in the gut microbiotas of EA versus W individuals. Significantly different (FDR < 0.1) features are highlighted in black and labeled by genus or the most specific taxonomic assignment. **(A-F)** n=22 EA and n=24 W individuals.

Figure 2. Metagenomic sequencing corroborates differences in the gut microbiota between ethnicities. (A) Principal coordinate analysis of Bray-Curtis distances reveals significant separation between ethnic groups (ADONIS test values shown). Each point represents a single individual's gut microbiota based upon shotgun sequencing. **(B)** PERMANOVA calculations for metadata variables on the x-axis with relation to variance in shotgun transformed species data with resulting effect size plotted on the y-axis. Variables are colored by metadata type (see inset; **p*<0.05, ADONIS). **(C)** Each point represents the average relative abundance for a given species within each ethnic group, connected with a line that is colored by the ethnic group with higher mean abundance of each species: EA (blue) and W (orange). Solid lines highlight four bacterial species that are significantly different between ethnicity (*p*<0.05, ALDEx2, also shown in panel D). (**D,E)** Volcano plot of ALDEx2 differential abundance testing on **(D)** genera and **(E)** species level shotgun data. Significantly different (*p*<0.05) features are highlighted in black and labeled by the most specific taxonomic assignment. **(A-E)** n=21 EA and n=24 W individuals. Data reflects metagenomic sequencing.

Figure 3. Metabolomics and targeted metabolite profiling highlight significant differences in bacterial fermentation end-products between ethnicities. (A) Global profiling of the stool metabolome by proton nuclear magnetic resonance (^{1}H NMR) revealed a significant separation in metabolomic profiles between EA and W individuals (ADONIS test values shown). **(B)** Representative stool metabolites contributing to the separation of stool metabolomic profiles between EA and W individuals (p<0.05, Wilcoxon rank-sum test). **(C-E)** Gas chromatography-mass spectrometry analysis of short-chain fatty acids (SCFAs) revealed significantly higher concentrations of acetate **(C)**, propionate **(D)** and isobutyrate **(E)** in the stool samples of EA compared to W individuals. p-values determined using Wilcoxon rank-sum tests. **(A-E)** n=10 EA and n=10 W individuals.

Figure 4. Ethnicity-associated differences in gut microbial diversity and community structure are more pronounced in lean individuals. (A,B) Bacterial richness is negatively correlated with (A) BMI and (B) percent body fat in W but not EA individuals (Spearman rank correlation coefficients and p-values are shown for each graph). **(C)** Microbial diversity metrics are more distinct between ethnic groups in lean relative to obese individuals. p-values determined using Wilcoxon rank-sum tests. **(D)** Principal coordinate analysis of PhILR Euclidean distances reveals significant separation between the gut microbiotas of EA and W lean individuals, with no separation in obese subjects (ADONIS test values shown). **(A-D)** n=12 EA lean, 10 EA obese, 11 W lean, and 13 W obese individuals. Data reflects 16S-seq. **(E)** Global profiling of the stool metabolome by proton nuclear magnetic resonance (^{1}H NMR) stratified by lean and obese individuals reveals a significant difference in the metabolomic profiles of lean EA and W individuals that is not detectable in obese individuals (ADONIS test values shown; n=5 individuals/group).

Figure 5. Ethnicity-associated bacterial taxa in lean and obese individuals. (A) 5/6 phyla that were differentially abundant between ethnicities (see **Fig. 1C**) were also significantly different between lean EA and W individuals. Three phyla were significantly different between obese EA and W individuals (p<0.05, Wilcoxon rank-sum test). **(B,C)** Volcano plot of ALDEx2 differential abundance testing on genera in stool microbiotas of lean (**B**) and obese (**C**) EA

versus W individuals, with significantly different genera highlighted (FDR<0.1, ALDEx2). **(A-C)** n=12 EA lean, 10 EA obese, 11 W lean, and 13 W obese individuals. Data reflects 16S-sequencing.

Figure 6. Ethnicity-associated differences in the gut microbiota of lean individuals correlate with geographic location. (A) Each symbol represents a subject's ZIP code. Symbols are colored by ethnicity with shape representing lean and obese subjects (n=44, data was unavailable for 2 subjects; **Supplementary File 1B**). **(B)** A subset of ZIP Code Tabulation Areas (ZCTAs) zoomed in to focus on San Francisco are colored by the proportion of each ethnicity (n=27 ZTCAs). The red star indicates a central point (latitude=37.7585102, longitude=-122.4539916) within San Francisco used for distances calculated in (C). **(C)** Distance to the center of San Francisco, which is indicated by a star in (B), for IDEO subjects stratified by ethnicity and BMI (n=9-13 individuals/group, *p*-values indicate Wilcoxon rank-sum test). **(D)** US census data for EA and W residents in ZCTAs from (B) is displayed by ethnic make-up (a total of 489,117 W and 347,200 Asian individuals in these areas). **(E,F)** PCoA principal coordinate axis 1 from PhILR Euclidean distances of the 16S-seq data is significantly correlated with (**E**) the distance of subject's ZIP code to the center of San Francisco and **(F)** the subject's duration of residence in the SF Bay Area (n=44 subjects; Spearman rank correlation). Data in E and F reflects 16S-sequencing.

Figure 7. Differences in the human gut microbiota between ethnicities are maintained following transplantation to germ-free mice. (A,F) Principal coordinate analysis of PhILR Euclidean distances of stool from germ-free recipient mice transplanted with stool microbial communities from lean EA or W donors and fed either a LFPP (**A**, combined results from two independent experiments; n=12 recipient mice per group) or HFHS (**F**, n=10 recipient mice per group) diet. Significance was assessed by ADONIS. Germ-free mice receiving the same donor sample are connected by a dashed line. Experimental designs are shown in **Figure 7-figure supplement 1**. **(B, G)** Principal coordinate analysis of PhILR Euclidean distances comparing donor input slurry (diamonds) and stool from recipient mice (circles) in the combined LFPP experiments (**B**, n=4 donors, 24 recipients) and HFHS experiment (**G**, n=10 donors, 20 recipients;for one donor sample, two separate slurries were prepared to inoculate the recipient

mice on separate days due to constraints on germ-free mice availability, resulting in 11 diamonds on the plot) respectively. See also donor metadata in **Supplementary File 1B,H**. **(C,H)** Bacterial richness is significantly higher in mice who received stool samples from W donors compared to EA donors on both the LFPP (**C**) and HFHS (**H**) diets. *p*-values determined using Wilcoxon rank-sum tests. **(D,I)** Volcano plot of ALDEx2 differential abundance testing on genera in the stool microbiomes between transplant groups. The x-axis represents the fold difference between EA (numerator) and W (denominator) subjects. The y-axis is proportional to the false discovery rate (FDR). Black dots indicate significantly different genera (FDR<0.1). *Bacteroides* and *Parabacteroides* (labelled in the volcano plots) are more abundant in mice that received stool samples from EA compared to W donors on both the LFPP (**D**) and HFHS (**I**) diets. See also **Supplementary File 1J** for the full list of significant genera. **(E,J)** Abundance of the *Bacteroides* genus in mice fed the LFPP (**E**) and HFHS (**J**) diets (ALDEx2 FDR shown). Data reflects 16S-seq.

Figure 8. Microbiome transplantation of samples from EA and W individuals differentially affects the body composition of genetically identical recipient mice. (A-C, E-G) Percent change in body weight **(A,E)**, fat mass **(B,F)**, and lean mass **(C,G)** relative to baseline are shown on the LFPP (**A-C**) and HFHS (**E-G**) diets. *p*-values determined using Wilcoxon rank-sum tests. **(D,H)** Glucose tolerance test results were not significantly different between groups on either diet. *p*-values determined using linear mixed effects models with mouse as a random effect. **(A-C)** n=12 recipient mice per group (combined data from two independent experiments). **(D)** n=6 recipient mice per group from a single experiment. **(E-H)** n=10 recipient mice per group. Experimental designs are shown in **Figure 7-figure supplement 1** and donor phenotypic data is in **Supplementary File 1H**.

SUPPLEMENTARY FIGURE LEGENDS

Figure 1-figure supplement 1. The gut microbiota can be used to predict ethnicity. (A) A phylogenetic tree of all ASVs generated from 16S-seq is shown. Leaves are colored by phyla. The inner circle indicates differential abundance (EA_{CLR}-W_{CLR}) between ethnicities. The outer circle is colored by significance (p<0.05, gray, FDR<0.1, black, Welch's t-test; labeled in **Fig. 1F**). **(B-D)** A random forest classifier was developed utilizing ASV data **(B,C)** and PhILR transformed ASV data (**D**) representing phylogenetic nodes on the tree visualized in panel A. 46 classifiers were trained on a subset of 45 individuals and then used to predict the remaining individual (leave-one-out cross-validation). **(B)** ASVs in the top 90th percentile for median mean decrease in Gini are plotted. Each dot represents the value for mean decrease in Gini for a given classifier (n=46 total classifiers made up of a subset of 45 samples). **(C-D)** Receiver operating characteristic curves for ASV data **(C)** and phylogenetic nodes obtained utilizing PhILR transformation **(D)** are plotted with values of area under the receiver operator curve (AUC) and 95% confidence intervals displayed. **(E)** CLR abundances of ASVs in the top 90th percentile of median mean decrease Gini, in the same order as shown in panel B (*p<0.05, Wilcoxon rank-sum test between ethnicity). **(F)** No significant difference in overall gut microbial colonization assessed by qPCR quantification of 16S rRNA gene copies per gram wet weight (n=13 EA, n=21 W, Wilcoxon rank-sum test).

Figure 3-figure supplement 1. Stool concentrations of branched chain amino acids and bile acids are comparable between East Asian and White subjects. We did not detect a significant difference in the concentrations of BCAAs **(A)** or bile acids **(B)** between EA (n=10) and W (n=10) stool samples. Statistical analyses performed using Wilcoxon rank-sum tests.

Figure 4-figure supplement 1. Microbial diversity metrics are consistently and negatively correlated with metabolic parameters in White individuals. **(A,B)** Faith's diversity is significantly correlated with (A) BMI and (B) percent body fat in W but not EA individuals. **(C,D)** Shannon diversity is significantly correlated with (C) BMI in both W and EA individuals, and with (D) percent body fat in W but not EA individuals. Spearman rank correlation

coefficients and *p*-values are shown for each graph (n=12 EA lean, 10 EA obese, 11 W lean, and 13 W obese individuals). Data reflects 16S-seq.

Figure 6-figure supplement 1. Variables associated with variance in microbial 16S-seq data. PERMANOVA calculations for metadata variables on the x-axis with relation to variance in PhILR-transformed 16S-seq data were calculated using the vegan package "adonis". The resulting effect size is plotted on the y-axis. Variables are colored by metadata type (see inset, *$p<0.05$, ADONIS).

Figure 6-figure supplement 2. Birth location of subjects. Symbols representing subjects' birth locations are plotted on a world map. The size, shape and color of the symbols represent the number, BMI and ethnicity of subjects at each location.

Figure 6-figure supplement 3. Identification of bacterial taxa associated with short-term dietary intake. Spearman's correlation was calculated between all 16S-seq ASVs and metagenomic species relative to ASA24 data for lean W **(A)** and lean EA **(B)** subjects. Colored boxes indicate correlations that meet an FDR < 0.1 cutoff and the direction and intensity of the Spearman's correlation are shown with correlation color indicated in the figure legend. ASVs are indicated by an ASV ID followed by the most specific taxonomic annotation available. No ASV-level associations were detected in lean EA subjects.

Figure 7-figure supplement 1. Experimental designs for gnotobiotic experiments. (A) LFPP1 experiment: Germ-free mice fed a LFPP diet received an aliquot of stool from a donor of either ethnicity and were monitored for 6 weeks (per donor n=6 recipient mice, 1 isolator, 2 cages). **(B)** LFPP2 experiment: Same experimental design as LFPP1 but colonization time was shortened to 3 weeks and two new donor samples were used (per donor n=6 recipient mice, 1 isolator, 2 cages). (**C**) HFHS experiment: 5 lean EA and 5 W donors' stool microbial communities were transplanted into 20 germ-free recipient mice fed a HFHS diet for 4 weeks prior to colonization and maintained on diet for another 3 weeks post-transplantation (per donor n=2 recipient mice, 1 IsoCage).

Figure 7-figure supplement 2. Engraftment efficiency is comparable between donor groups. There was no significant difference between groups in **(A)** the PhILR distance of recipient mice to their respective donors or **(B)** the proportion of donor ASVs detected in the recipient mice (Wilcoxon rank-sum test). **(C)** Relative abundance of shared and unique ASVs in each donor and the corresponding recipient mice.

Figure 7-figure supplement 3. Combined analysis of recipient mice reveals significant associations with donor ethnicity and recipient diet. A PhILR PCoA is plotted based on 16S-seq data from all gnotobiotic experiments. Individual mice are colored by **(A)** donor ethnicity or **(B)** the recipient's diet. Both ethnicity and diet were statistically significant contributors to variance (ADONIS *p*-values and estimated variance displayed using blocks restricted by donor identifiers to account for one donor going to multiple recipient mice). We also observed a trend for interaction between diet and ethnicity in this model (p=0.068, R^2=0.047, ADONIS).

Supplementary File 1. This file contains 10 supplementary tables that include detailed metadata, metabolomics data, and data visualized in the main text and supplemental figures for the manuscript.

MATERIALS AND METHODS

Key Resources Table				
Reagent type (species) or resource	**Designation**	**Source or reference**	**Identifiers**	**Additional information**
biological sample (*Homo Sapiens*)	Stool	This paper		n = 46 subjects (22 EA, 24 W)
commercial assay or kit	Wizard SV 96 Genomic DNA kit	Promega	Cat #A2370	
commercial assay or kit	ZymoBIOMICS 96 MagBead DNA Kit	Zymo Research	Cat #D4302	
software, algorithm	R	CRAN	v3.5.3, v4.0.2	r-project.org
software, algorithm	QIIME2	(Bolyen et al., 2019)	v2020.2	qiime2.org
software, algorithm	DADA2	(Callahan et al., 2016)		benjjneb.github.io/dada2
software, algorithm	MicrobeR	(Bisanz, 2017)	v0.3.2	github.com/jbisanz/MicrobeR
software, algorithm	qiime2R	(Bisanz, 2018)	v0.99.34	github.com/jbisanz/qiime2R
software, algorithm	MetaPhlAn2	(Truong et al., 2015)	v2.7.7	huttenhower.sph.harvard.edu/metaphlan2

software, algorithm	Vegan	(Oksanen et al., 2013)	v2.5-6	github.com/vegandevs/vegan
software, algorithm	APE	(Paradis and Schliep, 2019)	v5.3	ape-package.ird.fr
software, algorithm	Picante	(Kembel et al., 2010)	v1.8.1	github.com/skembel/picante
software, algorithm	PhILR	(Silverman et al., 2017)	v1.12.0	github.com/jsilve24/philr
software, algorithm	ALDEx2	(Fernandes et al., 2013)	v1.18.0	github.com/ggloor/ALDEx_bioc
software, algorithm	GGMaps	(Kahle and Wickham, 2013)	v3.0.0.902	https://github.com/dkahle/ggmap
software, algorithm	Open Street Maps	https://www.openstreetmap.org		https://wiki.openstreetmap.org/wiki/Main_Page
software, algorithm	Leaflet	https://www.openstreetmap.org	v1.6.0	rstudio.github.io/leaflet
software, algorithm	Imap	(Wallace, 2012)	v1.32	https://rdrr.io/cran/Imap
Strain, strain background (*mus musculus*)	C57BL/6J mice, germ-free	UCSF Gnotobiotics Core		
other	2018 U.S. Census data	data.census.gov		Table B02001: Race

555

Human subjects

The IDEO cohort was established to explore the pathogenesis of obesity and metabolic diseases in highly vulnerable segments of the population. It includes men and women of multiple ethnicities recruited from the general medicine, endocrinology, diabetes, general surgery, and bariatric surgery clinics at the University of California San Francisco (UCSF) and Zuckerberg San Francisco General Hospital and by public advertisements throughout the local San Francisco Bay Area. All study participants were part of the IDEO cohort, which has been previously described (Alba et al., 2018; Oguri et al., 2020). Briefly, IDEO consists of 25-65 year-old men and women of multiple ethnicities and across a wide BMI range (18.5–52 kg/m^2) living in the San Francisco Bay Area. Using IDEO, we recruited both lean and obese W and EA individuals into this study based on World Health Organization cut-offs: W/EA BMI≤24.9 kg/m^2 (lean); W BMI≥30 kg/m^2 (obese); and EA BMI≥27.5 kg/m^2 (obese) (Hsu et al., 2015; Jih et al., 2014; WHO Expert Consultation, 2004). To avoid bias towards non-English speaking participants, all documents including flyers, screening questionnaires and consents were available in Cantonese and Mandarin. Potential participants completed screening questionnaires and exclusion criteria were assessed in more detail. These included acute or chronic infections, current medications with a recognized impact on the immune system, recent antibiotic use, current smoking, recent changes in weight, active liver disease or liver failure, chronic kidney disease (eGFR<30 $mL/min/1.73m^2$), history of cancer and chemotherapy therapy within the past 5 years, psychiatric and neurological disorders, prior bariatric surgery, and weight >159 kg (the DXA scanner weight limit). Whereas exclusion criteria inherently lend bias towards healthy individuals, this is done to limit the confounding effects of a wide variety of chronic diseases and environmental exposures on the comparisons being made.

IDEO also limited bias by standardizing how individuals are asked to self-identify race/ethnicity. Individuals are asked to respond to two separate questions about ethnicity (e.g. “are you of Hispanic, Latino, or Spanish origin?”) and race (“What is your race?”). Hispanic/LatinX individuals were enrolled as part of a separate IDEO sub-study from the topic of this manuscript. Participants are also asked questions about their parents’ race and ethnic background. Each participant consented to take part in the study, which was approved by the University of California San Francisco (UCSF) Committee on Human Research. We utilized demographic, medical, dietary, and lifestyle metadata on each participant that were part of their

initial recruitment into IDEO, as previously reported (Alba et al., 2018; Oguri et al., 2020). Participants with Type 2 Diabetes (T2D) were classified in accordance with American Diabetes Association Standards of Medical Care guidelines (American Diabetes Association, 2019), defined by having glycated hemoglobin (HbA1c) ≥6.5% or the combination of a prior diagnosis of T2D and the active use of an antidiabetic medication. For stool sample collection, participants took home or were mailed a stool sample collection kit and detailed instructions on how to collect the specimen. All samples were collected at home, stored at room temperature, and brought to the UCSF Clinical Research Center by the participants within 24 hours of defecation. Samples were aliquoted and stored at -80°C.

Anthropometric and body composition measurements

We leveraged host phenotypic and demographic data from IDEO, which was the focus of two previous studies (Alba et al., 2018; Oguri et al., 2020). For the convenience of the reader, we restate our methods here. Height and weight were measured using a standard stadiometer and scale, and BMI (kg/m^2) was calculated from two averaged measurements. Waist and hip circumferences (to the nearest 0.5 cm) were measured using a plastic tape meter at the level of the umbilicus and of the greater trochanters, respectively, and waist-to-hip ratio (WHR) was calculated. Blood pressure was measured with a standard mercury sphygmomanometer on the left arm after at least 10 minutes of rest. Mean values were determined from two independent measurements. Blood samples were collected after an overnight fast and analyzed for plasma glucose, insulin, serum total cholesterol, high density lipoprotein (HDL) cholesterol, and triglycerides. Low density lipoprotein (LDL) cholesterol was estimated according to the Friedewald formula (Friedewald et al., 1972). Insulin resistance was estimated by the homeostatic model assessment of insulin resistance (HOMA-IR) index calculated from fasting glucose and insulin values (Matthews et al., 1985). Two obese subjects on insulin were included in the HOMA-IR analysis (1 EA, 1 W). Body composition of the subjects was estimated by Dual-Energy X-ray Absorptiometry (DEXA) using a Hologic Horizon/A scanner (3-minute whole-body scan, <0.1 G milligray) per manufacturer protocol. A single technologist analyzed all DEXA measurements using Hologic Apex software (13.6.0.4:3) following the International Society for Clinical Densitometry guidelines. Visceral adipose tissue (VAT) was estimated from a 5 cm-wide region across the abdomen just above the iliac crest, coincident with the fourth

lumbar vertebrae, to avoid interference from iliac crest bone pixels and matching the region commonly used to analyze VAT mass by CT scan (Bredella et al., 2013; Kaul et al., 2012; Neeland et al., 2016) . The short version of the International Physical Activity Questionnaire (IPAQ) was used to assess the habitual physical activity levels of the participants. The IPAQ total score is expressed in metabolic equivalent (MET)-minutes/week (Craig et al., 2003).

Dietary assessment

IDEO participants completed two dietary questionnaires, as previously described (Alba et al., 2018; Oguri et al., 2020), allowing for the assessment of usual total fiber intake and fiber from specific sources, as well as macronutrient, phytochemical, vitamin, and mineral uptake. The first instrument was an Automated Self-Administered 24-hour Dietary Assessment (ASA24) (McClung et al., 2018; Park et al., 2018; Timon et al., 2016), which queries intake over a 24-hour period. The 24-hour recalls and supplement data were manually entered in the ASA24 Dietary Assessment Tool (v. 2016), an electronic data collection and dietary analysis program. ASA24 employs research-based strategies to enhance dietary recall using a respondent-driven approach allowing initial recall to be self-defined. The second instrument was the National Cancer Institute's Diet History Questionnaire III (DHQIII) ("Diet History Questionnaire III (DHQ III)," n.d.; Millen et al., 2006). The DHQIII queries one's usual diet over the past month ("Diet History Questionnaire III (DHQ III)," n.d.). Completing the DHQIII was associated with participant survey fatigue and completion rates were accordingly only 42% after 1 phone-based administration of the instrument, although they improved to 79% by the 2nd session and reached 100% within four sessions over a 5-month period. Due to the effort needed to achieve DHQIII completion, we modified our protocol to request completion of the simpler ASA24 at three separate times, at appointments where there were computers and personnel assistance for online completion, in addition to completion of the DHQIII questionnaire. By combining both instruments, we were able to reliably obtain complete dietary information on all participants.

DNA extraction

Human stool samples were homogenized with bead beating for 5 min (Mini-Beadbeater-96, BioSpec) using beads of mixed size and material (Lysing Matrix E 2mL Tube, MP Biomedicals) in the digestion solution and lysis buffer of a Wizard SV 96 Genomic DNA kit (Promega). The

samples were centrifuged for 10 min at 16,000 g and the supernatant was transferred to the binding plate. The DNA was then purified according to the manufacturer's instructions. Mouse fecal pellets were homogenized with bead beating for 5 min (Mini-Beadbeater-96, BioSpec) using the ZR BashingBead lysis matrix containing 0.1 and 0.5 mm beads (ZR-96 BashingBead Lysis Rack, Zymo Research) and the lysis solution provided in the ZymoBIOMICS 96 MagBead DNA Kit (Zymo Research). The samples were centrifuged for 5 min at 3,000 g and the supernatant was transferred to 1 mL deep-well plates. The DNA was then purified using the ZymoBIOMICS 96 MagBead DNA Kit (Zymo Research) according to the manufacturer's instructions.

16S rRNA gene sequencing and analysis

For human samples, 16S rRNA gene amplification was carried out using GoLay-barcoded 515F/806R primers (Caporaso et al., 2012) targeting the V4 region of the 16S rRNA gene according to the methods of the Earth Microbiome Project (earthmicrobiome.org) (**Supplementary File 1B**). Briefly, 2 µL of DNA was combined with 25 µL of AmpliTaq Gold 360 Master Mix (Fisher Scientific), 5 µL of primers (2 µM each GoLay-barcoded 515/806R), and 18 µL H_2O. Amplification was as follows: 10 min 95°C, 30x (30s 95°C, 30s 50°C, 30s 72°C), and 7 min 72°C. Amplicons were quantified with PicoGreen (Quant-It dsDNA; Life Technologies) and pooled at equimolar concentrations. Aliquots of the pool were then column (MinElute PCR Purification Kit; Qiagen) and gel purified (QIAquick Gel Extraction Kit; Qiagen). Libraries were then quantified (KAPA Library Quantification Kit; Illumina) and sequenced with a 600 cycle MiSeq Reagent Kit (250x150; Illumina) with ~15% PhiX spike-in. For mouse samples, 16S rRNA gene amplification was carried out as per reference protocol and primers (Gohl et al., 2016). In brief, the V4 region of the 16S rRNA gene was amplified with 515F/806R primers containing common adaptor sequences, and then the Illumina flow cell adaptors and dual indices were added in a secondary amplification step (see **Supplementary File 1I** for index sequences). Amplicons were pooled and normalized using the SequalPrep Normalization Plate Kit (Invitrogen). Aliquots of the pool were then column (MinElute PCR Purification Kit, Qiagen) and gel purified (QIAquick Gel Extraction Kit, Qiagen). Libraries were then quantified and sequenced with a 600 cycle MiSeq Reagent Kit (270x270; Illumina) with ~15% PhiX spike-in.

Demultiplexed sequencing reads were processed using QIIME2 v2020.2 (Bolyen et al., 2019) with denoising by DADA2 (Callahan et al., 2016). Taxonomy was assigned using the DADA2 implementation of the RDP classifier (Wang et al., 2007) using the DADA2 formatted training sets for SILVA version 138 (benjjneb.github.io/dada2/assign.html). For Amplicon Sequence Variants (ASV) analysis, we utilized quality scores to set truncation and trim parameters. The reverse read of human 16S data suffered from low sequence quality and reduced the overall ASV counts, so we therefore analyzed only the forward reads, although a separate analysis using merged forward and reverse reads complemented the findings we report in this manuscript. For the manuscript, forward reads were truncated to 220 base pairs and underwent an additional 5 base pairs of trimming for 16S analysis of human stool. For gnotobiotic mice, forward and reverse reads were truncated to 200 and 150 base pairs respectively. ASVs were filtered such that they were present in more than one sample with at least a total of 10 reads across all samples. Alpha diversity metrics were calculated on subsampled reads using Vegan (Dixon, 2003) and Picante (Kembel et al., 2010) R packages. The PhILR Euclidean distance was calculated by first carrying out the phylogenetic isometric log ratio transformation (philr, PhILR (Silverman et al., 2017)) followed by calculating the Euclidean distance (vegdist, Vegan (Dixon, 2003)). Principal coordinates analysis was carried out using the pcoa function of APE (Paradis et al., 2004). ADONIS calculations were carried out (adonis, Vegan) with 999 replications on each distance metric. The permutational space for the adonis calculation for the 3 combined gnotobiotic experiments was restricted by donor identifier to account for multiple recipient mice for a given donor and applied to **Figure 7-figure supplement 3** using setblocks to define permutations and specifying these blocks in the command. Centered $\log_2$-ratio (CLR) normalized abundances were calculated using the Make.CLR function in MicrobeR package (Bisanz, 2017) with count zero multiplicative replacement (zCompositions; (Martín-Fernández et al., 2015)). ALDEx2 (Fernandes et al., 2013) was used to analyze differential abundances of count data, using features that represented at least 0.05% of total sequencing reads. Corrections for multiple hypotheses using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) were performed where applicable. Where described, a false discovery rate (FDR) indicates the Benjamini-Hochberg adjusted p-value for an FDR (0.1 unless otherwise specified). Analysis of distance matrices and alpha diversity mirror prior analyses developed in the Turnbaugh lab and were adapted to the current manuscript (Bisanz et al., 2019). Calculations of associations

between ASVs and ASA24 questionnaire data were completed by calculating a Spearman rank correlation and then adjusting the p-value for a Benjamini-Hochberg FDR using the cor_pmat function in the R package ggcorrplot (Kassambara and Kassambara, 2019) for all CLR transformed ASVs detected between ethnic groups. Shotgun data for each ethnicity was processed using Metaphlan2 and the species associations were calculated for relative abundance by ASA24 questionnaire data separate from the ASV data. The randomForest package (Liaw et al., 2002) was employed to generate random forest classifiers. Given the total number of samples (n=46) we generated 46 classifiers trained on a subset of 45 samples and used each classifier to predict the sample left out. AUCs are visualized utilizing the pROC (Robin et al., 2011) and ROCR (Sing et al., 2005) packages.

Metagenomic sequencing and analysis

Whole-genome shotgun libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina). Paired ends of all libraries were sequenced on the NovaSeq 6000 platform in a single sequencing run (n=45 subjects; see **Supplementary File 1B** for relevant metadata and statistics). Illumina reads underwent quality trimming and adaptor removal using fastp (Chen et al., 2018) and host read removal using BMTagger v1.1.0 (ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/) in the metaWRAP pipeline (github.com/bxlab/metaWRAP) (Uritskiy et al., 2018). Metagenomic samples were taxonomically profiled using MetaPhlan2 v2.7.7 (Truong et al., 2015) and functionally profiled using HUMAnN2 v0.11.2 (Franzosa et al., 2018), both with default parameters. Principal coordinates analysis on MetaPhlan2 species-level abundances was carried out using Bray Curtis distances and the pcoa function of APE (Paradis et al., 2004). Metaphlan2 abundance outputs were converted to counts and subsampled to even sample depth. Differences between groups were determined utilizing the Aldex2 package as described above. Tables of gene family abundances from HUMAnN2 were regrouped to KEGG orthologous groups using humann2_regroup_table. Functional pathways relating to short-chain fatty acid production were manually curated from the pathway outputs from HUMANn2 and normalized by the estimated genome equivalents in each microbial community obtained from MicrobeCensus (Nayfach and Pollard, 2015).

Quantification of bacterial load

Absolute 16S rRNA gene copy number was derived by adjustments for dilutions during DNA extraction and template normalization dividing by the total fecal mass used for DNA extraction in grams. Quantification of bacterial load was conducted using quantitative PCR (qPCR) given stool samples were frozen for the IDEO cohort as described above and bacterial lysis was achieved with a preparation including both bead beating and a detergent. Differences in 16S rRNA gene copy number between bacterial strains may have masked more subtle differences in colonization level. qPCR was performed on DNA extracted from the human stool samples. DNA templates were diluted 1:10 into a 96-well plate. Samples were aliquoted in a 384-well plate, and PCR primers and iTaq Universal Probes Supermix were added utilizing an Opentrons OT-2 instrument then analyzed on a BioRad CFX384 thermocycler with an annealing temperature of 60˚C. The following primers including a FAM labeled PCR probe was used for quantification: 891F, TGGAGCATGTGGTTTAATTCGA; 1003R, TGCGGGACTTAACCCAACA; 1002P, [6FAM]CACGAGCTGACGACARCCATGCA[BHQ1]. Absolute quantifications were determined against a standard curve of purified 8F/1542R amplified *Vibrio casei* DNA. Reactions identified as inappropriately amplified by the instrument were rejected, and the mean values were used for downstream analysis. Absolute 16S rRNA gene copy number was derived by adjustments for dilutions during DNA extraction and template normalization dividing by the total fecal mass used for DNA extraction in grams. Quantification of bacterial load was conducted using qPCR given stool samples were frozen for the IDEO cohort as described above and bacterial lysis was achieved with a preparation including both bead beating and a detergent.

Nuclear magnetic resonance (NMR) metabolomics

NMR spectroscopy was performed at 298K on a Bruker Avance III 600 MHz spectrometer configured with a 5 mm inverse cryogenic probe (Bruker Biospin, Germany) as previously described (Cai et al., 2017). Lean and obese EA and W individuals (n=20 total individuals, five in each group) were selected and matched based on body composition and metabolic parameters. Stool samples from these subjects were subjected to NMR-based metabolomics. 50 mg of human feces were extracted with 1 mL of phosphate buffer (K_2HPO_4/NaH_2PO_4, 0.1 M, pH 7.4, 50% v/v D_2O) containing 0.005% sodium 3-(trimethylsilyl) [2,2,3,3-2H4] propionate (TSP-d_4) as a chemical shift reference (δ 0.00). Samples were freeze-thawed three times with liquid nitrogen

and water bath for thorough extraction, then homogenized (6500 rpm, 1 cycle, 60 s) and centrifuged (11,180 g, 4 °C, 10 min). The supernatants were transferred to a new 2 mL tube. An additional 600 μL of PBS was added to the pellets, followed by the same extraction procedure described above. Combined fecal extracts were centrifuged (11,180 g, 4°C, 10 min), 600 μL of the supernatant was transferred to a 5 mm NMR tube (Norell, Morganton, NC) for NMR spectroscopy analysis. A standard one-dimensional NOESY pulse sequence noesypr1d (recycle delay-90°-t1-90°-tm-90°-acquisition) is used with a 90 pulse length of approximately 10s (-9.6 dbW) and 64 transients are recorded into 32k data points with a spectral width of 9.6 KHz. NMR spectra were processed as previously described (Cai et al., 2017). First, spectra quality was improved with Topspin 3.0 (Bruker Biospin, Germany) for phase and baseline correction and chemical shift calibration. AMIX software (version: 3.9.14, Bruker Biospin, Germany) was used for bucketing (bucket width 0.004 ppm), removal of interfering signal, and scaling (total intensity). Relative concentrations of identified metabolites were obtained by normalized peak area.

Targeted gas chromatography mass spectrometry (GC-MS) assays

Targeted analysis of short-chain fatty acids (SCFAs) and branched chain amino acids (BCAAs) was performed with an Agilent 7890A gas chromatograph coupled with an Agilent 5975 mass spectrometer (Agilent Technologies Santa Clara, CA) using a propyl esterification method as previously described (Cai et al., 2017). 50 mg of human fecal samples were pre-weighed, mixed with 1 mL of 0.005 M NaOH containing 10 μg/mL caproic acid-6,6,6-d3 (internal standard) and 1.0 mm diameter zirconia/silica beads (BioSpec, Bartlesville, OK). The mixture was thoroughly homogenized and centrifuged (13,200 g, 4°C, 20 min). 500 μL of supernatant was transferred to a 20 mL glass scintillation vial. 500 μL of 1-propanol/pyridine (v/v=3/2) solvent was added into the vial, followed by a slow adding of an aliquot of 100 μL of esterification reagent propyl chloroformate. After a brief vortex of the mixture for 1 min, samples were derivatized at 60°C for 1 hour. After derivatization, samples were extracted with hexane in a two-step procedure (300 μL + 200 μL) as described (Zheng et al., 2013). First, 300 μL of hexane was added to the sample, briefly vortexed and centrifuged (2,000g, 4°C, 5 min), and 300 μL of the upper layer was transferred to a glass autosampler vial. Second, an additional 200 μL of hexane was added to the sample, vortexed, centrifuged, and the 200 μL upper layer was transferred to the glass

autosampler vial. A combination of 500 µL of extracts were obtained for GC-MS analysis. A calibration curve of each SCFA and BCAA was generated with series dilution of the standard for absolute quantitation of the biological concentration of SCFAs and BCAAs in human fecal samples.

Targeted bile acid quantitation by UHPLC-MS/MS

Bile acid quantitation was performed with an ACQUITY ultra high pressure liquid chromatography (UHPLC) system using a Ethylene Bridged Hybrid C8 column (1,7 µm, 100 mm x 2.1 mm) coupled with a Xevo TQ-S mass spectrometer equipped with an electrospray ionization (ESI) source operating in negative mode (All Waters, Milford, MA) as previously described (Sarafian et al., 2015). Selected ion monitoring (SIM) for non-conjugated bile acids and multiple reaction monitoring (MRM) for conjugated bile acids was used. 50 mg of human fecal sample was pre-weighed, mixed with 1 mL of pre-cooled methanol containing 0.5 µM of stable-isotope-labeled bile acids (internal standards) and 1.0 mm diameter zirconia/silica beads (BioSpec, Bartlesville, OK), followed by thorough homogenization and centrifugation. Supernatant was transferred to an autosampler vial for analysis. 100 µL of serum was extracted by adding 200 µL pre-cooled methanol containing 0.5 µM deuterated bile acids as internal standards. Following centrifugation, the supernatant of the extract was transferred to an autosampler vial for quantitation. Calibration curves of individual bile acids were drafted with bile acid standards for quantitation of the biological abundance of bile acids.

Gnotobiotic mouse experiments

All mouse experiments were approved by the UCSF Institutional Animal Care and Use Committee and performed accordingly. Germ-free mice were maintained within the UCSF Gnotobiotic Core Facility and fed *ad libitum* autoclaved standard chow diet (Lab Diet 5021). Germ-free adult male C57BL/6J mice between 6-10 weeks of age were used for all the experiments described in this paper. 10 lean subjects in our IDEO cohort were selected as donors for the microbiota transplantation experiments, including 5 EA and 5 W donors. The selected donors for gnotobiotic experiments were matched for phenotypic data to the degree possible (**Supplementary File 1H**). Stool samples to be used for transplantation were resuspended in 10 volumes (by weight) of brain heart infusion media in an anaerobic Coy chamber. Each diluted

sample was vortexed for 1 min and left to settle for 5 min, and a single 200 µL aliquot of the clarified supernatant was administered by oral gavage into each germ-free mouse recipient. In experiments LFPP1 and LFPP2, microbiome transplantations were performed for 2 donors per experiment (1 W, 1 EA) with gnotobiotic mice housed in sterile isolators (CBC flexible, softwall isolator) and maintained on *ad libitum* standard chow also known as low-fat, high-plant-polysaccharide (LFPP) diet. In LFPP1, 6 germ-free mice per colonization group received an aliquot of stool from a donor of either ethnicity and body composition (measured using EchoMRI) were recorded on the day of colonization and at 6 weeks post-transplantation (per group n=6 recipient mice, 1 isolator, 2 cages). In LFPP2, we shortened the colonization time to 3 weeks and used two new donor samples. For the third experiment (HFHS experiment), mice were weaned onto an irradiated high-fat, high-sugar diet (HFHS, TD.88137, Envigo) for four weeks prior to colonization and housed in pairs in Tecniplast IsoCages. The same 4 donors from LFPP1 and LFPP2 were included in the HFHS experiment, in addition to 6 new donors (per donor n=2 recipient mice, 1 IsoCage). Body weight and body composition were recorded on the day of colonization and again at 3 weeks post-transplantation. Mice were maintained on the HFHS diet throughout the experiment. All samples were sequenced in a single pool (**Supplementary File 1I**). For comparisons between donors and recipient mice, donors and recipient mice were subsampled to even sequencing depth and paired between donor and recipient mice (range: 18,544-78,361 sequencing reads/sample).

Glucose tolerance tests

Food was removed from mice 10 hr (LFPP1 experiment) or 4 hr (HFHS experiment) prior to assessment of glucose tolerance. Mice received i.p. injections of D-glucose (2 mg/kg), followed by repeated collection of blood by tail nick and determination of glucose levels by handheld glucometer (Abbott Diabetes Care) over a 2-hour period.

Geographic analyses

Map tiles and distance data was obtained using GGMap (Kahle and Wickham, 2013), OpenStreet Maps (Fellows and Stotz, 2016), and the Imap R (Wallace, 2012) packages. GGMap was employed using a Google Cloud API key and the final map tiles were obtained in July 2020 (Kahle and Wickham, 2013). Spearman ranked correlation coefficients (*rho)* were calculated as

embedded in the ggpubr (Kassambara, 2018) R package. 2018 US Census data for EA and W subjects was obtained (B02001 table for race, data.census.gov) for the ZIP codes available in our study and using the leaflet (Cheng et al., 2018) package. The census data used is included as part of **Supplementary File 1B** to aid in reproduction. Each census region is plotted as a percentage of W individuals over a denominator of W and EA subjects. The leaflet package utilized ZIP Code Tabulation Areas (ZCTAs) from the 2010 census. We extracted all ZCTAs starting with 9, and the resulting 29 ZIP codes that overlap with IDEO subjects were analyzed (**Supplementary File 1B)**. Two ZCTAs (95687 and 95401) were primarily W when comparing W and EA subjects. There were two W subjects recruited from these ZTCAs. These ZIP codes are cut off based on the zoom magnification for that figure and as a result ZTCAs for 27 individuals are plotted. Distance to a central point in SF was calculated. The point of reference was latitude=37.7585102, longitude=-122.4539916.

Dietary questionnaire correlation analysis

DHQIII and ASA24 data were analyzed using a Euclidean distance matrix. These transformations were completed using the cluster package (Maechler et al., 2012). Subsequent analysis was completed using the vegan package (Dixon, 2003; Oksanen et al., 2013). Procrustes transformations were performed using 16S-seq data from human subjects, which was then subjected to a PhILR transformation. The resulting matrix was rotated against the distance matrix for ASA24 or DHQIII questionnaire data using the procrustes command in the vegan R package using 999 permutations. Mantel statistics were calculated utilizing the mantel command of the vegan package.

R packages used in this study

Picante (Kembel et al., 2010), PhILR (Silverman et al., 2017), MicrobeR (Bisanz, 2017), ALDEx2 (Fernandes et al., 2013), ggcorrplot (Kassambara and Kassambara, 2019), randomForest (Liaw et al., 2002), GGMap (Kahle and Wickham, 2013), OpenStreetMap (Fellows and Stotz, 2016), IMap (Wallace, 2012), ggpubr (Kassambara, 2018), leaflet (Cheng et al., 2018), cluster (Maechler et al., 2012), readxl (Wickham and Bryan, 2017), Rtsne (Krijthe, 2015), vegan (Dixon, 2003; Oksanen et al., 2013), ape (Paradis and Schliep, 2019), tigris (Walker, 2018), lmerTest (Kuznetsova et al., 2017), qiime2R (Bisanz, 2018), gghighlight

(Yutani, 2018), Phyloseq (McMurdie and Holmes, 2013), Janitor (Firke, 2018), table 1 (Rich, 2020), ggplot2 (Wickham, 2016).

Statistical analyses

Statistical analysis of the human data was performed using the table1 package in R (STATCorp LLC. College Station, TX). Human data were presented as mean ± SD. Unpaired independent Student's *t* tests were used to compare differences between the two groups in the case of continuous data and in the case of categorical data the χ^2 test was utilized for **Supplementary File 1A**. For a given lean or obese categories between ethnicity tests were adjusted for a Benjamini-Hochberg false discovery rate utilizing the command p.adjust in R, which is indicated as an adjusted *p*-value in the tables and none were significant as described in the table legend. In **Supplementary File 1G,H** no values met an adjusted *p*-value cutoff of < 0.1. In **Supplementary File 1A**, *p*-values indicated by numbers were pooled together for adjustments and those represented by symbols were separately pooled together for adjustment. All microbiome-related analyses were carried out in R version 3.5.3 or 4.0.2. Where indicated, Wilcoxon rank-sum tests were calculated. A Benjamini-Hochberg adjusted *p*-value (FDR) of 0.1 was used as the cutoff for statistical significance unless stated otherwise. Statistical analysis of glucose tolerance tests was carried out using linear mixed effects models with the lmerTest (Kuznetsova et al., 2017) R package and mouse as random effect. Graphical representation was carried out using ggplot2. Boxplots indicate the interquartile range (25th to 75th percentiles), with the center line indicating the median and whiskers representing 1.5x the interquartile range.

LIST OF ABBREVIATIONS

16S-seq: 16S rRNA gene sequencing

ASA24: Automated Self-Administered 24-Hour Dietary Assessment Tool

ASV: Amplicon Sequence Variant

BCAA: branched chain amino acid

BMI: body mass index

DHQIII: Diet History Questionnaire III

EA: East Asian

HFHS: high-fat, high-sugar

LFPP: low-fat, plant-polysaccharide rich

SCFA: short-chain fatty acid

W: White

ZCTA: ZIP Code Tabulation Area

DECLARATIONS

Ethics approval and consent to participate

Human stool samples were collected as part of a multi-ethnic clinical cohort study termed Inflammation, Diabetes, Ethnicity and Obesity (ClinicalTrials.gov identifier NCT03022682), consisting of 25- to 65-year-old men and women residing in Northern California and recruited from medical and surgical clinics at UCSF and the Zuckerberg San Francisco General Hospital, or through local public advertisements. The host phenotypic data from this cohort have been described in detail (Alba et al., 2018; Oguri et al., 2020). Informed consent was provided for all subjects participating in the study, which was approved by the UCSF Institutional Review Board (IRB #14-14248). Protocols for all experiments involving mice were approved by the University of California, San Francisco Institutional Animal Care and Use Committee, and performed accordingly (UCSF IACUC numbers AN183950 and AN184143).

Consent for publication

Not applicable.

Availability of data and materials

All 16S-seq and metagenomic sequencing data generated in the preparation of this manuscript have been deposited in NCBI's Sequence Read Archive under accession number PRJNA665061. Metabolomics results and metadata are available within this manuscript (**Supplementary File 1**). Code for our manuscript and a more comprehensive metadata table is available on GitHub (https://github.com/turnbaughlab/IDEO).

Competing interests

P.J.T. is on the scientific advisory board for Kaleido, Pendulum, and SNIPRbiome; there is no direct overlap between the current study and these consulting duties. All other authors declare that they have no competing interests.

Funding

This work was supported by the National Institutes of Health [R01HL122593, R01AR074500, R01DK114034 (PJT); R01DK11230401, R01DK11230403S1, P30DK098722 (SKK)]. DLA is supported by the American Diabetes Association (1-18-PMF-003). VU is supported by the National Institutes of Health T32HL007185. ADP is supported in part by the Pennsylvania Department of Health using Tobacco CURE funds, and the USDA National Institute of Food and Federal Appropriations under Project PEN04607 and Accession number 1009993. None of these funding bodies played a role in the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript.

Acknowledgements

We thank Jessie Turnbaugh and the other UCSF Gnotobiotics Core Facility staff and members of the Koliwad lab for help with the gnotobiotic mouse experiments. We thank Dr. Philip B. Smith from the Penn State Metabolomics Facility. We also thank the CZ Biohub Sequencing Platform for sequencing support, as well as all the subjects who participated in this study.

REFERENCES

Alba DL, Farooq JA, Lin MYC, Schafer AL, Shepherd J, Koliwad SK. 2018. Subcutaneous Fat Fibrosis Links Obesity to Insulin Resistance in Chinese Americans. *J Clin Endocrinol Metab* **103**:3194–3204.

American Diabetes Association. 2019. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2019. *Diabetes Care* **42**:S13–S28.

Basolo A, Hohenadel M, Ang QY, Piaggi P, Heinitz S, Walter M, Walter P, Parrington S, Trinidad DD, von Schwartzenberg RJ, Turnbaugh PJ, Krakoff J. 2020. Effects of underfeeding and oral vancomycin on gut microbiome and nutrient absorption in humans. *Nat Med* **26**:589–598.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* **57**:289–300.

Berding K, Donovan SM. 2018. Diet Can Impact Microbiota Composition in Children With Autism Spectrum Disorder. *Front Neurosci* **12**:515.

Bisanz JE. 2018. qiime2R: Importing QIIME2 artifacts and associated data into R sessions. *Version 0 99* **13**.

Bisanz JE. 2017. MicrobeR v 0.3.2: Handy functions for microbiome analysis in R. *GitHub*.

Bisanz JE, Upadhyay V, Turnbaugh JA, Ly K, Turnbaugh PJ. 2019. Meta-Analysis Reveals Reproducible Gut Microbiome Alterations in Response to a High-Fat Diet. *Cell Host Microbe* **26**:265–272.e4.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible

microbiome data science using QIIME 2. *Nat Biotechnol* **37**:852–857.

Bredella MA, Gill CM, Keating LK, Torriani M, Anderson EJ, Punyanitya M, Wilson KE, Kelly TL, Miller KK. 2013. Assessment of abdominal fat compartments using DXA in premenopausal women from anorexia nervosa to morbid obesity. *Obesity* **21**:2458–2464.

Brooks AW, Priya S, Blekhman R, Bordenstein SR. 2018. Gut microbiota diversity across ethnicities in the United States. *PLoS Biol* **16**:e2006842.

Brown K, Godovannyi A, Ma C, Zhang Y, Ahmadi-Vand Z, Dai C, Gorzelak MA, Chan Y, Chan JM, Lochner A, Dutz JP, Vallance BA, Gibson DL. 2016. Prolonged antibiotic treatment induces a diabetogenic intestinal microbiome that accelerates diabetes in NOD mice. *ISME J* **10**:321–332.

Cai J, Zhang J, Tian Y, Zhang L, Hatzakis E, Krausz KW, Smith PB, Gonzalez FJ, Patterson AD. 2017. Orthogonal comparison of GC-MS and 1H NMR spectroscopy for short chain fatty acid quantitation. *Anal Chem* **89**:7900–7906.

Cai J, Zhang L, Jones RA, Correll JB, Hatzakis E, Smith PB, Gonzalez FJ, Patterson AD. 2016. Antioxidant Drug Tempol Promotes Functional Metabolic Changes in the Gut Microbiota. *J Proteome Res* **15**:563–571.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**:581–583.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**:1621–1624.

Carmody RN, Gerber GK, Luevano JM Jr, Gatti DM, Somes L, Svenson KL, Turnbaugh PJ. 2015. Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host Microbe* **17**:72–84.

Cheng J, Karambelkar B, Xie Y. 2018. Leaflet: Create interactive web maps with the javascript'leaflet'library. *R package version* **2**.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884–i890.

Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF, Oja P. 2003. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc* **35**:1381–1395.

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly

and reproducibly alters the human gut microbiome. *Nature* **505**:559–563.

De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P. 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A* **107**:14691–14696.

Depommier C, Everard A, Druart C, Plovier H, Van Hul M, Vieira-Silva S, Falony G, Raes J, Maiter D, Delzenne NM, de Barsy M, Loumaye A, Hermans MP, Thissen J-P, de Vos WM, Cani PD. 2019. Supplementation with Akkermansia muciniphila in overweight and obese human volunteers: a proof-of-concept exploratory study. *Nat Med* **25**:1096–1103.

Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, Tremaroli V, Bakker GJ, Attaye I, Pinto-Sietsma S-J, van Raalte DH, Snijder MB, Nicolaou M, Peters R, Zwinderman AH, Bäckhed F, Nieuwdorp M. 2018. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nat Med* **24**:1526–1531.

Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, Archie EA, Turnbaugh PJ, Seed KD, Blekhman R, Aarestrup FM, Thomas BC, Banfield JF. 2019. Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat Microbiol* **4**:693–700.

Diet History Questionnaire III (DHQ III). n.d. https://epi.grants.cancer.gov/dhq3/

Dixon P. 2003. VEGAN, a package of R functions for community ecology. *J Veg Sci* **14**:927–930.

Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C, De Sutter L, Lima-Mendez G, D'hoe K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaudt L, Fu J, Henckaerts L, Zhernakova A, Wijmenga C, Raes J. 2016. Population-level analysis of gut microbiome variation. *Science* **352**:560–564.

Fellows I, Stotz JP. 2016. OpenStreetMap: Access to open street map raster images. R Package Version, 0.3. 3.

Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. 2013. ANOVA-like differential gene expression analysis of single-organism and meta-RNA-seq. *PLoS One* **8**:e67019.

Firke S. 2018. Janitor: Simple tools for examining and cleaning dirty data. *R package version* **1**.

Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, Prifti E, Vieira-Silva S, Gudmundsdottir V, Pedersen HK, Arumugam M, Kristiansen K, Voigt AY, Vestergaard H, Hercog R, Costea PI, Kultima JR, Li J, Jørgensen T, Levenez F, Dore J, MetaHIT consortium, Nielsen HB, Brunak S, Raes J, Hansen T, Wang J, Ehrlich SD, Bork P, Pedersen O. 2015. Disentangling type 2 diabetes and metformin treatment signatures in

the human gut microbiota. *Nature* **528**:262–266.

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**:962–968.

Friedewald WT, Levy RI, Fredrickson DS. 1972. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* **18**:499–502.

Garduño-Diaz SD, Husain W, Ashkanani F, Khokhar S. 2014. Meeting challenges related to the dietary assessment of ethnic minority populations. *J Hum Nutr Diet* **27**:358–366.

Gaulke CA, Sharpton TJ. 2018. The influence of ethnicity and geography on human gut microbiome composition. *Nat Med*.

Gehrig JL, Venkatesh S, Chang H-W, Hibberd MC, Kung VL, Cheng J, Chen RY, Subramanian S, Cowardin CA, Meier MF, O'Donnell D, Talcott M, Spears LD, Semenkovich CF, Henrissat B, Giannone RJ, Hettich RL, Ilkayeva O, Muehlbauer M, Newgard CB, Sawyer C, Head RD, Rodionov DA, Arzamasov AA, Leyn SA, Osterman AL, Hossain MI, Islam M, Choudhury N, Sarker SA, Huq S, Mahmud I, Mostafa I, Mahfuz M, Barratt MJ, Ahmed T, Gordon JI. 2019. Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. *Science* **365**. doi:10.1126/science.aau4732

Ghosh TS, Das M, Jeffery IB, O'Toole PW. 2020. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *Elife* **9**. doi:10.7554/eLife.50240

Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R, Knights D, Beckman KB. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat Biotechnol* **34**:942–949.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD, Project 1000 Genomes, Others. 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* **108**:11983–11988.

Gu D, He J, Duan X, Reynolds K, Wu X, Chen J, Huang G, Chen C-S, Whelton PK. 2006. Body weight and mortality among men and women in China. *JAMA* **295**:776–783.

Hehemann J-H, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. 2010. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**:908–912.

He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, Chen M-X, Chen Z-H, Ji G-Y, Zheng Z-D-X, Mujagond P, Chen X-J, Rong Z-H, Chen P, Lyu L-Y, Wang X, Wu C-B, Yu N, Xu Y-J, Yin J, Raes J, Knight R, Ma W-J, Zhou H-W. 2018. Regional variation limits

applications of healthy gut microbiome reference ranges and disease models. *Nat Med* **24**:1532–1535.

Hsu WC, Araneta MRG, Kanaya AM, Chiang JL, Fujimoto W. 2015. BMI cut points to identify at-risk Asian Americans for type 2 diabetes screening. *Diabetes Care* **38**:150–158.

Jih J, Mukherjea A, Vittinghoff E, Nguyen TT, Tsoh JY, Fukuoka Y, Bender MS, Tseng W, Kanaya AM. 2014. Using appropriate body mass index cut points for overweight and obesity among Asian Americans. *Prev Med* **65**:1–6.

Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, Kim AD, Shmagel AK, Syed AN, Personalized Microbiome Class Students, Walter J, Menon R, Koecher K, Knights D. 2019. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host Microbe* **25**:789–802.e5.

Johnson EL, Heaver SL, Walters WA, Ley RE. 2017. Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes. *J Mol Med* **95**:1–8.

Kahle D, Wickham H. 2013. ggmap: Spatial Visualization with ggplot2. *R J* **5**:144–161.

Kakar V, Voelz J, Wu J, Franco J. 2018. The Visible Host: Does race guide Airbnb rental rates in San Francisco? *J Hous Econ* **40**:25–40.

Kassambara A. 2018. ggpubr:"ggplot2" based publication ready plots. *R package version 0 1* **7**.

Kassambara A, Kassambara MA. 2019. Package "ggcorrplot." *R package version 0 1* **3**.

Kaul S, Rothney MP, Peters DM, Wacker WK, Davis CE, Shapiro MD, Ergun DL. 2012. Dual-Energy X-Ray Absorptiometry for Quantification of Visceral Fat. *Obesity*. doi:10.1038/oby.2011.393

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**:1463–1464.

Khine WWT, Zhang Y, Goie GJY, Wong MS, Liong M, Lee YY, Cao H, Lee Y-K. 2019. Gut microbiome of pre-adolescent children of two ethnicities residing in three distant cities. *Sci Rep* **9**:7831.

Kim KN, Yao Y, Ju SY. 2019. Short Chain Fatty Acids and Fecal Microbiota Abundance in Humans with Obesity: A Systematic Review and Meta-Analysis. *Nutrients* **11**. doi:10.3390/nu11102512

Krijthe JH. 2015. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R package version 0 13, URL https://github com/jkrijthe/Rtsne*.

Kuznetsova A, Brockhoff PB, Christensen RHB, Others. 2017. lmerTest package: tests in linear mixed effects models. *J Stat Softw* **82**:1–26.

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto J-M, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jørgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clément K, Doré J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker J-D, Raes J, Hansen T, MetaHIT consortium, Bork P, Wang J, Ehrlich SD, Pedersen O. 2013. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**:541–546.

Ley RE, Turnbaugh PJ, Klein S, Gordon JI. 2006. Human gut microbes associated with obesity. *Nature* **444**:1022–1023.

Liaw A, Wiener M, Others. 2002. Classification and regression by randomForest. *R news* **2**:18–22.

Liu R, Hong J, Xu X, Feng Q, Zhang D, Gu Y, Shi J, Zhao S, Liu W, Wang X, Xia H, Liu Z, Cui B, Liang P, Xi L, Jin J, Ying X, Wang X, Zhao X, Li W, Jia H, Lan Z, Li F, Wang R, Sun Y, Yang M, Shen Y, Jie Z, Li J, Chen X, Zhong H, Xie H, Zhang Y, Gu W, Deng X, Shen B, Xu X, Yang H, Xu G, Bi Y, Lai S, Wang J, Qi L, Madsen L, Wang J, Ning G, Kristiansen K, Wang W. 2017. Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat Med* **23**:859–868.

Lu Y, Fan C, Li P, Lu Y, Chang X, Qi K. 2016. Short Chain Fatty Acids Prevent High-fat-diet-induced Obesity in Mice by Regulating G Protein-coupled Receptors and Gut Microbiota. *Sci Rep* **6**:37589.

Lynch CJ, Adams SH. 2014. Branched-chain amino acids in metabolic signalling and insulin resistance. *Nat Rev Endocrinol* **10**:723–736.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K, Others. 2012. Cluster: cluster analysis basics and extensions. *R package version* **1**:56.

Martín-Fernández J-A, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat Modelling* **15**:134–158.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Ovreås L, Reysenbach A-L, Smith VH, Staley JT. 2006. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**:102–112.

Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. 1985. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* **28**:412–419.

McClung HL, Ptomey LT, Shook RP, Aggarwal A, Gorczyca AM, Sazonov ES, Becofsky K, Weiss R, Das SK. 2018. Dietary Intake and Physical Activity Assessment: Current Tools, Techniques, and Technologies for Use in Adult Populations. *Am J Prev Med* **55**:e93–e104.

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**:e61217.

Millen AE, Midthune D, Thompson FE, Kipnis V, Subar AF. 2006. The National Cancer Institute diet history questionnaire: validation of pyramid food servings. *Am J Epidemiol* **163**:279–288.

Müller M, Hernández MAG, Goossens GH, Reijnders D, Holst JJ, Jocken JWE, van Eijk H, Canfora EE, Blaak EE. 2019. Circulating but not faecal short-chain fatty acids are related to insulin sensitivity, lipolysis and GLP-1 concentrations in humans. *Sci Rep* **9**:12515.

Nayak RR, Alexander M, Deshpande I, Stapleton-Gray K, Rimal B, Patterson AD, Ubeda C, Scher JU, Turnbaugh PJ. 2021. Methotrexate impacts conserved pathways in diverse human gut bacteria leading to decreased host immune activation. *Cell Host Microbe*. doi:10.1016/j.chom.2020.12.008

Nayfach S, Pollard KS. 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* **16**:51.

Neeland IJ, Grundy SM, Li X, Adams-Huet B, Vega GL. 2016. Comparison of visceral fat mass measurement by dual-X-ray absorptiometry and magnetic resonance imaging in a multiethnic cohort: the Dallas Heart Study. *Nutr Diabetes* **6**:e221.

Oguri Y, Shinoda K, Kim H, Alba DL, Bolus WR, Wang Q, Brown Z, Pradhan RN, Tajima K, Yoneshiro T, Ikeda K, Chen Y, Cheang RT, Tsujino K, Kim CR, Greiner VJ, Datta R, Yang CD, Atabai K, McManus MT, Koliwad SK, Spiegelman BM, Kajimura S. 2020. CD81 Controls Beige Fat Progenitor Cell Growth and Energy Balance via FAK Signaling. *Cell* **182**:563–577.e20.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H, Others. 2013. Community ecology package. *R package version* 2–0.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**:289–290.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*.

Park Y, Dodd KW, Kipnis V, Thompson FE, Potischman N, Schoeller DA, Baer DJ, Midthune

D, Troiano RP, Bowles H, Subar AF. 2018. Comparison of self-reported dietary intakes from the Automated Self-Administered 24-h recall, 4-d food records, and food-frequency questionnaires against recovery biomarkers. *The American Journal of Clinical Nutrition*. doi:10.1093/ajcn/nqx002

Perry RJ, Peng L, Barry NA, Cline GW, Zhang D, Cardone RL, Petersen KF, Kibbey RG, Goodman AL, Shulman GI. 2016. Acetate mediates a microbiome–brain–β-cell axis to promote metabolic syndrome. *Nature* **534**:213–217.

Plovier H, Everard A, Druart C, Depommier C, Van Hul M, Geurts L, Chilloux J, Ottman N, Duparc T, Lichtenstein L, Myridakis A, Delzenne NM, Klievink J, Bhattacharjee A, van der Ark KCH, Aalvink S, Martinez LO, Dumas M-E, Maiter D, Loumaye A, Hermans MP, Thissen J-P, Belzer C, de Vos WM, Cani PD. 2017. A purified membrane protein from Akkermansia muciniphila or the pasteurized bacterium improves metabolism in obese and diabetic mice. *Nat Med* **23**:107–113.

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto J-M, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**:55–60.

Rich B. 2020. table1: Tables of Descriptive Statistics in HTML. *R package version 12*.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*.

Sarafian MH, Lewis MR, Pechlivanis A, Ralphs S, McPhail MJW, Patel VC, Dumas M-E, Holmes E, Nicholson JK. 2015. Bile acid profiling and quantification in biofluids using ultra-performance liquid chromatography tandem mass spectrometry. *Anal Chem* **87**:9662–9670.

Silverman JD, Washburne AD, Mukherjee S, David LA. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* **6**. doi:10.7554/eLife.21887

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics*.

Sordillo JE, Zhou Y, McGeachie MJ, Ziniti J, Lange N, Laranjo N, Savage JR, Carey V, O'Connor G, Sandel M, Strunk R, Bacharier L, Zeiger R, Weiss ST, Weinstock G, Gold DR, Litonjua AA. 2017. Factors influencing the infant gut microbiome at age 3-6 months: Findings from the ethnically diverse Vitamin D Antenatal Asthma Reduction Trial (VDAART). *J Allergy Clin Immunol* **139**:482–491.e14.

Timon CM, van den Barg R, Blain RJ, Kehoe L, Evans K, Walton J, Flynn A, Gibney ER. 2016. A review of the design and validation of web- and computer-based 24-h dietary recall tools. *Nutrition Research Reviews*. doi:10.1017/s0954422416000172

Tirosh A, Calay ES, Tuncman G, Claiborn KC, Inouye KE, Eguchi K, Alcala M, Rathaus M, Hollander KS, Ron I, Livne R, Heianza Y, Qi L, Shai I, Garg R, Hotamisligil GS. 2019. The short-chain fatty acid propionate increases glucagon and FABP4 production, impairing insulin action in mice and humans. *Sci Transl Med* **11**. doi:10.1126/scitranslmed.aav0120

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**:902–903.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009a. A core gut microbiome in obese and lean twins. *Nature* **457**:480–484.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**:1027–1031.

Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. 2009b. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med* **1**:6ra14.

Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**:158.

Vangay P, Johnson AJ, Ward TL, Al-Ghalith GA, Shields-Cutler RR, Hillmann BM, Lucas SK, Beura LK, Thompson EA, Till LM, Batres R, Paw B, Pergament SL, Saenyakul P, Xiong M, Kim AD, Kim G, Masopust D, Martens EC, Angkurawaranon C, McGready R, Kashyap PC, Culhane-Pera KA, Knights D. 2018. US Immigration Westernizes the Human Gut Microbiome. *Cell* **175**:962–972.e10.

Vieira-Silva S, Falony G, Belda E, Nielsen T, Aron-Wisnewsky J, Chakaroun R, Forslund SK, Assmann K, Valles-Colomer M, Nguyen TTD, Proost S, Prifti E, Tremaroli V, Pons N, Le Chatelier E, Andreelli F, Bastard J-P, Coelho LP, Galleron N, Hansen TH, Hulot J-S, Lewinter C, Pedersen HK, Quinquis B, Rouault C, Roume H, Salem J-E, Søndertoft NB, Touch S, MetaCardis Consortium, Dumas M-E, Ehrlich SD, Galan P, Gøtze JP, Hansen T, Holst JJ, Køber L, Letunic I, Nielsen J, Oppert J-M, Stumvoll M, Vestergaard H, Zucker J-D, Bork P, Pedersen O, Bäckhed F, Clément K, Raes J. 2020. Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* **581**:310–315.

Walker K. 2018. Tigris: Load census TIGER/Line Shapefiles. R package version 0.7.

Wallace JR. 2012. Imap: Interactive mapping. R package version 1.32.

Walter J, Armet AM, Finlay BB, Shanahan F. 2020. Establishing or Exaggerating Causality for the Gut Microbiome: Lessons from Human Microbiota-Associated Rodents. *Cell* **180**:221–232.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**:5261–5267.

Wen J, Rönn T, Olsson A, Yang Z, Lu B, Du Y, Groop L, Ling C, Hu R. 2010. Investigation of type 2 diabetes risk alleles support CDKN2A/B, CDKAL1, and TCF7L2 as susceptibility genes in a Han Chinese cohort. *PLoS One* **5**:e9153.

WHO Expert Consultation. 2004. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet* **363**:157–163.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis.

Wickham H, Bryan J. 2017. readxl: Read Excel Files. R package version 1.0. 0. *URL https://CRAN R-project org/package= readxl*.

Wu H, Esteve E, Tremaroli V, Khan MT, Caesar R, Mannerås-Holm L, Ståhlman M, Olsson LM, Serino M, Planas-Fèlix M, Xifra G, Mercader JM, Torrents D, Burcelin R, Ricart W, Perkins R, Fernàndez-Real JM, Bäckhed F. 2017. Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat Med* **23**:850–858.

Wu H, Tremaroli V, Schmidt C, Lundqvist A, Olsson LM, Krämer M, Gummesson A, Perkins R, Bergström G, Bäckhed F. 2020. The Gut Microbiota in Prediabetes and Diabetes: A Population-Based Cross-Sectional Study. *Cell Metab* **32**:379–390.

Xiang K, Wang Y, Zheng T, Jia W, Li J, Chen L, Shen K, Wu S, Lin X, Zhang G, Wang C, Wang S, Lu H, Fang Q, Shi Y, Zhang R, Xu J, Weng Q. 2004. Genome-wide search for type 2 diabetes/impaired glucose homeostasis susceptibility genes in the Chinese: significant linkage to chromosome 6q21-q23 and chromosome 1q21-q24. *Diabetes* **53**:228–234.

Xu J, Lawley B, Wong G, Otal A, Chen L, Ying TJ, Lin X, Pang WW, Yap F, Chong Y-S, Gluckman PD, Lee YS, Chong MF-F, Tannock GW, Karnani N. 2020. Ethnic diversity in infant gut microbiota is apparent before the introduction of complementary diets. *Gut Microbes* **11**:1362–1373.

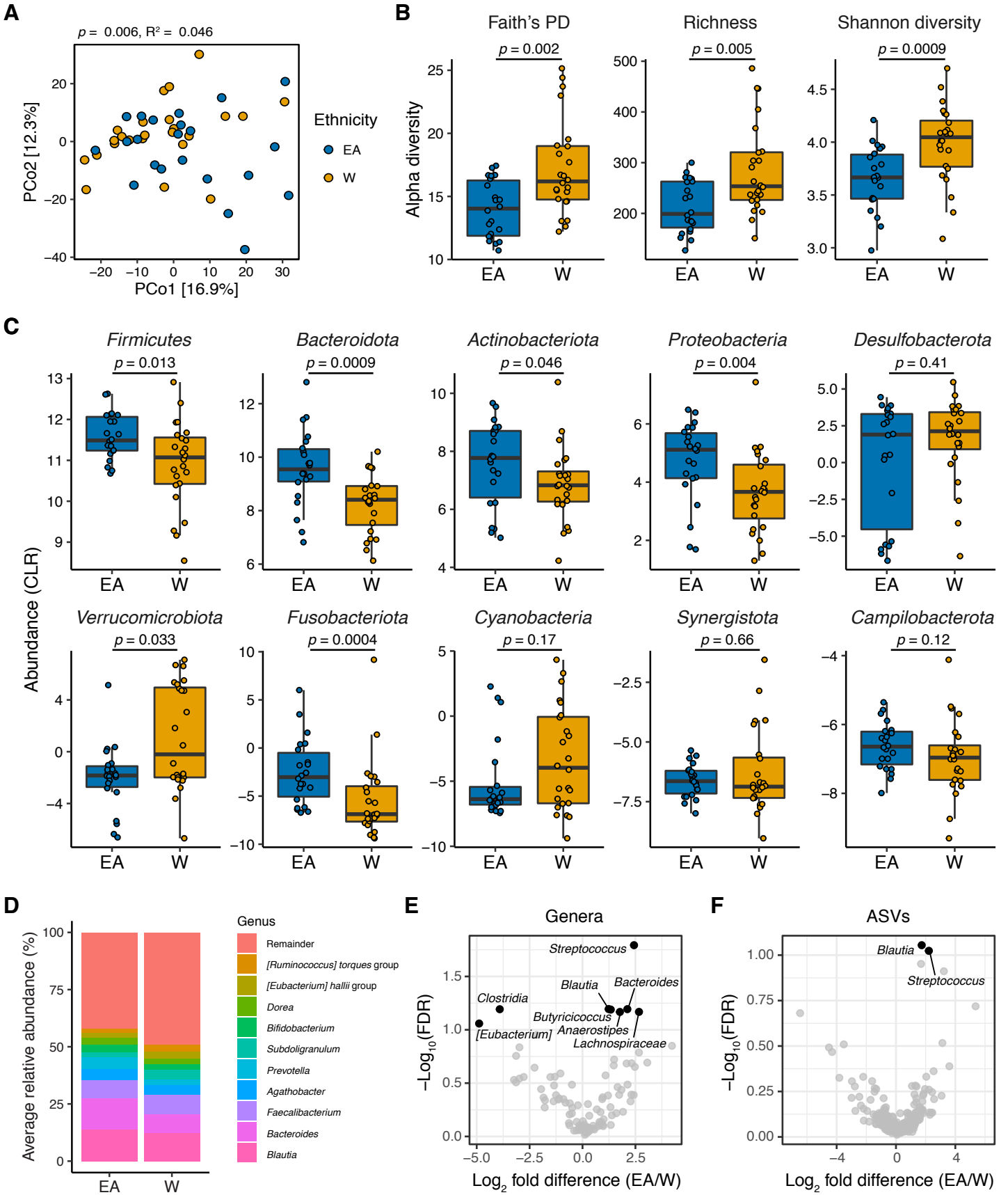
Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. *Nature* **486**:222–227.

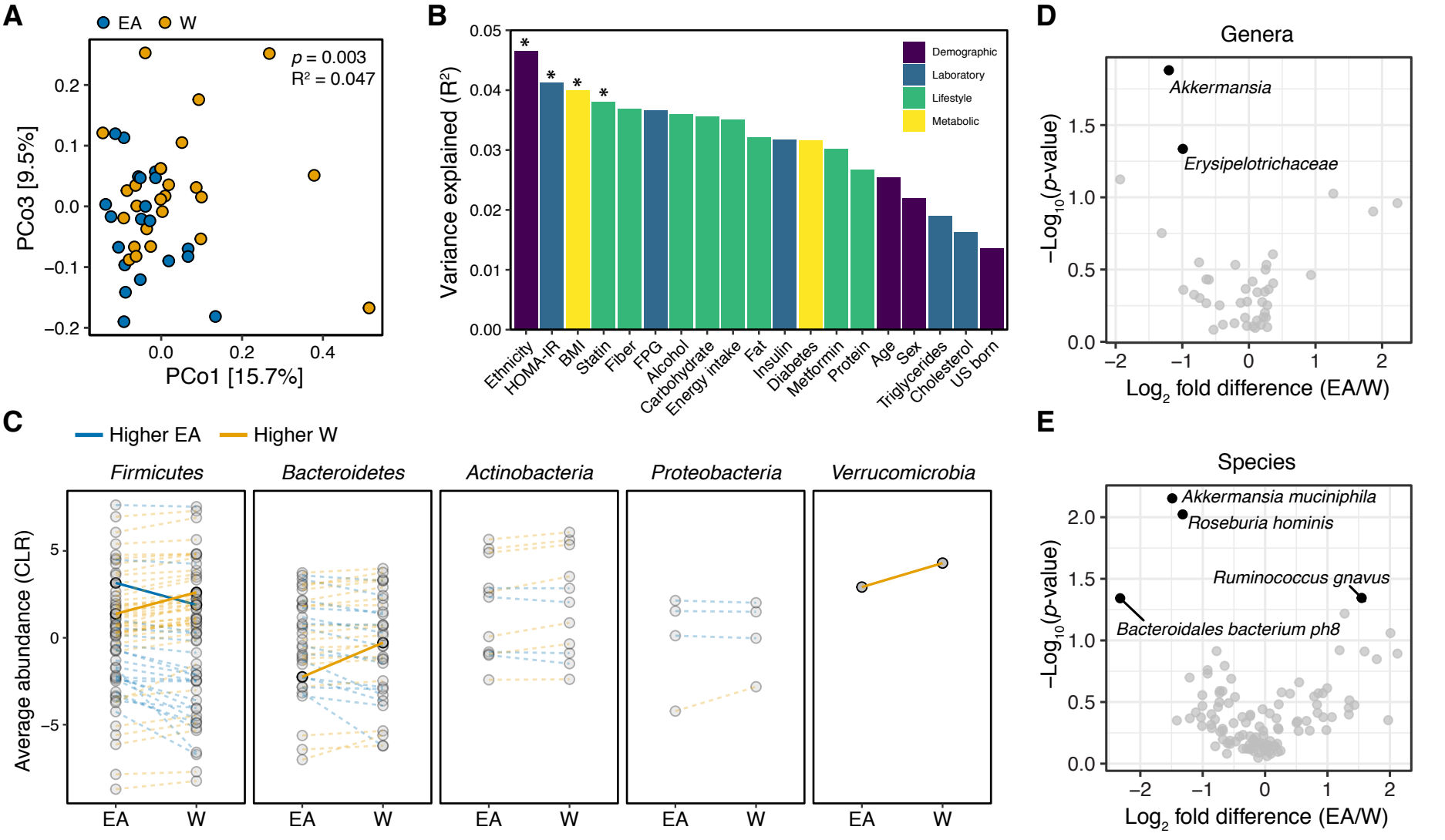
Yutani H. 2018. gghighlight: Highlight Lines and Points in “ggplot2.” *Manual available online at http://CRAN R-project org/package= gghighlight*.

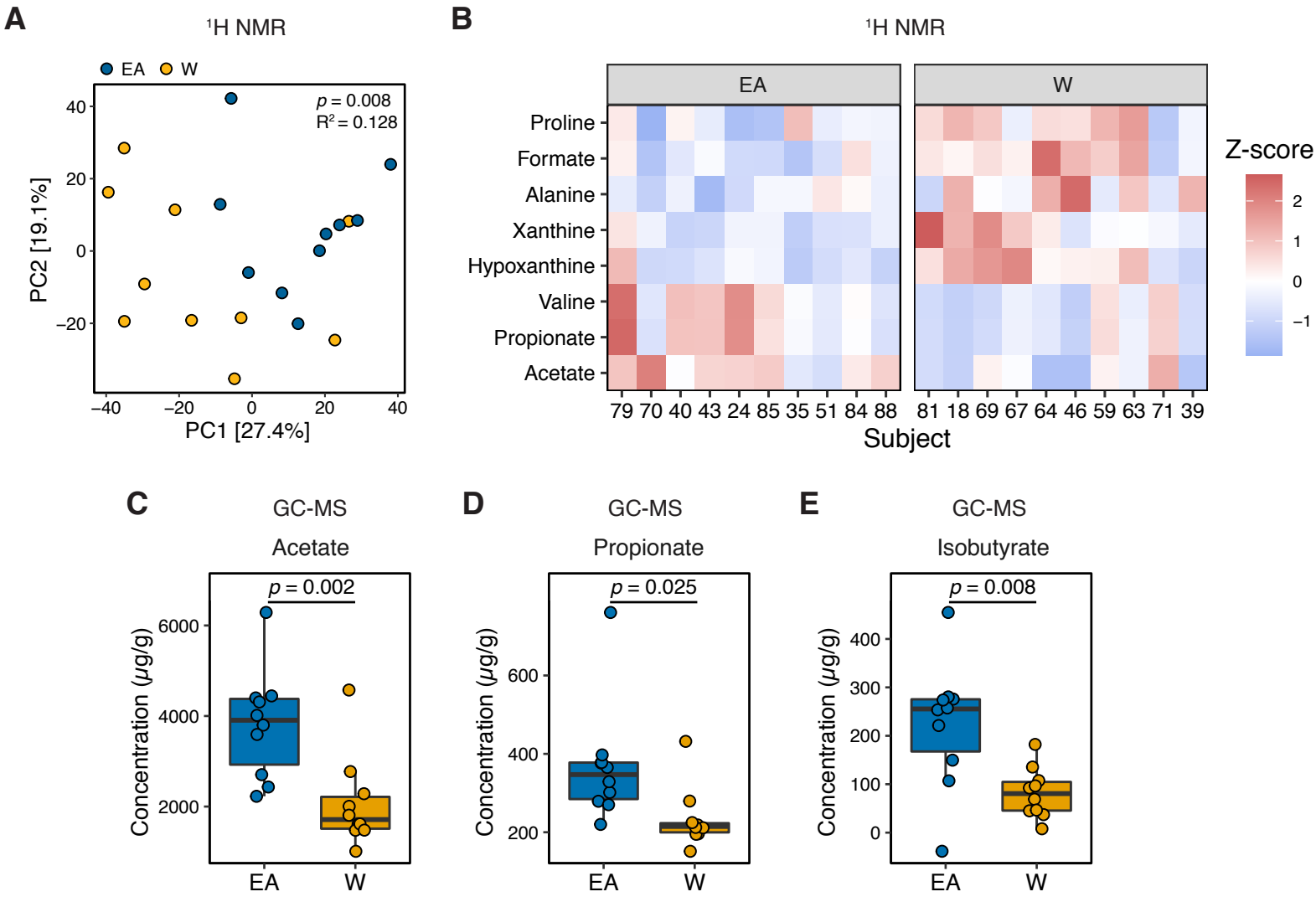
Zheng W, McLerran DF, Rolland B, Zhang X, Inoue M, Matsuo K, He J, Gupta PC, Ramadas K, Tsugane S, Irie F, Tamakoshi A, Gao Y-T, Wang R, Shu X-O, Tsuji I, Kuriyama S, Tanaka H, Satoh H, Chen C-J, Yuan J-M, Yoo K-Y, Ahsan H, Pan W-H, Gu D, Pednekar MS, Sauvaget C, Sasazuki S, Sairenchi T, Yang G, Xiang Y-B, Nagai M, Suzuki T, Nishino Y, You S-L, Koh W-P, Park SK, Chen Y, Shen C-Y, Thornquist M, Feng Z, Kang D, Boffetta P, Potter JD. 2011. Association between body-mass index and risk of death in more than 1 million Asians. *N Engl J Med* **364**:719–729.

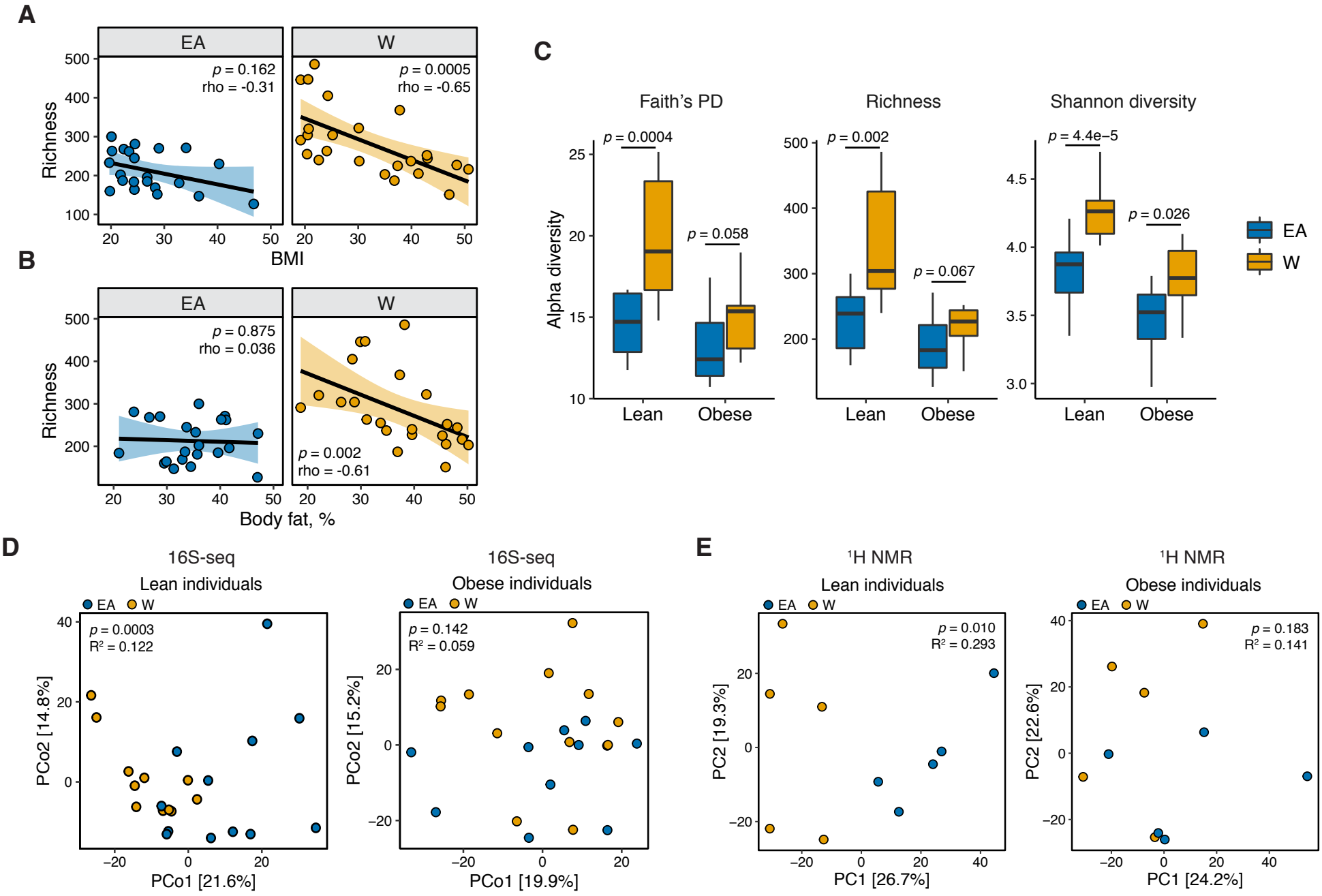
Zheng X, Qiu Y, Zhong W, Baxter S, Su M, Li Q, Xie G, Ore BM, Qiao S, Spencer MD, Zeisel SH, Zhou Z, Zhao A, Jia W. 2013. A targeted metabolomic protocol for short-chain fatty acids and branched-chain amino acids. *Metabolomics* **9**:818–827.

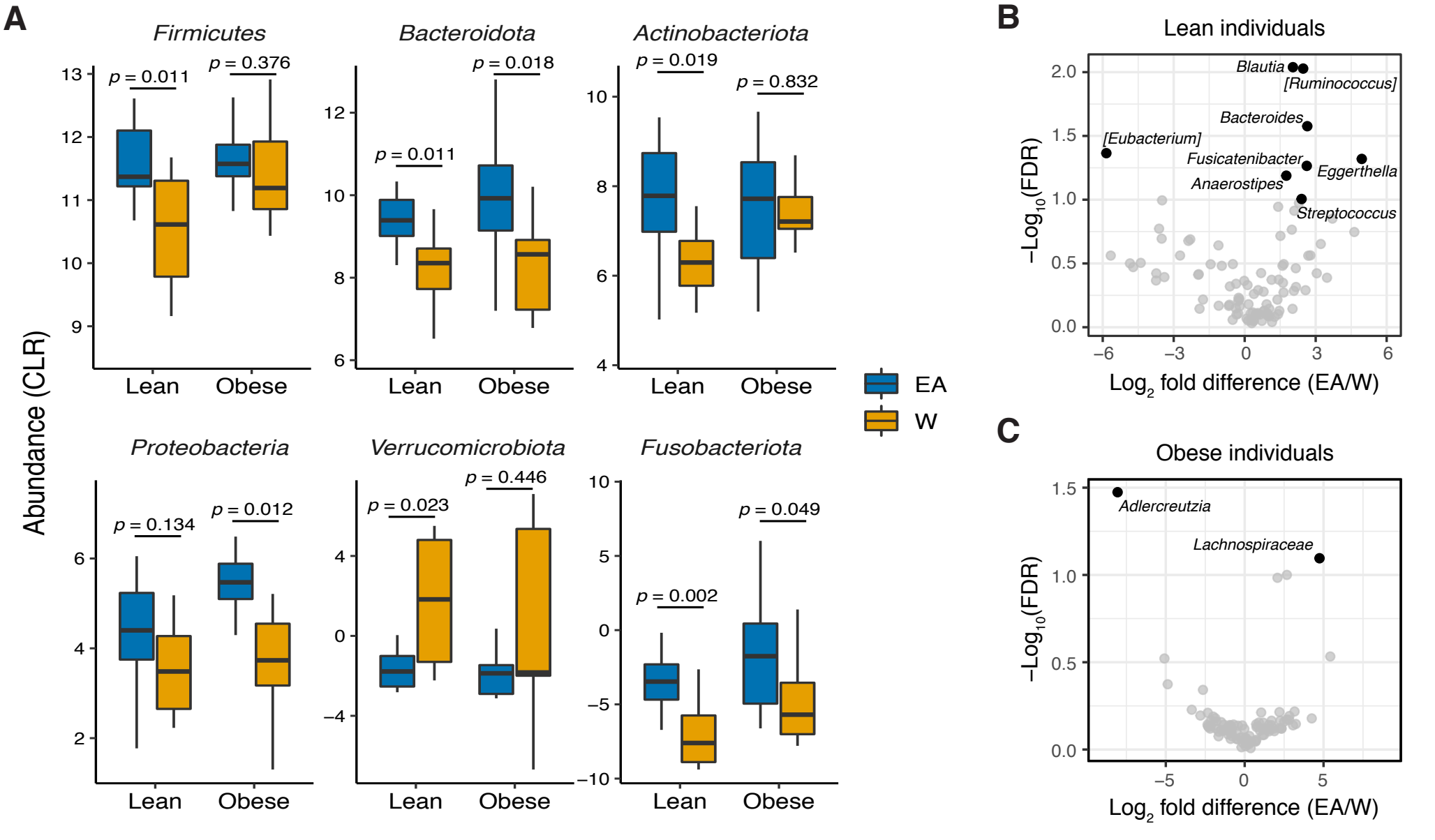
Zouiouich S, Loftfield E, Huybrechts I, Viallon V, Louca P, Vogtmann E, Wells PM, Steves CJ, Herzig K-H, Menni C, Jarvelin M-R, Sinha R, Gunter MJ. 2021. Markers of metabolic health and gut microbiome diversity: findings from two population-based cohort studies. *Diabetologia* **64**:1749–1759.

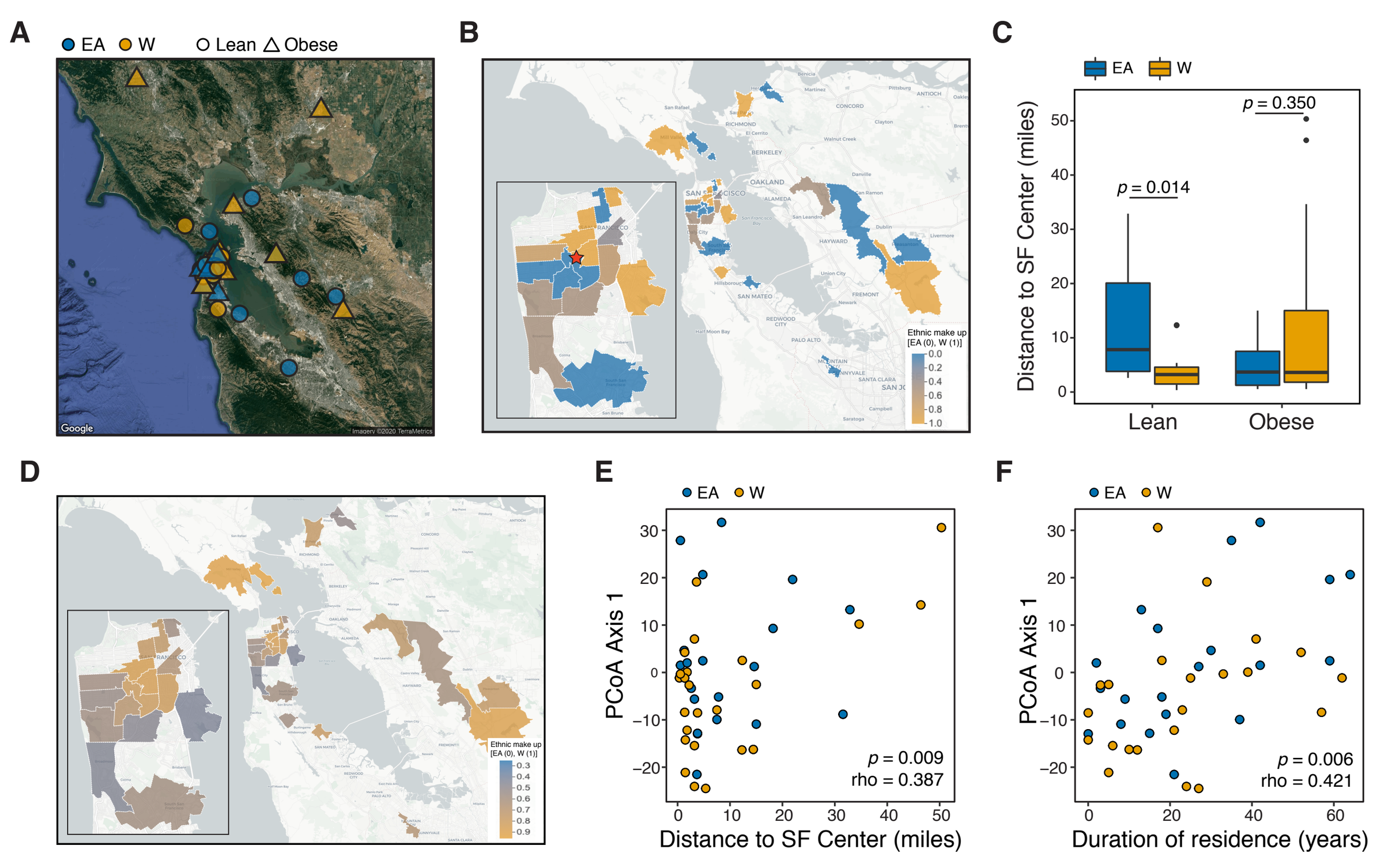

A
p = 0.006, R² = 0.046
PCo2 [12.3%]
PCo1 [16.9%]
Ethnicity
EA
W
B
Faith's PD
p = 0.002
Richness
p = 0.005
Shannon diversity
p = 0.0009
Alpha diversity
EA
W
C
Firmicutes
p = 0.013
Bacteroidota
p = 0.0009
Actinobacteriota
p = 0.046
Proteobacteria
p = 0.004
Desulfobacterota
p = 0.41
Verrucomicrobiota
p = 0.033
Fusobacteriota
p = 0.0004
Cyanobacteria
p = 0.17
Synergistota
p = 0.66
Campilobacterota
p = 0.12
Abundance (CLR)
D
Average relative abundance (%)
Genus
Remainder
[Ruminococcus] torques group
[Eubacterium] hallii group
Dorea
Bifidobacterium
Subdoligranulum
Prevotella
Agathobacter
Faecalibacterium
Bacteroides
Blautia
E
Genera
Streptococcus
Blautia
Bacteroides
Clostridia
Butyricicoccus
[Eubacterium]
Anaerostipes
Lachnospiraceae
−Log10(FDR)
Log2 fold difference (EA/W)
F
ASVs
Blautia
Streptococcus
−Log10(FDR)
Log2 fold difference (EA/W)

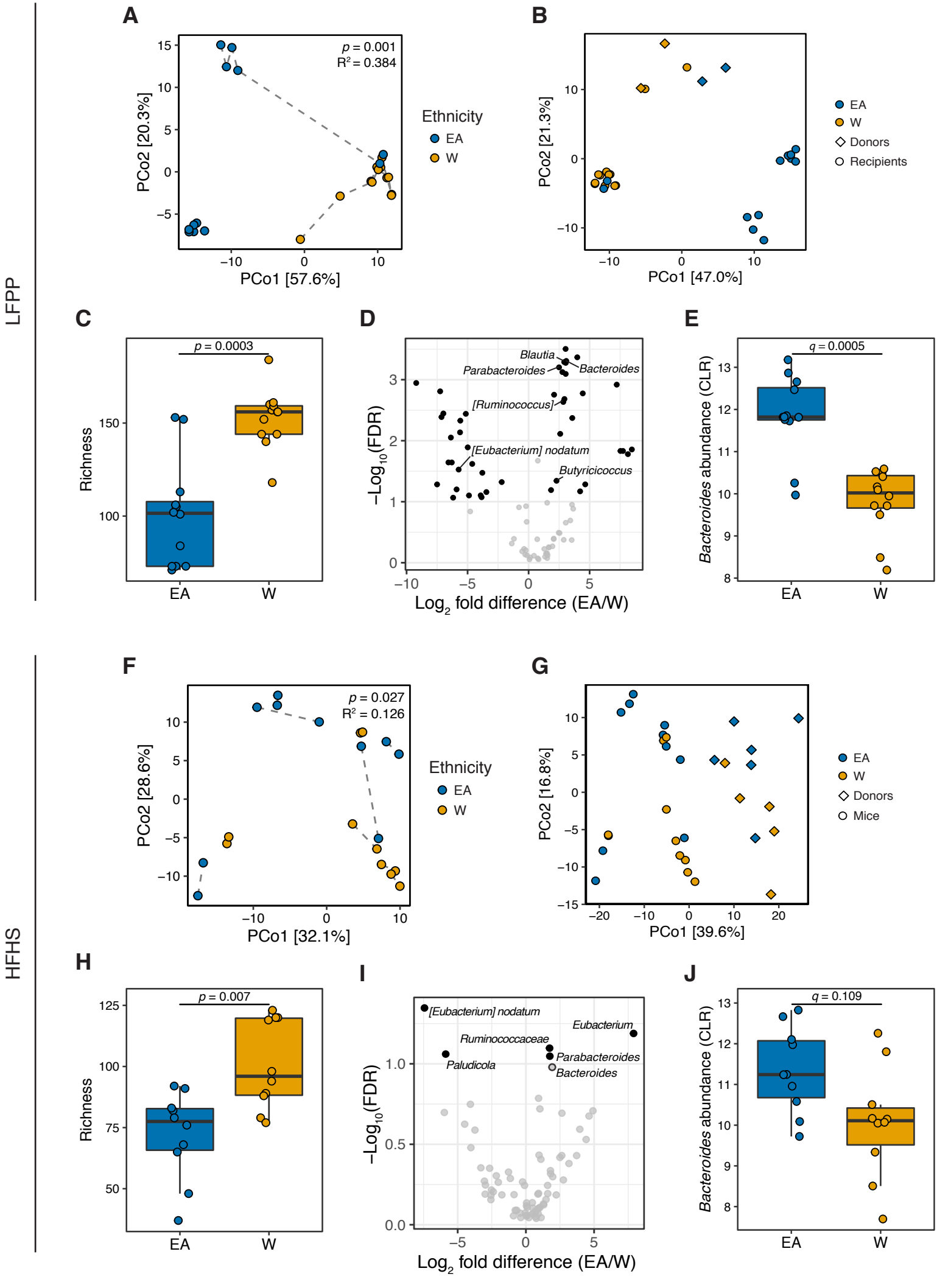

A
EA
W
p = 0.003
R² = 0.047
PCo3 [9.5%]
PCo1 [15.7%]
B
Variance explained (R²)
Demographic
Laboratory
Lifestyle
Metabolic
Ethnicity
HOMA-IR
BMI
Statin
Fiber
FPG
Alcohol
Carbohydrate
Energy intake
Fat
Insulin
Diabetes
Metformin
Protein
Age
Sex
Triglycerides
Cholesterol
US born
C
Higher EA
Higher W
Firmicutes
Bacteroidetes
Actinobacteria
Proteobacteria
Verrucomicrobia
Average abundance (CLR)
D
Genera
Akkermansia
Erysipelotrichaceae
−Log10(p-value)
Log2 fold difference (EA/W)
E
Species
Akkermansia muciniphila
Roseburia hominis
Ruminococcus gnavus
Bacteroidales bacterium ph8

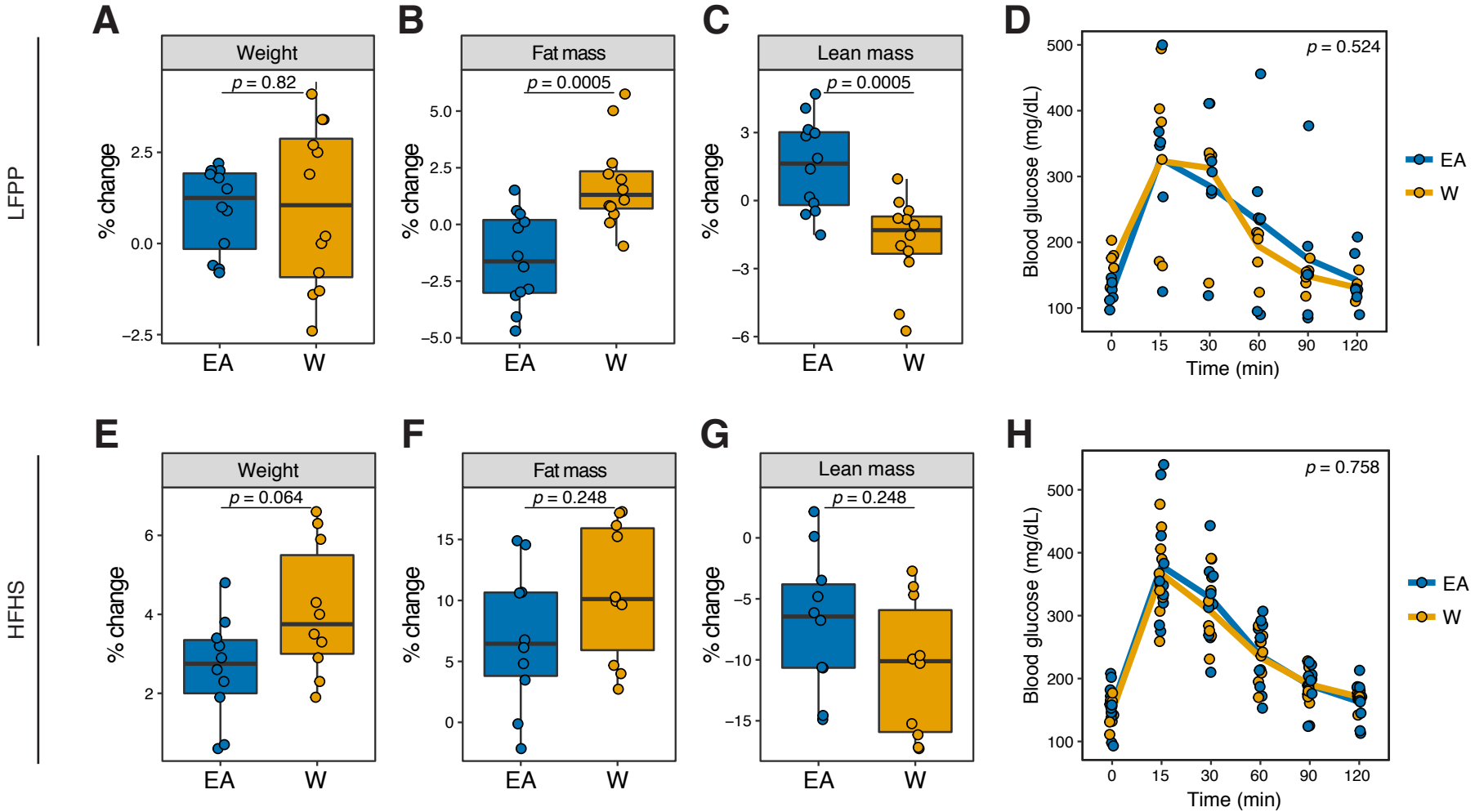

A
¹H NMR
EA
W
p = 0.008
R² = 0.128
PC2 [19.1%]
PC1 [27.4%]
B
¹H NMR
EA
W
Proline
Formate
Alanine
Xanthine
Hypoxanthine
Valine
Propionate
Acetate
79 70 40 43 24 85 35 51 84 88
81 18 69 67 64 46 59 63 71 39
Subject
Z-score
C
GC-MS
Acetate
p = 0.002
Concentration (μg/g)
D
GC-MS
Propionate
p = 0.025
Concentration (μg/g)
E
GC-MS
Isobutyrate
p = 0.008
Concentration (μg/g)


A
EA
W
p = 0.162
rho = -0.31
p = 0.0005
rho = -0.65
Richness
BMI
B
p = 0.875
rho = 0.036
p = 0.002
rho = -0.61
Body fat, %
C
Faith's PD
Richness
Shannon diversity
p = 0.0004
p = 0.058
p = 0.002
p = 0.067
p = 4.4e−5
p = 0.026
Alpha diversity
Lean
Obese
D
16S-seq
Lean individuals
Obese individuals
p = 0.0003
R² = 0.122
p = 0.142
R² = 0.059
PCo1 [21.6%]
PCo2 [14.8%]
PCo1 [19.9%]
PCo2 [15.2%]
E
¹H NMR
Lean individuals
Obese individuals
p = 0.010
R² = 0.293
p = 0.183
R² = 0.141
PC1 [26.7%]
PC2 [19.3%]
PC1 [24.2%]
PC2 [22.6%]

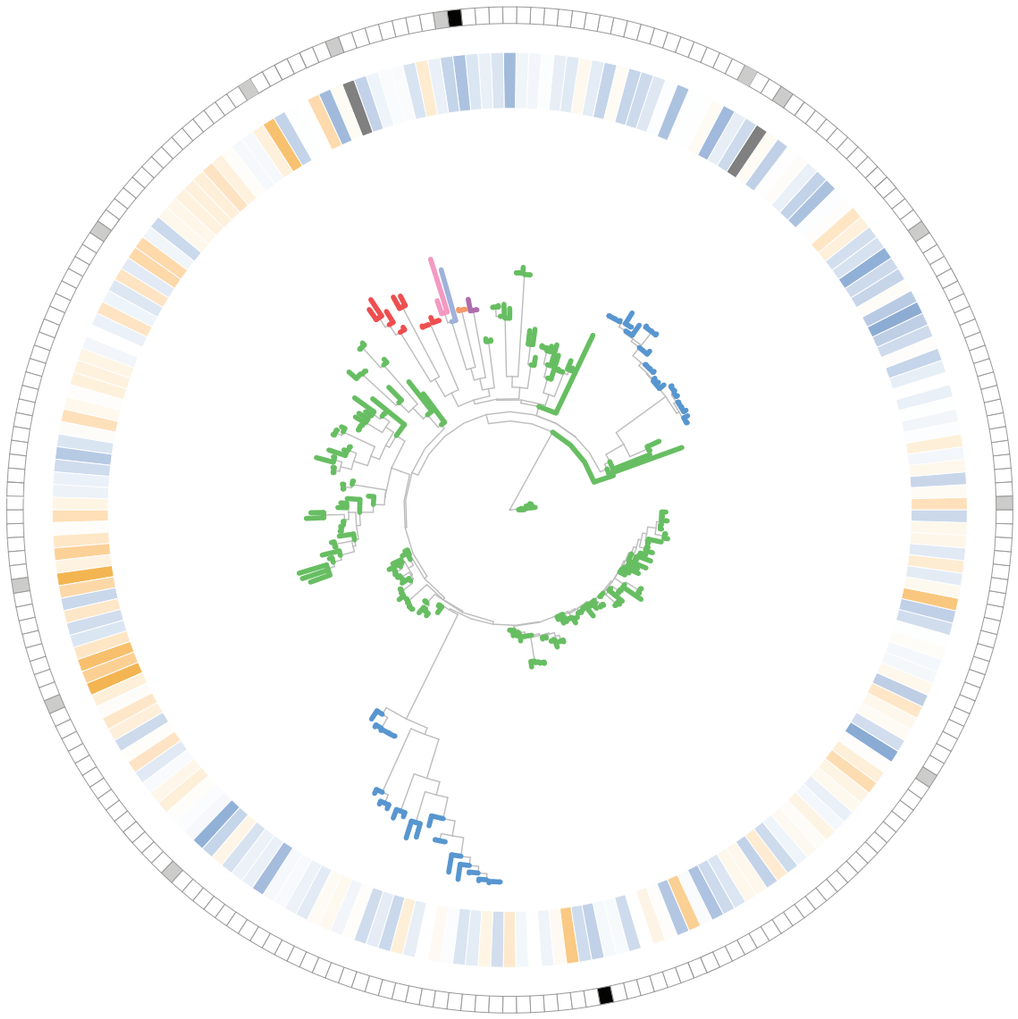

A
Firmicutes
Bacteroidota
Actinobacteriota
Proteobacteria
Verrucomicrobiota
Fusobacteriota
p = 0.011
p = 0.376
p = 0.011
p = 0.018
p = 0.019
p = 0.832
p = 0.134
p = 0.012
p = 0.023
p = 0.446
p = 0.002
p = 0.049
Lean
Obese
Abundance (CLR)
EA
W
B
Lean individuals
Blautia
[Ruminococcus]
Bacteroides
[Eubacterium]
Fusicatenibacter
Eggerthella
Anaerostipes
Streptococcus
−Log10(FDR)
Log2 fold difference (EA/W)
C
Obese individuals
Adlercreutzia
Lachnospiraceae
−Log10(FDR)
Log2 fold difference (EA/W)


A
EA
W
Lean
Obese
Google
Imagery ©2020 TerraMetrics
B
Ethnic make up
[EA (0), W (1)]
0.0
0.2
0.4
0.6
0.8
1.0
C
EA
W
p = 0.350
p = 0.014
Distance to SF Center (miles)
0
10
20
30
40
50
Lean
Obese
D
Ethnic make up
[EA (0), W (1)]
0.3
0.4
0.5
0.6
0.7
0.8
0.9
E
EA
W
PCoA Axis 1
p = 0.009
rho = 0.387
Distance to SF Center (miles)
F
EA
W
PCoA Axis 1
p = 0.006
rho = 0.421
Duration of residence (years)


LFPP
A
p = 0.001
R² = 0.384
PCo2 [20.3%]
PCo1 [57.6%]
Ethnicity
EA
W
B
PCo2 [21.3%]
PCo1 [47.0%]
EA
W
Donors
Recipients
C
p = 0.0003
Richness
EA
W
D
Blautia
Parabacteroides
Bacteroides
[Ruminococcus]
[Eubacterium] nodatum
Butyricicoccus
−Log10(FDR)
Log2 fold difference (EA/W)
E
q = 0.0005
Bacteroides abundance (CLR)
EA
W
HFHS
F
p = 0.027
R² = 0.126
PCo2 [28.6%]
PCo1 [32.1%]
Ethnicity
EA
W
G
PCo2 [16.8%]
PCo1 [39.6%]
EA
W
Donors
Mice
H
p = 0.007
Richness
EA
W
I
[Eubacterium] nodatum
Eubacterium
Ruminococcaceae
Parabacteroides
Paludicola
Bacteroides
−Log10(FDR)
Log2 fold difference (EA/W)
J
q = 0.109
Bacteroides abundance (CLR)
EA
W


LFPP
HFHS
A
Weight
p = 0.82
% change
EA
W
B
Fat mass
p = 0.0005
C
Lean mass
p = 0.0005
D
p = 0.524
Blood glucose (mg/dL)
Time (min)
E
Weight
p = 0.064
F
Fat mass
p = 0.248
G
Lean mass
p = 0.248
H
p = 0.758

A

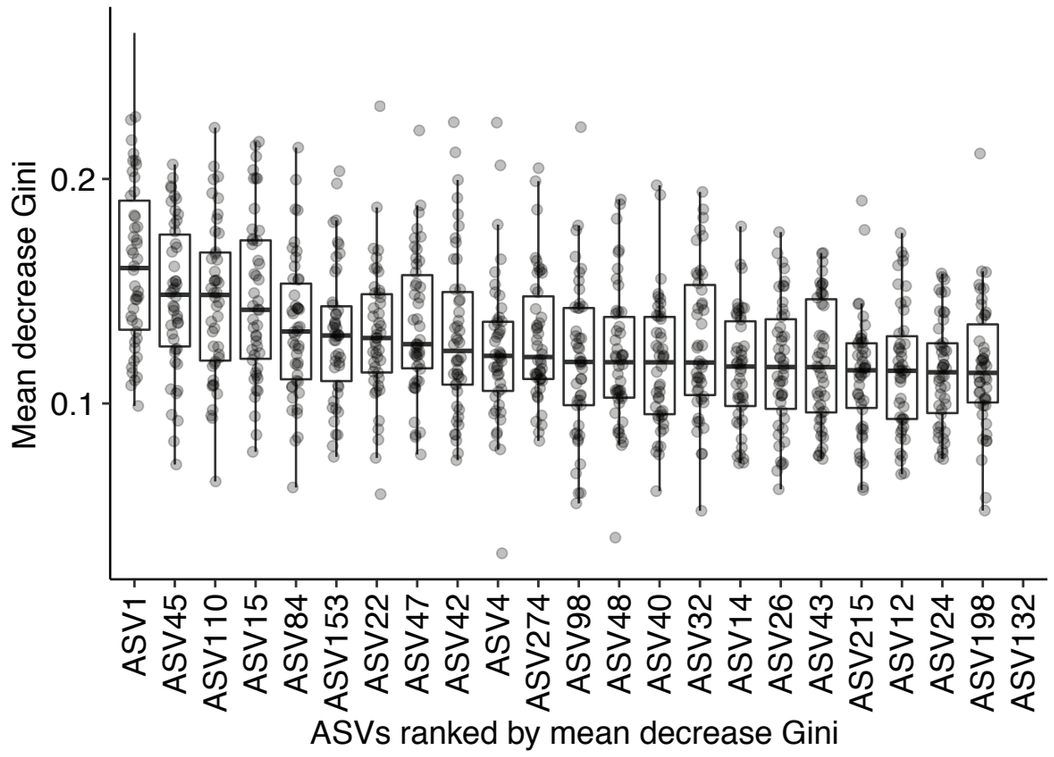


Phylum
- *Actinobacteriota*
- *Bacteroidota*
- *Desulfobacterota*
- *Firmicutes*
- *Fusobacteriota*
- *Proteobacteria*
- *Verrucomicrobiota*

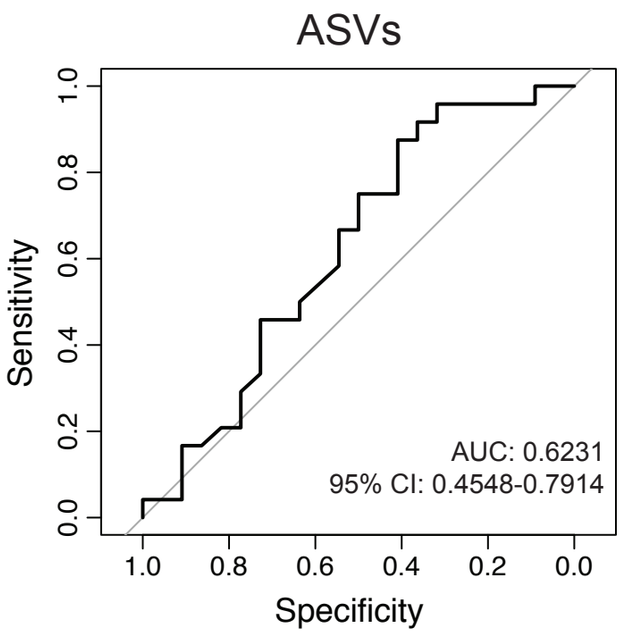
Log_2 fold difference

Significance
- $p > 0.05$
- $p < 0.05$
- FDR < 0.05

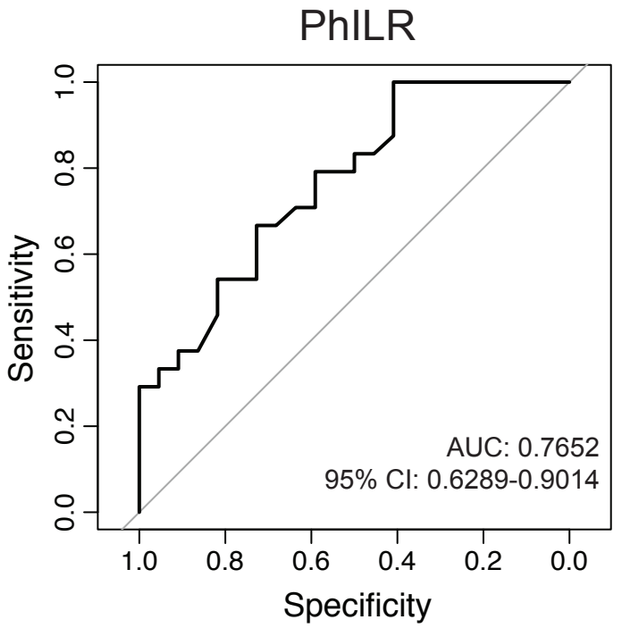
B



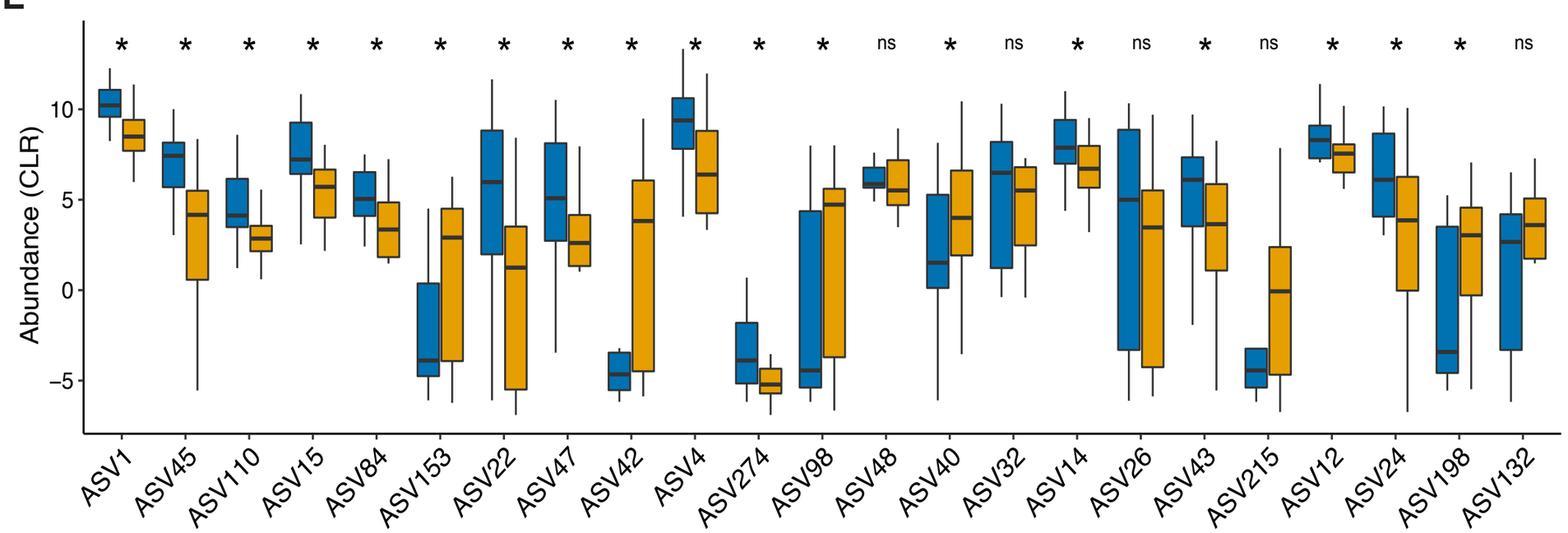
C



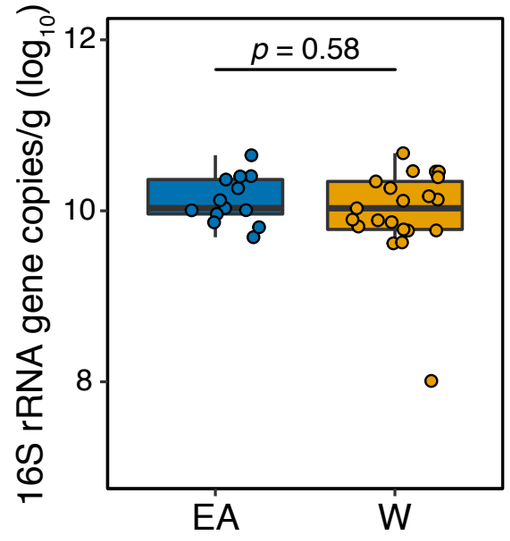
D

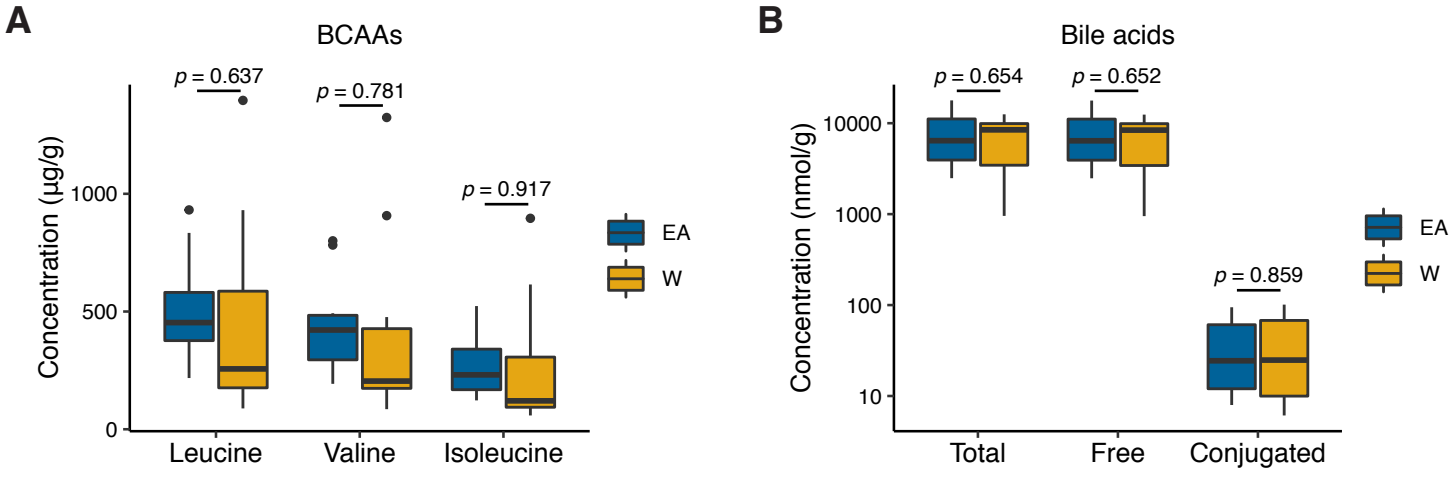


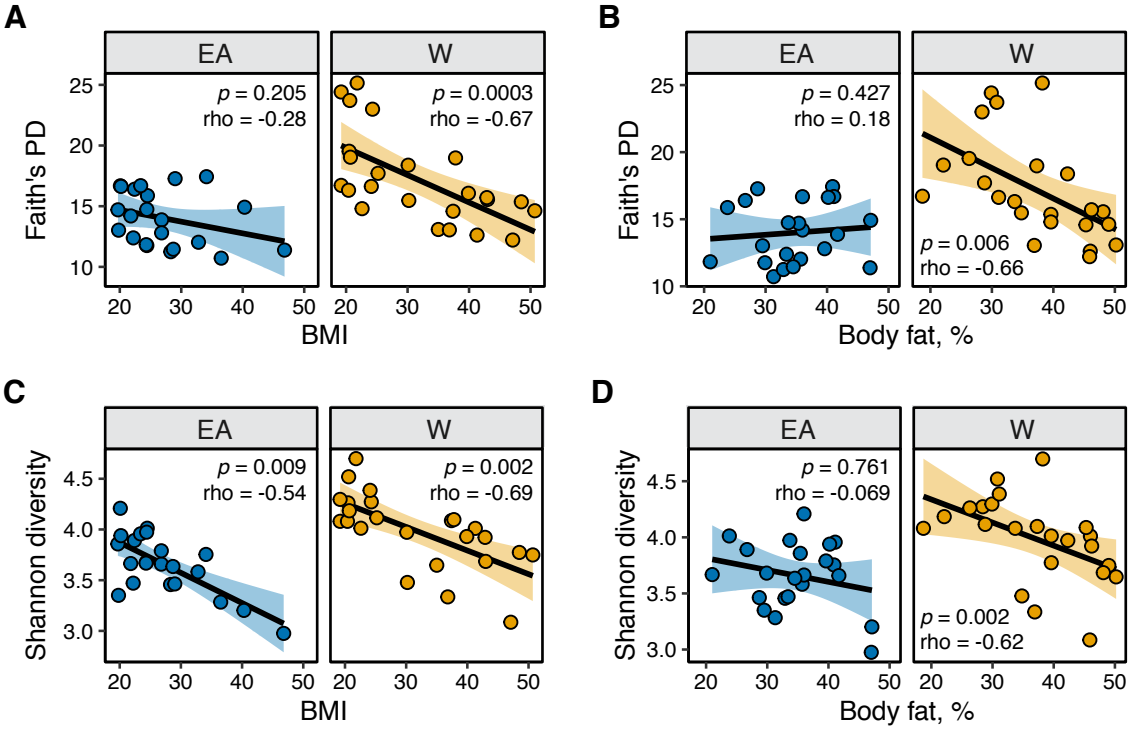
E

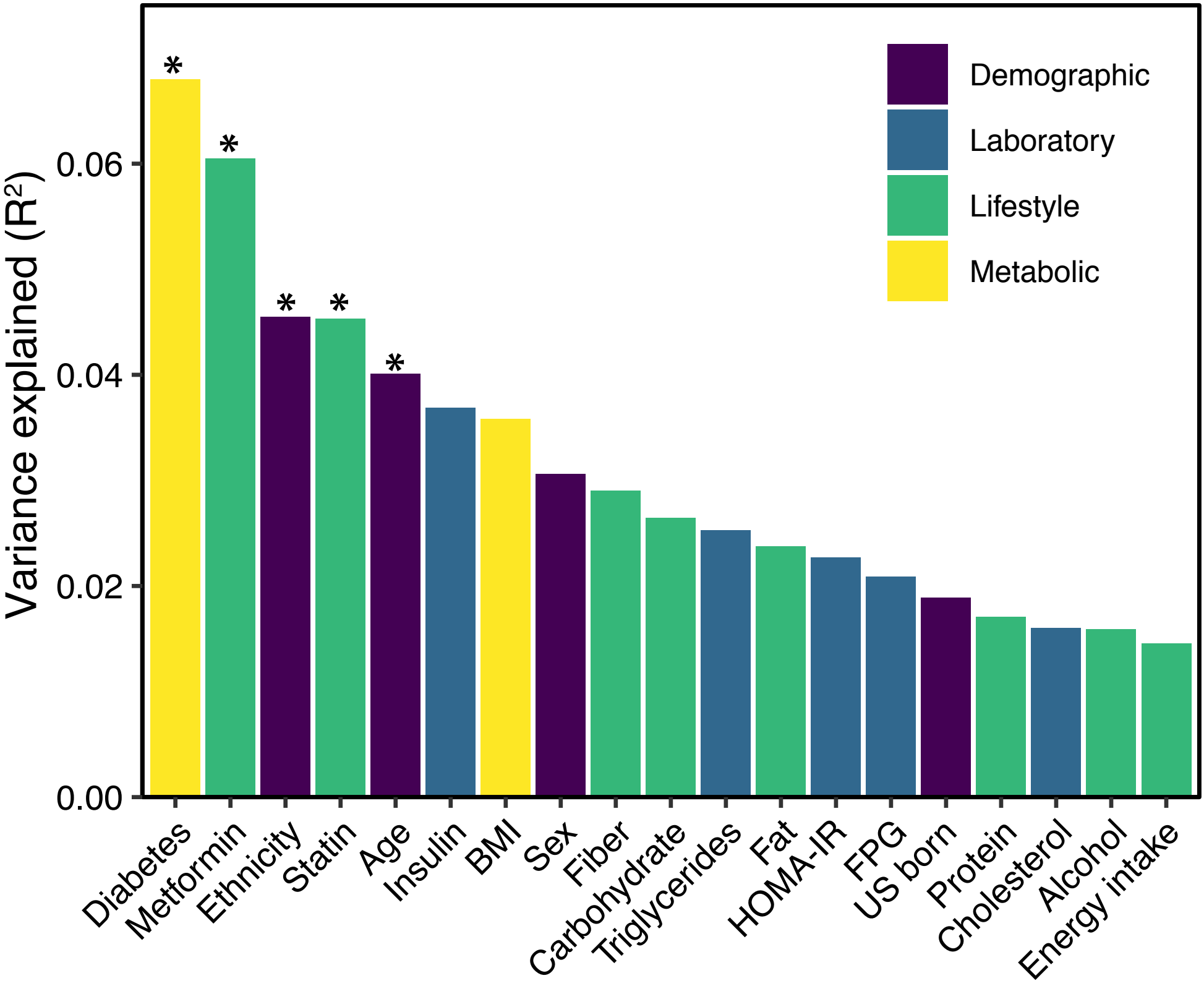


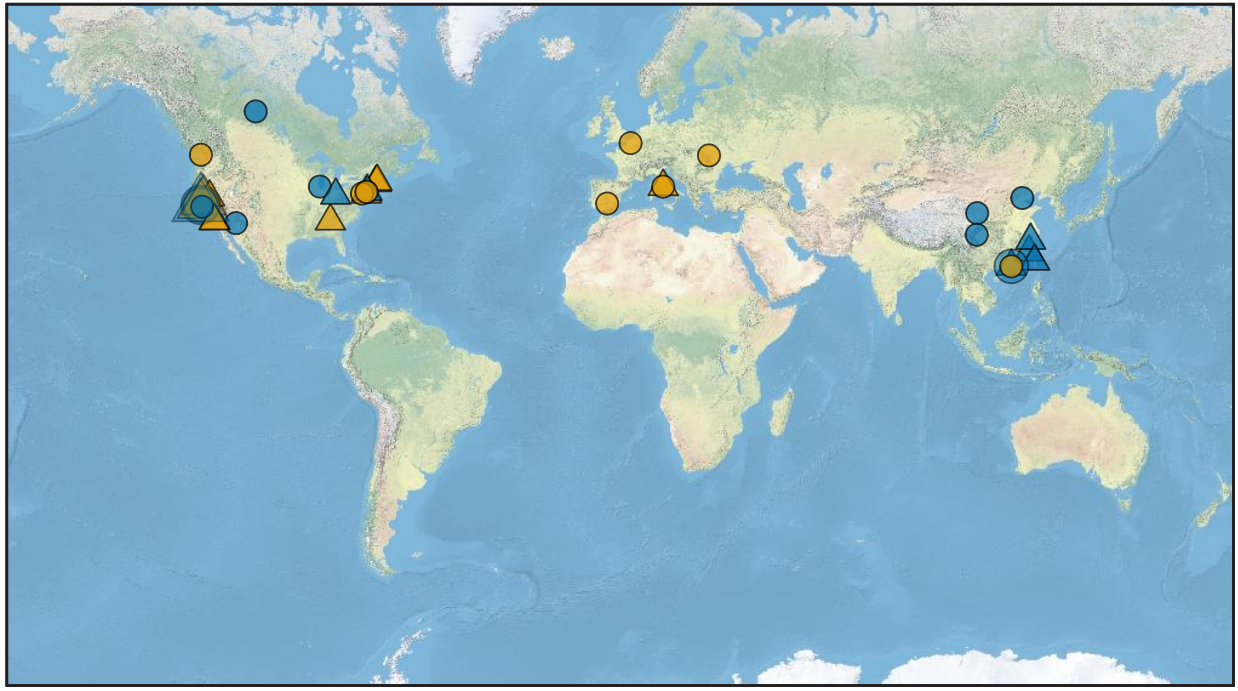
F

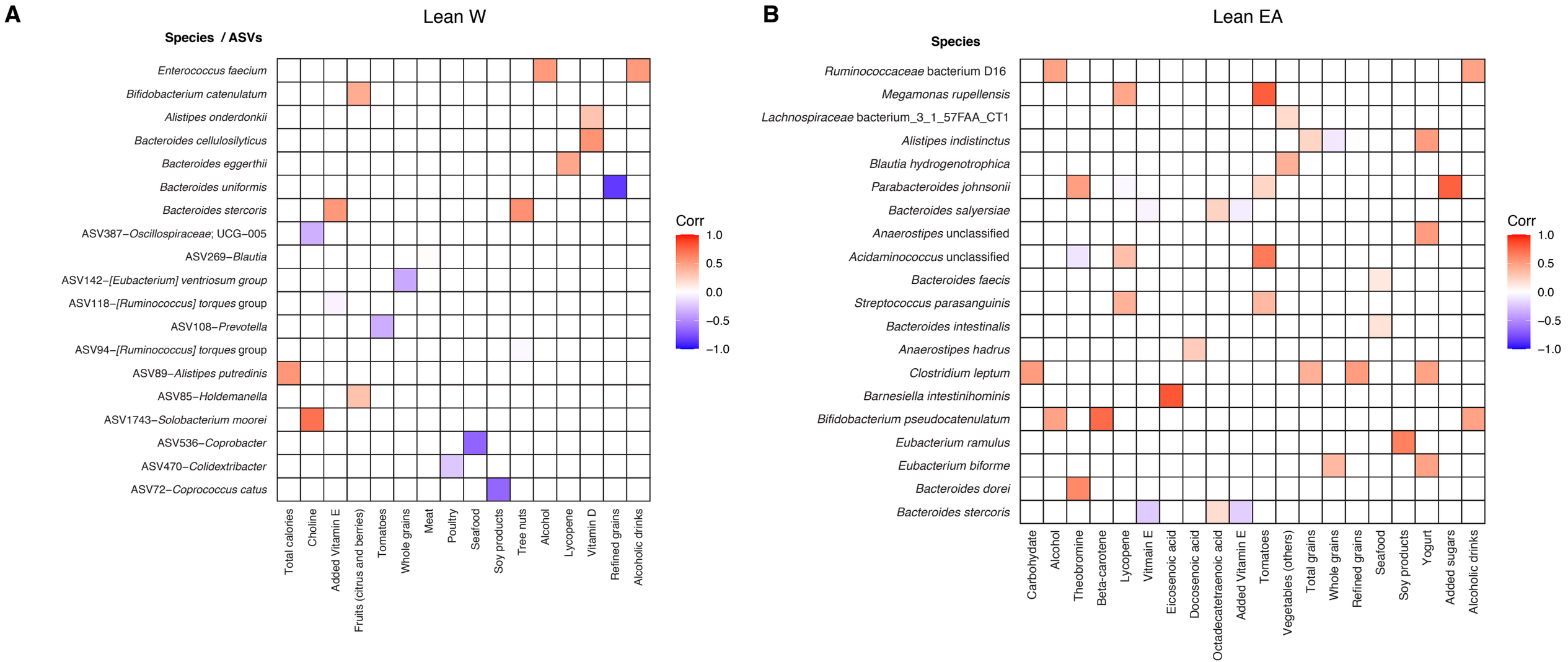


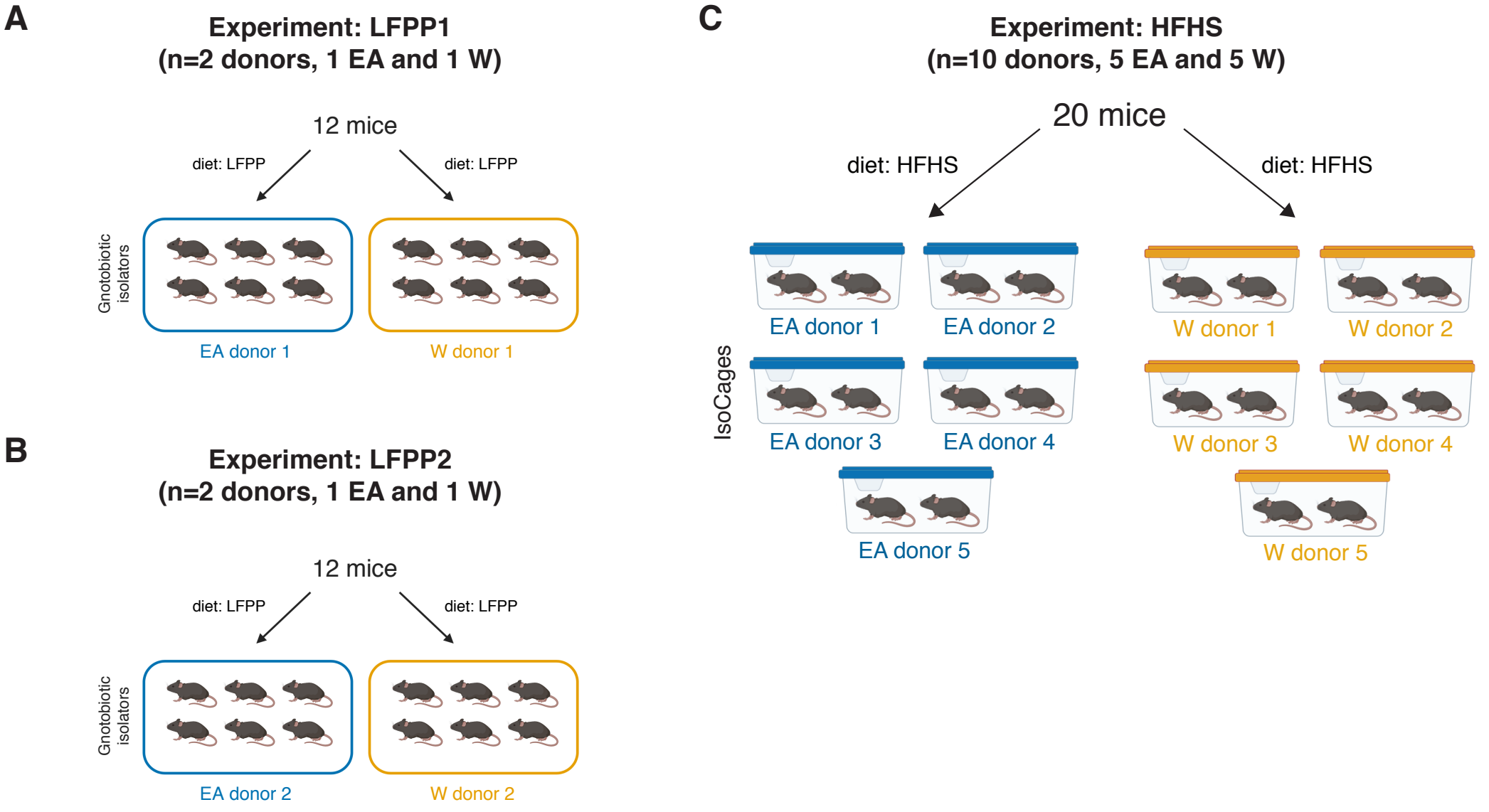

A
BCAAs
p = 0.637
p = 0.781
p = 0.917
Concentration (μg/g)
1000
500
0
Leucine
Valine
Isoleucine
EA
W
B
Bile acids
p = 0.654
p = 0.652
p = 0.859
Concentration (nmol/g)
10000
1000
100
10
Total
Free
Conjugated
EA
W

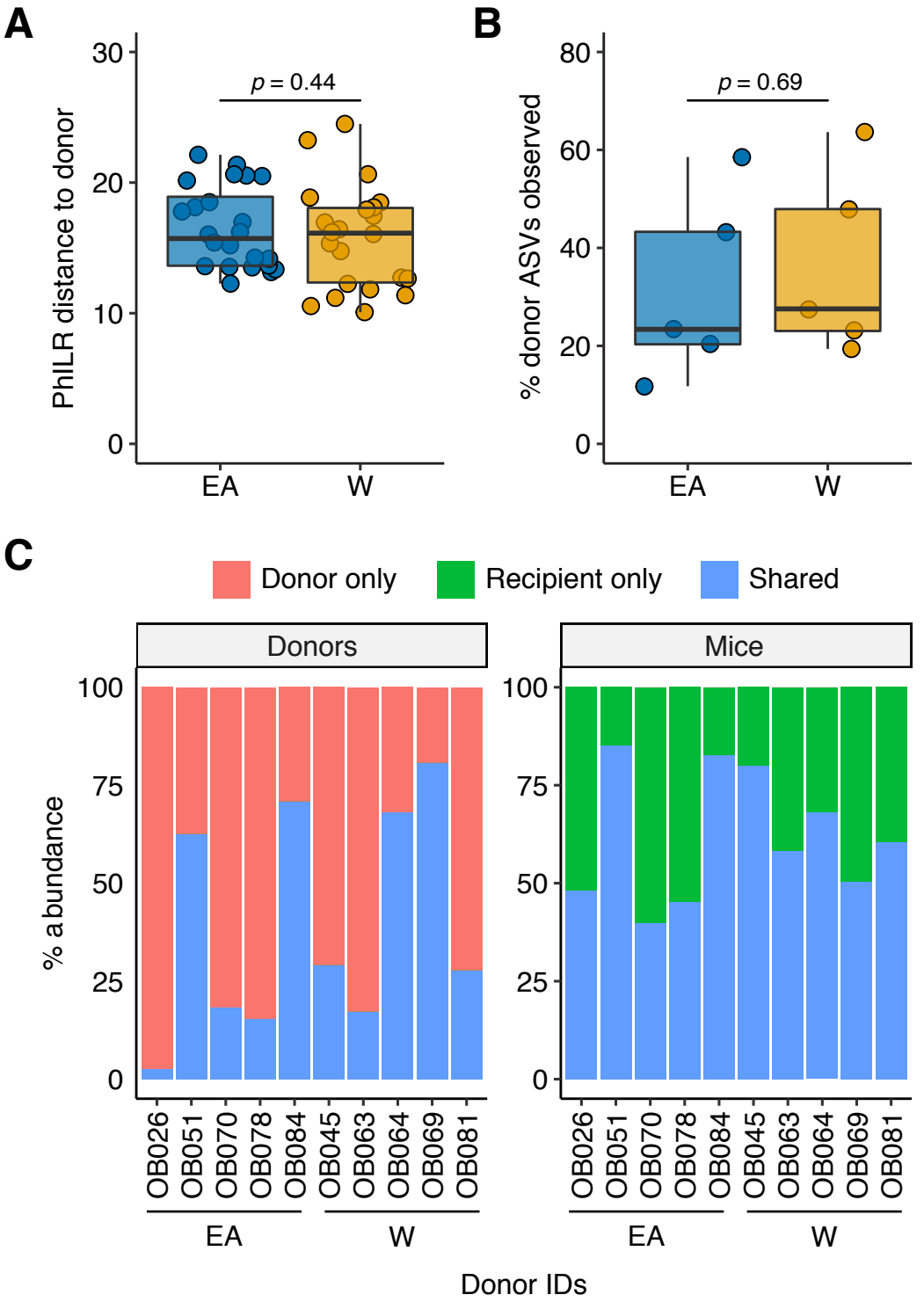

A
EA
W
p = 0.205
rho = -0.28
p = 0.0003
rho = -0.67
Faith's PD
BMI
B
EA
W
p = 0.427
rho = 0.18
p = 0.006
rho = -0.66
Faith's PD
Body fat, %
C
EA
W
p = 0.009
rho = -0.54
p = 0.002
rho = -0.69
Shannon diversity
BMI
D
EA
W
p = 0.761
rho = -0.069
p = 0.002
rho = -0.62
Shannon diversity
Body fat, %

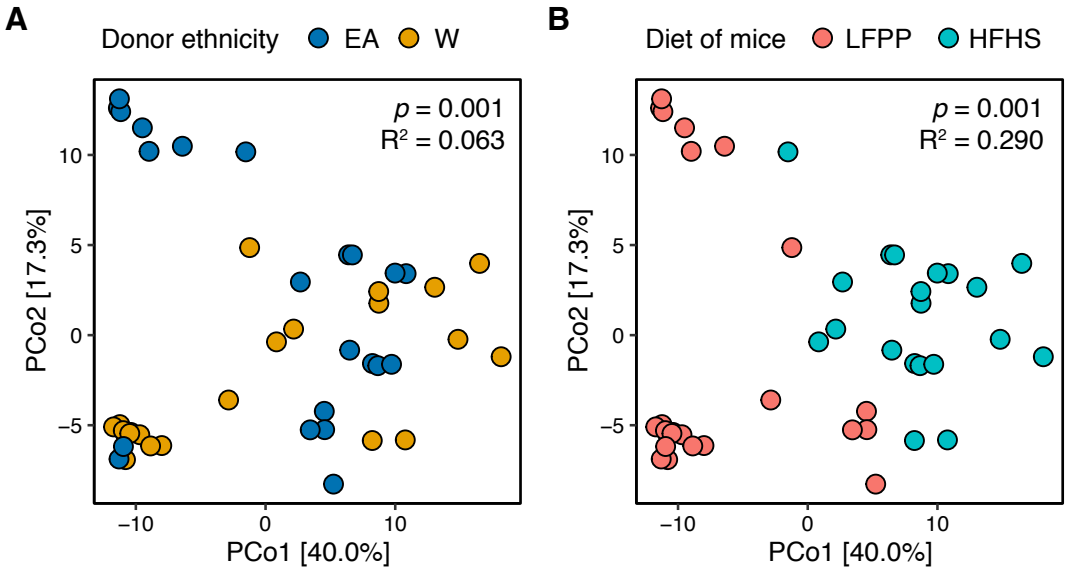

Variance explained (R^2)
0.00
0.02
0.04
0.06
Demographic
Laboratory
Lifestyle
Metabolic
*
*
*
*
*
Diabetes
Metformin
Ethnicity
Statin
Age
Insulin
BMI
Sex
Fiber
Carbohydrate
Triglycerides
Fat
HOMA-IR
FPG
US born
Protein
Cholesterol
Alcohol
Energy intake


Subjects (n)
1
2
3
4
BMI
Lean
Obese
Ethnicity
EA
W


A
Lean W
Species / ASVs
Enterococcus faecium
Bifidobacterium catenulatum
Alistipes onderdonkii
Bacteroides cellulosilyticus
Bacteroides eggerthii
Bacteroides uniformis
Bacteroides stercoris
ASV387–Oscillospiraceae; UCG–005
ASV269–Blautia
ASV142–[Eubacterium] ventriosum group
ASV118–[Ruminococcus] torques group
ASV108–Prevotella
ASV94–[Ruminococcus] torques group
ASV89–Alistipes putredinis
ASV85–Holdemanella
ASV1743–Solobacterium moorei
ASV536–Coprobacter
ASV470–Colidextribacter
ASV72–Coprococcus catus
Total calories
Choline
Added Vitamin E
Fruits (citrus and berries)
Tomatoes
Whole grains
Meat
Poultry
Seafood
Soy products
Tree nuts
Alcohol
Lycopene
Vitamin D
Refined grains
Alcoholic drinks
Corr
1.0
0.5
0.0
−0.5
−1.0
B
Lean EA
Species
Ruminococcaceae bacterium D16
Megamonas rupellensis
Lachnospiraceae bacterium_3_1_57FAA_CT1
Alistipes indistinctus
Blautia hydrogenotrophica
Parabacteroides johnsonii
Bacteroides salyersiae
Anaerostipes unclassified
Acidaminococcus unclassified
Bacteroides faecis
Streptococcus parasanguinis
Bacteroides intestinalis
Anaerostipes hadrus
Clostridium leptum
Barnesiella intestinihominis
Bifidobacterium pseudocatenulatum
Eubacterium ramulus
Eubacterium biforme
Bacteroides dorei
Bacteroides stercoris
Carbohydate
Alcohol
Theobromine
Beta-carotene
Lycopene
Vitmain E
Eicosenoic acid
Docosenoic acid
Octadecatetraenoic acid
Added Vitamin E
Tomatoes
Vegetables (others)
Total grains
Whole grains
Refined grains
Seafood
Soy products
Yogurt
Added sugars
Alcoholic drinks
Corr
1.0
0.5
0.0
−0.5
−1.0


A
Experiment: LFPP1
(n=2 donors, 1 EA and 1 W)
12 mice
diet: LFPP
diet: LFPP
Gnotobiotic isolators
EA donor 1
W donor 1
B
Experiment: LFPP2
(n=2 donors, 1 EA and 1 W)
12 mice
diet: LFPP
diet: LFPP
Gnotobiotic isolators
EA donor 2
W donor 2
C
Experiment: HFHS
(n=10 donors, 5 EA and 5 W)
20 mice
diet: HFHS
diet: HFHS
IsoCages
EA donor 1
EA donor 2
W donor 1
W donor 2
EA donor 3
EA donor 4
W donor 3
W donor 4
EA donor 5
W donor 5


A
PhILR distance to donor
30
20
10
0
p = 0.44
EA
W
B
% donor ASVs observed
80
60
40
20
0
p = 0.69
EA
W
C
Donor only
Recipient only
Shared
Donors
Mice
% abundance
100
75
50
25
0
OB026
OB051
OB070
OB078
OB084
OB045
OB063
OB064
OB069
OB081
EA
W
Donor IDs


A
Donor ethnicity
EA
W
p = 0.001
R² = 0.063
PCo2 [17.3%]
PCo1 [40.0%]
10
5
0
−5
−10
0
10
B
Diet of mice
LFPP
HFHS
p = 0.001
R² = 0.290
PCo2 [17.3%]
PCo1 [40.0%]