Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Human Diets & Microbiome Initiative: A five-country analysis of geographic and dietary drivers of gut microbiome composition and genomic variation

Andrew Bartko

abartko@ucsd.edu

University of California San Diego

Lora Khatib

UC San Diego https://orcid.org/0000-0002-6998-8278

Renee Oles

UC San Diego

Alejandra Hernández

UC San Diego

Tyler Myers

UC San Diego

Jianshu Zhao

UC San Diego

Megha Kumar

UC San Diego

Ana Garcia

Institute of Agrochemistry and Food Technology

Veronica Tolosa

Institute of Agrochemistry and Food Technology

Nobuko Matsunami

Institute of Science Tokyo

Daisuke Suzuki

Institute of Science Tokyo https://orcid.org/0000-0003-1750-8516

Nodoka Chiba

Institute of Science Tokyo https://orcid.org/0009-0002-2473-663X

Katharine Gilbert

UC San Diego

Promi Das

UC San Diego

Jon Sanders
UC San Diego
Lucas Patel
University of California San Diego https://orcid.org/0000-0001-8607-2782
Nicole Litwin
Crohn's & Colitis Foundation
Brent Nowinski
UC San Diego
Amanda Birmingham
UC San Diego
Sawyer Farmer
UC San Diego https://orcid.org/0009-0006-4160-9535
Caitriona Brennan
UC San Diego
MacKenzie Bryant
UC San Diego
Caitlin Tribelhorn
UC San Diego https://orcid.org/0009-0002-8788-471X
Karenina Sanders-Bodai
UC San Diego
Charles Cowart
UC San Diego
Oliver Nizet
Johns Hopkins University
Edgar Diaz
UC San Diego
Patrick Veiga
INRAE https://orcid.org/0000-0001-6318-8547
Sudarshan Shetty
Danone Research & Innovation
Jan Knol
University College Cork https://orcid.org/0000-0002-2103-5961
Christophe Lay
Danone Research & Innovation
Hana Koutnikova
Danone Research & Innovation
Soline Chaumont
Danone Research & Innovation
Julien Tap

UMR1319 MICALIS Institute
Aurélie Cotillard
Danone Research & Innovation
Muriel Derrien
KU Leuven https://orcid.org/0000-0001-8841-9153
Manolo Laiola
Danone Research & Innovation
Antonio González
UC San Diego
Armando Tovar
Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran
Nimbe Torres
Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran
Yolanda Sanz
Institute of Agrochemistry and Food Technology, Spanish National Research Council (IATA-CSIC)
https://orcid.org/0000-0002-1615-1976
Takuji Yamada
Institute of Science Tokyo
Daniel McDonald
UC San Diego
Se Jin Song
UC San Diego
Rob Knight
UC San Diego

Article

Keywords:

Posted Date: February 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8436332/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. Rob Knight is a scientific advisory board member, and consultant for BiomeSense, Inc., has equity and receives income. He is a scientific advisory board member and has equity in GenCirq. He has equity in and acts as a consultant for Cybele. He is a Vice President and board member of Microbiota Vault, Inc. He is a board member of N=1 IBS advisory board and receives income. He is a Senior Visiting Fellow of HKUST Jockey Club Institute for

Advanced Study. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. Nicole Litwin was employed by Center for Microbiome Innovation at the time of the project. Patrick Veiga, Jan Knol, Hana Koutnikova, Soline Chaumont, Aurélie Cotillard, Muriel Derrien, and Manolo Laiola were employees of Danone Research & Innovation at the time of the project. Daniel McDonald is a consultant for and has equity in BiomeSense, Inc. The terms of these arrangements have been reviewed and approved by the University of California, San Diego, in accordance with its conflict-of-interest policies. Andrew Bartko is a founder of Guilden Corporation and is an equity owner. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

The Human Diets & Microbiome Initiative: A five-country analysis of geographic and dietary drivers of gut microbiome composition and genomic variation

Lora Khatib[1,2], Renee Oles[1], Alejandra Ríos Hernández[1], Tyler Myers[3], Jianshu Zhao[1], Megha Kumar[1,4,5], Ana Larroya Garcia[6], Veronica Tolosa[6], Nobuko Matsunami[7], Daisuke Suzuki[7], Nodoka Chiba[7], Katharine E. Gilbert[3], Promi Das[3], Jon G Sanders[3], Lucas Patel[1,8,9], Nicole Litwin[10], Brent Nowinski[3], Amanda Birmingham[1], Sawyer Farmer[1], Caitriona Brennan[1], MacKenzie Bryant[1], Caitlin Tribelhorn[1], Karenina Sanders-Bodai[1], Charles Cowart[3], Oliver Nizet[11], Edgar A. Diaz[1], Patrick Veiga[12,13], Sudarshan A. Shetty[14], Jan Knol[14], Christophe Lay[15], Hana Koutnikova[16], Soline Chaumont[16], Julien Tap[13], Aurélie Cotillard[16], Muriel Derrien[17,18], Manolo Laiola[16], Antonio González[1], Armando R. Tovar[19], Nimbe Torres[19], Yolanda Sanz[6], Takuji Yamada[7], Daniel McDonald[1], Se Jin Song[3], Andrew Bartko[1,3,20], Rob Knight[1,3,20,21,22,23]

[1]Department of Pediatrics, University of California San Diego, La Jolla, California, USA
[2]Neurosciences Graduate Program, University of California San Diego, La Jolla, California, USA
[3]Center for Microbiome Innovation, University of California San Diego, La Jolla, California, USA
[4]Department of Mathematics, University of California San Diego, La Jolla, California, USA
[5]Department of Cognitive Science, University of California San Diego, La Jolla, California, USA
[6]Microbiome Innovation in Nutrition & Health Research Unit, Institute of Agrochemistry and Food Technology, Spanish National Research Council (IATA-CSIC), Paterna-Valencia, Spain
[7]School of Life Science and Technology, Institute of Science Tokyo, Tokyo, Japan
[8]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, USA
[9]Medical Scientist Training Program, University of California San Diego, La Jolla, CA, USA
[10]Crohn's & Colitis Foundation, 733 3rd Ave, New York, NY 10017
[11]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA
[12]Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France
[13]Université Paris-Saclay, INRAE, AgroParisTech, UMR1319 MICALIS Institute, Jouy-en-Josas, France
[14]Danone Research & Innovation, Utrecht, the Netherlands
[15]Precision Nutrition D-Lab, Danone Global Research & Innovation Center, Singapore
[16]Danone Research & Innovation, Gif sur Yvette, France
[17]VIB-KU Leuven Center for Microbiology, Leuven, Belgium
[18]Laboratory of Molecular Bacteriology, Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium
[19]Department of Physiology of Nutrition, Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran
[20]Shu Chien-Gene Lay Department of Bioengineering, University of California San Diego, La Jolla, CA, USA
[21]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
[22]Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA
[23]Hong Kong University of Science and Technology Jockey Club Institute for Advanced Study, Hong Kong, China

Abstract

Human diet and lifestyle vary widely across cultures, shaping global diversity in gut microbiome composition and function. However, comparable cross-population studies have been limited by methodological heterogeneity. The Human Diets and Microbiome Initiative (THDMI) integrates standardized recruitment, dietary assessment, and shotgun metagenomics across 1,976 participants from the United States, United Kingdom, Spain, Mexico, and Japan to disentangle biological from technical variation. We observed pronounced geographic differences in nutrient intake, food consumption, and microbial composition, with both shared and country-specific taxa linked to diet quality. Sparse canonical correlation and machine-learning analyses revealed that microbiome–diet associations were largely population-specific, with limited cross-country generalization. Strain-level analyses further uncovered significant genomic differentiation of gut microbes, reflecting localized dietary and ecological pressures. Together, these results show that while core microbial responses to healthy eating exist, the structure and function of the human gut microbiome are profoundly shaped by culture, geography, and long-term dietary practices.

Introduction

Patterns of human diet and lifestyle vary profoundly across the globe, shaped largely by geography, culture, and socioeconomic context, with genetic or ethnic factors contributing to a lesser extent.[1] These differences contribute to heterogeneity in dietary exposures, chronic disease risk, and health outcomes across populations. A growing body of work demonstrates that the gut microbiome mediates key effects of diet and environment on host physiology, influencing metabolic function, immune responses, and nutrient metabolism.[2–5] Yet, the extent to which diet-microbiome interactions are shared across cultures versus shaped by population-specific ecological and evolutionary pressures remains unresolved.

Early cross-population studies revealed striking geographic structuring of the gut microbiome, especially between drastically contrasting lifestyles, with agrarian or hunter-gatherer communities often exhibiting greater microbial diversity and distinct taxonomic and functional profiles compared to industrialized populations.[6–10] These differences are often attributed to dietary shifts from fiber-rich, minimally processed foods toward Westernized patterns high in refined grains, added sugars, and fats. Controlled dietary interventions further support a dominant role for diet, with extreme changes restructuring microbial communities within days, while modest modifications yield subtler responses.[11–14] Microbial signatures associated with immigration and acculturation additionally suggest that dietary convergence may coincide with loss of microbial diversity and functions linked to traditional diets.[15]

Despite these advancements, large-scale multi-country investigations with standardized microbiome profiling and harmonized diet and metadata collection remain rare.[16,17] Comparisons across independent cohorts are often confounded by technical variability (e.g., sampling, extraction, sequencing, and bioinformatics), which can obscure biological signals. Consequently, apparent cross-country differences may reflect methodological heterogeneity

rather than genuine ecological divergence.

To address these challenges, we present The Human Diets and Microbiome Initiative (THDMI), a coordinated multi-country study encompassing 1,976 participants from five industrialized nations (United States (US), United Kingdom (UK), Japan, Spain, and Mexico). THDMI integrates standardized recruitment, biospecimen processing, shotgun metagenomic sequencing and analysis, and harmonized dietary and lifestyle assessments, enabling directly comparable datasets across sites while retaining population-specific characteristics. Using these data, we aimed to (i) characterize cross-country variation in demographic, dietary, and microbial features, and (ii) determine whether microbial associations with host and dietary factors are shared or population-specific. Our findings reveal both conserved and context-dependent host-microbiome-diet interactions, offering insight into how cultural and dietary practices shape microbial ecology.

Results

Population Characteristics

After quality control, the analytic cohort included 1,976 participants across the US, Mexico, UK, Spain, and Japan (Figure 1A,B). Demographic variables differed significantly across countries (Table 1), most prominently in cosmetics use (75% Japan vs. ≤ 20%; $P < .001$), BMI (25.7 Mexico vs. 21.7 Japan; $P < .001$), and age distributions (52.1 UK vs. 42.0 Mexico; $P < .001$).

Dietary assessments used country-specific validated FFQs, which were harmonized to shared food codes and standardized food groupings (Table S1). Following harmonization, 29 nutrients and 60 aggregated food groups were retained. Nutrient and food intake varied significantly by country (Table S2-S3), with particularly large differences in dietary patterns consistent with regional staples (e.g., rice/pasta/noodles and sodium-rich foods).

To quantify cross-country structure across domains, we performed principal component analysis (PCA) on metadata, nutrients, and food groups, as well as robust PCA (RPCA) on metagenomic profiles (Figure 1C). All domains demonstrated strong clustering by country (PERMANOVA $P < .001$). Japan and Spain exhibited the greatest between-country dissimilarity across microbiome, and diet profiles (Figure S1C). Microbial separation was driven largely by obligate anaerobes within Clostridia and multiple Escherichia coli OGUs (operational genomic units). However, we were unable to determine whether E. coli OGUs represent distinct strains or reflect multimapping to a single strain. Metadata variables driving variation included BMI-related measures, clinical conditions, and lifestyle variables such as cosmetics/supplement use. Nutrients driving separation included fiber, iron, vitamin B6, carbohydrates, and fats. Important food drivers included green or string beans, tomatoes, salt, red meats, and poultry.

Associations Between the Microbiome, Metadata, Nutrients, and Foods

Next, we performed sparse canonical correlation analysis (sCCA)[19] to evaluate relationships

between metagenomic profiles and host metadata, nutrient intake, and food consumption across cohorts (Figure 2). sCCA revealed strong correlations between the microbiome and metadata ($r = 0.70$, $p < 0.0001$; Figure 2A), nutrients ($r = 0.83$, $p < 0.0001$; Figure 2B), and foods ($r = 0.72$, $p < 0.0001$; Figure 2C), driven largely by between-country heterogeneity. Canonical variate 1 (CV1) further confirmed large cross-country differences for all domains (Kruskal–Wallis $P < 0.0001$; Figures 2D–F), with Japan and Spain typically separating from Mexico, US, and UK.

Feature loadings highlighted microbial and host/diet drivers of canonical variates (Figures 2G–I). Among microbes, taxa such as Clostridium_A leptum, Butribacter intestini, Fontimonas thermaphila, and Mycobacterium cosmeticum showed strong contributions across all domains. Key metadata contributors included cosmetics use, constipation, probiotic use, and exercise. Diet contributors included protein, magnesium, niacin, and water-insoluble fiber, and foods such as garlic and rice/pasta/noodles.

Within-country sCCA revealed consistent correlations across domains (Figures S2–S4). Overlap of top microbial features across countries was limited. However, some taxa such as Enterocloster clostrioformis, Campylobacter_A concisus, Cloacibacillus evryensis, Acidaminococcus intestini, and Bacteroides_F pectinophilus overlapped between two to three countries. Host metadata features most strongly associated with metagenomic profiles included exercise frequency, bowel movement quality, and age. Cholesterol was a top nutrient along with MUFA and fiber. Oranges, grapefruit, and tangerines were among overlapping foods correlated with metagenomic profiles. Motivated by this pattern, we tested whether B. pectinophilus is associated with pectin intake and observed positive associations in three countries with pectin data present (Figure S4K; $Ps < .05$).

More broadly, many within-country associations reflected regionally variable or culturally salient foods, although exceptions suggested that frequency of intake alone does not fully explain association strength.

Associations with the Healthy Eating Index (HEI) differ across countries.

Overall dietary patterns have been shown to exert a stronger influence on the gut microbiome than individual foods.[20] To capture these broader dietary patterns, we used the Healthy Eating Index (HEI) 2015, which integrates food- and nutrient-based components of diet quality into a composite score for each participant[21] based on Dietary Guidelines for Americans.

Total HEI scores differed substantially across countries (Figure 3A; $F = 204.69$, $p < 0.0001$), with Spain exhibiting the highest (“healthiest”) average scores, followed by the UK, US, Mexico, and Japan. Across the “adequacy” components, those reflecting recommended food groups, Spain showed the highest scores for greens and beans, fatty acids (MUFA/PUFA), and seafood and plant proteins (Figure 3B), whereas Japan showed the lowest scores for greens and beans, fruit, and whole grains. For the “moderation” components, which reflect foods to limit, Spain again performed best, reporting the lowest intake of sodium, refined grains, and added sugars,

while Japan scored lowest (indicating higher intake) for sodium and refined grains (Figure 3C). However, Japan reported lower saturated fat intake compared to the other countries.

To examine microbiome-diet relationships, we performed PERMANOVA analyses within each country, testing associations between metagenomic profiles and both HEI total and component scores (Figure 3D). No HEI component was significantly associated with microbiome composition across all five countries, confirming substantial cross-cultural differences in microbiome-diet relationships, even at higher food classification levels. Significant associations with the total HEI score, as well as several component scores, were observed in the US, Mexico, and Spain, whereas Japan and the UK showed limited associations.

We next evaluated whether countries with significant HEI-microbiome associations shared similar predictive microbial features. Elastic Net regressors were trained within each country and then cross-applied to other countries (Figure 3E). Both root mean square error (RMSE) and R^2 declined markedly when models trained in one country were tested on another, compared to within-country testing, suggesting that distinct microbial features underlie diet-microbiome associations across populations. Interestingly, Spain and Mexico exhibited lower cross-country generalization compared to models involving the US, even though their microbiome profiles were more similar than to those of other countries (Figure S1A & C).

To test this further, we ran Bayesian Inferential Regression for Differential Microbiome Analysis (BIRDMAn)[22] against the HEI total score within each country. For each country, we calculated the log ratio of positively associated taxa to negatively associated taxa and then computed Spearman correlations between these log ratios and the HEI total scores (Figure 3F). This approach allowed us to evaluate whether taxa associated with healthy eating overlapped across countries. Interestingly, several countries showed overlapping microbial associations. In Japan, taxa associated with healthy eating were also significant in Mexico and the US, but not in Spain or the UK. In Mexico, similar features were associated with diet quality in the US and Spain, but not in the UK or Japan. In Spain and the UK, overlapping taxa were shared with all other countries except Japan. In the US, overlapping taxa linked to healthy eating were shared primarily with Spain and the UK.

Lastly, we compared the number of individual microbial taxa correlated with HEI across countries (Figure 3G). The majority of taxa associated with healthy eating were country-specific, providing further evidence towards strong geographic variation in microbiome-diet relationships. However, several taxa overlapped across countries, with 48 taxa shared across all five cohorts. Phylogenetic analysis of these shared taxa revealed a distinct clade within the Clostridia class that was positively correlated with healthy eating and generally depleted in Mexico compared to other countries (Figure 3H).

Prevalence and associations of microbial taxa across countries

We next examined the prevalence and associations of microbial taxa across the five countries (Figure S5). We tested for taxa with ≥60% prevalence in one country and ≤10% prevalence in

all others. Only three Bacillus subtilis OGUs met this criterion, all of which were found almost exclusively in Japan. Within Japan, these OGUs were consistently depleted in participants with thyroid conditions and enriched among those reporting daily multivitamin use. In addition, B. subtilis OGUs showed positive associations with several nutrients, including vitamin C, isoflavones, magnesium, potassium, and folate, and negative associations with cholesterol. All three genomes demonstrated strikingly elevated and continuous coverage within Japanese participants (Figure S6A). In contrast, samples from the US, UK, Mexico, and Spain showed sparse and discontinuous coverage. To evaluate prior reports that B. subtilis is enriched in the Japanese gut microbiome due to natto consumption,[24] we examined the association between B. subtilis abundance and fermented foods and among Japanese participants and observed a strong correlation with natto intake (Figure S6B; $r = 0.59$, $p = 4.9\times10^{-39}$).

We next identified enriched taxa, defined as those at least three times more prevalent in one country than in any other. This analysis yielded 48 enriched taxa (Figure S5). Although some showed correlations with metadata, nutrients, or foods, these taxa generally exhibited low coverage and prevalence across samples (Table S4).

Country-specific genomic differentiation of microbial species

To assess whether species exhibit country-specific distributions, we performed metagenome-assembled genome (MAG) assembly, compiled a quality-filtered MAG set (≥80% completeness, ≤10% contamination), and clustered assemblies at the species level (average nucleotide identity, ANI ≥95%). Within each sufficiently sampled species (≥30 genomes; n = 73), pairwise Mash distances were used to quantify genomic relatedness. We used PERMANOVA to test whether Mash genomic distances were significantly structured by country of origin per species. Thirty-five species met significance after FDR correction (FDR < 0.05), demonstrating non-random clustering of genetic lineages by country (Figure 4A). Several lineages within the Lachnospiraceae family and multiple Prevotella species showed consistently strong associations with country of origin. To identify the genetic basis of this structure, we performed unitig-based genome-wide association analyses using pyseer,[25] testing all pangenomic unitigs for enrichment in focal versus non-focal populations. Significant variants were mapped to genes, annotated with KEGG pathways, and summarized across species and countries to detect country-specific pathways (Figure 4B).

Population-specific enrichment of metabolic functions was widespread (Figure S7). Genetic potential to encode for enzymes involved in carbohydrate degradation, such as β-galactosidase (lacZ) and β-glucosidase (bglX), exhibited strong country- and species-dependent biases, as did nitrogen- and amino-acid–biosynthetic genes such as dapF (lysine/diaminopimelate pathway), carB/CPA2 (carbamoyl-phosphate synthase), and asnB (asparagine synthetase). Within each population, enrichment magnitudes varied widely among co-resident species rather than converging on a single KEGG term, underscoring that population-level dietary adaptation arises from a distributed, polyfunctional microbial network rather than uniform selection on identical loci.

We next analyzed gene presence/absence patterns in two gut commensals, Bifidobacterium longum and Faecalibacterium longum. In B. longum, principal-coordinate analysis of Mash distances revealed clear separation by country of origin (PERMANOVA $F = 2.24$, $p < 0.01$; Figure 4C). MAGs from Mexico exhibited a marked loss of galactose-metabolism genes (galC, galD) likely due to the high prevalence of lactose intolerance in the Mexican populations (31.1%)[9], retained in >70 % of non-Mexican genomes but in <10 % of genomes from Mexico (Figure 4D). Conversely, sacC, associated with sucrose and fructan metabolism, was enriched in genomes from Japan. In F. longum, genomes from Japan were depleted of histidine-biosynthetic genes (hisA–F; Figure 4E–G), whereas genomes from Mexico showed significant loss of putP, encoding a proline transporter. Together, these patterns indicate that differential nutrient availability or other environmental factors across human populations exert lineage-specific selective pressures on microbial metabolic repertoires.

Discussion

This study represents one of the first single-study, cross-cultural analyses of diet, lifestyle, and the gut microbiome in industrialized societies. Leveraging shotgun metagenomics, lifestyle and dietary intake data, and genomic resolution of microbial lineages, we examined 1,976 participants from the US, UK, Spain, Mexico, and Japan. We found pronounced differences across countries in metadata variables, dietary patterns, and microbiome composition. These differences were accompanied by country-specific and shared microbial signatures of diet, as well as clear genomic differentiation of microbial lineages, illustrating how cultural lifestyle and dietary practices shape the structure and function of the gut microbiome.

Aligned with prior cross-population studies,[6–9] we observed pronounced differences in microbiome composition and function, dietary intake, and lifestyle across countries. Several features driving this separation have been reported previously, such as a gradient toward lower BMI in Japan[26] and higher intake of olive oil in Spain.[27] Interestingly, Spain and Mexico harbored the most taxonomically similar gut microbiome profiles among the cohorts. However, cross-country generalization between them was unexpectedly low when compared to the US. This suggests that while the overall community composition may reflect shared genetic ancestry or historical population structure, consistent with prior studies linking host genetics to microbial taxa,[28,29] the functional capacities of these communities may have diverged in response to distinct dietary patterns and lifestyles.

Further, we find substantial heterogeneity in microbiome-diet associations across countries. Sparse canonical correlation analysis revealed pronounced cross-country differences in microbiome-diet canonical structures, indicating distinct context-dependent relationships between dietary intake and microbial composition. However, it is not possible to fully disentangle true biological variation from technical bias arising from differences in FFQs and nutrient calculation methodologies. These findings highlight a critical need for global nutrient harmonization to enable more accurate cross-population comparisons and to advance our understanding of host-microbiome relationships.

Here, no HEI component was consistently associated with microbiome composition in all cohorts, and machine-learning models trained to predict diet quality in one country generally failed to transfer across populations. These findings further suggest that diet-microbiome relationships are culturally contextual and highlight the need to consider alternative dietary scoring systems. It is also important to note that cross-country diet-microbiome comparisons assume sufficient overlap in detectable taxa across populations, which may not always hold.

Importantly, however, one subset of microbial taxa, largely within the class Clostridia, showed consistent associations with healthier eating patterns across countries, highlighting a potentially evolutionarily conserved microbial core responsive to diet quality. Yet, these findings largely challenge the assumption that universal microbiome markers of dietary health exist, and instead point to context-dependent microbial responses to diet. Cultural dietary substrates, culinary processing techniques, food availability, and habitual co-consumption patterns may contribute to population-specific nutritional ecotypes, underscoring the need for stratified, population-aware microbiome-targeted nutritional interventions.

Our strain-level analyses highlight geography-linked microbial evolution. Multiple species exhibited significant country-specific genomic differentiation, particularly within fiber-fermenting lineages such as Prevotella and Lachnospiraceae, extending prior observations in traditional versus Westernized populations.[6–10] Differential enrichment of carbohydrate-active enzymes and amino acid biosynthetic pathways across countries suggests that microbes adapt not only to host genetic or environmental factors but also to long-term population-level dietary exposures. These findings resonate with emerging evidence that fine-scale microbial population structure reflects localized ecological selection.[34,35]

Together, our results reveal that while core ecological principles link diet and microbial function globally, the microbial signatures of dietary quality are not universal. However, several limitations warrant acknowledgement. As with most crowdsourced initiatives, our populations are enriched in women and individuals with relatively high diet quality. Further efforts are needed to capture more representative populations. Although using country-specific FFQs is methodologically appropriate, a potential bias arises from harmonizing these instruments to a single nutrient database and applying standardized HEI scoring across countries. In addition, our analyses relied primarily on food-group-level intake, which may reduce sensitivity to detect certain food-microbiome associations. Finally, the analysis is cross-sectional, and causal mechanisms remain to be established. Interventional studies across diverse dietary cultures will be critical for validating country-specific and shared microbial determinants of nutritional responsiveness.

In sum, this work provides one of the most comprehensive examinations to date of global diet-microbiome relationships across distinct populations, demonstrating that while specific microbial signatures of diet quality are generic, microbiome-diet interactions are strongly shaped by culture, geography, and dietary patterns. As the field advances toward precision nutrition, these findings support a shift from universal microbiome dietary biomarkers to culturally informed, population-aware frameworks that capture the full breadth of human dietary ecology.

Methods

Sample collection

Adult participants were consented through the Microsetta Initiative platform from the US, UK, Mexico, Spain, and Japan in accordance with University of California San Diego Human Research Protection Program (Protocol #: 141853) and the Instituto Nacional de Ciencias Medicas y Nutrición Salvador Zubiran (INCMNSZ) in Mexico (Protocol #: 3853). Fecal samples were collected using the Microsetta fecal collection kit.[36] Specifically, swabs from used toilet paper were placed into tubes containing 2mL of 95% EtOH and stored at room temperature. In the US and UK, samples were shipped directly from participants to UC San Diego. In Mexico, Spain, and Japan, samples were first aggregated within the respective country before bulk shipped to UC San Diego. On receipt, samples were checked into the Microsetta system. Broken samples or samples which arrived without remaining EtOH were discarded. Only samples with complete information, including FFQ data, were used for sequencing and analysis.

Diet data collection and preparation

Participants in the US and UK completed the VioScreen™ food frequency questionnaire (FFQ) (Version 5, VioCare, Princeton, NJ)[37], while participants in Mexico completed Version 5-Mex (VioCare, Princeton, NJ), which was adapted to include foods common in Mexico. Participants in Spain completed the PREDIMED FFQ, and the data was collected by the Spanish National Research Council (CSIC) at the Institute of Agrochemistry and Food Technology in Spain.[38] Participants in Japan took a digitized version of the FFQ developed for the Next Generation Multipurpose Cohort Study (JPHC-NEXT).[39] Through these FFQs, participants reported their typical intake of a wide range of foods over the previous three months for Mexico, US, and UK, and previous year for Spain and Japan, which allowed for a comprehensive assessment of dietary patterns and nutrient intake. The FFQ provided detailed estimates of macronutrient (e.g., carbohydrates, proteins, fats) and micronutrient (e.g., vitamins, minerals) intake, as well as a breakdown of food categories (e.g., fruits, vegetables, meats). To improve comparability across cohorts, dietary data were harmonized using VioScreen food codes and collapsed to common food groupings. Original food names and their corresponding grouped categories are provided in Supplemental Table 2. Any nutrient or food columns that were not measured for an entire country were excluded. After cleaning, 60 food groups and 30 nutrient variables (including calories) remained. All food and nutrient values were standardized by total calorie intake prior to analysis.

The Healthy Eating Index–2015 (HEI–2015), previously validated for use in Japan,[40] was calculated for each country using the Food Patterns Equivalents Database (FPED), following code adapted from the sample SAS scripts available on the NIH HEI website (https://github.com/lkhatib/Calculating_HEI_Python). For Japan and Spain, FPED components were derived by matching foods in each country's dataset to their corresponding items in the Food and Nutrient Database for Dietary Studies (FNDDS; Table S5). The matched food items

were then converted to FPED component values through matrix multiplication.

DNA extraction and sequencing

DNA extraction was performed using the Matrix Method as described previously by Brennan et al.[41] Briefly, swabbed samples were transferred to Matrix Tubes™. 30 µL of 0.1, 0.5- and 1-mm zirconia-silica beads were added to each Matrix Tube using a LabTie bead dispenser (Molgen, Netherlands). 600 µL of Lysis buffer was added to each tube using The Biotek Multiflo bulk reagent dispenser (Agilent, CA, USA). The Matrix Tubes were capped using the Capit-All (ThermoFisher Scientific, MA, USA). Bead beating was performed on the SpexMiniG for 2 min at 1,200 x g. The remaining steps of the DNA extraction followed the manufacturer's instructions, with bead clean-ups performed using the automated KingFisher Flex Purification System (ThermoFisher Scientific, MA, USA). Extracted nucleic acids were stored at -80°C until further processing.

DNA was prepared for shotgun metagenomics sequencing as described previously.[42] Metagenomic libraries were normalized by iSeq (Illumina) read count distribution to generate a final pool that made sequencing on the NovaSeq more efficient.[43] The iSeq-normalized pools were run on a NovaSeq 6000 (Illumina, CA, USA). The reads are available through the European Bioinformatics Institute (EBI) (accession number: PRJEB11419).

Metagenomic data processing

After sequencing, BCL files were demultiplexed, converted to per-sample-FASTQs, and quality controlled and human sequence filtered using the Qiita[44] admin plugin: https://github.com/qiita-spots/qp-knight-lab-processing following the recommendations from Guccione et al.[45] In summary, adapter trimming was performed using fastp.[46] Each sample was then filtered against GRCh38.p14 and T2T-CHM13v2.0[47] via alignment using minimap2.[48] Reads aligning to either human reference genome were discarded. Next, a pangenome index was created using the 94 reference assemblies from the Human Pangenome Reference Consotrium release 1,[49] and reads with high pseudo-matching lengths to the pangenome were discarded. The resulting per-sample-FASTQs were uploaded into Qiita study ID #10317 (https://qiita.ucsd.edu/study/description/10317; Prep IDs: 16766, 16762, 16761, 16773, 16780, 16756, 16842, 16848, 16849, 16853, 16841, 16854, 16875) and processed with its default processing workflow. The metagenomic workflow maps paired-end reads using Bowtie2 v2.5.4[50] against Web of Life release 2[51] (WoL2) with the SHOGUN parameter set.[52] A feature table is then constructed using woltka v0.1.7.[53]

The listed Prep IDs were aggregated into a single analysis on Qiita.[44] Coverage values were calculated using MIcrobiome COVerage (micov)[23], and features with <1% cumulative coverage were excluded.

Quality control steps included removing blank and non-THDMI samples. Samples with invalid entries in any of the following metadata fields were excluded: thdmi_cohort, host_age, sex,

antibiotic_history, covid_level_of_wellbeing, and host_body_mass_index. We further removed samples missing dietary information. An additional 58 samples from Mexico were excluded due to technical bias introduced by re-sequencing following low read quality. Finally, samples with feature sums less than or equal to 1,000,000 reads were filtered out.

After these quality control procedures, the resulting feature table contained 1,976 samples (Japan: 410; Mexico: 397; Spain: 450; UK: 308; US: 411) and 3,375 microbial features.

Metadata data processing

Metadata were subjected to quality control and harmonization. Age values ≤18 or ≥100 were set to NaN and subsequently removed, while body mass index (BMI) values <12 or >70 were also set to NaN and excluded, consistent with previous work[9]. Categorical variables with fewer than two unique values within any country were dropped, as were binary columns with fewer than 10 individuals in one group per country.

Several variables were collapsed to harmonize categories across cohorts. Specifically, plant diversity was reduced to “Less than 10” and “More than 30”; and antibiotic history was reduced to “No use in the past year” and “Within the last month”, as previously described.[7] Clinical diagnosis columns were collapsed to “Diagnosed by a medical professional (doctor, physician assistant)” and “I do not have this condition”.[54] A binary “healthy” versus “unhealthy” variable was created following prior methodology.[18] Columns describing contraceptive and deodorant use were reduced to True/False. Home-cooked meals frequency was collapsed to “Daily” vs. “Not daily”, and drinks per session were simplified to “Less than 2” vs. “More than 2”.

Finally, columns with fewer than five samples in any group within a country were removed from the analysis.

MAG assembly, binning, and annotation

MaxBin 2 v2.27[55], MetaBAT 2 v2.18[56] and CONCOCT v1.10[57] were individually applied for contig binning with default parameters. The resulting MAGs were first refined (quality improvement) using DAS_Tool[58] with the searching engine USEARCH.[59] CheckM[60] lineage_wf was then used for MAG quality assessment with default parameters.[60] The quality score was defined as completeness - 5*contamination + Strain heterogeneity*0.5. Medium to high quality MAGs were defined as quality scores larger than 0.5. Dereplication of the MAGs was performed using dRep (v2.2.4) with an ANI threshold of 95% by the -fastANI option (minimum completeness 70%, maximum contamination 10% and quality score > 0.5 for filtering MAGs before dereplication).[61] Quality information from CheckM was passed to dRep to select the best quality MAG as representative of each resulting 95% ANI cluster. MAGs were classified against the Genome Taxonomy Database v214[62] using the GTDB-Tk v2.4.3 classify_wf workflow.[63]

Statistical analysis

Demographic, nutrient, and food summaries

Demographic characteristics, nutrient intake, and food group distributions were summarized across countries. Kruskal–Wallis tests were used for continuous variables and Chi-squared tests for categorical variables.

Principal component analyses (PCA)

Robust Principal Component Analysis (RPCA),[64] implemented in gemelli v0.0.9, was used to analyze microbiome composition to provide PCA-like embeddings comparable to those used for other data layers. For metadata, diet-related and COVID-19–related variables were excluded. Ordinal variables were encoded using OrdinalEncoder (scikit-learn v0.24.1), and non-ordinal categorical variables were one-hot encoded. Missing values were imputed using SimpleImputer with the mean strategy. Metadata, nutrient, and food data were standardized with StandardScaler. PCA was performed using sklearn.decomposition. Pairwise distance matrices were computed, and PERMANOVA tests were conducted using scikit-bio v0.5.7.[65]

Sparse canonical correlation analysis (sCCA)

Sparse canonical correlation analysis was conducted between metadata, foods, nutrients, and metagenomic data. Encoded and imputed metadata, and nutrient and food inputs from the PCA preprocessing described above were used. Metagenomic data were preprocessed with matrix_rclr and MatrixCompletion (gemelli.preprocessing). Sparse CCA was performed using SCCA_IPLS from cca-zoo v2.6.0.[19] Data were standardized within the function, and columns with zero variance were excluded. Tuning parameters (α and L1 ratio) were selected using 5-fold cross-validation and the one–standard error (1-SE) rule.

Prevalence analysis

For each country, the prevalence of each microbial taxon was calculated as the proportion of samples in which the taxon was detected (>0 counts). Ubiquitous taxa were defined as having >99% prevalence across all countries and excluded from comparative analyses. We first searched for “exclusive” taxa, defined as those with ≥80% prevalence in one country and 0% in all others, and no taxa met this criterion. We then lowered the threshold to identify taxa with ≥60% prevalence in one country and ≤10% prevalence in all others. Enriched taxa were defined as having a prevalence ratio ≥3 relative to the next-highest country.

Differential abundance was assessed using Bayesian Inferential Regression for Differential Microbiome Analysis (BIRDMAn)[22] (v0.1.0), testing metadata, nutrient, and food variables with consistent variance across countries. Variables to test were selected based on those that showed the greatest variation across countries (Figure 1). All associations were adjusted for thdmi_cohort. Metadata variables tested included: host_age, bowel_movement_quality_renamed, unhealthy, host_body_mass_index, sex, cosmetics_frequency, vitamin_d_supplement_frequency, vitamin_b_supplement_frequency,

multivitamin, autoimmune_reduced, probiotic_frequency, thyroid_reduced, drinks_per_session_reduced. Nutrients tested included: isoflavones_total, iron, fiber, fiber_water, protein, carbohydrates, monounsaturated fat, polyunsaturated fat, cholesterol, vitamin B6, potassium, magnesium, thiamin, niacin_equivalents, zinc, folate_total, vitamin C, insoluble fiber, and total fat. Foods tested included: garlic, rice/pasta/noodles, yogurt, processed meats, salt, soups/creams.

The WoL2 reference phylogeny was pruned to include only the enriched OGUs and uploaded to Interactive Tree Of Life (iTOL) version 7.[66] In iTOL, branch-length scaling was disabled to emphasize taxonomic topology rather than evolutionary distance. BIRDMAn association results were encoded as upward or downward arrows to indicate the direction of association and were added as annotation tracks in iTOL.

HEI across countries

Healthy Eating Index (HEI) total scores were compared across countries using analysis of variance (ANOVA) from scipy.stats v1.10.1.[67] Associations between HEI scores and metagenomic profiles were assessed using the adonis2[68] function in R, adjusting for sex, age, and BMI.

To test whether microbiome features associated with HEI in one country could generalize to other countries (US, Mexico, Spain), we applied an Elastic Net regression framework for both within and cross-country prediction. Prior work has shown that using presence–absence representations for Elastic Net regression performs as well as, or better than, abundance-based transformations for predictive tasks.[69] Therefore, metagenomic features were converted to presence–absence. To remove noisy or redundant features, taxa were retained only if they were present in at least 5% of samples. To prioritize features associated with dietary quality, we conducted a supervised univariate screen using Pearson correlations between each taxon and Healthy Eating Index (HEI) scores, followed by false discovery rate (FDR) correction ($q \leq 0.25$). To address redundancy, we applied correlation pruning to remove features with pairwise $|\rho| \geq 0.95$, keeping one representative per correlated group.

We trained Elastic Net regression models (sklearn.linear_model.ElasticNetCV) using 5-fold cross-validation to optimize model hyperparameters, including the L1 ratio and regularization strength (α). Prior to fitting, features were standardized to unit variance. Models were first evaluated within countries (e.g., 5-fold cross-validation within the Mexico, US, and Spain cohorts). To assess cross-cohort generalization, we additionally trained models on one country and evaluated predictive performance in another (e.g., train on Mexico, test on US). The same feature selection and filtering steps were applied based solely on the training cohort, and the resulting feature set was applied to the test cohort (with missing features set to zero).

BIRDMAn models were then run separately within each country with HEI total scores as the outcome. For each country, the log ratio of positively versus negatively associated taxa was computed, and Spearman correlations were calculated between these log ratios and the total

HEI score within each country. Upset plots were generated to compare the number of taxa associated with HEI across countries.

The WoL2 reference phylogeny was pruned to include the taxa that were significantly correlated with HEI in all five countries and uploaded to iTOL[66]. Spearman correlations were computed between CLR-transformed abundances and HEI total scores. Mean CLR-transformed values and prevalence for each country were then calculated and visualized on the phylogenetic tree.

Country-specific genomic differentiation of microbial species

We assembled and quality-filtered metagenome-assembled genomes (MAGs) at ≥80% completeness and ≤10% contamination. MAGs were assigned to species-level bins using an ANI-based map, and all downstream analyses were performed within each species bin. Panpiper-kit was used to compute association between genomic distances within a species to country of origin. For each pair of MAGs within a species bin, we computed pairwise Mash[70] distances (1-ANI) using k=18 and s=5000 (panpiper-kit), which sketch genomes and provide a fast, MinHash-based approximation to overall sequence divergence. Using each species' Mash distance matrix, we tested whether MAG distances were associated with country. We used PERMANOVA (9999 permutations) for the categorical factor ("country") and report the test statistic and permutation p-value.

For species where a country level had ≥30 genomes, we performed unitig-based GWAS with pyseer[25] v1.4.0, using unitigs generated by unitig-caller. Unitigs capture fine-scale signals across the pangenome, including accessory genes and intergenic/regulatory variants. We ran binary contrasts between a focal country vs all others and controlled multiple testing with Benjamini–Hochberg FDR. Significant unitigs were collapsed by overlapping coordinates and mapped to genes on a representative genome; genes were annotated with Bakta[71], and results were summarized by gene and pathway for interpretation.

For specific species of interest, we reduced the feature space to gene presence/absence (PAV) to maintain power with small sample sizes. For each gene, we tested the presence in the focal country vs others (2×2 table) and controlled FDR across genes. This yields interpretable pathway-level signals with a far lower multiple-testing burden than unitig GWAS.
All distance- and GWAS-level steps were orchestrated with panpiper-kit (ppk-run CLI). Key settings were: --perms 999, --mash-k 18, --mash-s 5000, and --kmer 31. The toolkit writes lineage-level association tables (PERMANOVA results) and per-species pyseer outputs (BH-adjusted) as plain TSVs for downstream statistics and figures.

Data processing and statistical analysis scripts will be made available at:
https://github.com/lkhatib/Five-Country-THDMI-Manuscript-Scripts

Conflict of Interest

Rob Knight is a scientific advisory board member, and consultant for BiomeSense, Inc., has

equity and receives income. He is a scientific advisory board member and has equity in GenCirq. He has equity in and acts as a consultant for Cybele. He is a Vice President and board member of Microbiota Vault, Inc. He is a board member of N=1 IBS advisory board and receives income. He is a Senior Visiting Fellow of HKUST Jockey Club Institute for Advanced Study. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. Nicole Litwin was employed by Center for Microbiome Innovation at the time of the project. Patrick Veiga, Jan Knol, Hana Koutnikova, Soline Chaumont, Aurélie Cotillard, Muriel Derrien, and Manolo Laiola were employees of Danone Research & Innovation at the time of the project. Daniel McDonald is a consultant for and has equity in BiomeSense, Inc. The terms of these arrangements have been reviewed and approved by the University of California, San Diego, in accordance with its conflict-of-interest policies. Andrew Bartko is a founder of Guilden Corporation and is an equity owner. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

References

1. Parizadeh, M. & Arrieta, M.-C. The global human gut microbiome: genes, lifestyles, and diet. Trends Mol. Med. 29, 789–801 (2023).
2. Wahlström, A., Sayin, S. I., Marschall, H.-U. & Bäckhed, F. Intestinal Crosstalk between Bile Acids and Microbiota and Its Impact on Host Metabolism. Cell Metab. 24, 41–50 (2016).
3. Sonnenburg, J. L. & Bäckhed, F. Diet–microbiota interactions as moderators of human metabolism. Nature 535, 56–64 (2016).
4. Belkaid, Y. & Hand, T. W. Role of the Microbiota in Immunity and Inflammation. Cell 157, 121–141 (2014).
5. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. Cell 165, 1332–1345 (2016).
6. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. Nature 486, 222–227 (2012).
7. McDonald, D. et al. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems 3, e00031-18 (2018).
8. Senghor, B., Sokhna, C., Ruimy, R. & Lagier, J.-C. Gut microbiota diversity according to dietary habits and geographical provenance. Hum. Microbiome J. 7–8, 1–9 (2018).
9. Khatib, L. et al. A three-country analysis of the gut microbiome indicates taxon associations with diet vary by taxon resolution and population. mSystems 10, e00544-25 (2025).
10. Schnorr, S. L. et al. Gut microbiome of the Hadza hunter-gatherers. Nat. Commun. 5, 3654 (2014).
11. O'Keefe, S. J. D. et al. Fat, fibre and cancer risk in African Americans and rural Africans. Nat. Commun. 6, 6342 (2015).
12. David, L. A. et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505, 559–563 (2014).

13. Li, F. et al. Cardiometabolic benefits of a non-industrialized-type diet are linked to gut microbiome modulation. Cell 188, 1226-1247.e18 (2025).
14. Vanegas, S. M. et al. Substituting whole grains for refined grains in a 6-wk randomized trial has a modest effect on gut microbiota and immune and inflammatory markers of healthy adults. Am. J. Clin. Nutr. 105, 635–650 (2017).
15. Vangay, P. et al. US Immigration Westernizes the Human Gut Microbiome. Cell 175, 962-972.e10 (2018).
16. Berry, S. et al. Personal metabolic responses to food predicted using multi-omics machine learning in 1,100 twins and singletons: The PREDICT I Study. Proc. Nutr. Soc. 79, E143 (2020).
17. Berry, S. E. et al. Human postprandial responses to food and potential for precision nutrition. Nat. Med. 26, 964–973 (2020).
18. Bao, Z. et al. Predicting host health status through an integrated machine learning framework: insights from healthy gut microbiome aging trajectory. Sci. Rep. 14, 31143 (2024).
19. Chapman, J. & Wang, H.-T. CCA-Zoo: A collection of Regularized, Deep Learning based, Kernel, and Probabilistic CCA methods in a scikit-learn style framework. J. Open Source Softw. 6, 3823 (2021).
20. Cotillard, A. et al. A posteriori dietary patterns better explain variations of the gut microbiome than individual markers in the American Gut Project. Am. J. Clin. Nutr. 115, 432–443 (2022).
21. Krebs-Smith, S. M. et al. Update of the Healthy Eating Index: HEI-2015. J. Acad. Nutr. Diet. 118, 1591–1602 (2018).
22. Rahman, G. et al. BIRDMAn: A Bayesian differential abundance framework that enables robust inference of host-microbe associations. Preprint at https://doi.org/10.1101/2023.01.30.526328 (2023).
23. Weng, Y. et al. Calculating fast differential genome coverages among metagenomic sources using micov. Commun. Biol. 8, 1624 (2025).
24. Tomofuji, Y. et al. Prokaryotic and viral genomes recovered from 787 Japanese gut metagenomes revealed microbial features linked to diets, populations, and diseases. Cell Genomics 2, 100219 (2022).
25. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics 34, 4310–4312 (2018).
26. Ministry of Health, Labour, and Welfare (Japan). National Health and Nutrition Survey 2017. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/eiyou/h29-houkoku.html.
27. Moreiras-Varela, O. The Mediterranean diet in Spain. Eur. J. Clin. Nutr. 43 Suppl 2, 83–87 (1989).
28. Goodrich, J. K. et al. Human Genetics Shape the Gut Microbiome. Cell 159, 789–799 (2014).
29. Blekhman, R. et al. Host genetic variation impacts microbiome composition across human body sites. Genome Biol. 16, 191 (2015).
30. Baldeon, A. D., McDonald, D., Gonzalez, A., Knight, R. & Holscher, H. D. Diet Quality

and the Fecal Microbiota in Adults in the American Gut Project. J. Nutr. 153, 2004–2015 (2023).
31. Bowyer, R. C. E. et al. Use of dietary indices to control for diet in human gut microbiota studies. Microbiome 6, 77 (2018).
32. Maskarinec, G. et al. Fecal Microbial Diversity and Structure Are Associated with Diet Quality in the Multiethnic Cohort Adiposity Phenotype Study. J. Nutr. 149, 1575–1584 (2019).
33. Wastyk, H. C. et al. Gut-microbiota-targeted diets modulate human immune status. Cell 184, 4137-4153.e14 (2021).
34. Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J. & Polz, M. F. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. Cell 178, 820-834.e14 (2019).
35. Costea, P. I. et al. Subspecies in the global human gut microbiome. Mol. Syst. Biol. 13, 960 (2017).
36. Marotz, C. et al. Evaluation of the Effect of Storage Methods on Fecal, Saliva, and Skin Microbiome Composition. mSystems 6, 10.1128/msystems.01329-20 (2021).
37. Kristal, A. R. et al. Evaluation of Web-Based, Self-Administered, Graphical Food Frequency Questionnaire. J. Acad. Nutr. Diet. 114, 613–621 (2014).
38. Fernández-Ballart, J. D. et al. Relative validity of a semi-quantitative food-frequency questionnaire in an elderly Mediterranean population of Spain. Br. J. Nutr. 103, 1808–1816 (2010).
39. Sawada, N. et al. The Japan Public Health Center-based Prospective Study for the Next Generation (JPHC-NEXT): Study Design and Participants. J. Epidemiol. 30, 46–54 (2020).
40. Murakami, K., Livingstone, M. B. E., Fujiwara, A. & Sasaki, S. Application of the Healthy Eating Index-2015 and the Nutrient-Rich Food Index 9.3 for assessing overall diet quality in the Japanese context: Different nutritional concerns from the US. PLOS ONE 15, e0228318 (2020).
41. Brennan, C. et al. Clearing the plate: a strategic approach to mitigate well-to-well contamination in large-scale microbiome studies. mSystems 9, e00985-24 (2024).
42. Sanders, J. G. et al. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. Genome Biol. 20, 226 (2019).
43. Brennan, C. et al. Maximizing the potential of high-throughput next-generation sequencing through precise normalization based on read count distribution. mSystems e00006-23 (2023) doi:10.1128/msystems.00006-23.
44. Gonzalez, A. et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat. Methods 15, 796–798 (2018).
45. Knight, R. et al. Incomplete human reference genomes can drive false sex biases and expose patient-identifying information in metagenomic data. Preprint at https://doi.org/10.21203/rs.3.rs-4721159/v1 (2024).
46. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890 (2018).
47. Rhie, A. et al. The complete sequence of a human Y chromosome. Nature 621, 344–354 (2023).
48. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–

3100 (2018).
49. Liao, W.-W. et al. A draft human pangenome reference. Nature 617, 312–324 (2023).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
51. Zhu, Q. et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat. Commun. 10, 5477 (2019).
52. Hillmann, B. et al. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems 3, 10.1128/msystems.00069-18 (2018).
53. Zhu, Q. et al. Phylogeny-Aware Analysis of Metagenome Community Ecology Based on Matched Reference Genomes while Bypassing Taxonomy. mSystems 7, e00167-22 (2022).
54. Strandwitz, P. et al. GABA-modulating bacteria of the human gut microbiota. Nat. Microbiol. 4, 396–403 (2018).
55. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607 (2016).
56. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7, e7359 (2019).
57. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146 (2014).
58. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. 3, 836–843 (2018).
59. Zhou, Y., Liu, Y.-X. & Li, X. USEARCH 12: Open-source software for sequencing analysis in bioinformatics and microbiome. iMeta 3, e236 (2024).
60. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055 (2015).
61. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 11, 2864–2868 (2017).
62. Parks, D. H. et al. GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. Nucleic Acids Res. gkaf1040 (2025) doi:10.1093/nar/gkaf1040.
63. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927 (2020).
64. Martino, C. et al. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. mSystems 4, 10.1128/msystems.00016-19 (2019).
65. Aton M., McDonald D., Cañardo Alastuey J., Azom R., Batra P., Bezshapkin V., Bolyen E., Cagle A., Caporaso J.G., Debelius J.W., Gorlick K., Hamsanipally N., Hunger L., Keluskar A., Liao D., Lu Y.Y., Navas-Molina J.A., Pitman A., Rideout J., Sazonov A., Sathappan B., Schwarzberg Lipson K., Sfiligoi I., Tapo C., Vázquez-Baeza Y., Wu Z., Xu Z.Z., Ye M.S., Zhao J., Knight R., Morton J.T., Zhu Q & Cañardo Alastuey J. Scikit-bio: a fundamental Python library for biological omic data analysis. Nat. Methods In press,.
66. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the

phylogenetic tree display and annotation tool. Nucleic Acids Res. 52, W78–W82 (2024).
67. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272 (2020).
68. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. vegan: Community Ecology Package. (2022).
69. Karwowska, Z. et al. Effects of data transformation and model selection on feature importance in microbiome classification data. Microbiome 13, 2 (2025).
70. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132 (2016).
71. Schwengers, O. et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. Microb. Genomics 7, (2021).

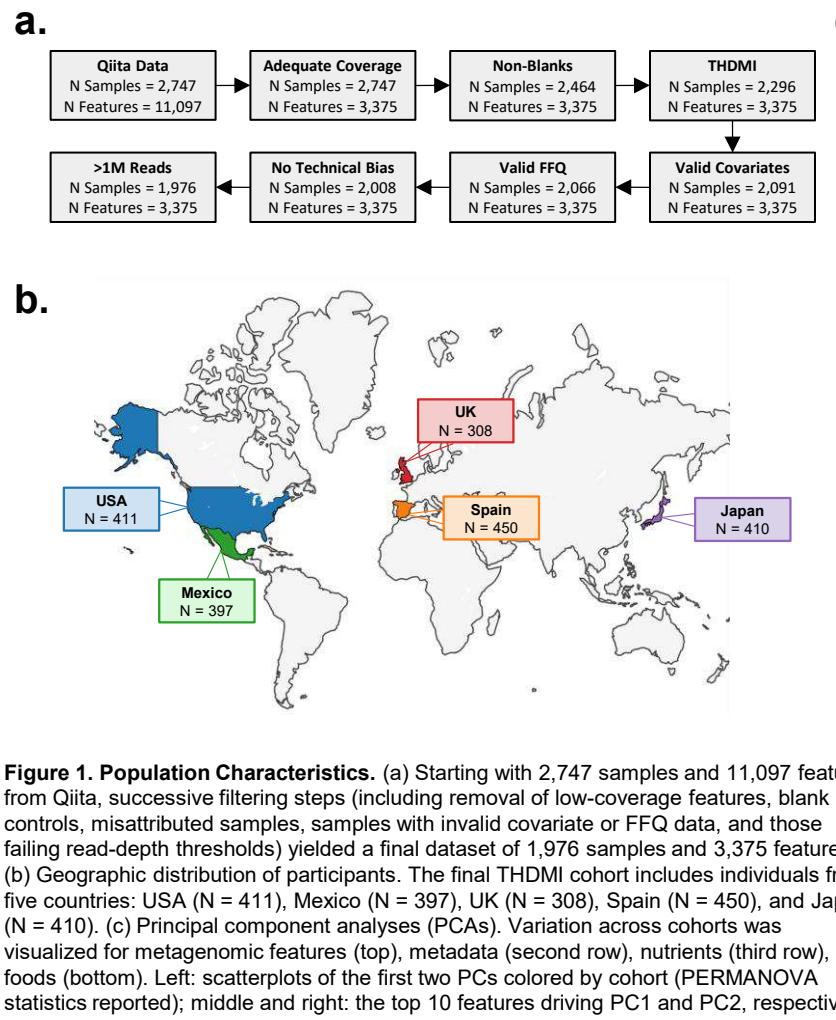


Figure 1. Population Characteristics. (a) Starting with 2,747 samples and 11,097 features from Qiita, successive filtering steps (including removal of low-coverage features, blank controls, misattributed samples, samples with invalid covariate or FFQ data, and those failing read-depth thresholds) yielded a final dataset of 1,976 samples and 3,375 features. (b) Geographic distribution of participants. The final THDMI cohort includes individuals from five countries: USA (N = 411), Mexico (N = 397), UK (N = 308), Spain (N = 450), and Japan (N = 410). (c) Principal component analyses (PCAs). Variation across cohorts was visualized for metagenomic features (top), metadata (second row), nutrients (third row), and foods (bottom). Left: scatterplots of the first two PCs colored by cohort (PERMANOVA statistics reported); middle and right: the top 10 features driving PC1 and PC2, respectively.

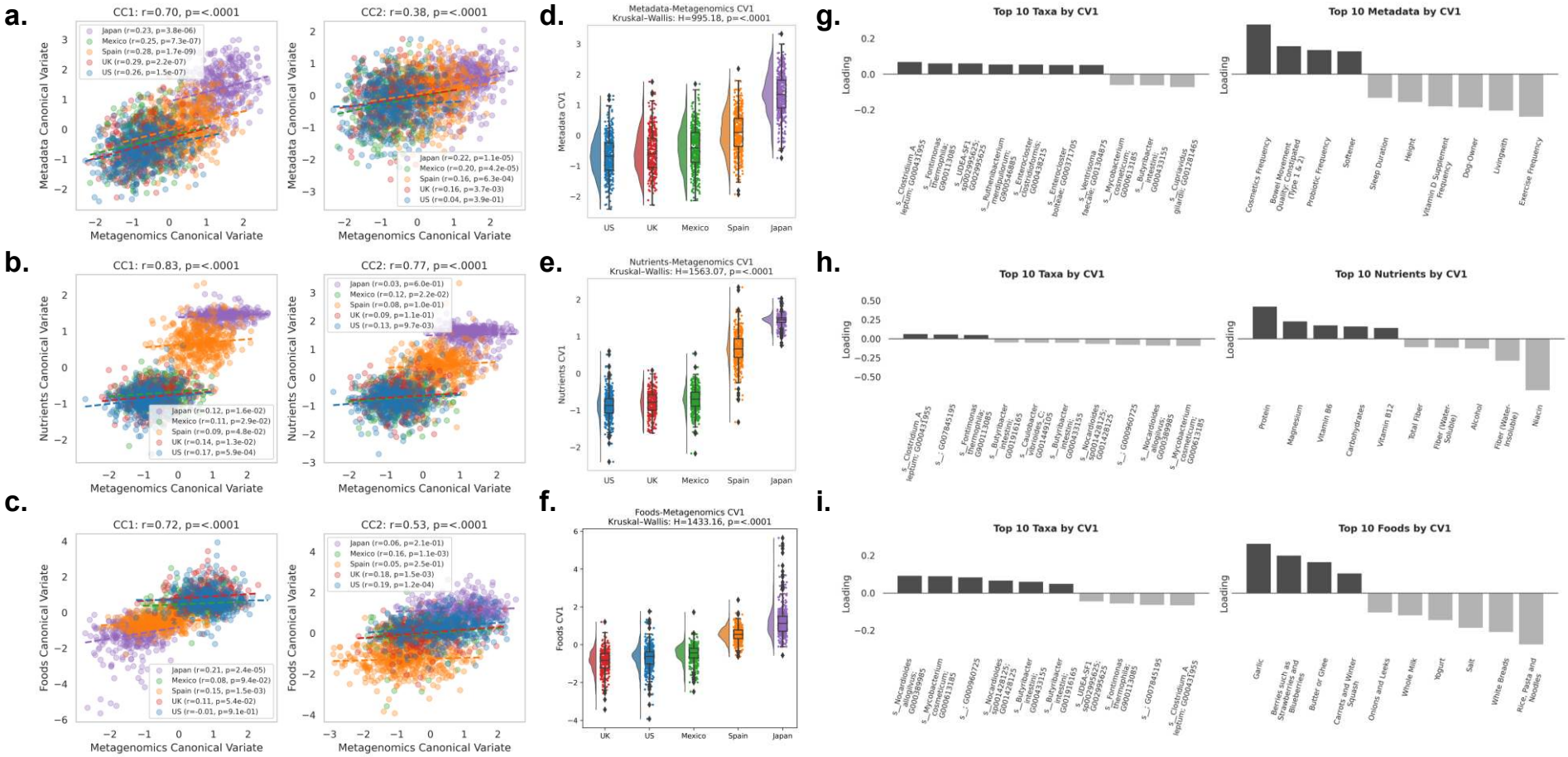


Figure 2. Sparse canonical correlation analysis (sCCA) of metagenomics with metadata, nutrients, and foods. (a–c) Scatterplots of canonical variates (CVs) from sCCA across five countries, showing correlations between metagenomics and (a) metadata, (b) nutrients, and (c) foods. Each point represents an individual, colored by country, with best-fit regression lines for each cohort. Pearson correlation coefficients (r) and p-values are shown for each country and overall. (d–f) Raincloud/violin plots of the first canonical variate (CV1) scores by country for (d) metadata, (e) nutrients, and (f) foods, illustrating cross-country distributional differences. For panel f, CV1 values were multiplied by −1 to facilitate direct visual comparison with panels d and e. Kruskal–Wallis tests confirm significant group differences ($p < 0.0001$ for all comparisons). (g–i) Barplots of the top 10 features with the largest absolute loadings on CV1 for each view. Positive and negative bars indicate direction of association. (g) Top metagenomic taxa and metadata variables, (h) top metagenomic taxa and nutrient variables, and (i) top metagenomic taxa and food variables.

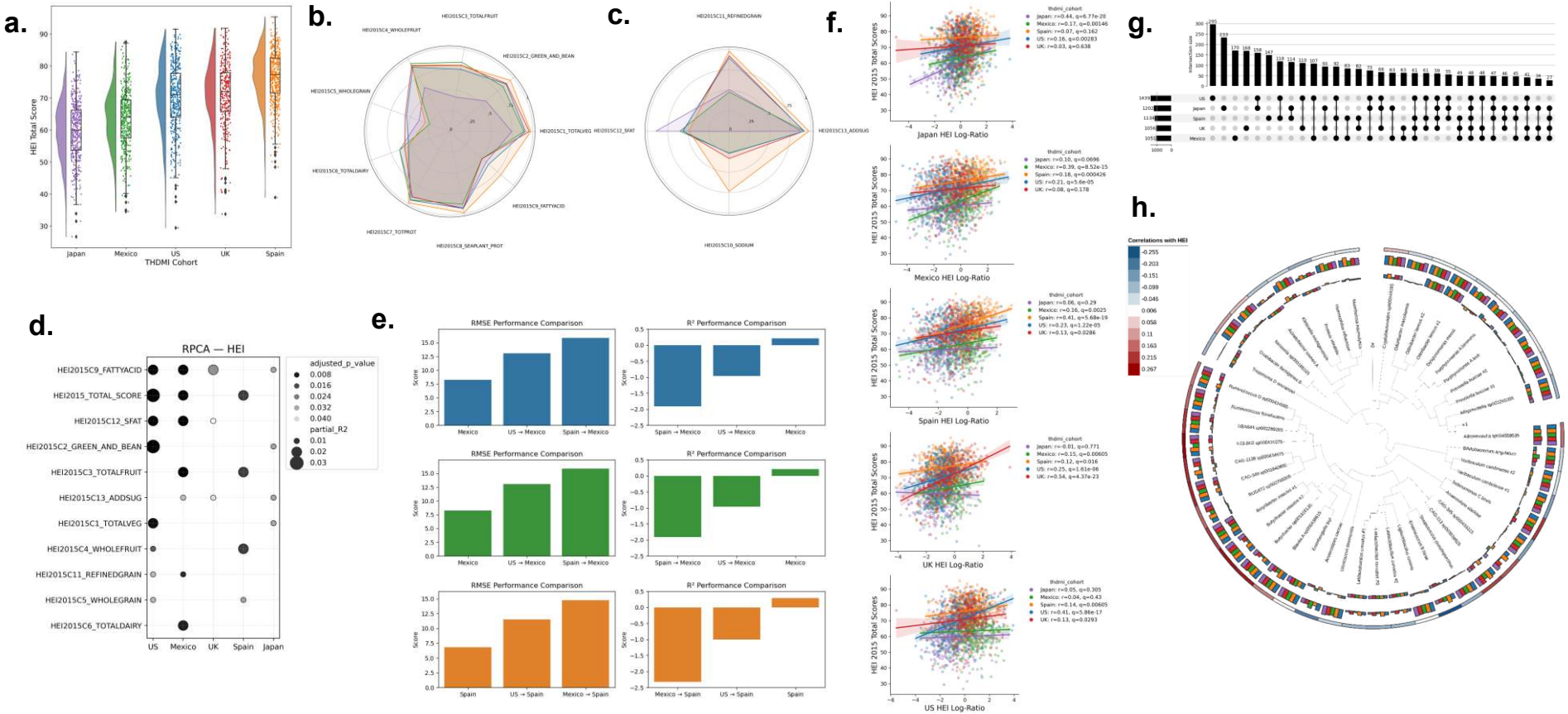


Figure 3. Healthy Eating Index (HEI) across countries. Healthy Eating Index (2015) scores were calculated for all participants. (a) Distribution of HEI total scores across the five THDMI cohorts. Boxplots show medians, interquartile ranges, and variability within each country. (b–c) Radar plots of average HEI component scores by country. (d) Robust principal component analysis (RPCA) results linking HEI components to gut microbiome variation across countries. Each circle represents a HEI component, with color denoting adjusted p-value (FDR-corrected), and circle size corresponding to the partial R^2 (effect size) of the association. (e) Bar plots showing Root Mean Square Error (RMSE) (left) and R^2 (right) from Elastic Net regressions trained on the HEI-2015 Total Score. Models were trained using 5-fold cross-validation within each country and then evaluated across countries. (f) Scatter plots show the relationship between HEI total scores and the log-ratio of positively to negatively associated taxa identified by BIRDMAn within each country. Each panel corresponds to a model trained in one country (Japan, Mexico, Spain, UK, US) and tested across all five THDMI cohorts. Lines represent fitted regression slopes with 95% confidence intervals, and points are colored by country. Spearman correlation coefficients (r) and associated q-values. (g) UpSet plot showing overlap in HEI-associated taxa across countries. (h) Phylogenetic distribution of the 48 taxa associated with HEI total scores across five countries. Inner ring: Centered log-ratio–transformed abundance of each taxon, stratified by country. Middle ring: Prevalence of each taxon by country. Outer ring: Spearman correlations between taxa and HEI total scores. Correlation strength is represented by a continuous color scale (left), with positive correlations in red and negative correlations in blue. Zero or near-zero correlations are shown in white. Branch lengths were disabled during visualization to emphasize taxon relationships rather than evolutionary distances.

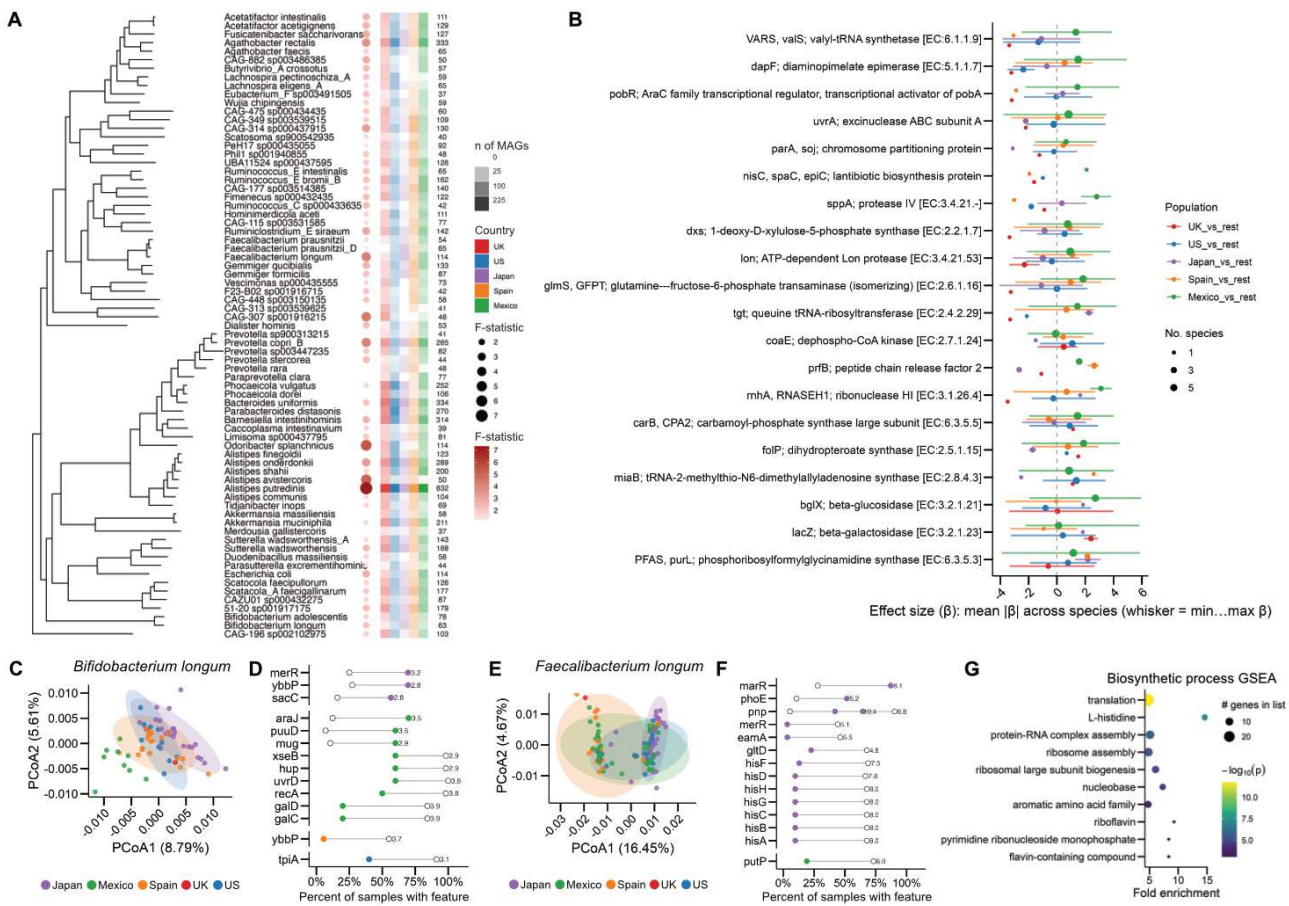


Figure 4. Country-specific genomic differentiation of microbial species. Metagenome-assembled genomes (MAGs) were quality-filtered (≥80% completeness, ≤10% contamination) and clustered at the species level. For each species, we computed all-vs-all Mash distances and tested whether genomic content correlated with country.

(A) Phylogenetic distribution of representative genomes of species with ≥30 members. The adjacent F-statistic represents the significance of the Mash distance matrix by country using PERMANOVA. The number next to the heatmap represents number of MAGs per species.
(B) Cross-population meta-analysis of functionally annotated genes significantly enriched in at least one cohort. Points represent the mean absolute effect size (β) across species for each KEGG-annotated gene; horizontal whiskers denote the range (min–max β) observed across species, and point size indicates the number of species contributing to the estimate.
(C–F) Within-species principal-coordinate analyses (PCoA) of Mash distance matrices for two representative taxa—(C) Bifidobacterium longum and (E) Faecalibacterium longum—highlighting country-specific population structure. Points are colored by population, and 95 % confidence ellipses are shown. Panels (D) and (F) display the frequency of top enriched genes within these same species, plotted as the proportion of isolates carrying each feature (dot color = country of interest against all other countries).
(G) Gene-set enrichment analysis (GSEA) summarizing biological processes over-represented among population-associated loci. Points represent Gene Ontology biological-process terms; color indicates significance ($-\log_{10}$ p), size indicates the number of contributing genes, and x-axis values correspond to fold enrichment.

Parameter	Japan	Mexico	Spain	UK	US	Statistic	Q-value
No. of participants	410 (20.7%)	397 (20.1%)	450 (22.8%)	308 (15.6%)	411 (20.8%)		
Age	47.549 ± 9.449	41.987 ± 15.198	42.567 ± 6.612	52.149 ± 13.260	45.009 ± 15.202	160.013	4.84E-3
BMI	21.731 ± 3.556	25.654 ± 5.083	23.417 ± 4.753	25.083 ± 5.698	25.232 ± 5.308	204.476	2.05E-4
Sex: Female	354 (86.3%)	269 (67.8%)	374 (83.1%)	223 (72.4%)	262 (63.7%)	83.886	4.36E-1
Unhealthy	151 (36.8%)	212 (53.4%)	251 (55.8%)	195 (63.3%)	249 (60.6%)	66.958	1.24E-1
Bowel Movement Quality: Normal	245 (61.3%)	301 (77.8%)	235 (55.6%)	219 (75.3%)	300 (76.5%)	74.516	3.60E-1
Antibiotic History: Within the Last Month	38 (13.1%)	47 (19.0%)	38 (12.8%)	11 (4.3%)	11 (3.8%)	34.445	6.04E-0
Types of Plants: More than 30	32 (14.8%)	17 (7.2%)	19 (8.8%)	53 (55.2%)	57 (32.0%)	61.227	1.78E-1
Types of Plants: Less than 10	184 (85.2%)	220 (92.8%)	197 (91.2%)	43 (44.8%)	121 (68.0%)	152.538	1.45E-3
Cosmetics Frequency: Daily	308 (75.1%)	50 (12.6%)	70 (15.6%)	61 (19.8%)	61 (14.8%)	580.531	2.53E-12
Exercise Frequency: Daily	33 (8.0%)	35 (8.8%)	40 (8.9%)	86 (27.9%)	79 (19.2%)	90.519	2.04E-1

Table 1. Demographics of participants

Parameter	Japan	Mexico	Spain	UK	US	Statistic	Q-value
Vitamin C	0.068 ± 0.013	0.098 ± 0.066	0.071 ± 0.029	0.095 ± 0.051	0.082 ± 0.049	78.419	3.77E-16
Alcohol	0.003 ± 0.005	0.003 ± 0.005	0.002 ± 0.002	0.006 ± 0.008	0.005 ± 0.007	91.891	5.40E-19
Calories	1757.161 ± 166.074	2289.243 ± 898.876	2075.074 ± 796.898	2011.354 ± 690.252	2018.798 ± 678.109	100.87	6.88E-21
Vitamin D	0.005 ± 0.001	0.004 ± 0.002	0.005 ± 0.003	0.004 ± 0.002	0.003 ± 0.002	275.461	2.36E-58
Protein	0.040 ± 0.001	0.042 ± 0.008	0.049 ± 0.011	0.038 ± 0.008	0.040 ± 0.009	318.149	1.52E-67
Cholestrol	0.165 ± 0.010	0.163 ± 0.081	0.195 ± 0.118	0.128 ± 0.087	0.131 ± 0.095	319.537	7.92E-68
Potassium	1.581 ± 0.160	1.696 ± 0.471	2.015 ± 0.402	1.898 ± 0.457	1.666 ± 0.405	351.861	8.69E-75
Folate	0.223 ± 0.030	0.202 ± 0.057	0.169 ± 0.049	0.224 ± 0.067	0.219 ± 0.064	356.91	7.37E-76
Saturated Fat	0.010 ± 0.001	0.012 ± 0.003	0.014 ± 0.003	0.013 ± 0.005	0.013 ± 0.004	390.409	4.48E-83
Riboflavin	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	395.794	3.22E-84
Phosphorus	0.609 ± 0.042	0.763 ± 0.121	0.676 ± 0.117	0.697 ± 0.115	0.693 ± 0.125	426.166	9.25E-91
Polyunsaturated Fat	0.007 ± 0.000	0.008 ± 0.002	0.010 ± 0.003	0.009 ± 0.002	0.009 ± 0.002	452.375	2.10E-96
Zinc	0.005 ± 0.000	0.006 ± 0.001	0.006 ± 0.001	0.005 ± 0.001	0.006 ± 0.001	498.137	2.82E-106
Thiamin	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	583.686	9.28E-125
Fiber (Water-Insolubl	0.006 ± 0.001	0.010 ± 0.003	0.010 ± 0.004	0.011 ± 0.004	0.011 ± 0.004	621.725	5.77E-133
Calcium	0.334 ± 0.052	0.557 ± 0.158	0.492 ± 0.150	0.565 ± 0.165	0.545 ± 0.195	683.741	2.31E-146
Lycopene	1.033 ± 0.255	3.586 ± 1.832	1.712 ± 1.223	2.736 ± 1.716	2.634 ± 1.836	720.076	3.36E-154
Selenium	0.042 ± 0.001	0.059 ± 0.012	0.044 ± 0.012	0.055 ± 0.012	0.057 ± 0.013	720.917	2.38E-154
Total Fiber	0.008 ± 0.001	0.014 ± 0.005	0.014 ± 0.005	0.016 ± 0.006	0.015 ± 0.005	721.97	1.52E-154
Magnesium	0.164 ± 0.014	0.190 ± 0.046	0.267 ± 0.067	0.208 ± 0.046	0.198 ± 0.045	740.732	1.44E-158
Total Fat	0.032 ± 0.002	0.039 ± 0.007	0.048 ± 0.009	0.040 ± 0.010	0.042 ± 0.009	791.702	1.44E-169
Vitamin B12	0.004 ± 0.000	0.002 ± 0.001	0.002 ± 0.001	0.002 ± 0.001	0.002 ± 0.001	911.163	2.12E-195
Sodium	2.160 ± 0.083	1.881 ± 0.499	1.295 ± 0.368	1.818 ± 0.581	1.914 ± 0.602	911.19	2.12E-195
Fiber (Water-Soluble	0.002 ± 0.000	0.004 ± 0.001	0.004 ± 0.001	0.005 ± 0.002	0.004 ± 0.002	945.569	9.55E-203
Carbohydrates	0.143 ± 0.004	0.122 ± 0.020	0.090 ± 0.022	0.119 ± 0.026	0.114 ± 0.025	953.314	2.34E-204
Vitamin B6	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	973.57	1.14E-208
Monounsaturated Fa	0.012 ± 0.000	0.014 ± 0.004	0.024 ± 0.006	0.015 ± 0.005	0.016 ± 0.005	984.099	7.47E-211
Iron	0.005 ± 0.001	0.007 ± 0.001	0.009 ± 0.002	0.007 ± 0.002	0.007 ± 0.002	1016.046	1.19E-217
Niacin	0.010 ± 0.001	0.019 ± 0.004	0.015 ± 0.003	0.018 ± 0.003	0.020 ± 0.005	1167.736	2.36E-250
Isoflavones	0.013 ± 0.009	0.001 ± 0.002	0.000 ± 0.000	0.004 ± 0.006	0.005 ± 0.009	1458.249	<1.00E-300

Table S3. Nutrient profiles per country, standardized by total calorie intake

Table S4. Food profiles per country, standardized by total calorie intake

Parameter	Japan	Mexico	Spain	UK	US	Statistic	Q-value
Ketchup	0.000 ± 0.001	0.001 ± 0.003	0.000 ± 0.001	0.001 ± 0.003	0.001 ± 0.004	20.244	0.000447
100% Fruit Juice such as Apple, Grape and Cran...	0.003 ± 0.016	0.001 ± 0.007	0.002 ± 0.008	0.004 ± 0.019	0.005 ± 0.021	25.04	5.02E-05
Corn, Soybean, Peanut, Sesame and Sunflower Oil	0.001 ± 0.002	0.001 ± 0.002	0.000 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	26.626	2.45E-05
Margarine	0.001 ± 0.003	0.000 ± 0.001	0.000 ± 0.001	0.001 ± 0.002	0.000 ± 0.001	36.031	3.00E-07
Eggs	0.021 ± 0.039	0.021 ± 0.022	0.015 ± 0.012	0.015 ± 0.019	0.016 ± 0.023	37.249	1.72E-07
Jam, Jelly, Honey, Syrup or Sugar	0.002 ± 0.004	0.003 ± 0.006	0.002 ± 0.004	0.002 ± 0.004	0.002 ± 0.004	41.37	2.46E-08
Oranges, Grapefruit and Tangerines (Not Juice)	0.015 ± 0.026	0.015 ± 0.030	0.020 ± 0.030	0.020 ± 0.035	0.015 ± 0.026	46.12	2.58E-09
Potato, Tortilla Chips, Corn Chips and Puffs	0.004 ± 0.006	0.002 ± 0.005	0.002 ± 0.003	0.003 ± 0.005	0.004 ± 0.006	46.65	2.04E-09
Beer	0.043 ± 0.100	0.034 ± 0.083	0.026 ± 0.040	0.025 ± 0.071	0.037 ± 0.080	48.015	1.08E-09
Bacon and Breakfast Sausage	0.001 ± 0.002	0.001 ± 0.001	0.001 ± 0.002	0.001 ± 0.003	0.001 ± 0.002	63.62	5.91E-13
Grapes	0.003 ± 0.004	0.005 ± 0.008	0.010 ± 0.015	0.008 ± 0.015	0.005 ± 0.009	76.163	1.36E-15
Liquor and Mixed Drinks	0.006 ± 0.019	0.002 ± 0.007	0.000 ± 0.001	0.003 ± 0.009	0.004 ± 0.010	97.155	4.86E-20
Coffee	0.096 ± 0.117	0.091 ± 0.160	0.029 ± 0.031	0.119 ± 0.196	0.128 ± 0.184	97.42	4.35E-20
Cookies and Cakes	0.003 ± 0.004	0.004 ± 0.005	0.002 ± 0.005	0.005 ± 0.010	0.004 ± 0.007	99.657	1.49E-20
Low and Nonfat Milk	0.017 ± 0.082	0.026 ± 0.068	0.049 ± 0.078	0.031 ± 0.055	0.021 ± 0.061	100.939	8.10E-21
100% Orange Juice and Grapefruit Juice	0.004 ± 0.011	0.009 ± 0.026	0.006 ± 0.017	0.008 ± 0.023	0.004 ± 0.017	106.754	4.78E-22
Apples, Applesauce and Pears	0.011 ± 0.016	0.023 ± 0.028	0.027 ± 0.034	0.034 ± 0.039	0.023 ± 0.029	119.045	1.17E-24
Organ Meats	0.001 ± 0.001	0.000 ± 0.001	0.001 ± 0.005	0.000 ± 0.002	0.000 ± 0.001	125.331	5.41E-26
Onions and Leeks	0.019 ± 0.020	0.012 ± 0.015	0.022 ± 0.021	0.022 ± 0.022	0.016 ± 0.019	127.109	2.31E-26
Bananas	0.009 ± 0.015	0.018 ± 0.019	0.019 ± 0.019	0.021 ± 0.023	0.019 ± 0.025	127.654	1.81E-26
Ice Cream and Milkshakes	0.009 ± 0.012	0.004 ± 0.009	0.005 ± 0.016	0.005 ± 0.010	0.009 ± 0.016	129.092	9.14E-27
Canned Fish and Shellfish	0.003 ± 0.004	0.006 ± 0.010	0.006 ± 0.007	0.004 ± 0.006	0.005 ± 0.008	134.626	6.14E-28
Sweets and Confectionery	0.005 ± 0.007	0.002 ± 0.004	0.004 ± 0.006	0.005 ± 0.006	0.003 ± 0.005	142.877	1.08E-29
Carrots and Winter Squash	0.011 ± 0.013	0.012 ± 0.015	0.015 ± 0.015	0.026 ± 0.027	0.019 ± 0.025	172.248	5.58E-36
Mayonnaise	0.001 ± 0.003	0.001 ± 0.002	0.000 ± 0.001	0.001 ± 0.002	0.001 ± 0.002	179.048	1.99E-37
Red Wine	0.006 ± 0.025	0.011 ± 0.026	0.004 ± 0.008	0.025 ± 0.045	0.014 ± 0.037	186.072	6.34E-39
Lunch Meats	0.001 ± 0.002	0.005 ± 0.007	0.003 ± 0.004	0.002 ± 0.004	0.003 ± 0.007	189.404	1.26E-39
White, Rosé, Muscatel and Rice Wine	0.002 ± 0.009	0.002 ± 0.006	0.003 ± 0.007	0.014 ± 0.030	0.009 ± 0.026	197.939	1.89E-41
Leafy Greens	0.006 ± 0.006	0.015 ± 0.026	0.019 ± 0.019	0.018 ± 0.021	0.021 ± 0.024	217.557	1.18E-45
Regular Soft Drinks (Not Diet)	0.042 ± 0.114	0.021 ± 0.055	0.005 ± 0.027	0.002 ± 0.010	0.007 ± 0.049	222.501	1.05E-46
Commercial Fruit Juices	0.020 ± 0.038	0.000 ± 0.003	0.003 ± 0.021	0.001 ± 0.007	0.004 ± 0.022	239.099	2.90E-50
Berries such as Strawberries and Blueberries	0.004 ± 0.006	0.011 ± 0.018	0.004 ± 0.004	0.028 ± 0.038	0.026 ± 0.039	244.292	2.28E-51
Cheese	0.005 ± 0.011	0.008 ± 0.009	0.016 ± 0.013	0.011 ± 0.012	0.012 ± 0.014	248.566	2.84E-52
Shellfish	0.002 ± 0.002	0.002 ± 0.003	0.004 ± 0.003	0.001 ± 0.003	0.001 ± 0.003	265.597	6.30E-56
Mushrooms	0.006 ± 0.006	0.004 ± 0.006	0.014 ± 0.017	0.008 ± 0.010	0.005 ± 0.009	277.964	1.41E-58
Butter or Ghee	0.001 ± 0.004	0.001 ± 0.003	0.001 ± 0.001	0.003 ± 0.004	0.003 ± 0.003	290.597	2.77E-61
Muffins, Scones, Croissants and Biscuits	0.004 ± 0.006	0.007 ± 0.011	0.007 ± 0.010	0.004 ± 0.008	0.002 ± 0.004	290.95	2.43E-61
Yogurt	0.099 ± 0.156	0.018 ± 0.026	0.027 ± 0.032	0.030 ± 0.034	0.019 ± 0.029	294.359	4.66E-62
Cruciferous Vegetables	0.025 ± 0.025	0.010 ± 0.017	0.017 ± 0.022	0.028 ± 0.028	0.022 ± 0.029	295.242	3.14E-62
Watermelon and Red Melon	0.002 ± 0.005	0.005 ± 0.013	0.009 ± 0.014	0.001 ± 0.004	0.004 ± 0.014	321.672	6.53E-68
Cherries, Picotas, Plums, Peaches, Nectarines,...	0.001 ± 0.002	0.005 ± 0.016	0.015 ± 0.018	0.004 ± 0.011	0.005 ± 0.013	328.598	2.19E-69
White Fish	0.007 ± 0.012	0.005 ± 0.010	0.012 ± 0.019	0.005 ± 0.008	0.003 ± 0.008	335.745	6.62E-71
Whole Milk	0.074 ± 0.166	0.027 ± 0.063	0.017 ± 0.048	0.012 ± 0.034	0.008 ± 0.031	346.447	3.42E-73
Summer Squash, Zucchini, Eggplants, Cucumbers	0.006 ± 0.009	0.007 ± 0.014	0.017 ± 0.017	0.007 ± 0.012	0.007 ± 0.012	349.74	7.04E-74
Potatoes	0.005 ± 0.005	0.008 ± 0.010	0.019 ± 0.016	0.024 ± 0.026	0.012 ± 0.023	364.756	4.28E-77
Dark Fish	0.008 ± 0.008	0.003 ± 0.007	0.011 ± 0.010	0.008 ± 0.010	0.005 ± 0.008	374.808	3.08E-79
Peppers Cooked and Raw	0.002 ± 0.002	0.011 ± 0.019	0.015 ± 0.015	0.019 ± 0.022	0.013 ± 0.021	384.82	2.27E-81
Tea	0.085 ± 0.116	0.075 ± 0.166	0.015 ± 0.026	0.309 ± 0.287	0.113 ± 0.183	391.844	7.42E-83
Processed Meats	0.003 ± 0.009	0.001 ± 0.002	0.002 ± 0.003	0.000 ± 0.002	0.001 ± 0.004	422.577	1.84E-89
Chicken and Turkey	0.011 ± 0.011	0.011 ± 0.012	0.033 ± 0.027	0.008 ± 0.010	0.011 ± 0.014	423.204	1.47E-89
Garlic	0.000 ± 0.000	0.001 ± 0.001	0.000 ± 0.000	0.001 ± 0.001	0.001 ± 0.001	444.895	3.31E-94
Fresh Tomatoes	0.007 ± 0.011	0.029 ± 0.035	0.038 ± 0.032	0.036 ± 0.040	0.018 ± 0.022	467.956	3.80E-99
Beef, Pork, Ham and Lamb	0.025 ± 0.045	0.015 ± 0.015	0.022 ± 0.017	0.007 ± 0.011	0.007 ± 0.012	519.566	2.94E-110
Green or String Beans	0.000 ± 0.001	0.003 ± 0.006	0.015 ± 0.015	0.007 ± 0.012	0.004 ± 0.008	560.016	5.97E-119
White Breads	0.030 ± 0.049	0.004 ± 0.007	0.009 ± 0.014	0.005 ± 0.010	0.004 ± 0.010	565.587	4.34E-120
Peanut Butter, Peanuts and Other Nuts and Seeds	0.001 ± 0.005	0.001 ± 0.003	0.013 ± 0.012	0.007 ± 0.011	0.009 ± 0.010	738.181	2.26E-157
Olive Oil	0.001 ± 0.002	0.002 ± 0.004	0.012 ± 0.008	0.004 ± 0.006	0.004 ± 0.007	824.332	6.18E-176
Soups and Creams	0.067 ± 0.060	0.045 ± 0.040	0.002 ± 0.004	0.029 ± 0.039	0.034 ± 0.045	861.841	6.17E-184
Salt	0.002 ± 0.001	0.000 ± 0.001	0.001 ± 0.001	0.000 ± 0.001	0.000 ± 0.001	885.491	6.95E-189
Rice, Pasta and Noodles	0.053 ± 0.044	0.000 ± 0.001	0.015 ± 0.013	0.015 ± 0.014	0.019 ± 0.021	1006.472	8.47E-215

Enriched Taxa	Coverage	Japan CLR	Japan RA	Japan Prev	Mexico CLR	Mexico RA	Mexico Prev	Spain CLR	Spain RA	Spain Prev	UK CLR	UK RA	UK Prev	US CLR	US RA	US Prev
s__Bacillus subtilis;G000747645	99.731573	1.457557	1.42E-04	0.658537	-2.370008	1.86E-08	0.09068	-2.28096	2.35E-08	0.064444	-2.006602	1.06E-07	0.068182	-1.955243	8.27E-08	0.082725
s__Bacillus subtilis;G000696615	85.18573	1.129196	8.32E-05	0.631707	-2.369036	1.92E-08	0.088161	-2.288121	2.02E-08	0.057778	-2.001871	8.40E-08	0.074675	-1.959881	5.80E-08	0.080292
s__Bacillus subtilis;G900110305	83.013829	1.114895	8.23E-05	0.631707	-2.368315	1.85E-08	0.09068	-2.287681	2.01E-08	0.06	-2.004564	8.04E-08	0.068182	-1.955957	5.63E-08	0.075426
s__Bacillus spizizenii;G000740655	42.959272	0.11002	1.39E-05	0.519512	-2.42405	5.77E-10	0.005038	-2.330059	3.66E-09	0.02	-2.051267	1.25E-08	0.029221	-1.945883	4.21E-08	0.026764
s__Bacillus cabrialesii;G001663575	38.641209	0.049235	1.27E-05	0.490244	-2.418166	2.12E-09	0.010076	-2.325665	3.83E-09	0.013333	-2.061579	9.98E-09	0.016234	-1.945813	4.07E-08	0.026764
s__Bacillus velezensis;G001687195	4.584745	-1.040785	1.30E-06	0.358537	-2.421441	1.55E-09	0.010076	-2.328664	7.53E-09	0.022222	-2.085212	3.25E-09	0.022727	-1.979354	8.08E-09	0.029197
s__Bacillus amyloliquefaciens;G000196735	3.999323	-1.331898	5.17E-07	0.326829	-2.420454	2.91E-09	0.025189	-2.317726	2.96E-08	0.051111	-2.095376	1.43E-09	0.019481	-2.001547	3.36E-09	0.034063
s__Bacillus velezensis;G001461825	4.433534	-1.378359	4.64E-07	0.3	-2.423788	1.47E-09	0.012594	-2.337965	2.47E-09	0.02	-2.087991	2.32E-09	0.016234	-1.975097	8.47E-09	0.026764
s__Bacillus velezensis;G001517185	4.309493	-1.498164	3.18E-07	0.273171	-2.424433	5.23E-10	0.005038	-2.329759	4.58E-09	0.017778	-2.084802	2.44E-09	0.012987	-1.978092	7.26E-09	0.021898
s__Bacillus velezensis;G001617995	4.458047	-1.474869	3.38E-07	0.273171	-2.425044	2.41E-10	0.002519	-2.33583	2.54E-09	0.015556	-2.087015	1.35E-09	0.006494	-1.978996	7.57E-09	0.024331
s__Bacillus velezensis;G001587325	4.137925	-1.512635	2.99E-07	0.268293	-2.4239	5.89E-10	0.005038	-2.337081	2.93E-09	0.02	-2.087015	1.35E-09	0.006494	-1.982304	7.37E-09	0.026764
s__Bacillus velezensis;G900156245	4.307204	-1.545037	2.61E-07	0.258537	-2.425227	4.47E-10	0.005038	-2.335157	2.17E-09	0.013333	-2.083942	1.73E-09	0.006494	-1.978379	7.29E-09	0.021898
s__Bacillus velezensis;G000740715	4.326891	-1.518239	3.03E-07	0.253659	-2.426868	3.48E-10	0.005038	-2.333979	2.30E-09	0.013333	-2.086462	1.11E-09	0.003247	-1.975732	7.72E-09	0.024331
s__Bacillus velezensis;G001440465	4.438185	-1.61769	1.93E-07	0.246341	-2.425204	6.70E-10	0.007557	-2.335105	1.98E-09	0.011111	-2.085528	1.48E-09	0.003247	-1.967412	1.28E-08	0.024331
s__Bacillus velezensis;G001975995	4.018299	-1.608399	1.91E-07	0.246341	-2.424097	7.94E-10	0.007557	-2.335233	2.85E-09	0.013333	-2.086462	1.11E-09	0.003247	-1.975189	8.02E-09	0.024331
s__Bacillus velezensis;G000746035	4.311094	-1.579116	2.36E-07	0.243902	-2.425044	2.41E-10	0.002519	-2.334641	2.81E-09	0.017778	-2.087015	1.35E-09	0.006494	-1.980163	7.75E-09	0.026764
s__Bacillus glycinifermentans;G900093775	1.270987	-1.628991	2.07E-07	0.234146	-2.399334	1.05E-08	0.032746	-2.338327	2.47E-09	0.024444	-2.083755	3.44E-09	0.022727	-1.997678	1.58E-09	0.012165
s__Bacillus pumilus;G001578205	1.122992	-1.797288	9.52E-08	0.182927	-2.423633	8.24E-10	0.007557	-2.341384	1.36E-09	0.017778	-2.087898	1.53E-09	0.012987	-1.99777	1.52E-09	0.007299
s__Novosphingobium nitrogenifigens;G000192575	1.863711	-1.701684	1.18E-07	0.204878	-2.378013	1.92E-08	0.052897	-2.304233	1.77E-08	0.055556	-2.064347	1.51E-08	0.022727	-1.955102	3.33E-08	0.048662
s__Agrobacterium arsenijevicii;G000949895	12.925085	-1.958698	3.33E-08	0.056098	-1.468126	5.56E-07	0.518892	-2.25924	3.78E-08	0.104444	-2.081619	4.75E-09	0.032468	-1.964911	3.73E-08	0.068127
s__Limnospira fusiformis;G000307915	76.347293	-1.926659	1.22E-07	0.05122	-1.223752	1.44E-06	0.556675	-2.114148	1.43E-07	0.173333	-1.877389	3.87E-06	0.097403	-1.85517	1.51E-07	0.087591
s__Enterococcus_B faecalis_A;G000415185	1.514966	-1.918499	6.15E-08	0.065854	-1.928837	3.34E-07	0.299748	-2.283472	2.24E-08	0.057778	-2.048101	1.81E-08	0.032468	-1.969367	1.88E-08	0.034063
s__Enterococcus_J nangangensis;G005405245	1.054169	-1.968118	2.28E-08	0.058537	-2.097068	1.23E-07	0.304786	-2.308235	2.11E-08	0.08	-2.084745	8.46E-09	0.055195	-1.980748	8.87E-09	0.051095
s__Enterococcus_A hermanniensis;G001885945	2.964787	-1.850196	1.46E-07	0.102439	-1.933269	3.49E-07	0.312343	-2.293712	2.42E-08	0.066667	-2.047152	2.95E-08	0.051948	-1.963804	1.81E-08	0.038929
s__Atopococcus tabaci;G000429585	1.261123	-1.990977	1.59E-08	0.031707	-2.156448	1.01E-07	0.2267	-2.313896	2.72E-08	0.042222	-2.049602	1.69E-08	0.038961	-1.96078	2.36E-08	0.051095
s__Lacticaseibacillus pantheris;G001311175	69.942188	-1.996443	1.18E-08	0.043902	-2.142944	1.30E-07	0.229219	-2.305214	1.34E-08	0.068889	-2.088245	2.54E-09	0.022727	-1.964232	4.40E-06	0.051095
s__Lacticaseibacillus pantheris;G001435155	69.860403	-2.003085	9.70E-09	0.036585	-2.139303	1.31E-07	0.231738	-2.300313	1.46E-08	0.075556	-2.087243	3.17E-09	0.025974	-1.965085	4.45E-06	0.048662
s__Weissella minor;G001437425	1.244814	-2.031405	8.97E-10	0.007317	-2.275902	8.15E-08	0.120907	-2.310899	9.11E-09	0.031111	-2.079587	7.39E-09	0.006494	-1.999214	8.82E-10	0.004866
s__Companilactobacillus farciminis;G000184535	2.790678	-2.022132	3.66E-09	0.012195	-2.301548	4.28E-08	0.088161	-2.338533	4.71E-10	0.006667	-2.087868	2.81E-09	0.012987	-1.992698	1.02E-08	0.024331
s__Ochrobactrum anthropi;G001586735	6.195265	-2.016687	4.28E-09	0.019512	-2.373645	1.54E-08	0.075567	-2.329868	5.75E-09	0.022222	-2.091058	2.70E-10	0.003247	-2.000779	2.38E-09	0.019465
s__UBA1436 sp902794895;G902794895	1.024035	-2.020284	4.80E-08	0.002439	-2.165651	2.89E-06	0.060453	-2.271193	4.21E-07	0.013333	-2.069504	2.11E-08	0.00974	-1.990634	2.16E-08	0.002433
s__Helicobacter_D apodemus;G000765745	1.00366	-2.033485	0.00E+00	0	-2.367952	2.98E-08	0.040302	-2.335872	0.00E+00	0	-2.090881	0.00E+00	0	-2.003634	1.03E-10	0.002433
s__GCF-000787395 sp000787395;G002021855	1.040128	-2.012961	9.02E-08	0.007317	-2.341437	2.88E-07	0.042821	-2.336745	1.32E-10	0.002222	-2.095577	6.71E-11	0.003247	-2.003637	3.19E-10	0.004866
s__Acinetobacter marinus;G900096915	2.236579	-2.023623	2.95E-09	0.007317	-2.36718	2.07E-08	0.047859	-2.333065	2.31E-09	0.015556	-2.080637	9.32E-08	0.012987	-2.00069	1.55E-09	0.014599
s__Salinivibrio socompensis;G000513715	1.475546	-2.011759	1.46E-07	0.009756	-2.369638	3.88E-08	0.042821	-2.333059	1.55E-09	0.008889	-2.09745	8.88E-10	0.012987	-1.995416	6.96E-09	0.004866
s__Aliiglaciecola lipolytica;G000314975	1.195774	-2.009596	3.53E-07	0.004878	-2.409856	1.04E-08	0.025189	-2.325119	1.26E-08	0.004444	-2.093326	1.34E-10	0.003247	-1.998426	1.52E-09	0.007299
s__GCF-000787395 sp000787395;G002021105	1.229098	-2.018581	9.65E-08	0.002439	-2.423167	1.12E-09	0.010076	-2.335205	2.64E-10	0.002222	-2.096326	5.33E-11	0.003247	-2.001871	0.00E+00	0
s__Bdellovibrio bacteriovorus_F;G000317895	3.039442	-2.033485	0.00E+00	0	-2.384087	1.18E-08	0.042821	-2.069325	2.22E-07	0.157778	-2.017145	3.57E-08	0.042208	-1.984276	6.08E-09	0.024331
s__Bdellovibrio bacteriovorus;G000196175	4.200875	-2.033485	0.00E+00	0	-2.378278	1.49E-08	0.04534	-2.035482	3.01E-07	0.162222	-2.032977	2.90E-08	0.048701	-1.98902	5.53E-09	0.024331
s__Bdellovibrio bacteriovorus;G000691605	4.005324	-2.033485	0.00E+00	0	-2.388201	1.29E-08	0.035264	-2.041044	3.01E-07	0.162222	-2.030406	3.27E-08	0.051948	-1.991322	4.95E-09	0.024331
s__Sedimenibacterium sp000515315;G000515315	1.203025	-2.036407	6.46E-11	0.002439	-2.425259	0.00E+00	0	-2.253211	4.73E-08	0.108889	-2.091183	7.40E-10	0.006494	-1.999813	1.03E-09	0.007299
s__GCA-2401445 sp002401445;G002401445	1.115239	-2.031627	4.59E-10	0.002439	-2.425667	1.88E-10	0.002519	-2.303421	1.67E-08	0.053333	-2.089093	4.94E-10	0.003247	-2.001871	0.00E+00	0
s__Methanoculleus sp002508705;G002508805	76.882064	-2.033485	0.00E+00	0	-2.425259	0.00E+00	0	-2.335872	0.00E+00	0	-2.031794	3.70E-05	0.016234	-1.999741	5.13E-10	0.002433
s__;G900660535	1.275495	-2.033485	0.00E+00	0	-2.423731	4.06E-10	0.002519	-2.335548	5.81E-10	0.004444	-2.053376	5.71E-08	0.016234	-2.002251	1.83E-10	0.002433
s__Buchnera aphidicola_A;G000225445	2.866839	-2.031127	5.63E-10	0.002439	-2.412554	3.43E-08	0.002519	-2.335872	0.00E+00	0	-2.097475	1.32E-09	0.016234	-2.007079	1.61E-10	0.004866
s__Moheibacter sediminis;G900176425	1.479833	-2.025794	3.25E-09	0.017073	-2.402706	7.52E-09	0.047859	-2.334656	4.04E-09	0.028889	-2.074404	1.34E-08	0.045455	-1.951601	3.38E-08	0.158151
s__Soonwooa buanensis;G900167905	1.510648	-1.929218	1.21E-07	0.036585	-2.406372	5.34E-09	0.027708	-2.341918	1.52E-09	0.02	-2.092835	3.94E-09	0.032468	-2.003785	2.13E-08	0.114355
s__Aliarcobacter faecis;G000585155	1.058198	-2.033485	0.00E+00	0	-2.417423	2.99E-09	0.007557	-2.338069	2.52E-10	0.004444	-2.088441	8.39E-10	0.006494	-1.973429	1.23E-08	0.024331

Table S5. Genome Coverage, Mean CLR, Mean Relative Abundance, and Prevalence of Enriched Taxa Per Country

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaltables.xlsx
- supplementaryfigures.pdf