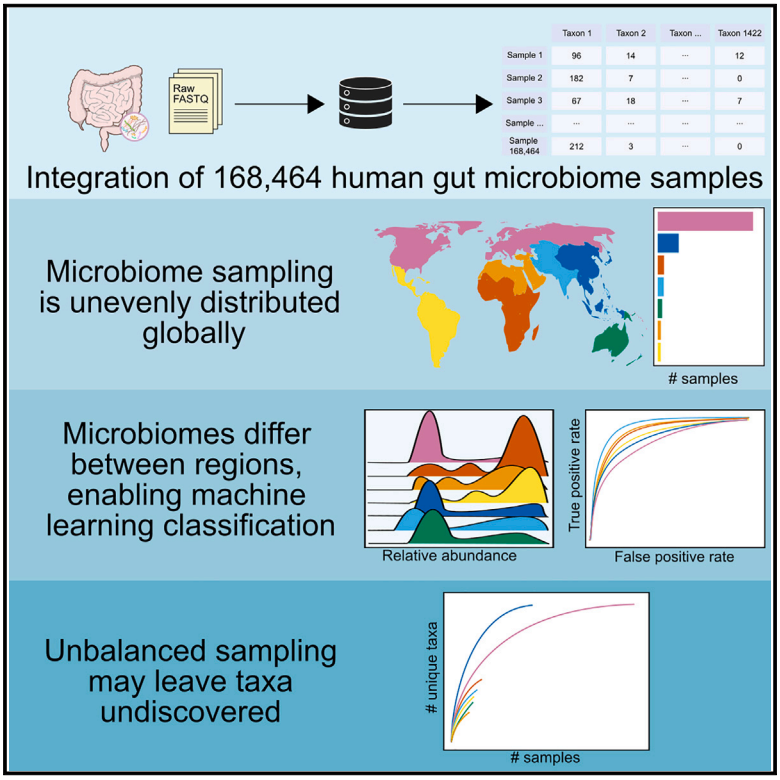
Cell

Resource

Integration of 168,000 samples reveals global patterns of the human gut microbiome

Graphical abstract



Highlights

- 168,000 public 16S gut microbiome samples processed and integrated at microbiomap.org
- Microbiome composition and diversity differ widely between world regions
- Technical factors like amplicon choice associate with compositional differences
- Classifiers trained on compendium data can infer world regions from composition

Authors

Richard J. Abdill, Samantha P. Graham, Vincent Rubinetti, ..., Casey S. Greene, Sean Davis, Ran Blekhman

Correspondence

blekhman@uchicago.edu

In brief

The Human Microbiome Compendium is a unified database of publicly available human gut microbiome 16S samples, built with the integrated data from hundreds of independent projects. The compendium is used to quantify variation in gut microbiome composition across the globe.

Abdill et al., 2025, Cell *188*, 1100–1118
February 20, 2025
https://doi.org/10.1016/j.cell.2024.12.017

CellPress

CellPress
OPEN ACCESS

Cell

Resource

Integration of 168,000 samples reveals global patterns of the human gut microbiome

Richard J. Abdill,[1,7] Samantha P. Graham,[2,7] Vincent Rubinetti,[3,4] Mansooreh Ahmadian,[5] Parker Hicks,[3] Ashwin Chetty,[1] Daniel McDonald,[6] Pamela Ferretti,[1] Elizabeth Gibbons,[1] Marco Rossi,[1] Arjun Krishnan,[3,5] Frank W. Albert,[2] Casey S. Greene,[3,4] Sean Davis,[3,4] and Ran Blekhman[1,8,*]

[1]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA
[2]Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN, USA
[3]Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA
[4]Center for Health Artificial Intelligence (CHAI), University of Colorado School of Medicine, Aurora, CO, USA
[5]Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, School of Public Health, Aurora, CO, USA
[6]Department of Pediatrics, School of Medicine, University of California, San Diego, La Jolla, CA, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: blekhman@uchicago.edu
https://doi.org/10.1016/j.cell.2024.12.017

SUMMARY

The factors shaping human microbiome variation are a major focus of biomedical research. While other fields have used large sequencing compendia to extract insights requiring otherwise impractical sample sizes, the microbiome field has lacked a comparably sized resource for the 16S rRNA gene amplicon sequencing commonly used to quantify microbiome composition. To address this gap, we processed 168,464 publicly available human gut microbiome samples with a uniform pipeline. We use this compendium to evaluate geographic and technical effects on microbiome variation. We find that regions such as Central and Southern Asia differ significantly from the more thoroughly characterized microbiomes of Europe and Northern America and that composition alone can be used to predict a sample's region of origin. We also find strong associations between microbiome variation and technical factors such as primers and DNA extraction. We anticipate this growing work, the Human Microbiome Compendium, will enable advanced applied and methodological research.

INTRODUCTION

The human microbiome is an important factor in understanding health and disease. Systematic differences have been observed between the composition of the microbiome in healthy individuals and those with microbiota-linked conditions such as colorectal cancer[1–3] and inflammatory bowel disease,[4] informing the importance of understanding the determinants of variation in the microbiome. Studies have shown this variation is driven by factors including host genetics[5] and ethnicity,[6–8] many of which are tied to geographic region.[9] For example, dietary fiber and the consumption of processed foods vary between countries,[10] as does the use of antibiotics,[11] both of which are known to impact gut microbiota. Microbiome composition links location, culture, and human health, a dynamic observed in the compositional shifts experienced by individuals immigrating to the United States from Thailand,[12,13] Latin America, and Korea.[14]

Despite the importance of understanding microbiome variation between world regions, cultures, and social groups,[15–18] many populations are practically excluded from the literature. In our previous work, we demonstrated that high-income countries, such as the United States, are dramatically overrepresented in public databases, while others, such as countries in eastern Asia, are under-sampled relative to their population.[19] As with genome-wide association studies,[20] a limited range of subjects raises the question of how broadly we can apply established links between the microbiome and human health.[21,22] The many publicly available microbiome datasets could be useful in quantifying differences between the most thoroughly studied world regions and those that are still comparatively uncharacterized.

Such analysis is complicated by the numerous technical factors implicated in alterations to observed microbiome composition, including sample collection,[23] preservation,[24] and storage,[25–27] sequencing platform,[28] and the concentration of extracted DNA.[29] Extraction kit has been a factor of particular emphasis,[30] especially the inclusion of a bead-beating step in which the enzymatic degradation of bacterial cells is supplemented or replaced by a mechanical process of agitating and breaking the cells with a set of small beads to obtain DNA. Many studies have found this mechanical lysis step can greatly increase the number of observed taxa and the relative

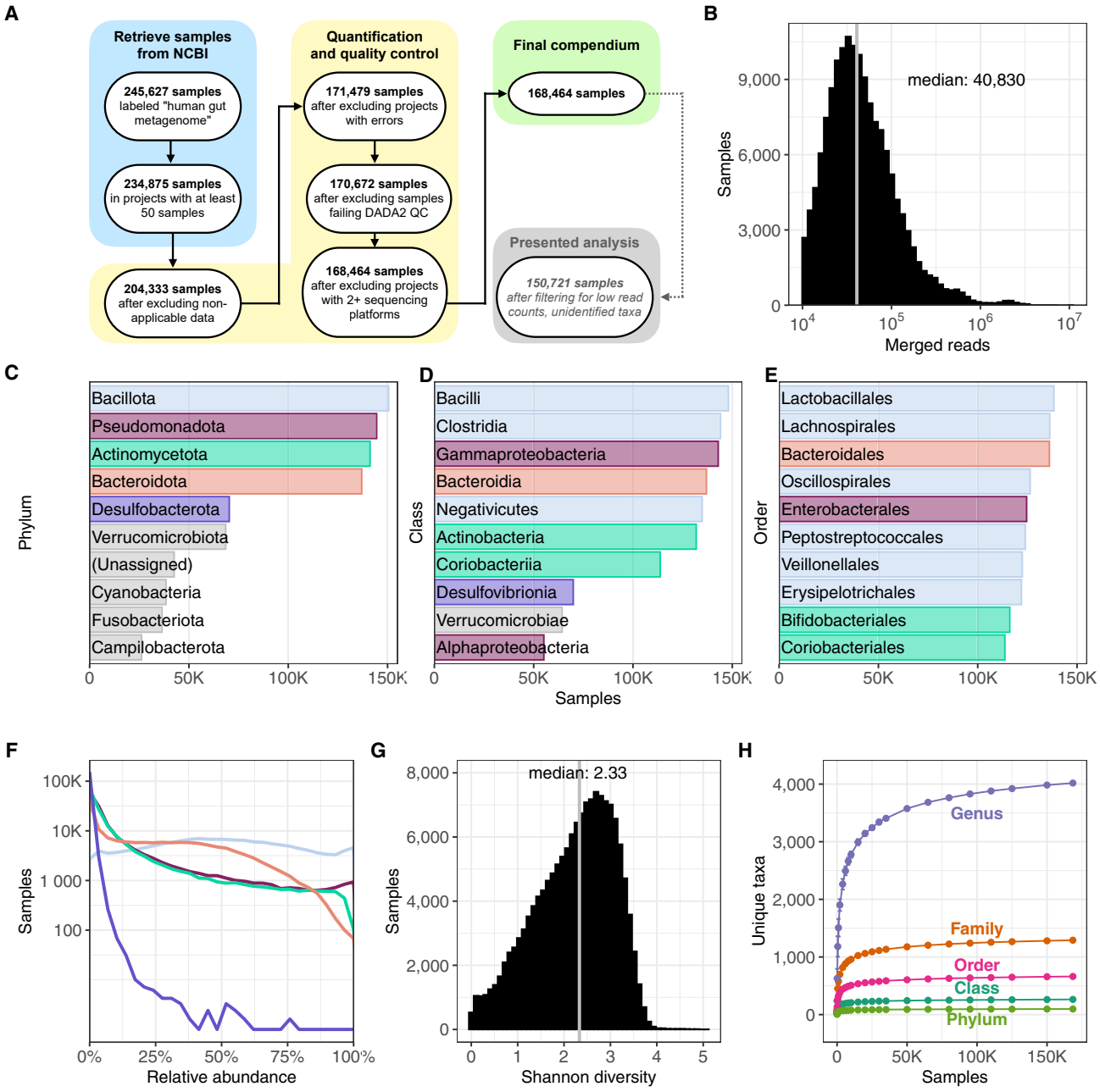


Figure 1. Overview of the Human Microbiome Compendium

(A) A list of the general steps in the data pipeline and how many samples completed each step. See STAR Methods for more details about each process.

(B) A histogram illustrating the distribution of reads that were classified in each sample. The x axis indicates the number of reads in each sample, and the y axis indicates the number of samples with that number of reads. The vertical gray line indicates the median read count.

(C–E) The most prevalent taxa observed in the compendium. The reads in each sample are assigned the most specific taxonomic name possible, down to the genus level. Each panel illustrates results when these assignments are consolidated at the three highest taxonomic levels. In each, the y axis lists the 10 most prevalent taxa at that level, and the x axis indicates the number of samples in which that taxon was observed. (C) indicates the most prevalent phyla, and the top five are each assigned a color. These colors are used in the remaining two panels to indicate the phylum of each taxon. (D) indicates the most prevalent classes of bacteria observed in the dataset, and (E) indicates the most prevalent orders. Lower taxonomic orders are illustrated in Figure S1.

(F) A density plot illustrating the relative abundance of phyla across the compendium. Each line represents one of the five most prevalent phyla in the dataset, using the same colors as (C). The gray line indicates all other phyla. The x axis indicates the relative abundance of a given phylum in a single sample, and the y axis (on a pseudo-log scale) indicates how many samples were observed to have that abundance of the given taxon.

(legend continued on next page)

abundance of gram-positive microbes,[31] including common gut-associated genera such as *Bifidobacterium*[32] and *Blautia*.[33,34] Primer choice—that is, which of the 16S hypervariable regions is amplified for sequencing—has also been linked to alpha[35] and beta diversity.[36]

In an environment as complex as the human gut, important patterns may only become apparent after collecting thousands of samples. Large compendia such as recount3[37,38] for transcriptomic analysis have revealed strain-level differences in complex microbial gene expression patterns[39] and human gene expression modules that can be used to enhance transcriptome-wide association studies.[40] To address the need for a large-scale microbiome resource, we present here the Human Microbiome Compendium, a collection of more than 168,000 publicly available human gut microbiome samples from 68 countries. All samples were homogeneously reprocessed from sequence data and combined into a single dataset available in multiple formats described below and at microbiomap.org. We demonstrate the value of the compendium by evaluating patterns in microbiome composition around the world, quantifying gaps in our current knowledge of the human gut microbiome, and identifying the impact of common technical factors on microbiome quantification.

RESULTS

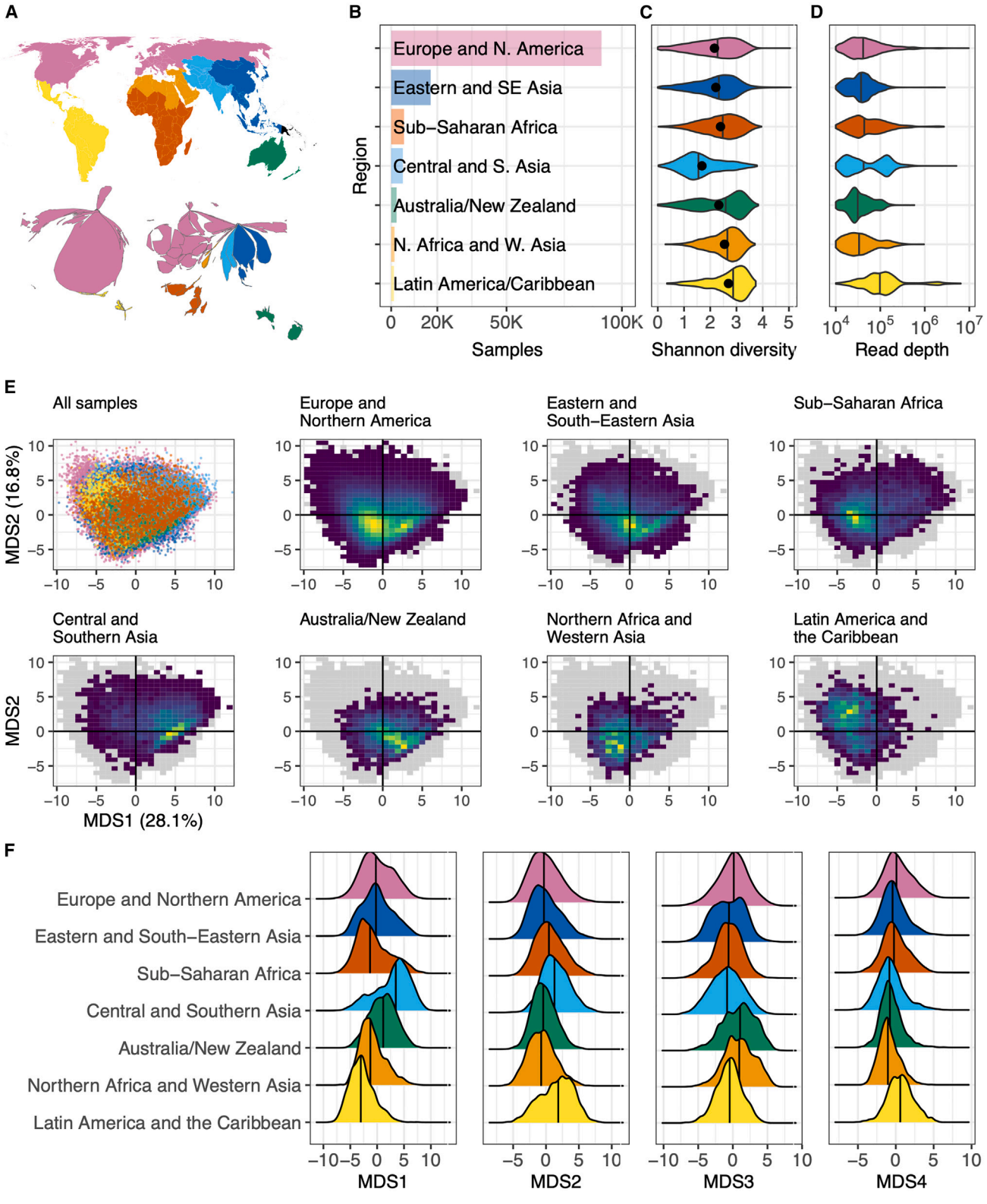
Uniform processing enables quantification of large-scale global microbiome diversity

To generate the Human Microbiome Compendium, we identified 245,627 samples of 16S rRNA gene amplicon sequencing available in the BioSample database maintained by the U.S. National Center for Biotechnology Information (NCBI). The samples are organized into studies (“BioProjects”). We focused on Illumina-based assays and discarded BioProjects reporting pyrosequencing and long-read sequencing data. Divisive Amplicon Denoising Algorithm 2 (DADA2)[41] was used to generate a taxonomic table for each BioProject in which each row is a sample and each column is a single taxon. To integrate the data across BioProjects, we processed and quantified each BioProject’s amplicon sequence variants (ASVs), each representing a distinct sequence observed in the samples in the BioProject. Then, each ASV was classified as specifically as possible down to the genus level using the same reference (in this work, SILVA v138.0; see STAR Methods for other classifications available in the dataset). The final results quantified the number of reads in each sample that were assigned to each taxon. We repeated this for the 482 BioProjects in the dataset, which resulted in a full compendium of 168,464 samples from 68 nations encompassing 5.57 terabases of sequencing data processed using a uniform pipeline (Figure 1A). See STAR Methods for a comprehensive description of the pipeline and quality control.

We then developed automated annotation processes to infer country of origin and technical factors such as extraction kit and amplicon choice. This, combined with manual curation of publication information and study population, allowed us to quantify patterns in gut microbiome composition on a global scale and will enable users to subset this uniformly processed data to more closely fit their own research questions. The processed data, the Human Microbiome Compendium, is freely available in multiple formats at microbiomap.org (see Resource availability), where users can browse, visualize, filter, and download the data and metadata. We also created an R package, MicroBioMap, to facilitate further analysis of the data. We believe this large compendium is well-suited to the expanding repertoire of machine learning tools with microbial ecology applications, following on the work of projects such as GMrepo[42] and mBodyMap[43] to provide expanded access to uniformly processed 16S human gut samples in a single taxonomic table (Table S1).

For further analyses, we created a filtered compendium of 150,721 samples containing at least 10,000 reads each after excluding rare taxa (see STAR Methods) to filter out low-quality samples with insufficient data on composition and microbes that are too rare to compare between BioProjects or world regions. The median sample contained 40,830 reads, after trimming, quality filtering, and merging of paired reads, and 9.6% of samples had more than 150,000 reads (Figure 1B). As observed even in early sequencing assays of the human gut microbiome,[44] we find the Bacillota phylum (formerly “Firmicutes”)[45] is by far the most prevalent (Figure 1C), found in 150,540 of 150,721 samples (99.9%), followed by Pseudomonadota (formerly “Proteobacteria”; 144,489 samples; 95.9%), Actinomycetota (formerly “Actinobacteria”; 93.7%), and Bacteroidota (formerly “Bacteroidetes”; 90.9%), before a sharp drop-off to phyla such as Desulfobacterota and Verrucomicrobiota. Bacillota also contains three of the five most prevalent classes (Figure 1D) and six of the 10 most prevalent orders (Figure 1E; Table S2). The prevalence of the Bacteroidota phylum, long a focus of analysis (e.g., Mariat et al.[46]) is due almost entirely to the Bacteroidales order in our data (136,085 samples; 99.3% of Bacteroidota-positive samples; Figure 1E), particularly the Bacteroidaceae family (Figure S1). Visual inspection of the phylum-level relative abundances found in the compendium shows a surprisingly uniform distribution for the abundance of Bacillota (Figure 1F). We also find the relative abundance distribution of Pseudomonadota closely resembles that of Actinomycetota, and that Desulfobacterota, despite being the fifth-most prevalent phylum, is found at relative abundances lower than 1% in 88.8% of those samples (Figure 1F).

(G) A histogram illustrating the distribution of Shannon diversity observed in the compendium. The x axis indicates a given sample’s alpha diversity, as measured by Shannon diversity index. The y axis indicates the number of samples that were observed to have that score. Please note that samples were not rarefied for this calculation, and this figure reflects diversity observed in the samples with varying quantities of processed reads.
(H) The results of a rarefaction analysis in which a simulated compendium of various sizes was generated repeatedly and evaluated for taxonomic richness. The x axis indicates the number of microbiome samples in the simulated compendium, and the y axis indicates the mean count of unique taxa were observed in repeated subsamplings. Each line indicates the number of observed taxa at successively specific taxonomic levels.
See also Figure S1.



(legend on next page)

We observed a wide range of alpha diversity (a measure of taxonomic richness within samples), with a median Shannon diversity of 2.33 and values as high as 5.07 (Figure 1G), consistent with ranges identified in previous meta-analyses of alpha diversity across multiple microbiome studies.[47,48] To estimate the completeness of this census, we performed a sample-based rarefaction analysis,[49] in which we selected random subsets of samples of different sizes without replacement from the full compendium and evaluated the number of unique taxa observed in each subsample (see STAR Methods for details). The discovery rate for taxa approached zero after 25,000 samples for all levels except genus, the most specific. Between subsamples of 150,000 samples and the full dataset of 168,464, we observed only one new genus every 4,831 samples (Figure 1H). This suggests the reference database saturates at most ranks after about 25,000 samples, though we would also observe this effect with a perfectly comprehensive reference after all taxa had actually been observed. This also suggests the compendium currently captures all but the rarest taxa represented by the reference database in the populations covered by the dataset, given the current distribution of reads per sample. Overall, we find broad variation in the composition of human gut samples (Figure S1), but within a limited selection of microbial taxa drawn mostly from the Bacillota phylum (Figure 1E).

World regions harbor unique microbiome signatures

Though much of the public metadata available for BioSamples is inconsistently reported,[50] the "geo_loc_name" identifier was available for 92.4% of samples in the filtered compendium, which we used to associate each sample with its country of origin (see STAR Methods), and we then consolidated these countries into eight world regions defined by the United Nations Sustainable Development Goals (SDG) program (Figure 2A, top). As in previous work,[19] we found most samples were from Europe and Northern America (91,144 samples; 60.5%), with the Eastern and South-Eastern Asia region a distant second at 17,086 samples (11.3%; Figure 2B). Sub-Saharan Africa was the third-most represented (5,538 samples; 3.7%), followed by Central and Southern Asia (5,046 samples; 3.4%), though much of the world remains underrepresented (Figure 2A, bottom).

We observed the highest alpha diversity among the 1,195 samples from Latin America and the Caribbean (median Shannon index = 2.69; Figure 2C). Samples from Central and Southern Asia exhibited the lowest average diversity (median Shannon index = 1.68), though the diversity of understudied world regions is likely underestimated because of gaps in reference databases used for taxonomic assignment (see Discussion). Comparison here is complicated by differences in read depth between regions: the Australia/New Zealand region had the lowest median reads per sample (30,415 reads; Figure 2D) compared with the average of 98,641 reads from Latin America and the Caribbean, which has the most deeply sequenced samples, but a rarefaction analysis shows regional differences in alpha diversity persist even after controlling for samples per region and reads per sample (Figure S1; see STAR Methods). We also evaluated differences in the phylogenetic diversity of the taxa observed in each region, using Faith's Phylogenetic Diversity (PD) score to measure the length of all branches of the reference tree containing the observed taxa in a region (Figure S2).[51,52] When evaluating the tree of Europe and Northern America as the baseline, we find adding the taxa of Eastern and South-Eastern Asia increases total branch length by 68.6%, followed by Sub-Saharan Africa (18.9%) and Latin America and the Caribbean (16.9%). This highlights another potential pitfall of drawing broad conclusions from samples taken only from Europe and Northern America, the most overrepresented region. Even in regions with similar levels of diversity (Figure S1), the taxa present may be from entirely different clades.

To explore differences in microbiome composition across world regions, we visualized the dataset by applying principal coordinates analysis (PCoA) using the robust Aitchison distance[53] (Figure 2E; see STAR Methods). We plotted the samples from each region using the first two dimensions, which together account for 44.9% of variation (Figure S3). Though this approach meant the analysis was heavily weighted in favor of variance observed in Europe and Northern America, using these axes

Figure 2. Regional structure

(A) The top map illustrates the categorization into seven world regions by color, and below it, the map was distorted to reflect each country's proportion of samples in the dataset. Oceania is represented here in black, though this region was excluded from these analyses because too few samples remained in the filtered dataset used here.

(B) A bar plot illustrating the number of samples analyzed from each world region. The x axis illustrates total samples, and the colors used here are the same as those used in (A).

(C) A violin plot illustrating the distribution of observed Shannon index values assigned to samples from each world region. The x axis indicates the Shannon index value, as calculated using all unique taxonomic identifications in samples from each world region as in Figure 1G. Colors indicate the region (same as in A), and the y axis for each violin indicates the relative frequency with which diversity of a given magnitude was observed. The vertical lines in each violin indicate the median value. The black points within each violin indicate the mean Shannon diversity as determined by rarefaction analysis, as illustrated in Figure S1. See Figure S2 for an evaluation of differences in diversity between regions.

(D) A violin plot organized in the same manner as (C), with the x axis indicating the number of merged reads that were included in the filtered taxonomic table.

(E) A series of plots illustrating the results of a principal coordinates analysis of samples from all world regions. The top-left plot is a scatter plot in which each point is a single sample, and the color indicates the sample's region, using the scheme described in (B). The x axis is the first PCoA axis, which explains the most variation across the dataset, and the y axis is the PCoA axis explaining the second-most variation. The seven other plots use the same axes, but each includes only samples from a single world region. These plots use a heatmap design, rather than a scatter plot, to help evaluate areas with many overlapping points—yellow areas indicate portions of the space with a higher concentration of samples, and dark blue areas indicate portions with few (but not zero) samples. The gray shadow indicates the area occupied by all points from all world regions. See also Figure S3.

(F) A series of density plots illustrating the distributions of the first four axes of variation determined by the ordination analysis displayed in (E). Each panel illustrates a single factor, and the x axis indicates the value of that factor, and the y axis indicates the relative frequency of the value in the given world region. See also Figures S1, S2, and S3.

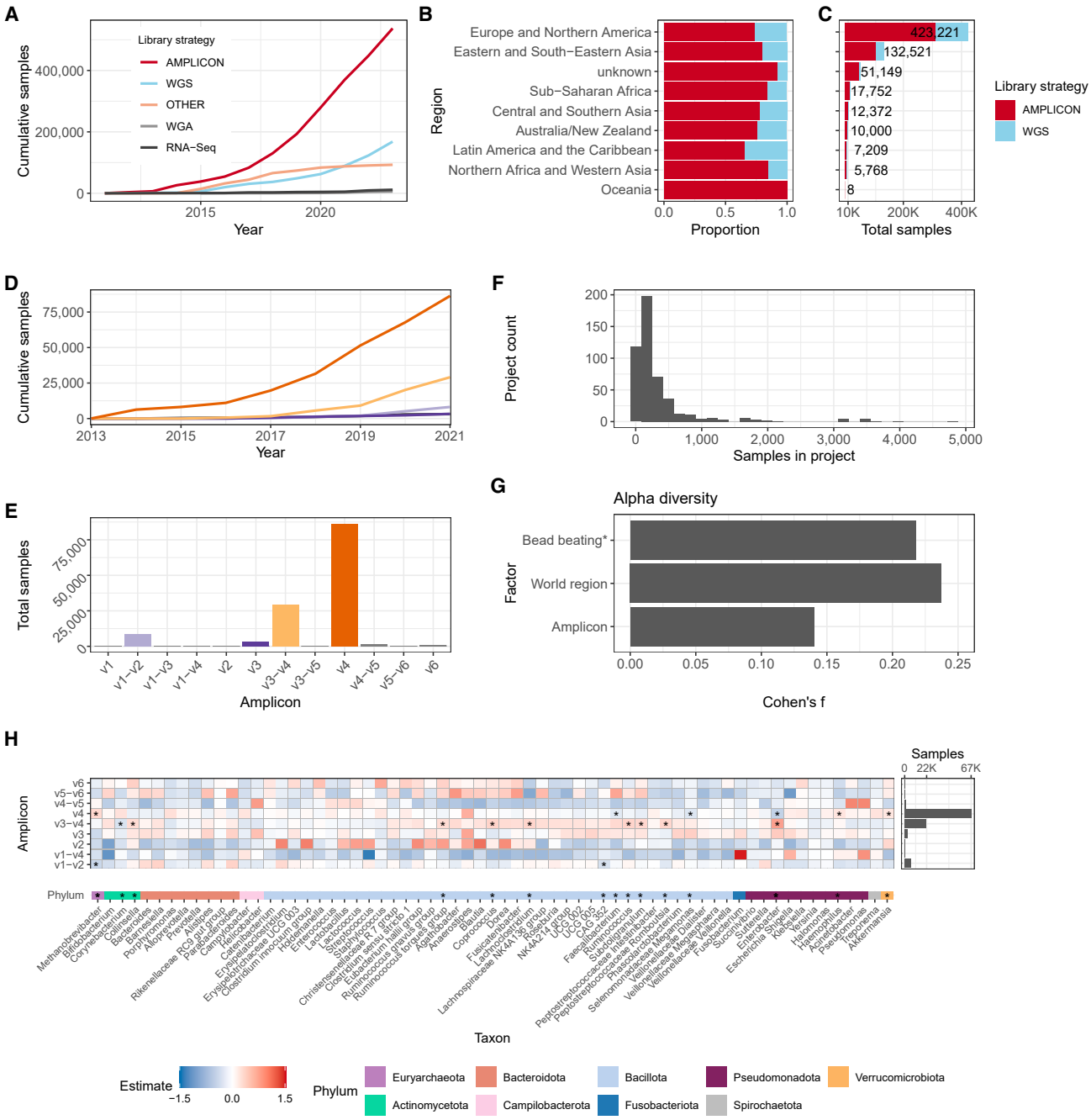


Figure 3. Technical factors

(A) Reported sequencing library strategies over time. The y axis indicates the cumulative number of samples, and the x axis indicates the year of publication, and each line indicates the most common reported values for library strategy.

(B and C) Library strategies by world region. This shows the same data as (A) but using only the data for amplicon and shotgun ("WGS") samples. (B) illustrates the proportion of samples from each region that are classified as either 16S or shotgun samples. (C) indicates the total samples. See also Figure S4.

(D) Amplicon choice over time: the x axis indicates the year in which a sample was published in the BioSample database, and the y axis indicates the cumulative number of samples using a given amplicon, indicated by the color scale. This and all following panels refer to samples in the compendium, rather than the BioSample database.

(E) Overall amplicon choice: this illustrates the same data as (D) with the same color scale, showing the total samples using each amplicon. The x axis lists each amplicon, and the y axis indicates the sample count.

(F) Samples per project. The x axis indicates the number of samples from a given project have been included in the compendium, and the y axis indicates the number of BioProjects that fall into each size segment. Please note that in some cases, projects were included but had some samples filtered out.

(legend continued on next page)

helps to illustrate whether samples from the rest of the world differ from those of the most thoroughly characterized region. Even in these two dimensions, we observe systematic differences between world regions. Samples from Australia/New Zealand appear to cluster in subsets of the main areas occupied by Europe and Northern America, but others, such as those from Latin America and the Caribbean, occupy areas of the projected space that are farther from the samples of other regions (Figure 2F), and, using all eight axes extracted from the dataset, clusters defined by world region are much more compact and distinct than would be expected by chance (Davies-Bouldin index = 6.93; $p < 4 \times 10^{-6}$; Figure S3). Together, these results demonstrate the relationship between geography and microbiome composition.

Technical factors affect global microbiome composition

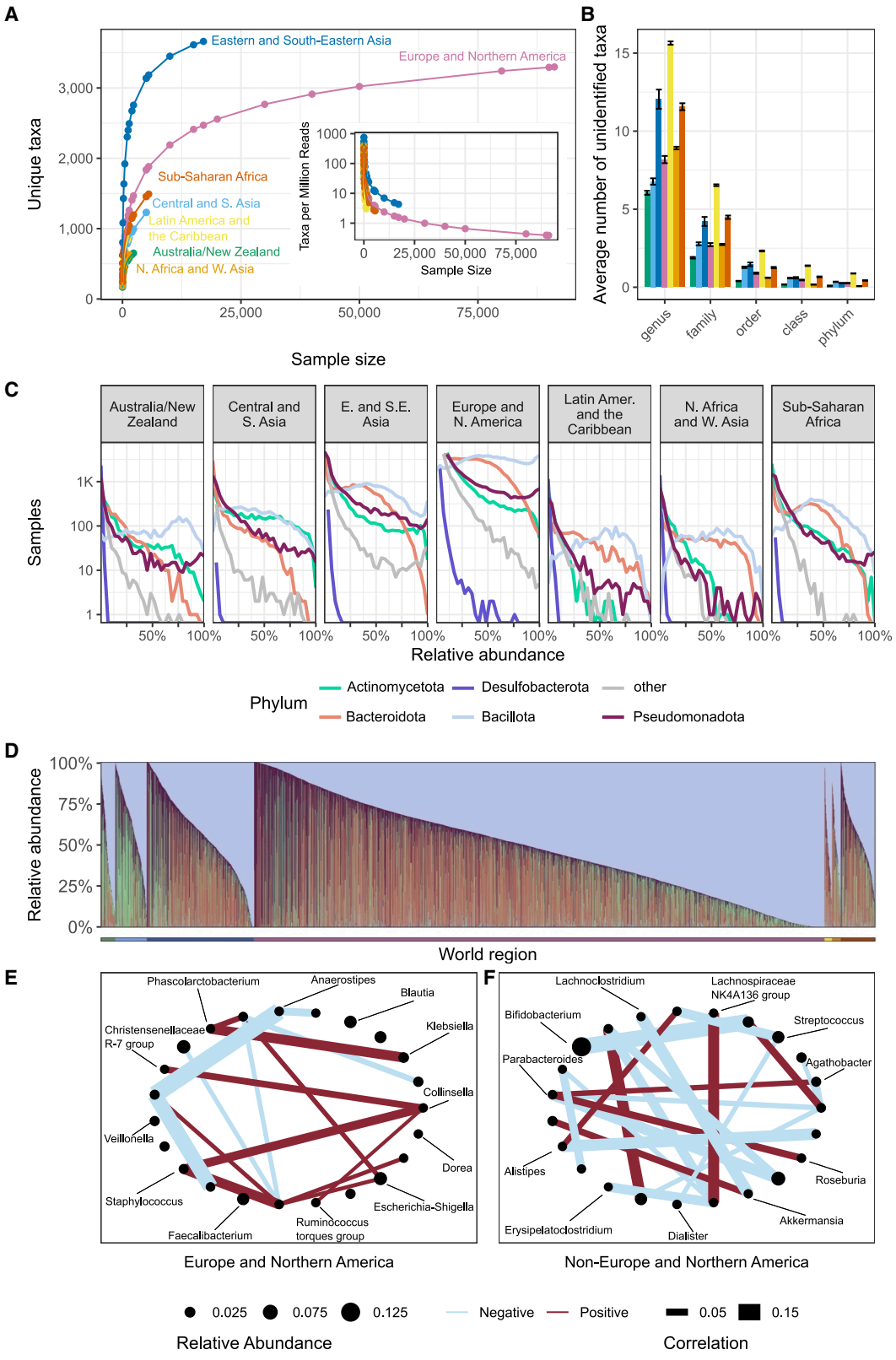
Our dataset contains the processed data from a subset of samples shared in the BioSample database. To obtain a better understanding of how the Human Microbiome Compendium data compares to the field's sequencing data at large, we retrieved metadata for all human gut microbiome samples available in that database, removing our filters for properties such as assay type or sequencing platform (see STAR Methods). We were able to obtain metadata for 814,916 human gut microbiome samples from 2011 through 2023. We found 65.9% reported using amplicon sequencing, followed by 20.6% using shotgun metagenomic sequencing. RNA sequencing (RNA-seq) data accounts for another 1.4% of samples, among others (Figure 3A). We see a similar pattern at the project level, where there are three times more amplicon sequencing projects than shotgun sequencing projects (Figure S4), suggesting amplicon sequencing's relative popularity is not solely attributable to those projects being larger. Annotating these samples with their geographic origin revealed that the majority of samples in every world region were assayed using amplicon sequencing (Figure 3B) and that 76.4% of shotgun sequencing samples were from Europe and Northern America (Figure 3C), though their residents make up less than 15% of the world population.[19] This metadata also show the Illumina MiSeq instrument has been the dominant 16S rRNA gene sequencing platform since 2014 (405,907 samples, 75.6%; Figure S4). These results highlight that amplicon sequencing continues to be a popular and widespread approach to microbiome quantification—in many world regions, it is the only type of human gut microbiome sequencing data available at any practical scale.

After evaluating the information available for all human gut microbiome BioSamples, we then turned to the samples processed for the Human Microbiome Compendium, which we annotated with information on peer-reviewed publications, extraction kit, mechanical lysis, and amplicon choice (Table S3; see STAR Methods). We found that the V4 region was a common choice even early in the field's development (Figure 3D), with V3–V4 amplicons growing in popularity after 2017 but still trailing far behind (Figure 3E). Of the 144,154 samples for which lysis information could be inferred, 113,228 samples (78.6%) were from a study that either reported using bead-beating explicitly in a publication or reported using an extraction kit that includes a mechanical lysis step (Table S3). The median project size is 164 samples, with a positively skewed distribution (Figure 3F), and of the 482 BioProjects in the compendium, 348 projects (72.2%) have fewer than 300 samples.

When examining alpha diversity (Faith's phylogenetic diversity,[54] calculated for each sample after rarefaction to 1,000 sequences), we found that world region exhibited the largest effect size (Cohen's $f = 0.237$; Figure 3G), followed by bead beating (estimated Cohen's $f = 0.218$; see STAR Methods) and amplicon choice (Cohen's $f = 0.140$). A PERMANOVA analysis[55] evaluating the relationship between technical factors and microbiome composition found that all evaluated factors had a significant effect ($p \leq 0.002$ for all factors by permutation; see STAR Methods), with world region having a stronger effect than technical factors (world region $R^2 = 0.026$, amplicon $R^2 = 0.009$, and bead beating $R^2 = 0.004$), and we note that these coefficients sum to far less than 1.0, indicating a broad amount of unexplained variation remains after controlling for these effects. Importantly, the effect of the interaction between world region and amplicon ($R^2 = 0.010$; $p \leq 0.002$) was larger than that of amplicon evaluated separately, indicating that the influence of amplicon choice may have larger impacts on microbiota specific to individual regions.

Relatively early microbiome studies on amplification bias in 16S rRNA libraries found some taxa were affected by the choice of hypervariable region used in the amplification process.[56] Our differential abundance analysis identified 15 genera with significant differences when comparing samples from each hypervariable region to all other samples, mostly within the V4 and V3–V4 hypervariable regions (Figure 3H). (We note that there are wide differences in sample sizes between amplicons, and all significant shifts were observed among the three most prevalent amplicons.) The affected taxa include nine members of the Bacillota phylum (eight of the class Clostridia), plus one taxon each from five other phyla. *Akkermansia* has a higher relative abundance in studies using the V4 hypervariable region ($q = 0.019$). Two genera exhibit significant shifts in multiple hypervariable regions: The relative abundance of *Enterobacter* was in higher relative abundance when the V3–V4 hypervariable region was used for sequencing ($q = 1.30 \times 10^{-18}$) and depleted with use of the V4 hypervariable region ($q = 1.42 \times 10^{-11}$), and relative abundance of the archaean *Methanobrevibacter* was lower with use of

(G) Technical factor effect size. The y axis lists the factors evaluated, and the x axis indicates their estimated effect on sample alpha diversity. Note that the "Cohen's *f*" value reported for bead beating is a Cohen's *d* value that has been halved for more accurate comparison; see STAR Methods for details.
(H) A heatmap of differential abundance by amplicon. The x axis lists all evaluated taxa, and the y axis lists all amplicons with samples from at least two world regions. Each cell indicates the effect size of an amplicon on the relative abundance of a given taxon, and significant ($q < 0.05$) differences are marked with an asterisk both in the relevant cell of the heatmap and along the bottom row. The row of colored cells at the bottom of the plot indicates the phylum of each taxon, which are ordered alphabetically so each genus appears near its closest relatives. The horizontal bar plot on the right indicates total samples using each amplicon.
See also Figure S4.



(legend on next page)

the V1–V2 amplicon (q = 0.00476) and increased with V4 ($q = 1.36 \times 10^{-9}$), a difference that may be affected by the design of bacteria-specific forward primers for the V1 region.[57]

Undersampling suggests unobserved taxa in most world regions

To investigate microbiome diversity in world regions, we repeatedly subsampled each region and identified the number of unique microbial taxa present in the selected microbiome samples (Figures 4A and S5). For this analysis, we used all 4,018 taxa quantified by the initial DADA2 assignment and all samples with a world region assignment. Notably, more taxa are observed in samples from Eastern and South-Eastern Asia than any other region, despite having around 70,000 fewer samples than the largest region. 3,662 of the 4,018 taxa are present in samples from Eastern and South-Eastern Asia, and only 874 taxa are observed in Latin America and the Caribbean. We note that the rate of discovery drops for each world region after the first few thousand samples: 2,190 unique taxa were identified in the first 10,000 samples from Europe and Northern America, but the subsequent 81,144 samples uncovered an average of only 1,109 new taxa. From another perspective, the rate of discovery for Europe and Northern America falls below 1 taxon per million reads before 30,000 samples are assayed (Figure 4A, inset). By comparison, we continue to observe an average of 8.19 new taxa per million reads in Northern Africa and Western Asia even when all available samples have been evaluated. In Latin America and the Caribbean, the discovery rate is 3.03 new taxa per million reads when all samples from the region are assayed. Given previous findings that current references are biased toward taxa in the western microbiome,[58,59] we evaluated whether ASVs with missing classifications for at least one rank would be found more frequently outside of Europe and Northern America. We find the highest number of unidentified taxa in Latin America and the Caribbean (Figure 4B), a trend that holds at every taxonomic level. In general, regions with high sample-level alpha diversity tend to have more unidentified taxa. Though we would expect the discovery rate in all regions to eventually reach zero, this relationship to alpha diversity, combined with our observation that the discovery rate in many regions is well above zero even when available samples are exhausted, suggests that additional samples from most regions could still add thousands of entries to the list of taxa present there.

While the top phyla remain consistent across world regions, their abundance and prevalence differ (Figure 4C): in Europe and Northern America, the relative abundance of Bacillota is approximately uniformly distributed. In Sub-Saharan Africa, however, the distribution of Bacillota peaks at approximately 40%, with higher relative abundances becoming less and less common. While Bacillota (Figure 4D, light blue segments) is dominant in each region, samples from Central and Southern Asia have higher relative abundances of Actinomycetota (green segments; Wilcoxon test, $p < 2.2 \times 10^{-16}$ for Central and Southern Asia versus other) than the other regions, plus correspondingly lower abundances of *Bacteroides* (orange segments; Wilcoxon test, $p < 2.2 \times 10^{-16}$ for Central and Southern Asia versus other; Figure 4D).

Next, we examined the relationship between microbes, hypothesizing that there may be different relationships between microbial taxa in different world regions. We focused on evaluating co-abundance patterns that were unique to samples from Europe and Northern America (see STAR Methods), and we found 16 significant ($p < 0.001$ by permutation) pairwise correlations that are observed only among samples from this region (Figure 4E); inversely, we found 22 significant correlations (between 23 taxa) that are observed only among samples from outside that region (Figure 4F), providing context when considering the generalizability of findings from a single population.

Global microbiome composition differs from most commonly studied countries

To quantify how specific microbial taxa vary between world regions, we performed differential abundance analysis using a linear mixed model that accounts for BioProject, use of bead beating, and amplicon used for sequencing (Table S3), as these experimental factors may bias results. We focused our analysis on the 65 genera that had a minimum of 1% prevalence and 0.5% relative abundance in at least one world region (see STAR

Figure 4. Geographic regions vary in microbiome composition

(A) The number of unique taxa discovered in subsamples of varying size from each world region. Each point represents the average number of unique taxa identified in a subsample from a given region over 1,000 repetitions. The x axis indicates the number of microbiome samples selected, the y axis the number of unique taxa identified in those samples, and the color indicates the world region being sampled. The inset uses the same x axis and color scheme but displays the average number of taxa discovered per million reads on the y axis. See also Figure S5.

(B) Bar chart showing the number of unidentified taxa in each world region at every taxonomic level. Each bar indicates one world region (following the same color scheme as A), and bars are grouped by taxonomic rank. Error bars indicate the standard deviation.

(C) Histograms illustrating the distribution of the relative abundance of the most prevalent phyla in the compendium. Each panel visualizes all samples from a single world region. The x axis indicates the relative abundance of the taxon, and the y axis indicates the number of samples (on a log scale) with the indicated relative abundance. Each line illustrates the results for a single phylum, indicated by line color.

(D) A stacked bar chart showing the relative abundance of the five most prevalent phyla in the compendium. Each column is a sample, and the colored segments indicate the relative abundance of a given phylum in that sample. Phylum color follows the same color scheme as (B). Samples are ordered first by world region (indicated by the colored bar below the x axis) and then by relative abundance of the 5 most prevalent phyla. World region color follows the same color scheme as (A). See also Figure S1.

(E) Taxon correlation network showing significant correlations between taxa within samples from Europe and Northern America. Node size indicates the average relative abundance of the taxon, edge width indicates the strength of the correlation, blue edges indicate negative correlations, and red edges indicate positive correlations.

(F) Taxon correlation network showing significant correlations between samples from regions other than Europe and Northern America. Node size, edge width, and edge color follow the same legend as (E).

See also Figures S1 and S5.

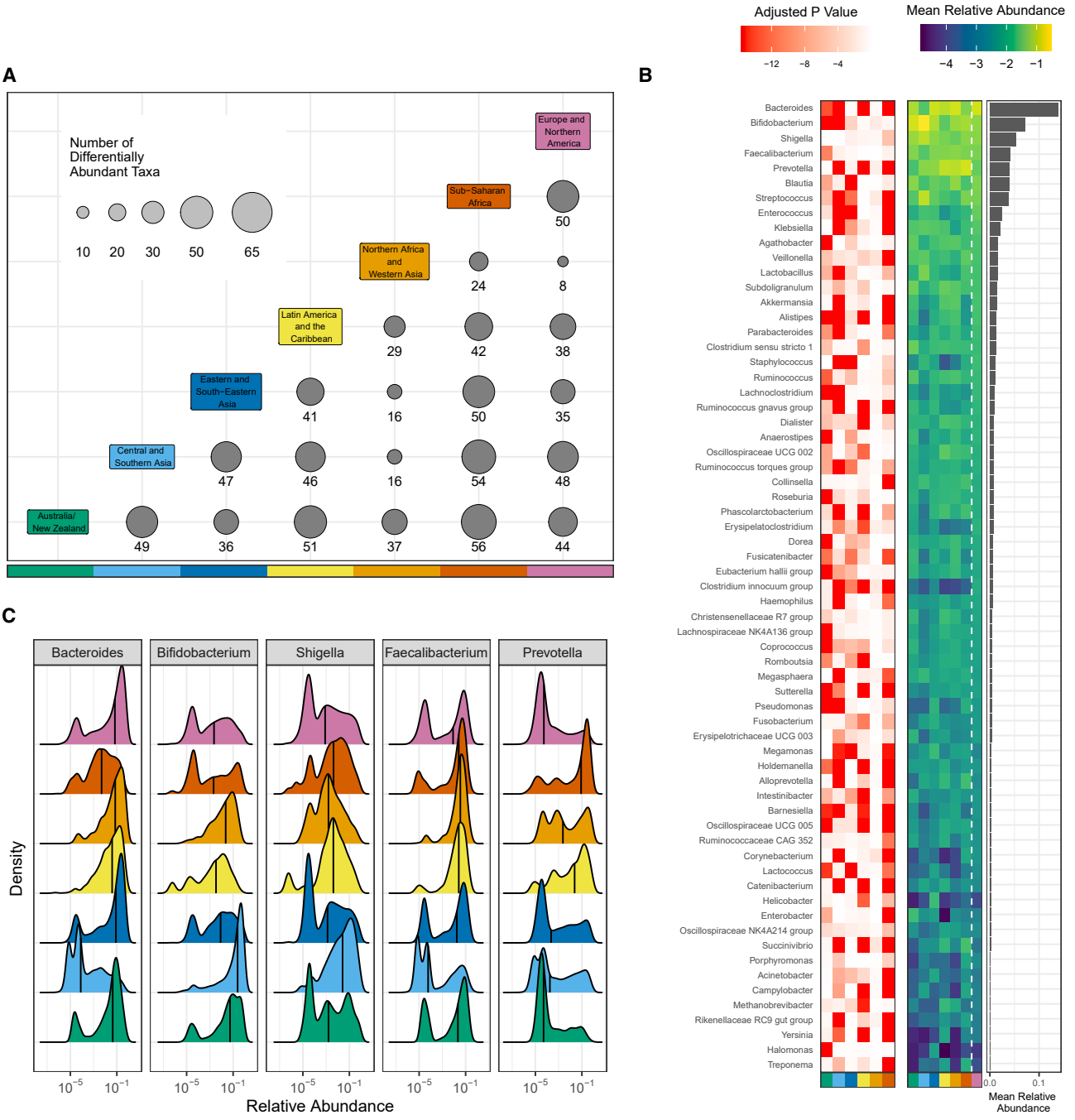
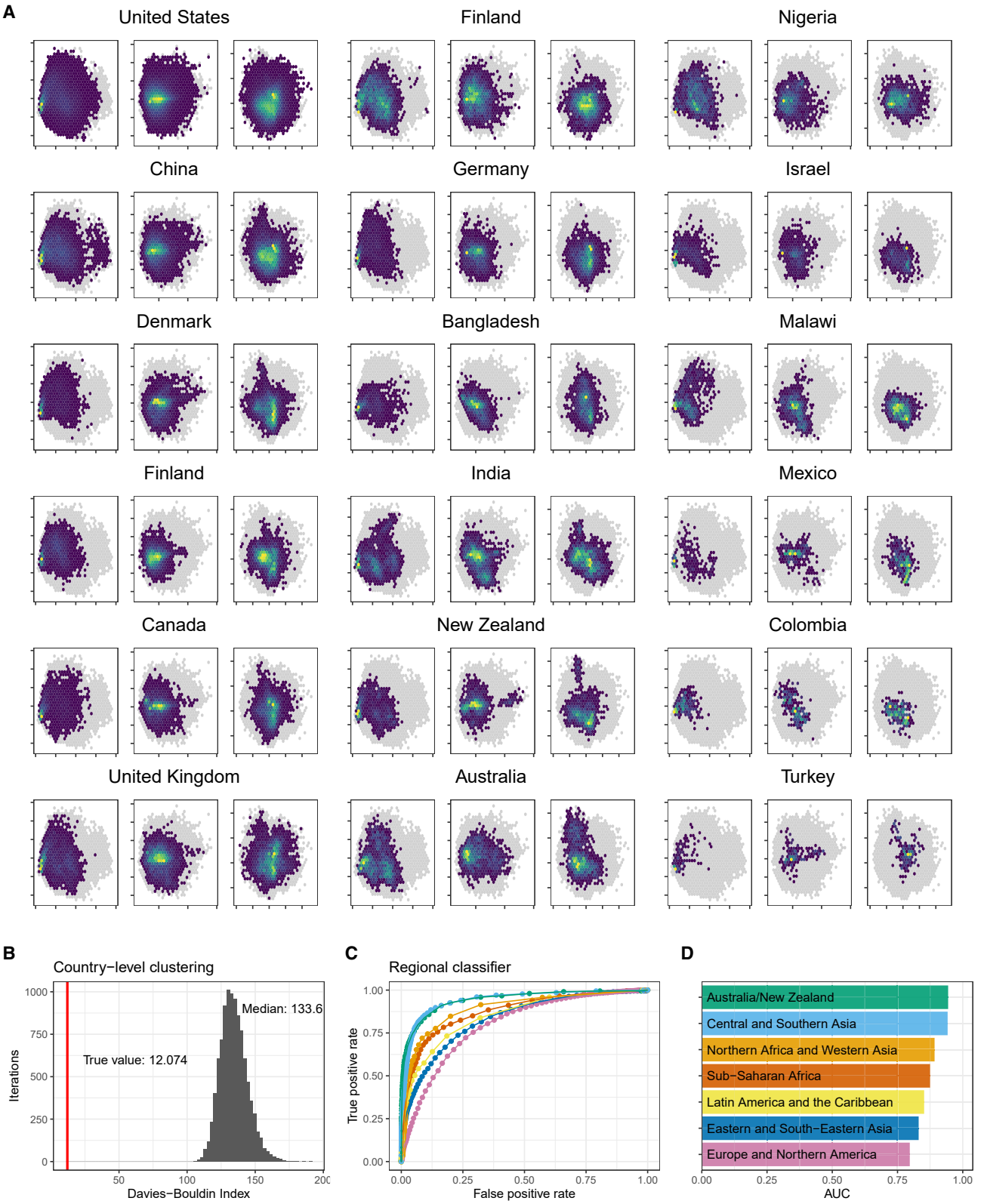


Figure 5. Taxa are differentially abundant between world regions

(A) 65 taxa were selected to be tested for differential abundance between regions. The x and y axes are each colored by world region, and at each intersection, the size of the circle and the number underneath it indicates the number of taxa that were significantly different between the two regions listed.

(B) The red-white heatmap illustrates adjusted *p* values for regional differences when each world region is compared with Europe and Northern America. The y axis lists all evaluated genera, the x axis lists each region (using the same color scale as A), and each cell represents the strength of the differential abundance result for that taxon. The blue-green heatmap illustrates mean relative abundance (log 10) of each taxon in each world region, as indicated by the x axis. The bar chart illustrates the mean relative abundance of each taxon across all regions.

(C) Each panel illustrates the relative abundance (log 10) of one of the 5 most abundant taxa. Each colored area indicates the distribution from a single world region, using the same colors as (A). The x axis indicates (log 10) relative abundance of the specified genus, and the y axis indicates the relative frequency with which that abundance is observed in the specified region. Black vertical lines indicate the median.



(legend on next page)

Methods). Pairwise comparison of each region revealed distinct differences among these genera, and all 65 taxa tested were found to be significantly differentially abundant between at least one pair of regions (Figure 5A). The highest number of differentially abundant taxa (n = 56) were found between Sub-Saharan Africa and Australia/New Zealand, while the fewest differentially abundant taxa (n = 8) were found between samples from Europe and Northern America and those from Northern Africa and Western Asia.

As samples from Europe and Northern America make up over half of the compendium, subsequent analyses focused on the differences found between that region and each of the others to evaluate the most broadly observed differences relative to the most sampled region. Concordant with prior literature,[60] the relative abundance of *Bacteroides* is higher in Europe and Northern America than in Sub-Saharan Africa (Figure 5B; $q < 2.2 \times 10^{-16}$), Latin America and the Caribbean ($q < 2.2 \times 10^{-16}$), Central and Southern Asia ($q < 2.2 \times 10^{-16}$), and Australia/New Zealand ($q = 7.50 \times 10^{-14}$). Conversely, the relative abundance of *Prevotella* is lower in Europe and Northern America compared with Sub-Saharan Africa ($q < 2.2 \times 10^{-16}$), Latin America and the Caribbean ($q < 2.2 \times 10^{-16}$), and Central and Southern Asia ($q < 2.2 \times 10^{-16}$).

Closer examination of the distribution of relative abundances of different genera across world regions reveals more specific patterns. Most regions have many samples with a high relative abundance of *Bacteroides*, as evidenced by the strong peak close to 1 in Figure 5C, which is evident for regions other than Central and Southern Asia and Sub-Saharan Africa. More than 20% of samples from Europe and Northern America contain at least 30% *Bacteroides* by relative abundance, and only 2.6% and 1.6% of samples from Sub-Saharan Africa and Central and Southern Asia, respectively, contain as much *Bacteroides*. *Prevotella*, commonly associated with positive health outcomes and non-western microbiomes, has higher relative abundance in Sub-Saharan Africa ($q < 2.2 \times 10^{-16}$) and Latin America and the Caribbean ($q < 2.2 \times 10^{-16}$) when compared with Europe and Northern America. Only 17.5% of samples from Europe and Northern America have more than 1% *Prevotella*—visible in the strong peak near 0 in Figure 5C—compared with 42% of samples from Northern Africa and Western Asia, 59% of Latin America and the Caribbean, and 63% of Sub-Saharan African samples.

We then sought to define region-specific signatures of gut microbiomes by identifying the taxa most closely linked to the overall variance observed in principal-component analysis (PCA), performed separately for each region (termed here the "variance score"; see STAR Methods). We find variability in Europe and Northern America is best represented by the relative abundances of *Escherichia/Shigella* (variance score = 0.98; Table S4), *Enterococcus* (0.97), *Lactobacillus* (0.95), *Akkermansia* (0.94), and *Bifidobacterium* (0.93), while the microbiomes of Northern Africa and Western Asia are defined by the genera *Prevotella* (0.85), *Escherichia/Shigella* (0.81), *Akkermansia* (0.77), *Dialister* (0.57), and *Bacteroides* (0.52). Using the top 10 taxa in each regional signature, we found that six genera appear in the signatures of all world regions, all of which are corroborated by the literature: *Bifidobacterium*,[61] *Bacteroides*,[62,63] *Prevotella*,[63,64] *Streptococcus*,[65–67] *Veillonella*,[65,66,68] and *Escherichia/Shigella*, which together form what we could consider the core taxa most useful for explaining global variation in the human gut microbiome—not necessarily the most prevalent, but the taxa that vary most widely in all regions. These genera also reflect the core taxa of the five "enterosignatures" defined by Frioux et al.[69] as the small number of functional "guilds" that combine to make up healthy microbiomes in their analysis. Oppositely, there are five genera that appear in the top 10 taxa for only a single region: *Staphylococcus* (Central and Southern Asia), *Megamonas* (Eastern and South-Eastern Asia), *Dialister* (Northern Africa and Western Asia), *Collinsella* (Sub-Saharan Africa), and *Alistipes* (Latin America and the Caribbean). Many of these genera, including the region-specific taxa *Dialister*, *Megamonas*, and *Alistipes*, were also found to vary significantly between global enterotypes identified in shotgun data by Keller et al.[70] (Please note that the *Lactobacillus* genus in this work does not reflect changes to the Lactobacillaceae family[71]; see STAR Methods for details.)

Region classification and country-level ordination

To evaluate patterns in beta diversity reflecting geographic origin, we next generated PCoA plots to visualize the overall differences in microbiome composition between samples at the country level (Figure 6A). We observed that the United States and China, the countries with the highest number of samples in the dataset, exhibited a large amount of overlap in the first several axes when using Aitchison distance. Other countries exhibited different patterns in the ordination, sometimes apparent only at higher PCs. We found the country-level clustering was much more compact than would be expected by chance (Davies-Bouldin index = 12.074; $p < 10^{-4}$; Figure 6B).

Given the presence of significant differences in microbiome composition between world regions, we trained random forest

Figure 6. Country-level ordination

(A) A principal-component analysis plot of compendium samples, segmented by country of origin. Each country is represented by three heatmaps, directly under the country name, which together illustrate the six axes that explain the most variation in all samples considered. The left panel of each country plots PC1 (x axis) against PC2 (y axis). The middle panel plots PC3 against PC4, and the right panel plots PC5 against PC6. These plots use a heatmap design, rather than a scatterplot, to help evaluate areas with many overlapping points—yellow areas indicate portions of the space with a higher concentration of samples, and dark blue areas indicate portions with few (but not zero) samples. The gray shadow indicates the area occupied by all points in the analysis.

(B) A histogram illustrating the Davies-Bouldin index of cluster strength for 10,000 iterations in which the country labels were randomly permuted. The vertical red line indicates the observed value (12.07), which is substantially lower than any permuted value (minimum = 105.3).

(C) Receiver operating characteristic (ROC) curves for the one-versus-all classifiers. Each line illustrates the ROC curve for a single region, using the color scheme labeled in (D). The x axis indicates the false-positive rate of the model for a given threshold, and the y axis indicates the true-positive rate at the same threshold.

(D) A bar plot illustrating the area under the ROC curve (AUC) as calculated for each region.

See also Figure S6.

classifiers to determine whether region could be inferred from an individual microbiome sample. We trained binary one-versus-all classifiers for each of the remaining seven regions (Figure 6C; see STAR Methods). The highest prediction accuracy was observed for identifying which samples were from Australia/New Zealand (area under the receiver operating characteristic [ROC] curve [AUC] = 0.944; Figure 6D), and the lowest was observed for Europe and Northern America (0.797). When evaluating the taxa that were most influential in each region's model, we found each relied on fewer than 100 taxa (Figure S6) and that taxa that were differentially abundant between regions (Figure 5B) were significantly overrepresented in those that contributed most heavily to the accuracy of the classification models ($q < 0.1$; see STAR Methods). We also trained region-versus-region random forest classifiers between each pair of regions (Figure S6). Between regions, the highest classification accuracy was observed for distinguishing between microbiomes from Eastern and South-Eastern Asia and microbiomes from Central and Southern Asia (AUC = 0.945). The least accurate distinctions were drawn between Europe and Northern America and microbiomes from Latin America and the Caribbean (AUC = 0.704).

DISCUSSION

Here, we integrated data from 168,464 publicly available 16S rRNA gene amplicon sequencing samples from 482 BioProjects to evaluate global variation in the human gut microbiome. We found the majority of available samples were from Europe and Northern America, which have been so extensively sampled that most known microbiota present in the region's gut microbiomes have likely already been observed, while further samples from other regions may uncover up to 20 times as many new taxa per million reads. Thousands of taxa have also been observed in Eastern and South-Eastern Asia, but samples show such remarkable diversity that there are likely many more yet to be uncovered. Though practically all taxa are shared to some degree between world regions, we found that each region occupies a unique niche within the ordination space defined via multidimensional scaling (Figure 2E) and cluster in ways that are detectable by multiple machine learning classification approaches (Figure 6). We also identified many microbiome patterns associated with technical factors that will help contextualize findings and inform study design going forward: despite falling costs, we did not find a departure from amplicon sequencing in favor of shotgun sequencing (Figure 3A), for example, and we quantified how popular the V4 and V3–V4 amplicons have become in the assays used to quantify the microbiome (Figure 3D), a dynamic that could have implications for which taxa are observed and in what proportions (Figure 3H).

Others have articulated the vital importance of studying microbiomes from diverse populations[21,22] and evaluating the potential consequences of inaction.[16,72] For example, we found that variance in Europe and Northern America, by far the most thoroughly sampled region, is closely tied to the relative abundance of *Lactobacillus* (variance score = 0.95), which has been linked to obesity in the United States[73] and bipolar disorder in Austria.[74] It remains to be answered how these results should be interpreted in Latin America and the Caribbean, where *Lactobacillus* is practically absent from the regional signature (variance score = 0.03), though not absent from the microbiomes there. We also found many taxa exhibit significant differences in abundance when comparing world regions to Europe and Northern America (Figure 5B). This includes highly abundant genera such as *Bacteroides*, *Bifidobacterium*, and *Prevotella*, the abundance and proportions of which may play a role in inflammation,[75] obesity,[76] inflammatory bowel disease,[64] and the development of the pediatric gut microbiome.[77] Together, this provides more detailed evidence of the difficulty in generalizing microbiome findings to populations that may not encounter even the same bacterial families.

The compendium provides a broad quantification of human gut microbiome variation that can help us understand the boundaries and patterns that constrain the composition of all microbiomes, healthy or not. Still, there is reason for caution in drawing strong conclusions from such a wide range of samples collected for different reasons in hundreds of projects. First, world region may be confounded with why the samples were collected, and the data currently do not have consistent information about host health or the disease status of specific samples within a study. Diet information is missing, for example, though food and drink have a complex relationship to microbiome composition and host health.[78] Ethnicity information is also unavailable, and though some of this may be captured by geographic information, studies have found ethnic background to have a strong relationship to microbiome composition even for subjects within a single city.[7] Relatedly, reference databases may have less coverage of taxa that appear more commonly outside of Europe and North America,[58,79] which would result in more unidentified taxa and deflated diversity estimates in samples from other regions of the world.

Regarding our analysis, any combination of studies raises concerns about batch effects, or artifactual findings that are caused by technical details but appear to be of biological origin.[24,80] We believe that these effects are minimized in large-scale analyses—in our previous work integrating more than 100 individual datasets, we found batch correction of large compendia is ineffective when there are many "batches"—in this case, projects. As the number of batches grows, the disparate project-level effects are overshadowed by the legitimate biological signal, which is more consistent across studies,[81] and another study found large phenotypic effects generally outweighed study-level batch effects even among a collection of 12 studies.[35] This compendium contains 482 studies, annotated with technical information that could be used for further modeling.

Lastly, the dataset compiled for this project does not resolve the broad issue of representational imbalances in global human microbiome research,[19,82,83] though there are many ongoing projects that seek to increase the diversity of microbiome research—projects such as the African Microbiome Program are expanding work in human and agricultural microbiomes,[84] and initiatives like H3ABioNet made strides addressing the structural challenges to expanding the populations under study.[85,86] We are optimistic about the Human Microbiome Compendium's utility in providing context for the data from these regions and better utilizing existing resources: The National

Institutes of Health have directly invested more than $1 billion in human microbiome research,[87] and raw data for tens of thousands of microbiome samples are uploaded every year (Figure 3A) to the databases of the International Nucleotide Sequence Database Collaboration (INSDC), which includes NCBI, the European Nucleotide Archive and the DNA Data Bank of Japan.[88] In the microbiome space, the Sequence Read Archive estimates the taxonomic content of all deposited samples,[89,90] analysis platforms such as MGnify[91] and Qiita[92,93] enable users to process and download samples from their own projects and those of others, and resources such as MicrobiomeHD,[94] GMrepo,[42] and curatedMetagenomicData[95,96] provide access to pre-generated tables from multiple studies (Table S1). Machine learning approaches that rely on large compendia and have been readily applied in fields that have access to such data,[97–99] and we believe the Human Microbiome Compendium, with a taxonomic table several times larger than those currently available, represents a key resource for bringing these groundbreaking techniques to the microbiome field. To facilitate this, we will continue processing and curating microbiome data, including expanded annotations demonstrated here and the inclusion of additional body sites.

In summary, we present here a large-scale collection of human gut microbiome data. We use this compendium to study microbiome variation at a global scale, comparing world regions and showing that some regions likely have many taxa that remain undiscovered due to undersampling. We expect this compendium will be a valuable resource for the community and enable expanded insights into the microbial ecology of the human gut.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ran Blekhman (blekhman@uchicago.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- The full Human Microbiome Compendium dataset is available for download from Zenodo at https://doi.org/10.5281/zenodo.8186993.
- The open-source MicroBioMap R package is available on GitHub at https://github.com/seandavi/MicroBioMap.
- An interactive website at microbiomap.org includes summary data for the compendium, links for downloads, and information about ongoing updates to the project.
- The code used for data processing of the raw data, plus the code used to generate the figures in this manuscript, is available on Zenodo at https://doi.org/10.5281/zenodo.13733483.
- Ongoing development for the Human Microbiome Compendium can be tracked on GitHub at https://github.com/blekhmanlab/compendium.
- Text mining code used to annotate extraction kit information is available on GitHub at https://github.com/krishnanlab/microbiome-annotation.

ACKNOWLEDGMENTS

We thank the members of the Blekhman lab for helpful discussion. We also thank Chad L. Myers, R. Stephanie Huang, Anna Selmecki, William Harcombe (University of Minnesota), and Benjamín Valderrama (University College Cork) for their feedback. This work was completed with resources provided by the University of Chicago's Research Computing Center and the Minnesota Supercomputing Institute. This work was supported by NIH grant R01LM013863 (to R.B. and C.S.G.).

AUTHOR CONTRIBUTIONS

Conceptualization, R.J.A., S.P.G., R.B., and F.W.A.; data curation, R.J.A., S.P.G., V.R., M.A., and P.H.; formal analysis, R.J.A., S.P.G., M.A., D.M., A.C., P.F., E.G., and M.R.; software, R.J.A., V.R., and S.D.; supervision, A.K., F.W.A., C.S.G., S.D., and R.B.; writing – original draft, R.J.A., S.P.G., and V.R.; writing – review and editing, F.W.A., C.S.G., S.D., and R.B.

DECLARATION OF INTERESTS

D.M. is a consultant for BiomeSense, Inc., has equity, and receives income. 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.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Sample selection
- METHOD DETAILS
  - Data processing
  - Annotation and filtering
  - Microbiome analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Calculation of Shannon diversity
  - Country metadata accuracy estimate
  - World region alpha diversity comparison
  - Rarefaction analysis, regional alpha diversity
  - Principal coordinates analysis
  - Analysis of variance
  - Differential abundance analysis for amplicon choice
  - Differential abundance analysis
  - Country-level ordination plots
  - Cluster strength evaluation
- ADDITIONAL RESOURCES
  - Website for high-level exploration
  - R package implementation
  - Software tools

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.12.017.

Received: October 2, 2023
Revised: September 9, 2024
Accepted: December 13, 2024
Published: January 22, 2025

REFERENCES

1. Bullman, S., Pedamallu, C.S., Sicinska, E., Clancy, T.E., Zhang, X., Cai, D., Neuberg, D., Huang, K., Guevara, F., Nelson, T., et al. (2017). Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. Science *358*, 1443–1448. https://doi.org/10.1126/science.aal5240.
2. Hale, V.L., Jeraldo, P., Chen, J., Mundy, M., Yao, J., Priya, S., Keeney, G., Lyke, K., Ridlon, J., White, B.A., et al. (2018). Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient

mismatch repair colorectal cancers. Genome Med. *10*, 78. https://doi.org/10.1186/s13073-018-0586-6.

3. Burns, M.B., Montassier, E., Abrahante, J., Priya, S., Niccum, D.E., Khoruts, A., Starr, T.K., Knights, D., and Blekhman, R. (2018). Colorectal cancer mutational profiles correlate with defined microbial communities in the tumor microenvironment. PLoS Genet. *14*, e1007376. https://doi.org/10.1371/journal.pgen.1007376.
4. Matsuoka, K., and Kanai, T. (2015). The gut microbiota and inflammatory bowel disease. Semin. Immunopathol. *37*, 47–55. https://doi.org/10.1007/s00281-014-0454-4.
5. Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T., et al. (2014). Human genetics shape the gut microbiome. Cell *159*, 789–799. https://doi.org/10.1016/j.cell.2014.09.053.
6. Brooks, A.W., Priya, S., Blekhman, R., and Bordenstein, S.R. (2018). Gut microbiota diversity across ethnicities in the United States. PLoS Biol. *16*, e2006842. https://doi.org/10.1371/journal.pbio.2006842.
7. Deschasaux, M., Bouter, K.E., Prodan, A., Levin, E., Groen, A.K., Herrema, H., Tremaroli, V., Bakker, G.J., Attaye, I., Pinto-Sietsma, S.-J., et al. (2018). Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat. Med. *24*, 1526–1531. https://doi.org/10.1038/s41591-018-0160-1.
8. Mallott, E.K., Sitarik, A.R., Leve, L.D., Cioffi, C., Camargo, C.A., Jr., Hasegawa, K., and Bordenstein, S.R. (2023). Human microbiome variation associated with race and ethnicity emerges as early as 3 months of age. PLoS Biol. *21*, e3002230. https://doi.org/10.1371/journal.pbio.3002230.
9. Porras, A.M., and Brito, I.L. (2019). The internationalization of human microbiome research. Curr. Opin. Microbiol. *50*, 50–55. https://doi.org/10.1016/j.mib.2019.09.012.
10. De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., and 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. USA *107*, 14691–14696. https://doi.org/10.1073/pnas.1005963107.
11. Langdon, A., Crook, N., and Dantas, G. (2016). The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. Genome Med. *8*, 39. https://doi.org/10.1186/s13073-016-0294-z.
12. Vangay, P., Johnson, A.J., Ward, T.L., Al-Ghalith, G.A., Shields-Cutler, R.R., Hillmann, B.M., Lucas, S.K., Beura, L.K., Thompson, E.A., Till, L.M., et al. (2018). US immigration westernizes the human gut microbiome. Cell *175*, 962–972.e10. https://doi.org/10.1016/j.cell.2018.10.029.
13. Le Bastard, Q., Vangay, P., Batard, E., Knights, D., and Montassier, E. (2020). US Immigration Is Associated With Rapid and Persistent Acquisition of Antibiotic Resistance Genes in the Gut. Clin. Infect. Dis. *71*, 419–421. https://doi.org/10.1093/cid/ciz1087.
14. Peters, B.A., Yi, S.S., Beasley, J.M., Cobbs, E.N., Choi, H.S., Beggs, D.B., Hayes, R.B., and Ahn, J. (2020). US nativity and dietary acculturation impact the gut microbiome in a diverse US population. ISME J. *14*, 1639–1650. https://doi.org/10.1038/s41396-020-0630-6.
15. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227. https://doi.org/10.1038/nature11053.
16. Gupta, V.K., Paul, S., and Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Front. Microbiol. *8*, 1162. https://doi.org/10.3389/fmicb.2017.01162.
17. Vujkovic-Cvijin, I., Sklar, J., Jiang, L., Natarajan, L., Knight, R., and Belkaid, Y. (2020). Host variables confound gut microbiota studies of human disease. Nature *587*, 448–454. https://doi.org/10.1038/s41586-020-2881-9.
18. Porras, A.M., Shi, Q., Zhou, H., Callahan, R., Montenegro-Bethancourt, G., Solomons, N., and Brito, I.L. (2021). Geographic differences in gut microbiota composition impact susceptibility to enteric infection. Cell Rep. *36*, 109457. https://doi.org/10.1016/j.celrep.2021.109457.
19. Abdill, R.J., Adamowicz, E.M., and Blekhman, R. (2022). Public human microbiome data are dominated by highly developed countries. PLoS Biol. *20*, e3001536. https://doi.org/10.1371/journal.pbio.3001536.
20. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164. https://doi.org/10.1038/538161a.
21. Amato, K.R., Arrieta, M.-C., Azad, M.B., Bailey, M.T., Broussard, J.L., Bruggeling, C.E., Claud, E.C., Costello, E.K., Davenport, E.R., Dutilh, B.E., et al. (2021). The human gut microbiome and health inequities. Proc. Natl. Acad. Sci. USA *118*, e2017947118. https://doi.org/10.1073/pnas.2017947118.
22. Shanahan, F., Ghosh, T.S., and O'Toole, P.W. (2023). Human microbiome variance is underestimated. Curr. Opin. Microbiol. *73*, 102288. https://doi.org/10.1016/j.mib.2023.102288.
23. Chanderraj, R., Brown, C.A., Hinkle, K., Falkowski, N., Woods, R.J., and Dickson, R.P. (2022). The bacterial density of clinical rectal swabs is highly variable, correlates with sequencing contamination, and predicts patient risk of extraintestinal infection. Microbiome *10*, 2. https://doi.org/10.1186/s40168-021-01190-y.
24. Blekhman, R., Tang, K., Archie, E.A., Barreiro, L.B., Johnson, Z.P., Wilson, M.E., Kohn, J., Yuan, M.L., Gesquiere, L., Grieneisen, L.E., et al. (2016). Common methods for fecal sample storage in field studies yield consistent signatures of individual identity in microbiome sequencing data. Sci. Rep. *6*, 31519. https://doi.org/10.1038/srep31519.
25. Panek, M., Čipčić Paljetak, H., Barešić, A., Perić, M., Matijašić, M., Lojkić, I., Vranešić Bender, D., Krznarić, Ž., and Verbanac, D. (2018). Methodology challenges in studying human gut microbiota - effects of collection, storage, DNA extraction and next generation sequencing technologies. Sci. Rep. *8*, 5143. https://doi.org/10.1038/s41598-018-23296-4.
26. Yeoh, Y.K., Chen, Z., Hui, M., Wong, M.C.S., Ho, W.C.S., Chin, M.L., Ng, S.C., Chan, F.K.L., and Chan, P.K.S. (2019). Impact of inter- and intra-individual variation, sample storage and sampling fraction on human stool microbial community profiles. PeerJ *7*, e6172. https://doi.org/10.7717/peerj.6172.
27. Marotz, C., Cavagnero, K.J., Song, S.J., McDonald, D., Wandro, S., Humphrey, G., Bryant, M., Ackermann, G., Diaz, E., and Knight, R. (2021). Evaluation of the effect of storage methods on fecal, saliva, and skin microbiome composition. mSystems *6*, e01329-20. https://doi.org/10.1128/mSystems.01329-20.
28. Whon, T.W., Chung, W.-H., Lim, M.Y., Song, E.-J., Kim, P.S., Hyun, D.-W., Shin, N.-R., Bae, J.-W., and Nam, Y.-D. (2018). The effects of sequencing platforms on phylogenetic resolution in 16 S rRNA gene profiling of human feces. Sci. Data *5*, 180068. https://doi.org/10.1038/sdata.2018.68.
29. Multinu, F., Harrington, S.C., Chen, J., Jeraldo, P.R., Johnson, S., Chia, N., and Walther-Antonio, M.R. (2018). Systematic bias Introduced by Genomic DNA Template Dilution in 16S rRNA Gene-Targeted Microbiota Profiling in Human Stool Homogenates. mSphere *3*, e00560-17. https://doi.org/10.1128/mSphere.00560-17.
30. Kennedy, N.A., Walker, A.W., Berry, S.H., Duncan, S.H., Farquarson, F.M., Louis, P., Thomson, J.M., UK IBD Genetics Consortium, Satsangi, J., and Flint, H.J. (2014). The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoS One *9*, e88982. https://doi.org/10.1371/journal.pone.0088982.
31. Wagner Mackenzie, B., Waite, D.W., and Taylor, M.W. (2015). Evaluating variation in human gut microbiota profiles due to DNA extraction method

and inter-subject differences. Front. Microbiol. *6*, 130. https://doi.org/10.3389/fmicb.2015.00130.

32. Lim, M.Y., Song, E.-J., Kim, S.H., Lee, J., and Nam, Y.-D. (2018). Comparison of DNA extraction methods for human gut microbial community profiling. Syst. Appl. Microbiol. *41*, 151–157. https://doi.org/10.1016/j.syapm.2017.11.008.

33. Santiago, A., Panda, S., Mengels, G., Martinez, X., Azpiroz, F., Dore, J., Guarner, F., and Manichanh, C. (2014). Processing faecal samples: a step forward for standards in microbial community analysis. BMC Microbiol. *14*, 112. https://doi.org/10.1186/1471-2180-14-112.

34. Yang, F., Sun, J., Luo, H., Ren, H., Zhou, H., Lin, Y., Han, M., Chen, B., Liao, H., Brix, S., et al. (2020). Assessment of fecal DNA extraction protocols for metagenomic studies. GigaScience *9*, giaa071. https://doi.org/10.1093/gigascience/giaa071.

35. Lozupone, C.A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., Jansson, J.K., Gordon, J.I., and Knight, R. (2013). Meta-analyses of studies of the human microbiota. Genome Res. *23*, 1704–1714. https://doi.org/10.1101/gr.151803.112.

36. Ragan-Kelley, B., Walters, W.A., McDonald, D., Riley, J., Granger, B.E., Gonzalez, A., Knight, R., Perez, F., and Caporaso, J.G. (2013). Collaborative cloud-enabled tools allow rapid, reproducible biological insights. ISME J. *7*, 461–464. https://doi.org/10.1038/ismej.2012.123.

37. Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S.E., Taub, M.A., Hansen, K.D., Jaffe, A.E., Langmead, B., and Leek, J.T. (2017). Reproducible RNA-seq analysis using recount2. Nat. Biotechnol. *35*, 319–321. https://doi.org/10.1038/nbt.3838.

38. Wilks, C., Zheng, S.C., Chen, F.Y., Charles, R., Solomon, B., Ling, J.P., Imada, E.L., Zhang, D., Joseph, L., Leek, J.T., et al. (2021). recount3: summaries and queries for large-scale RNA-seq expression and splicing. Genome Biol. *22*, 323. https://doi.org/10.1186/s13059-021-02533-6.

39. Lee, A.J., Doing, G., Neff, S.L., Reiter, T., Hogan, D.A., and Greene, C.S. (2023). Compendium-wide analysis of Pseudomonas aeruginosa Core and accessory genes reveals transcriptional patterns across strains PAO1 and PA14. mSystems *8*, e0034222. https://doi.org/10.1128/msystems.00342-22.

40. Pividori, M., Lu, S., Li, B., Su, C., Johnson, M.E., Wei, W.-Q., Feng, Q., Namjou, B., Kiryluk, K., Kullo, I.J., et al. (2023). Projecting genetic associations through gene expression patterns highlights disease etiology and drug mechanisms. Nat. Commun. *14*, 5562. https://doi.org/10.1038/s41467-023-41057-4.

41. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods *13*, 581–583. https://doi.org/10.1038/nmeth.3869.

42. Dai, D., Zhu, J., Sun, C., Li, M., Liu, J., Wu, S., Ning, K., He, L.-J., Zhao, X.-M., and Chen, W.-H. (2022). GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. Nucleic Acids Res. *50*, D777–D784. https://doi.org/10.1093/nar/gkab1019.

43. Jin, H., Hu, G., Sun, C., Duan, Y., Zhang, Z., Liu, Z., Zhao, X.-M., and Chen, W.-H. (2022). mBodyMap: a curated database for microbes across human body and their associations with health and diseases. Nucleic Acids Res. *50*, D808–D816. https://doi.org/10.1093/nar/gkab973.

44. Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. Science *308*, 1635–1638. https://doi.org/10.1126/science.1110591.

45. Oren, A., and Garrity, G.M. (2021). Valid publication of the names of forty-two phyla of prokaryotes. Int. J. Syst. Evol. Microbiol. *71*. https://doi.org/10.1099/ijsem.0.005056.

46. Mariat, D., Firmesse, O., Levenez, F., Guimarães, V., Sokol, H., Doré, J., Corthier, G., and Furet, J.P. (2009). The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. BMC Microbiol. *9*, 123. https://doi.org/10.1186/1471-2180-9-123.

47. Sze, M.A., and Schloss, P.D. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. mBio *7*, e01018-16. https://doi.org/10.1128/mBio.01018-16.

48. Pinart, M., Dötsch, A., Schlicht, K., Laudes, M., Bouwman, J., Forslund, S.K., Pischon, T., and Nimptsch, K. (2021). Gut Microbiome Composition in Obese and Non-Obese Persons: A Systematic Review and Meta-Analysis. Nutrients *14*, 12. https://doi.org/10.3390/nu14010012.

49. Gotelli, N.J., and Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. *4*, 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x.

50. Kasmanas, J.C., Bartholomäus, A., Corrêa, F.B., Tal, T., Jehmlich, N., Herberth, G., von Bergen, M., Stadler, P.F., Carvalho, A.C.P., and Nunes da Rocha, U. (2021). HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. Nucleic Acids Res. *49*, D743–D750. https://doi.org/10.1093/nar/gkaa1031.

51. McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018). American gut: an open platform for citizen science microbiome research. mSystems *3*, e00031-18. https://doi.org/10.1128/mSystems.00031-18.

52. Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. https://doi.org/10.1038/nature11234.

53. Martino, C., Morton, J.T., Marotz, C.A., Thompson, L.R., Tripathi, A., Knight, R., and Zengler, K. (2019). A novel sparse compositional technique reveals microbial perturbations. mSystems *4*, e00016–e00019. https://doi.org/10.1128/mSystems.00016-19.

54. Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. Biol. Conserv. *61*, 1–10. https://doi.org/10.1016/0006-3207(92)91201-3.

55. Li, S., Vogtmann, E., Graubard, B.I., Gail, M.H., Abnet, C.C., and Shi, J. (2022). fast.adonis: a computationally efficient non-parametric multivariate analysis of microbiome data for large-scale studies. Bioinform. Adv. *2*, vbac044. https://doi.org/10.1093/bioadv/vbac044.

56. Kuczynski, J., Lauber, C.L., Walters, W.A., Parfrey, L.W., Clemente, J.C., Gevers, D., and Knight, R. (2011). Experimental and analytical tools for studying the human microbiome. Nat. Rev. Genet. *13*, 47–58. https://doi.org/10.1038/nrg3129.

57. Yang, N., Tian, C., Lv, Y., Hou, J., Yang, Z., Xiao, X., and Zhang, Y. (2022). Novel primers for 16S rRNA gene-based archaeal and bacterial community analysis in oceanic trench sediments. Appl. Microbiol. Biotechnol. *106*, 2795–2809. https://doi.org/10.1007/s00253-022-11893-3.

58. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell *176*, 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

59. Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. Nature *568*, 505–510. https://doi.org/10.1038/s41586-019-1058-x.

60. Gorvitovskaia, A., Holmes, S.P., and Huse, S.M. (2016). Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome *4*, 15. https://doi.org/10.1186/s40168-016-0160-7.

61. Ladeira, R., Tap, J., and Derrien, M. (2023). Exploring Bifidobacterium species community and functional variations with human gut microbiome structure and health beyond infancy. Microbiome Res. Rep. *2*, 9. https://doi.org/10.20517/mrr.2023.01.

62. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011).

Enterotypes of the human gut microbiome. Nature *473*, 174–180. https://doi.org/10.1038/nature09944.

63. Zhou, X., Shen, X., Johnson, J.S., Spakowicz, D.J., Agnello, M., Zhou, W., Avina, M., Honkala, A., Chleilat, F., Chen, S.J., et al. (2024). Longitudinal profiling of the microbiome at four body sites reveals core stability and individualized dynamics during health and disease. Cell Host Microbe *32*, 506–526.e9. https://doi.org/10.1016/j.chom.2024.02.012.
64. Prasoodanan P K, V., Sharma, A.K., Mahajan, S., Dhakan, D.B., Maji, A., Scaria, J., and Sharma, V.K. (2021). Western and non-western gut microbiomes reveal new roles of Prevotella in carbohydrate metabolism and mouth-gut axis. NPJ Biofilms Microbiomes *7*, 77. https://doi.org/10.1038/s41522-021-00248-x.
65. Santoru, M.L., Piras, C., Murgia, A., Palmas, V., Camboni, T., Liggi, S., Ibba, I., Lai, M.A., Orrù, S., Blois, S., et al. (2017). Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients. Sci. Rep. *7*, 9523. https://doi.org/10.1038/s41598-017-10034-5.
66. Bajer, L., Kverka, M., Kostovcik, M., Macinga, P., Dvorak, J., Stehlikova, Z., Brezina, J., Wohl, P., Spicak, J., and Drastich, P. (2017). Distinct gut microbiota profiles in patients with primary sclerosing cholangitis and ulcerative colitis. World J. Gastroenterol. *23*, 4548–4558. https://doi.org/10.3748/wjg.v23.i25.4548.
67. Lim, M.Y., Hong, S., Bang, S.-J., Chung, W.-H., Shin, J.-H., Kim, J.-H., and Nam, Y.-D. (2021). Gut microbiome structure and association with host factors in a Korean population. mSystems *6*, e0017921. https://doi.org/10.1128/mSystems.00179-21.
68. Alsulaiman, R.M., Al-Quorain, A.A., Al-Muhanna, F.A., Piotrowski, S., Kurdi, E.A., Vatte, C., Alquorain, A.A., Alfaraj, N.H., Alrezuk, A.M., Robinson, F., et al. (2023). Gut microbiota analyses of inflammatory bowel diseases from a representative Saudi population. BMC Gastroenterol. *23*, 258. https://doi.org/10.1186/s12876-023-02904-2.
69. Frioux, C., Ansorge, R., Özkurt, E., Ghassemi Nedjad, C., Fritscher, J., Quince, C., Waszak, S.M., and Hildebrand, F. (2023). Enterosignatures define common bacterial guilds in the human gut microbiome. Cell Host Microbe *31*, 1111–1125.e6. https://doi.org/10.1016/j.chom.2023.05.024.
70. Keller, M.I., Nishijima, S., Podlesny, D., Kim, C.Y., Robbani, S.M., Schudoma, C., Fullam, A., Richter, J., Letunic, I., Akanni, W., et al. (2024). Refined Enterotyping reveals dysbiosis in global fecal metagenomes. Preprint at bioRxiv. https://doi.org/10.1101/2024.08.13.607711.
71. Zheng, J., Wittouck, S., Salvetti, E., Franz, C.M.A.P., Harris, H.M.B., Mattarelli, P., O'Toole, P.W., Pot, B., Vandamme, P., Walter, J., et al. (2020). A taxonomic note on the genus Lactobacillus: description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. Int. J. Syst. Evol. Microbiol. *70*, 2782–2858. https://doi.org/10.1099/ijsem.0.004107.
72. 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., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat. Med. *24*, 1532–1535. https://doi.org/10.1038/s41591-018-0164-x.
73. Peters, B.A., Shapiro, J.A., Church, T.R., Miller, G., Trinh-Shevrin, C., Yuen, E., Friedlander, C., Hayes, R.B., and Ahn, J. (2018). A taxonomic signature of obesity in a large study of American adults. Sci. Rep. *8*, 9749. https://doi.org/10.1038/s41598-018-28126-1.
74. Painold, A., Mörkl, S., Kashofer, K., Halwachs, B., Dalkner, N., Bengesser, S., Birner, A., Fellendorf, F., Platzer, M., Queissner, R., et al. (2019). A step ahead: exploring the gut microbiota in inpatients with bipolar disorder during a depressive episode. Bipolar Disord. *21*, 40–49. https://doi.org/10.1111/bdi.12682.
75. Iljazovic, A., Roy, U., Gálvez, E.J.C., Lesker, T.R., Zhao, B., Gronow, A., Amend, L., Will, S.E., Hofmann, J.D., Pils, M.C., et al. (2021). Perturbation of the gut microbiome by Prevotella spp. enhances host susceptibility to mucosal inflammation. Mucosal Immunol. *14*, 113–124. https://doi.org/10.1038/s41385-020-0296-4.
76. Dong, T.S., Guan, M., Mayer, E.A., Stains, J., Liu, C., Vora, P., Jacobs, J.P., Lagishetty, V., Chang, L., Barry, R.L., et al. (2022). Obesity is associated with a distinct brain-gut microbiome signature that connects Prevotella and Bacteroides to the brain's reward center. Gut Microbes *14*, 2051999. https://doi.org/10.1080/19490976.2022.2051999.
77. de Goffau, M.C., Jallow, A.T., Sanyang, C., Prentice, A.M., Meagher, N., Price, D.J., Revill, P.A., Parkhill, J., Pereira, D.I.A., and Wagner, J. (2022). Gut microbiomes from Gambian infants reveal the development of a non-industrialized Prevotella-based trophic network. Nat. Microbiol. *7*, 132–144. https://doi.org/10.1038/s41564-021-01023-6.
78. Ross, F.C., Patangia, D., Grimaud, G., Lavelle, A., Dempsey, E.M., Ross, R.P., and Stanton, C. (2024). The interplay between diet and the gut microbiome: implications for health and disease. Nat. Rev. Microbiol. *22*, 671–686. https://doi.org/10.1038/s41579-024-01068-4.
79. Thomas, A.M., and Segata, N. (2019). Multiple levels of the unknown in microbiome research. BMC Biol. *17*, 48. https://doi.org/10.1186/s12915-019-0667-z.
80. Wesolowska-Andersen, A., Bahl, M.I., Carvalho, V., Kristiansen, K., Sicheritz-Pontén, T., Gupta, R., and Licht, T.R. (2014). Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome *2*, 19. https://doi.org/10.1186/2049-2618-2-19.
81. Lee, A.J., Park, Y., Doing, G., Hogan, D.A., and Greene, C.S. (2020). Correcting for experiment-specific variability in expression compendia can remove underlying signals. GigaScience *9*, giaa117. https://doi.org/10.1093/gigascience/giaa117.
82. Bah, S.Y., Morang'a, C.M., Kengne-Ouafo, J.A., Amenga-Etego, L., and Awandare, G.A. (2018). Highlights on the Application of Genomics and Bioinformatics in the Fight Against Infectious Diseases: Challenges and Opportunities in Africa. Front. Genet. *9*, 575. https://doi.org/10.3389/fgene.2018.00575.
83. Allali, I., Abotsi, R.E., Tow, L.A., Thabane, L., Zar, H.J., Mulder, N.M., and Nicol, M.P. (2021). Human microbiota research in Africa: a systematic review reveals gaps and priorities for future research. Microbiome *9*, 241. https://doi.org/10.1186/s40168-021-01195-7.
84. Makhalanyane, T.P., Bezuidt, O.K.I., Pierneef, R.E., Mizrachi, E., Zeze, A., Fossou, R.K., Kouadjo, C.G., Duodu, S., Chikere, C.B., Babalola, O.O., et al. (2023). African microbiomes matter. Nat. Rev. Microbiol. *21*, 479–481. https://doi.org/10.1038/s41579-023-00925-y.
85. Mulder, N.J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., Akanle, B., Alibi, M., Armstrong, D.L., Aron, S., et al. (2017). Development of bioinformatics infrastructure for genomics research. Glob. Heart *12*, 91–98. https://doi.org/10.1016/j.gheart.2017.01.005.
86. Shaffer, J.G., Mather, F.J., Wele, M., Li, J., Tangara, C.O., Kassogue, Y., Srivastav, S.K., Thiero, O., Diakite, M., Sangare, M., et al. (2019). Expanding Research Capacity in Sub-Saharan Africa Through Informatics, Bioinformatics, and Data Science Training Programs in Mali. Front. Genet. *10*, 331. https://doi.org/10.3389/fgene.2019.00331.
87. NIH; Human; Microbiome; Portfolio; Analysis Team (2019). A review of 10 years of human microbiome research activities at the US National Institutes of Health, fiscal years 2007–2016. Microbiome *7*, 31. https://doi.org/10.1186/s40168-019-0620-y.
88. International Nucleotide Sequence Database Collaboration. Global participation. https://www.insdc.org/global-participation/.
89. Katz, K.S., Shutov, O., Lapoint, R., Kimelman, M., Brister, J.R., and O'Sullivan, C. (2021). STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. Genome Biol. *22*, 270. https://doi.org/10.1186/s13059-021-02490-0.

90. National Center for Biotechnology Information. SRA taxonomy analysis tool. [Internet]. https://www.ncbi.nlm.nih.gov/sra/docs/sra-taxonomy-analysis-tool/.

91. Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M.L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L.J., et al. (2023). MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res. *51*, D753–D759. https://doi.org/10.1093/nar/gkac1080.

92. Gonzalez, A., Navas-Molina, J.A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A.D., Orchanian, S.B., et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. Nat. Methods *15*, 796–798. https://doi.org/10.1038/s41592-018-0141-9.

93. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A., and Alm, E.J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat. Commun. *8*, 1784. https://doi.org/10.1038/s41467-017-01973-8.

94. McDonald, D., Kaehler, B., Gonzalez, A., DeReus, J., Ackermann, G., Marotz, C., Huttley, G., and Knight, R. (2019). redbiom: a Rapid Sample Discovery and Feature Characterization System. mSystems *4*, e00215–e00219. https://doi.org/10.1128/msystems.00215-19.

95. Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D.T., Beghini, F., Malik, F., Ramos, M., Dowd, J.B., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. Nat. Methods *14*, 1023–1024. https://doi.org/10.1038/nmeth.4468.

96. Schiffer, L., and Waldron, L. (2023). curatedMetagenomicData. https://waldronlab.io/curatedMetagenomicData/articles/curatedMetagenomicData.html.

97. Nishijima, S., Stankevic, E., Aasmets, O., Schmidt, T.S.B., Nagata, N., Keller, M.I., Ferretti, P., Juel, H.B., Fullam, A., Robbani, S.M., et al. (2025). Fecal microbial load is a major determinant of gut microbiome variation and a confounder for disease associations. Cell *188*, 1–15. https://doi.org/10.1016/j.cell.2024.10.022.

98. Tan, J., Doing, G., Lewis, K.A., Price, C.E., Chen, K.M., Cady, K.C., Perchuk, B., Laub, M.T., Hogan, D.A., and Greene, C.S. (2017). Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. Cell Syst. *5*, 63–71.e6. https://doi.org/10.1016/j.cels.2017.06.003.

99. Banerjee, J., Taroni, J.N., Allaway, R.J., Prasad, D.V., Guinney, J., and Greene, C. (2023). Machine learning in rare disease. Nat. Methods *20*, 803–814. https://doi.org/10.1038/s41592-023-01886-z.

100. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. *41*, D590–D596. https://doi.org/10.1093/nar/gks1219.

101. McDonald, D., Jiang, Y., Balaban, M., Cantrell, K., Zhu, Q., Gonzalez, A., Morton, J.T., Nicolaou, G., Parks, D.H., Karst, S.M., et al. (2024). Greengenes2 unifies microbial data in a single reference tree. Nat. Biotechnol. *42*, 715–718. https://doi.org/10.1038/s41587-023-01845-1.

102. "fasterq-dump" (GIThub). SRA Toolkit. https://github.com/ncbi/sra-tools.

103. Wei, C.-H., Allot, A., Lai, P.-T., Leaman, R., Tian, S., Luo, L., Jin, Q., Wang, Z., Chen, Q., and Lu, Z. (2024). PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. Nucleic Acids Res. *52*, W540–W546. https://doi.org/10.1093/nar/gkae235.

104. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. *3*, 1–23. https://doi.org/10.1145/3458754.

105. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857. https://doi.org/10.1038/s41587-019-0209-9.

106. scikit-bio, [Version 0.5.8] (2022). http://scikit-bio.org.

107. Watts, S.C., Ritchie, S.C., Inouye, M., and Holt, K.E. (2019). FastSpar: rapid and scalable correlation estimation for compositional data. Bioinformatics *35*, 1064–1066. https://doi.org/10.1093/bioinformatics/bty734.

108. Wright, M.N., and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw. *77*, 1–17. https://doi.org/10.18637/jss.v077.i01.

109. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1. https://doi.org/10.18637/jss.v028.i05.

110. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

111. Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szoecs, E., et al. (2022). Vegan: Community Ecology Package.

112. Delicado, P., and Pachon-Garcia, C. (2020). Multidimensional scaling for big data. Preprint at arXiv. https://doi.org/10.48550/arXiv.2007.11919.

113. Rahman, G., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., Jiang, L., Casals-Pascual, C., Peddada, S., Hakim, D., Dilmore, A.H., Nowinski, B., et al. (2022). Scalable power analysis and effect size exploration of microbiome community differences with Evident. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.19.492684.

114. Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. J. Stat. Softw. *67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

115. Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. (2017). lmerTest package: tests in linear mixed effects models. J. Stat. Softw. *82*, 1–26. https://doi.org/10.18637/jss.v082.i13.

116. Walesiak, M., and Dudek, A. (2020). The choice of variable normalization method in cluster analysis. In Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development during Global Challenges, K.S. Soliman, ed. (International Business Information Management Association (IBIMA)), pp. 325–340.

117. Massicotte, P., and South, A. (2017). World Map Data from Natural Earth [R package rnaturalearth version 0.1.0]. https://cran.r-project.org/package=rnaturalearth.

118. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer Science & Business Media) https://doi.org/10.1007/978-0-387-98141-3.

119. Pedersen, T.L. (2024). Patchwork: the Composer of Plots. https://patchwork.data-imaginist.com.

120. Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2024). dplyr: A Grammar of Data Manipulation. https://github.com/tidyverse/dplyr.

121. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) *2020*, baaa062. https://doi.org/10.1093/database/baaa062.

122. DADA2 (2018). Frequently asked questions. Internet. https://benjjneb.github.io/dada2/faq.html.

123. Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. *75*, 7537–7541. https://doi.org/10.1128/AEM.01541-09.

124. Westcott, S., and Schloss, P.D. (2021). sracommand.cpp" (GitHub). In Mothur (Version ba42f8d). https://github.com/mothur/mothur/blob/ba42f8ddea4f30e8cf261633dd0ea5133f1f559a/source/commands/sracommand.cpp#L80.

125. Lee, M. (2019). Happy Belly Bioinformatics: an open-source resource dedicated to helping biologists utilize bioinformatics. J. Open Source Educ. *2*, 53. https://doi.org/10.21105/jose.00053.

126. Callahan, B. A DADA2 workflow for Big Data (1.4 or later). https://benjjneb.github.io/dada2/bigdata.html.

127. Hitch, T.C.A., Bisdorf, K., Afrizal, A., Riedel, T., Overmann, J., Strowig, T., and Clavel, T. (2022). A taxonomic note on the genus Prevotella: description of four novel genera and emended description of the genera Hallella and Xylanibacter. Syst. Appl. Microbiol. *45*, 126354. https://doi.org/10.1016/j.syapm.2022.126354.

128. Bik, E.M., Bird, S.W., Bustamante, J.P., Leon, L.E., Nieto, P.A., Addae, K., Alegría-Mera, V., Bravo, C., Bravo, D., Cardenas, J.P., et al. (2019). A novel sequencing-based vaginal health assay combining self-sampling, HPV detection and genotyping, STI detection, and vaginal microbiome analysis. PLoS One *14*, e0215945. https://doi.org/10.1371/journal.pone.0215945.

129. Lim, M.Y., Park, Y.-S., Kim, J.-H., and Nam, Y.-D. (2020). Evaluation of fecal DNA extraction protocols for human gut microbiome studies. BMC Microbiol. *20*, 212. https://doi.org/10.1186/s12866-020-01894-5.

130. PowerSoil, M.A. DNA kit. https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/microbial-dna/magattract-powersoil-dna-isolation-kit.

131. Cao, Q., Sun, X., Rajesh, K., Chalasani, N., Gelow, K., Katz, B., Shah, V.H., Sanyal, A.J., and Smirnova, E. (2020). Effects of rare microbiome taxa filtering on statistical analysis. Front. Microbiol. *11*, 607325. https://doi.org/10.3389/fmicb.2020.607325.

132. Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584. https://doi.org/10.7717/peerj.2584.

133. Armstrong, G., Cantrell, K., Huang, S., McDonald, D., Haiminen, N., Carrieri, A.P., Zhu, Q., Gonzalez, A., McGrath, I., Beck, K.L., et al. (2021). Efficient computation of Faith's phylogenetic diversity with applications in characterizing microbiomes. Genome Res. *31*, 2131–2137. https://doi.org/10.1101/gr.275777.121.

134. McDonald, D., Vázquez-Baeza, Y., Koslicki, D., McClelland, J., Reeve, N., Xu, Z., Gonzalez, A., and Knight, R. (2018). Striped UniFrac: enabling microbiome analysis at unprecedented scale. Nat. Methods *15*, 847–848. https://doi.org/10.1038/s41592-018-0187-8.

135. Sfiligoi, I., Armstrong, G., Gonzalez, A., McDonald, D., and Knight, R. (2022). Optimizing UniFrac with OpenACC yields greater than one thousand times speed increase. mSystems *7*, e0002822. https://doi.org/10.1128/msystems.00028-22.

136. Taxonomy browser (human gut metagenome). https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=408170.

137. Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2016). GenBank. Nucleic Acids Res. *44*, D67–D72. https://doi.org/10.1093/nar/gkv1276.

138. Ehresmann, C., Stiegler, P., Fellner, P., and Ebel, J.P. (1972). The determination of the primary structure of the 16S ribosomal RNA of Escherichia coli. 2. Nucleotide sequences of products from partial enzymatic hydrolysis. Biochimie *54*, 901–967. https://doi.org/10.1016/s0300-9084(72)80007-5.

139. Zhao, M., Lee, W.-P., Garrison, E.P., and Marth, G.T. (2013). SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. PLoS One *8*, e82138. https://doi.org/10.1371/journal.pone.0082138.

140. Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods *69*, 330–339. https://doi.org/10.1016/j.mimet.2007.02.005.

141. Sanders, H.L. (1968). Marine benthic diversity: A comparative study. Am. Nat. *102*, 243–282. https://doi.org/10.1086/282541.

142. Li, K., Bihan, M., Yooseph, S., and Methé, B.A. (2012). Analyses of the microbial diversity across the human microbiome. PLoS One *7*, e32118. https://doi.org/10.1371/journal.pone.0032118.

143. Friedman, J., and Alm, E.J. (2012). Inferring correlation networks from genomic survey data. PLoS Comput. Biol. *8*, e1002687. https://doi.org/10.1371/journal.pcbi.1002687.

144. Janitza, S., Celik, E., and Boulesteix, A.-L. (2015). A Computationally Fast Variable Importance Test for Random Forests for High-Dimensional Data. Paper Available from Universitätsbibliothek Der Ludwig-Maximilians-Universität München. https://doi.org/10.5282/UBM/EPUB.25587.

145. Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

146. Naing, L., Winn, T., and Rusli, B.N. (2006). Practical issues in calculating the sample size for prevalence studies. Arch. Orofac. Sci., 9–14.

147. Jovanovic, B.D., and Levy, P.S. (1997). A look at the rule of three. Am. Stat. *51*, 137–139. https://doi.org/10.1080/00031305.1997.10473947.

148. Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biom. Bull. *1*, 80–83. https://doi.org/10.2307/3001968.

149. Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika *75*, 800–802. https://doi.org/10.2307/2336325.

150. Schloss, P.D. (2023). Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses. Preprint at bioRxiv. https://doi.org/10.1101/2023.06.23.546313.

151. Borg, I., and Groenen, P.J.F. (2005). Modern Multidimensional Scaling: Theory and Applications (Springer Science & Business Media) https://doi.org/10.1007/0-387-28981-x.

152. Davies, D.L., and Bouldin, D.W. (1979). A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. *1*, 224–227. https://doi.org/10.1109/tpami.1979.4766909.

153. Huang, R., Soneson, C., Ernst, F.G.M., Rue-Albrecht, K.C., Yu, G., Hicks, S.C., and Robinson, M.D. (2020). TreeSummarizedExperiment: a S4 class for data with hierarchical structure. F1000Res *9*, 1246. https://doi.org/10.12688/f1000research.26669.2.

154. Šavrič, B., Patterson, T., and Jenny, B. (2019). The Equal Earth map projection. Int. J. Geogr. Inf. Sci. *33*, 454–465. https://doi.org/10.1080/13658816.2018.1504949.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Processed data integrated from 482 studies, plus sample- and project-level annotations	This paper	https://doi.org/10.5281/zenodo.10452633
Software and algorithms		
SILVA reference database, release 138 (Ref NR 99)	Quast et al.[100]	RRID:SCR_006423; https://www.arb-silva.de/no_cache/download/archive/release_138/ARB_files/
Greengenes2 reference database, release 2022.10	McDonald et al.[101]	https://greengenes2.ucsd.edu/
SRA Toolkit	National Center for Biotechnology Information[102]	RRID:SCR_024350; https://github.com/ncbi/sra-tools
DADA2, version 1.14.0	Callahan et al.[41]	RRID:SCR_023519; https://benjjneb.github.io/dada2/index.html
Docker image with DADA2 and dependencies	This paper	https://hub.docker.com/repository/docker/blekhmanlab/dada2/general
R programming language, versions 3.6, 4.2.2 and 4.3.1	R Core Team	RRID:SCR_001905; https://www.R-project.org/
PubTator 3.0	Wei et al.[103]	https://www.ncbi.nlm.nih.gov/research/pubtator3/
BioMedBERT	Gu et al.[104]	https://github.com/BioMedBERT/biomedbert
QIIME2, version 2023.9 (amplicon distribution)	Bolyen et al.[105]	RRID:SCR_021258; https://qiime2.org/
scikit-bio Python package, version 0.5.8	Rideout et al.[106]	https://scikit.bio/
FastSpar, version 1.0.0	Watts et al.[107]	https://github.com/scwatts/FastSpar
ranger R package, version 0.16.0	Wright and Ziegler[108]	RRID:SCR_022521; https://github.com/imbs-hl/ranger
caret R package, version 6.0-94	Kuhn[109]	RRID:SCR_021138; https://topepo.github.io/caret/
scikit-learn Python package, version 1.2.1	Pedregosa et al.[110]	RRID:SCR_002577; https://scikit-learn.org/stable/
vegan R package, version 2.6.4	Oksanen et al.[111]	RRID:SCR_011950; https://github.com/vegandevs/vegan
bigmds R package, version 2.0.1	Delicado and Pachón-García[112]	https://github.com/pachoning/bigmds
evident R package, version 0.4.0	Rahman et al.[113]	https://github.com/biocore/evident/
lme4 R package	Bates et al.[114]	RRID:SCR_015654; https://github.com/lme4/lme4
lmerTest R package	Kuznetsova et al.[115]	RRID:SCR_015656; https://github.com/runehaubo/lmerTestR
clusterSim R package, version 0.51-3	Walesiak and Dudek[116]	RRID:SCR_023743; https://doi.org/10.32614/CRAN.package.clusterSim
rnaturalearth R package	Massicotte and Smith[117]	https://ropensci.github.io/rnaturalearth/
ggplot2 R package	Wickham[118]	RRID:SCR_014601; https://ggplot2.tidyverse.org/
patchwork R package	Pedersen[119]	RRID:SCR_024826; https://patchwork.data-imaginist.com/
dplyr R package	Wickham et al.[120]	https://github.com/tidyverse/dplyr

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Sample selection

All sequencing data processed for the compendium and described here were publicly available from the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/). We retrieved metadata for all SRA samples categorized in the NCBI Taxonomy[121] under "human gut metagenome" on 9 October 2021. After removing samples that could not be associated with a BioProject or sequencing run, we

selected only those for which the library source was "genomic" or "metagenomic" (excluding values such as "metatranscriptomic," "synthetic" and "other"). Of these, we then selected only samples with a "library strategy" value of "amplicon" (and not values such as "WGS" and "RNA-Seq"). This left 245,627 samples across 1,437 BioProjects (Figure 1A). We then excluded samples from BioProjects that contained less than 50 samples meeting our criteria, leaving us with a list of 234,875 samples in 811 BioProjects. We did not attempt to infer longitudinal subject information from available metadata to account for hosts with multiple samples: the "host_subject_id" field is unavailable for 74% of samples; the remaining tags are inconsistently specified but are included in the compendium metadata.

Because pyrosequencing technologies developed by companies such as 454 Life Sciences and Ion Torrent require different processing parameters,[122] we then removed BioProjects containing pyrosequencing data and other sequencing instruments that use processes dissimilar from Illumina sequencing, such as MinION. We used the SRA Toolkit APIs maintained by NCBI to retrieve sequencing instrument information for each sample and evaluated BioProjects that reported using 454 or Ion Torrent instruments. We found 19 such instruments ("454 GS FLX Titanium," "Ion Torrent PGM," "454 GS FLX," etc.), but manual review revealed one instrument was consistently mislabeled: In many projects (BioProject: PRJNA685914, PRJNA605031, for example) the sequencing instrument was reported as "454 GS" even though the authors report elsewhere in the BioProject that they used sequencers such as Illumina's MiSeq, a platform we wanted to include. We found many of these BioProjects performed their analysis using Mothur,[123] a popular microbiome analysis tool, and "454 GS" is Mothur's default entry for the "instrument" field when uploading to SRA.[124] To avoid removing relevant samples, we removed BioProjects reporting using any of the pyrosequencing instruments but kept those reporting "454 GS."

METHOD DETAILS

Data processing

Sequencing data retrieval

Samples were retrieved one BioProject at a time. Each BioProject had a file listing all accessions for runs associated with samples meeting the criteria described above. This file was used as the input for the "fasterq-dump" tool[102] from the SRA Toolkit, which downloads and converts data files from the Sequence Read Archive (SRA) for downstream processing. We used samples from all available INSDC members; of the completed samples, 126,452 were from SRA (74%), 38,971 were from the European Nucleotide Archive, and 5249 were from the DNA Data Bank of Japan. We did not include projects in which more than 10 samples failed to download, which was generally caused by files missing from SRA.

Amplicon processing

If the number of files for forward reads matched the number of files for reverse reads, we processed the BioProject as paired-end sequencing. If there was a mismatch, or there were no reverse reads, we processed the BioProject as single-end data. In both cases, we used DADA2 v1.14.0 to process the data,[41] inside a container running R 3.6. We used general settings that we believed would be effective across many BioProjects,[125] aiming to maintain as many samples as possible while excluding low-quality data: We did not trim a set number of bases from either end, nor did we limit the maximum length of a read. We removed reads shorter than 20 nucleotides, reads with any ambiguous ("N") base calls, and any reads that aligned to the phiX genome (if present, almost certainly as a control in Illumina sequencing runs). We also disabled quality-based truncation of reads. Paired-end reads were merged with a minimum overlap of 20 bases. In some cases, the process of merging reads failed, and close to zero forward reads were merged with their paired reverse read, likely due to sequencing strategies that involve non-overlapping reads or reads with very minimal overlap. For BioProjects in which fewer than 50% of forward reads were merged successfully, the reverse reads were removed, and these BioProjects were re-processed as single-end data. If the number of forward reads for a sample did not match its number of reverse reads, we used DADA2 to detect the sequence identifier field in the FASTQ files to match the samples that could be salvaged. If this was unsuccessful, we removed the reverse reads and reprocessed these as single-end data as well. We believe this was the most reliable way to process this data, given the lack of information about project-level sequencing strategies. The merging process in paired-end datasets would be much more effective with more knowledge of study design, particularly in cases where the amplicon length was greater than the read length and the paired-end reads did not overlap. In addition, DADA2 recommends building separate error models for each sequencing run,[126] but only BioProject could be reliably inferred, which means ASV inference at the study level may not capture run-level patterns. Taxonomic assignment was performed by DADA2 using the SILVA database release 138.0.[100] Phylum names in this work were updated to reflect the 2021 changes to the names of bacterial phyla such as Firmicutes (now "Bacillota").[45] The most recent version of the Human Microbiome Compendium (1.1.0; see resource availability) now includes a taxonomic table with classifications using the more recent SILVA version 138.2, which includes more recent changes such as the separation of the Lactobacillaceae family[72] and *Prevotella* genus[127] into distinct clades. When comparing ASVs that were classified in a different phylum in the most recent version, we found 2,226,534 of 2,294,534 ASVs (97.0%) were due to nomenclature updates. We also include classifications using the Greengenes2 reference (see "Phylogenetic analysis," below).

Though we removed obviously non-applicable samples (mycobiome assays using the 18S rRNA gene rather than 16S, for example), we did not pursue more stringent filtering. To minimize the number of legitimate samples removed from the compendium, we avoided making discretionary decisions about removing samples or BioProjects from the compendium that we encountered during analysis—for example, 199 samples from BioProject: PRJEB25853 are included because they were deposited in the "human gut

metagenome" category, though manual review of the metadata shows the samples are actually from a diagnostic assay of the vaginal microbiome.[128] During a manual review of project-linked publications, we found 6 projects (1.2%) containing 2488 samples (1.5%) that deposited data in the "human gut metagenome" category but reported samples taken from non-human sources such as mice or incubators. Another 7 projects (1.5%) with 2086 samples (1.2%) reported sequencing human microbiome samples drawn from somewhere other than the gut (Table S3).

Pipeline success

Due to resource constraints, we did not attempt to process BioProjects with fewer than 50 samples, which accounted for 10,752 out of 245,627 samples initially reviewed for inclusion (Figure 1A). Of the 234,875 samples we processed, we found 31,887 samples (13.6%) contained non-applicable data—BioProjects that targeted fungi or archaea, for example, and mislabeled BioProjects that used shotgun or nanopore sequencing instead of 16S rRNA gene amplicon sequencing. Another 31,509 samples (13.4%) were excluded because extracting acceptable results, if possible, would have required manual intervention and more knowledge of the sequencing strategy: DADA2 identified excessive chimeric reads in some BioProjects, for example, and we excluded any BioProject in which at least five of the first 10 samples processed contained more than 25% chimeric reads. Several BioProjects were also excluded because they contained samples associated with multiple sequencing runs or were associated with samples that could not be downloaded. 807 samples were removed from BioProjects because all reads were filtered out.

Annotation and filtering

Publication links

BioProjects were manually annotated with what evaluators believed to be the original publication associated with each project—in general, this was the earliest publication that mentions the BioProject accession (Table S3). For each project, we first used Google Scholar to search for the BioProject accession (e.g. "PRJNA682076") and evaluated the resulting papers to see the context in which the project is referenced: If there were no results, or the only papers available referred to the BioProject as third-party data, we then searched for other accession codes that may have been used to identify the same project: Some were cited in publications using the BioProject's associated SRA study (e.g. "SRP295653"), and several were identified by the "submission" code (e.g. "SRA1166126"). We were able to associate 351 BioProjects with 345 unique publications.

Extraction kit inference

To retrieve text describing extraction kits from project-associated publications, we used the PubTator API[103] to download the full texts and extracted relevant descriptions, typically found in the Materials and Methods section. In cases where the extraction kit information was in the supplementary text, it was retrieved manually. The extracted text snippets were then preprocessed by removing stop words, numbers, and punctuation, and converting all words to lowercase.

Next, we developed a semantic matching approach to identify the presence of the names of extraction kits within these descriptions. Given an extraction kit name (with n words) and a study description (from the literature; with m words), for each kit word e_i in $\{e_1, e_2,..., e_n\}$, we recorded its similarity to the closest study word among $\{d_1, d_2,..., d_m\}$. Similarly, for each study word d_j, we recorded its similarity to the closest kit word. Similarity between a pair of words was defined as the cosine similarity between their word embeddings generated using BioMedBERT.[104] The overall similarity of the kit-study pair was calculated by averaging these best word-pair similarities, weighted by each word's "informativeness" (quantified using its term-frequency inverse document frequency; TF-IDF). Finally, we used the Stouffer's z-score method to correct the similarity score for each kit-study pair to account for background signals. Specifically, the corrected kit-study score is a combination of two z-scores of the original similarity score calculated based on the μ and σ of two distributions: that kit's similarity to all studies and that study's similarity to all kits. Finally, each study description was annotated to the extraction kit with which it had the highest corrected similarity score. To ensure accuracy, we manually reviewed the descriptions with very low similarity scores (the bottom 10%).

We also annotated each project with an indicator of whether the extraction protocol included a bead-beating step, in light of the findings that mechanical lysis can affect observed abundances, particularly of gram-positive bacteria.[129] A project was flagged with bead-beating if the relevant snippet of the methods section included the words "bead" or "mechanical," or if the inferred extraction kit explicitly includes mechanical lysis in its protocol (e.g. the Qiagen MagAttract PowerSoil DNA Kit[130]). In cases where we were unable to make a confident automated inference about extraction kit, both kit and bead-beating status were annotated manually.

Country and region inference

We were able to obtain a "geo_loc_name" tag from the NCBI BioSample database (https://www.ncbi.nlm.nih.gov/biosample/) for 155,584 of 168,464 samples (92.4%). These tags held 455 unique values, which were associated with countries by manually reviewing the values and associating them with a country. Most tags contained enough context to confidently assign a country name: The most common tag value was "usa:new york" found in 14,142 samples from six BioProjects, followed by "usa" (12,510 samples in 46 BioProjects). However, the third-most common tag was "missing" (5532 samples; 27 BioProjects), and other tags such as "not applicable" (4317 samples) and "not available" (3481 samples) appeared many times. Overall, we were able to assign a country to 447 of the 455 unique values (98.2%) representing 153,152 samples (90.9%). In total, we found 68 countries represented in the geo_loc_name values—to simplify comparisons, we consolidated these assignments into eight world regions defined by the United Nations Sustainable Development Goals (SDG) program (Figure 2A).

Dataset for analysis

We combined BioProject-level taxonomy tables into one large matrix containing 168,464 samples from 482 BioProjects for rows and 4018 unique taxonomic identifiers for columns. For most of the analysis reported in this paper, we applied additional quality control steps: First, we removed 16,781 samples (10%) with fewer than 10,000 reads. To reduce sparsity introduced by exceedingly rare taxa, we then removed 2018 taxonomic entries (50%) with fewer than 1000 total reads across all remaining samples, and a further 578 taxa (14% of the original total) that were detected in fewer than 100 samples.[131] After these taxa were removed, we removed another 19 samples that now had less than 10,000 reads. (The removal of these 19 samples did not push any more taxa below the above thresholds.) We then evaluated the proportion of reads in each sample for which a taxonomic assignment could not be assigned at least the phylum level. We removed 943 samples for which more than 10% of the sample's reads had an unassigned phylum. This left us with 150,721 samples and 1422 taxa. Unless otherwise noted, this subset of the compendium was used in all presented analyses.

Microbiome analysis

Phylogenetic analysis

To generate the alpha- and beta-diversity measures used in the effect-size estimation (Figure 3G) and PERMANOVA analysis, ASVs were mapped against the Greengenes2[101] 2022.10 backbone using the q2-greengenes2 2023.9 "non-v4-16s" QIIME 2 2023.9 (amplicon distribution)[105] action at 99% similarity, which in turn executes a closed-reference clustering using q2-vsearch 2023.9.[132] Taxonomy was derived using the q2-greengenes2 "taxonomy-from-table" action with the clustered table. To mitigate differences in read count between regions, the clustered table was rarefied to 1000 sequences per sample with q2-feature-table 2023.9 "rarefy" action. Faith's PD was computed with q2-diversity's "alpha-phylogenetic" using SFPhD[133] implementation in the Striped UniFrac library.[134] Weighted normalized UniFrac was computed with single-precision floating point values using the cache-optimized modifications of Striped UniFrac,[135] with 32 CPU cores. For the phylogenetic gain analysis in Figure S2, we selected 1000 random samples from each region (to mitigate differences in sample counts) and combined each region's samples into a single "regional" sample. We recalculated the Faith's PD score for each of these regional samples, to serve as a baseline. We then recalculated for each pairwise combination of regions. Subtracting each region's baseline score from these combinations yields an estimate of the evolutionary distance between taxa appearing in one region but not another. In short, each region's baseline measures the length of all branches of the phylogenetic tree encompassing all taxa observed in its samples, and the gain observed between two regions is how much larger a region's tree becomes when it is combined with the tree of a second region. We repeated this subsampling and comparison 1000 times for each pair to obtain the mean gain (Figure S2). For each, we then evaluated the proportional gain by dividing the gain by the region's baseline, to establish each region's percentage increase when combined with a second region.

World region signature determination

We used principal component analysis to extract the taxa that are most closely linked to overall variance observed in the microbiomes of each region, which results in a heuristic we refer to here as the variance score. This score ranges from 0, indicating no relationship, to 1, indicating that the major axes of variation in the region also perfectly explain the variation observed in that taxon. We started with the taxonomic table of each region, applied the robust centered log ratio transformation to the read counts,[53] then applied PCA to each region separately. We kept as many principal components (PCs) as was required to account for at least 50% of variance in each region's data. We then summed the resulting eigenvectors for each taxon among the selected PCs. These were used to calculate a score for each taxon observed in that region that indicates how much variance of that taxon was explained by the selected PCs. These proportions range between 0 and 1 and are used as the variance score for each taxon. The key assumption is that if a subset of principal components explains the majority of variance in the dataset, and those same PCs explain a high proportion of variance for a single taxon, then that taxon is more strongly linked to overall variability in the dataset than taxa that are poorly captured by the selected PCs.

Corpus evaluation

When evaluating the human gut microbiome samples available from BioSample, visualized in Figures 3A–3C, we first pulled available metadata for all sequencing runs deposited in SRA under the "human gut metagenome" category ("txid408170") as of 6 March 2024.[136] We used the "library_strategy" property to infer the type of sequencing performed: we assume the "AMPLICON" library strategy indicates a library using one of the common amplification-based approaches such as 16S and 18S rRNA gene sequencing. The "WGS" library strategy is assumed to be shotgun metagenomic sequencing, though there appears to be at least one single-cell sequencing project here as well. We then used the NCBI eUtils API to annotate these samples with information from each run's associated BioSample, which included the "geo_loc_name" field used to infer region of origin. One project, BioProject: PRJNA803937, was excluded from the "WGS" sample collection because it contained more than 50,000 samples (more than all other projects combined) of non-applicable sequencing.

Amplicon inference

We used the amplicon sequence variants (ASVs) generated by DADA2 for each BioProject to infer the sequencing strategy used by each BioProject, primarily determining which of the nine hypervariable regions were targeted for amplification (Figures 3D and 3E), but also the size of the amplicon targeted. To do this for each BioProject, we retrieved the sequence of all ASVs detected in that BioProject. We aligned each ASV to the sequence of the full *E. coli* 16S rRNA gene sequence, obtained from GenBank[137] under accession J01859.1,[138] using the striped Smith–Waterman library[139] integrated into scikit-bio v0.5.8.[106] If the optimal alignment

covered 70% or less of the full ASV sequence, the alignment was discarded and the ASV was classified as unknown. For the remaining ASVs, the coordinates of the alignment were used to determine which of the nine hypervariable regions were covered by the ASV, as defined by Chakravorty et al.[140] A region was considered to be covered if more than half of its length was covered by the ASV—for example, if an ASV's alignment starts just before the beginning of V3 and ends 60 bases into the 107-base V4 region, that ASV would be classified as "V3-V4." If the same alignment ended only 20 bases into the V4 region, that ASV would be classified only as "V3." Beginning region and ending region (i.e. "V3" and "V4" from this example) were tallied separately. If more than half of all ASVs were categorized in a single starting region, that region was determined to be the starting region for the entire BioProject. If more than half of all ASVs were categorized in the same ending region, that region was determined to be the ending region for the entire BioProject. In situations where the threshold was met for only the starting or ending region (generally because of wide variation in ASV length), the opposite region was determined using the known region and the average ASV length. In situations where the ending region was determined to be before the starting region, the assignments were discarded under the assumption that this indicated multiple sets of primers were used.

Rarefaction analysis, taxon discovery rate

For the analysis described in Figure 1H, we estimated the relationship between compendium size (in number of samples) and total taxa observed by building a sample-based rarefaction curve for each taxonomic rank.[49,51,52] Specifically, we built simulated compendia of various sizes between 1 and 150,000 by subsampling the filtered analysis dataset and counting the number of unique taxa observed in each subsample at each taxonomic level. We repeated each compendium size simulation 150 times and plotted the mean observed taxa at each taxonomic level; this allowed us to build curves plotting observed taxa against compendium size.[141,142] Generally, the x-axis of this curve would plot *total reads*, rather than *total samples*, to account for variation in the number of observations (i.e. a single read from a single microbe) in different size compendia. Here, we visualized total samples instead to incorporate the differences in read depth actually present in the data. The trade-off is that these metrics will likely underestimate *future* taxa observed if used to extrapolate forward into larger compendium sizes, if the distribution of reads per sample (Figure 1B) trends upward with falling sequencing costs.

World region taxon discovery rate

To generate the region-level taxon discovery rate curves shown in Figure 4A, we randomly selected *n* microbiome samples from the region and recorded the number of unique taxa present in these samples. We repeated this subsampling 1,000 times for each sample size. The data illustrated in Figure 4A (inset) was generated using the same sampling strategy. After recording the total number of reads in the selected samples for each iteration, we calculated the number of taxa discovered per one million reads for each repetition and reported the mean. When measuring the number of unidentified taxa in each region (Figure 4B), we controlled for sequencing depth by rarefying all samples to 10,000 reads and counting the number of taxa with an "NA" assignment at each taxonomic level in every sample. We repeated this 1,000 times and reported the mean and standard deviation.

Microbe correlation networks

To generate the networks shown in Figures 4E and 4F, we first split the dataset into two groups: samples from Europe and Northern America, and samples with a region identifier other than Europe and Northern America. Samples without a region identifier were excluded. We then selected genera with an average relative abundance greater than 0.5% in the compendium to analyze. We then used FastSpar,[107] a C++ implementation of SparCC,[143] to calculate pairwise correlations between each pair of taxa. We used FastSpar's bootstrapping method with 1000 repetitions to estimate p-values. Any correlation with an estimated p-value greater than 0.001 (the minimum p-value with 1,000 permutations) was considered non-significant.

Region-level classifiers

The classifiers described in Figures 6C and 6D were trained using the ranger R package (version 0.16.0).[108] Samples with a region listed as "unknown" (n=28,192) or "Oceania" (n=4) were excluded from the analysis. Taxa were filtered based on their minimum prevalence (proportion of samples where the taxon has more than zero reads) and their minimum abundance (mean read count across samples for a given taxon) among each of the seven world regions considered. Filtering and random forest parameters were chosen by grid search. The values considered for each parameter passed to ranger were: min.node.size (1, 5, 10), sample.fraction (1, 0.75, 0.5), mtry (90,120, 150, 180, 210, 240), num.trees (800, 1000, 2000). The values considered for filtering were for the abundance threshold (0, 0.05, 0.1, 0.2, 0.5) and prevalence threshold (0, 0.01, 0.1, 0.2, 0.3). The objective function of the grid search was the p-value given by the "confusionMatrix" function in the caret R package (version 6.0-94),[109] which represents a binomial test that H0: the classifier is as accurate as the no-information-rate, which is the accuracy of a hypothetical classifier that always classifies samples to be the most-represented class in the dataset, in this case "Europe and Northern America." After hyperparameter tuning, the values used were: min.node.size=1, sample.fraction=1, mtry=210, num.trees=2000, prevalence threshold=0.01, and abundance threshold=0.05. After filtering and data were CLR transformed, 1050 taxa remained for all of the subsequent classifiers. Samples for all classifiers were randomly split into 80% training and 20% test sets. Samples were given weights in training inversely proportional to the representation of their region in the overall dataset. The weight given to samples of a particular class was $w_i = \frac{1/N_i}{\sum_{j=1}^{n} 1/N_j}$, where N_i is the number of samples in class i, and n is the total number of classes.

ROC curves were generated by plotting the false-positive-rate against the true-positive-rate for the range of possible classification thresholds from 0 to 1 with a step of 0.01. For each classification threshold, votes for each tree in the random forest were counted by setting predict.all=TRUE in the predict.ranger function, and a classifier returned a positive result if the proportion of votes was greater

than or equal to the classification threshold. AUC (area under the ROC curve) was computed by using the trapezoidal rule to integrate the ROC curve. An overall classifier was first constructed by predicting the "region" metadata field from microbiome abundances. Region–region classifiers were constructed by training a binary classifier on samples only from the two regions being considered. One-vs-all classifiers were constructed by training a binary classifier that predicted whether a given sample was either in or not in the given region. Variable importance was computed using the method defined by Janitza et al.[144] as implemented in the ranger package. We used Fisher's exact tests to determine the overlap between the top features used in the model and the taxa that were identified as being differentially abundant between regions with an adjusted p-value less than 0.1. For each region (n=7), we constructed a contingency table in which each row was a taxon that was differentially abundant in that region compared to any other region. Each column was a taxon from the top *N* (either 50 or 100) most important features in the random forest one-versus-all model, as computed with ranger. For each of these 14 contingency tables, we used the fisher.test function in R to determine the significance of overlap. To determine the relative contributions of the taxa with the highest importance scores to model performance, we trained one-versus-all classifiers for each region using the first *N* taxa, where *N* ranged from only the 5 most important taxa for that region to 1050 (all taxa after filtering) and determined the AUC for each classifier. We trained logistic regression one-versus-all models using the scikit-learn Python package (version 1.2.1),[110] using the same abundance and prevalence filtering hyperparameters as with the random forest model, on the CLR-transformed microbiome abundances. AUC was computed with an 80%-20% train-test split.

QUANTIFICATION AND STATISTICAL ANALYSIS

Calculation of Shannon diversity

To calculate the Shannon index for individual samples, illustrated in Figures 1G and 2C, we used the "diversity" function in the vegan R package v2.6.4,[111,145] using natural logarithms and all columns in the filtered dataset—that is, ASVs were consolidated if their taxonomic assignments matched exactly, but counts were not consolidated at any single taxonomic level. This data was also used to calculate each sample's Simpson's Index and species count via vegan's "diversity" and "specnumber" functions respectively.

Country metadata accuracy estimate

To assess the accuracy of our process for associating samples with their country of origin, we selected a random sample of microbiomes and manually determined the country of origin (Table S5), primarily by finding publications referencing the data but also using other metadata associated with the samples and parent BioProjects. We found the "geo_loc_name" tags to reliably include country names, which gave us confidence in our ability to infer country from tag, but the additional factor of interest is whether samples were mislabeled by the original authors. Between these two factors, we assumed 95% accuracy for our sample-size calculation, which was designed to detect this level of accuracy at a precision of ±5% at a 95 percent confidence interval.[146] This results in a sample size of 73.0; after accounting for a 25% "dropout" rate (samples with a "geo_loc_name" tag but no other means with which to verify it), our revised sample size was 97.3. Because some BioProjects are 60 times larger than others, we wanted to mitigate the effect of selecting many samples from one large, single-country BioProject by first selecting 100 random BioProjects, then selecting one sample from each project to evaluate. To verify the accuracy of our inferences, we first looked for an explicit statement of the study's country of origin in the project description in the BioProject database. If this did not yield an answer, we looked to any publications linked to the BioProject and searching Google Scholar for several factors indicating a link to the BioProject: First, we searched the BioProject accession, then its corresponding SRA accession (such as "ERP006059" for BioProject PRJEB6518), then its ID number (such as "bioproject 589558" for BioProject PRJDB6499), then any unique phrases from the BioProject title or description. If the paper did not explicitly state a country of origin for the subjects, we considered the classification confirmed if the paper included ethical approval from an institutional review board in that country. If none of these steps could confirm a country, we classified it as a dropout. Of 100 samples evaluated, eight were dropouts. Two were deemed not applicable because the papers described samples that were incubated prior to sequencing. Of the remaining 90, we were able to validate that all 90 had world region assignments that were confirmed by either the BioProject description or a publication associated with the data. With a 100% success rate, we can then use the rule of three[147] to estimate that the lower bound of the confidence interval (at our original 95% confidence level) is 96.67% accuracy.

World region alpha diversity comparison

We compared alpha diversity measurements (Figure 2C) between all regions using the Wilcoxon rank-sum test,[148] with multiple test correction done using Hochberg's method[149] as implemented in the R "stats" package's "wilcox.text" and "p.adjust" functions respectively. The calculations were performed using the filtered compendium dataset used in other analyses but was limited to samples with a world region annotation other than "unknown."

Rarefaction analysis, regional alpha diversity

To account for variation in sampling between regions, we performed a rarefaction analysis that allowed us to determine an average Shannon diversity while controlling for both reads per sample and samples per region (Figures 2C and S1). We estimated regional alpha diversity (in the filtered analysis dataset) by selecting 1000 random samples from each world region (enough that each region could provide all samples without replacement), then rarefied each of these samples down to 10,000 randomly selected reads each.

From these rarefied samples, we determined the mean Shannon diversity for each region, then repeated the entire process 1000 times.[150]

Principal coordinates analysis

We began by using the matrix of read counts to build a distance matrix between all samples using the robust Aitchison distance.[53] We then performed multidimensional scaling (Figure 2E) using the "divide and conquer" approach described by Delicado and Pachon-Garcia[112] and implemented in the "bigmds" R package version 2.0.1. We extracted 8 principal coordinates and used 16 points as the overlap between partitions.[151]

Analysis of variance

For the results reported in Figure 3G, we used the Evident package v0.4.0[113] to estimate the effect size of technical factors on alpha diversity. Because Cohen's *d*, used for comparisons between two groups, is not directly comparable to Cohen's *f* (used for more than two groups), we estimate that an equivalent *f* for bead-beating would be half of the observed *d*, or about 0.218. The PERMANOVA analysis of technical factors was performed on a subset of the compendium data used in other analyses; the 150,721 samples in the analysis dataset were further filtered to remove samples with an unknown value for either world region, amplicon or bead beating; samples from Oceania were also excluded because of the impractically small sample size (final n=84,782). We used the weighted UniFrac matrix described above as the input to the "fast.adonis" R package.[55]

Differential abundance analysis for amplicon choice

For the analysis presented in Figure 3H, we further filtered the dataset used for analysis for genera with a mean relative abundance of at least 0.5% and prevalence of at least 1% in any world region, resulting in 65 remaining genera. This resulted in analysis of 123,762 samples and 65 genera. We used a linear mixed model using the lme4 and lmerTest R packages[114,115] to model amplicon choice as a fixed effect and random effects for BioProject and world region. We ran a model for each taxon and each amplicon choice, in order to compare all samples that use a single amplicon to all samples that use any other amplicon.

$$\text{taxon abundance} \sim \text{amplicon} + (1 \mid \text{BioProject}) + (1 \mid \text{world region})$$

Taxon abundances were centered log-ratio transformed after adding a pseudocount of 1 to any 0 values, and subsequently scaled to a mean of 0 and standard deviation of 1. The p-values from each model were multiple-test corrected using the Benjamini–Hochberg method.

Differential abundance analysis

For the analyses illustrated in Figure 5, we further filtered the dataset used for analysis for genera with a mean relative abundance of at least 0.5% and prevalence of at least 1% in any world region, resulting in 65 remaining genera. We then filtered the samples to include only those with at least 1000 reads in the 65 genera. This resulted in analysis of 123,346 samples and 65 genera. To test these taxa for differential abundance, we used a linear mixed model using the lme4 and lmerTest R packages. We modeled world region as a fixed effect, and to account for technical artifacts included BioProject, use of bead beating, and amplicon as random effects. We ran a single model for each taxon, running each model 7 times so that each world region could be the reference variable, to enable pairwise comparison between each region–region pair as follows:

$$\text{taxon abundance} \sim \text{region} + (1 \mid \text{BioProject}) + (1 \mid \text{amplicon}) + (1 \mid \text{bead beating})$$

Taxon abundances were centered log-ratio transformed after adding a pseudocount of 1 to any 0 values, and subsequently scaled to a mean of 0 and standard deviation of 1. The p-values from each model were corrected for multiple testing using the Benjamini–Hochberg method. In Figure 5B, p-values lower than 2.2×10^{-16} were illustrated as 2.2×10^{-16}. The full results of this analysis are reported in the source data for Figure 5.

Country-level ordination plots

The PCA plots illustrated in Figure 6A were generated using a random subset of 33% of the compendium dataset, excluding samples either from Oceania or an unknown region, using the sklearn package in Python 3. Relative abundance data were transformed using the centered log-ratio transformation (CLR) after a pseudocount of 1 was added. The first eight principal components were plotted.

Cluster strength evaluation

We evaluated the clustering found in Figures 2E and 6A using the Davies–Bouldin Index[116,152] of the principal component coordinates calculated in the methods described above. The index was calculated using the clusterSim R package[116] (v0.51-3) We used permutation to estimate how the observed index compared to scores observed for random clusters. For the regional clustering, we generated 250,000 scores using the same data but with regional labels that were shuffled without replacement. For the country-level clustering, we generated 10,000 scores (Figure 6B). In both cases, the observed value was lower than the minimum score observed in the shuffled data.

ADDITIONAL RESOURCES

Website for high-level exploration

We designed a website to serve as an entry point to the Human Microbiome Compendium, hosted at https://microbiomap.org. The website displays important links to downloads and materials, provides a brief overview of the project and data, and features controls and visualizations for answering basic questions about the data. The website was implemented in TypeScript as a React single-page application, scaffolded and bundled with Vite to allow for cleaner implementation of interactive features. The D3 library was used for visualizations and interactions. The website's data is generated by a set of pre-processing "compile" scripts, which automatically download the latest Human Microbiome Compendium data files from the place of record, then restructure and pare down data into a more practical, static subset for efficient display on the web. These scripts run before any build of the website and can also be run on a schedule or on-demand.

R package implementation

We implemented an R package, MicroBioMap (https://github.com/seandavi/MicroBioMap), that provides convenient access to compendium data. Data are loaded into a Bioconductor TreeSummarizedExperiment object,[153] providing opportunities to use our compendium data with extensive Bioconductor microbiome analysis and visualization tools. The package includes documentation and example use cases.

Software tools

Most analyses were performed using R 4.2.2. Analyses requiring a high-performance computing environment, the rarefaction curves in Figure 1 and the multidimensional scaling analysis from Figure 2, used R 4.3.1. Maps use the Equal Earth projection[154] and the rnaturalearth R package.[117] Plots were generated using ggplot2,[118] patchwork,[119] and dplyr.[120]

CellPress
OPEN ACCESS

Supplemental figures

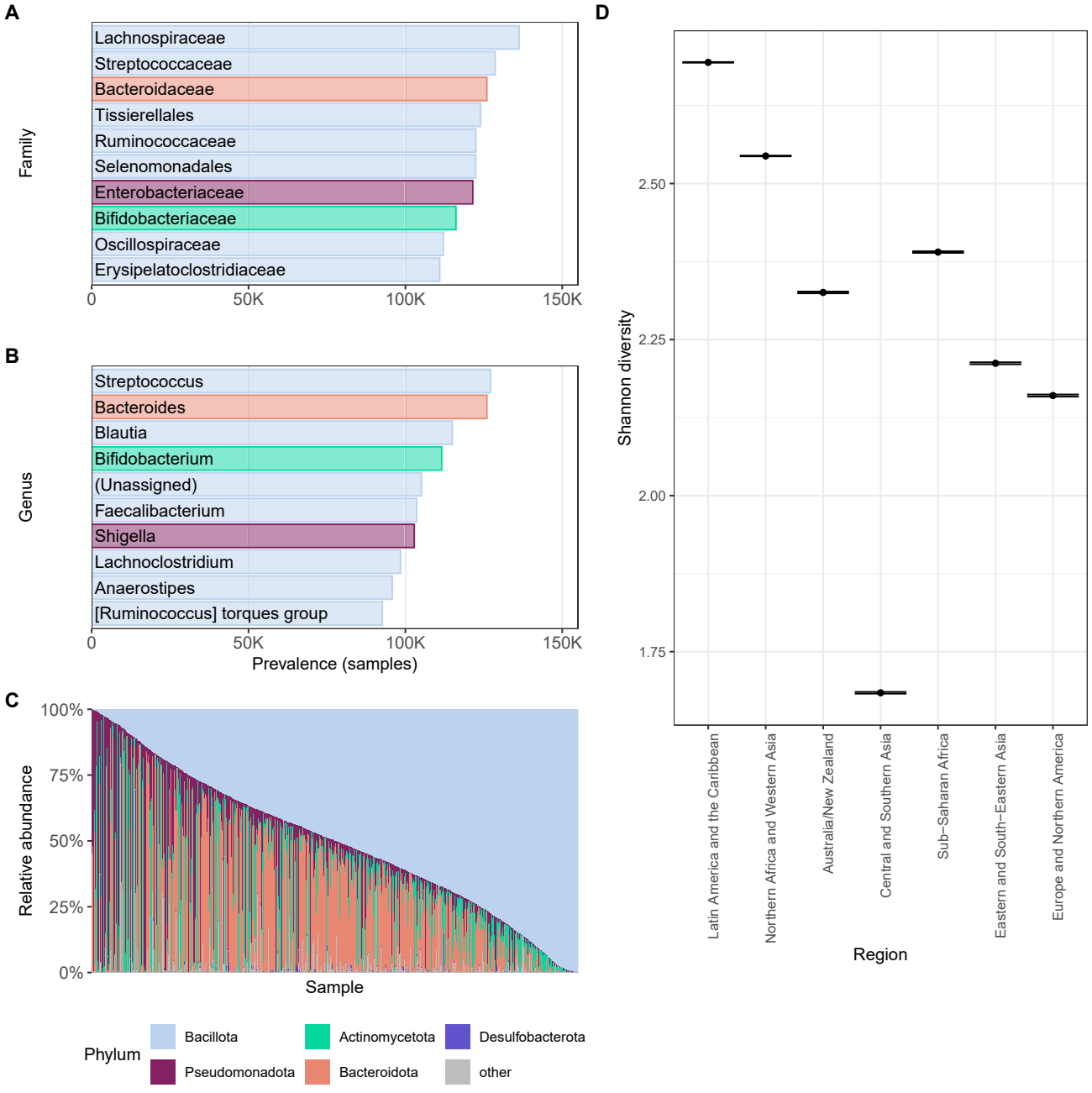


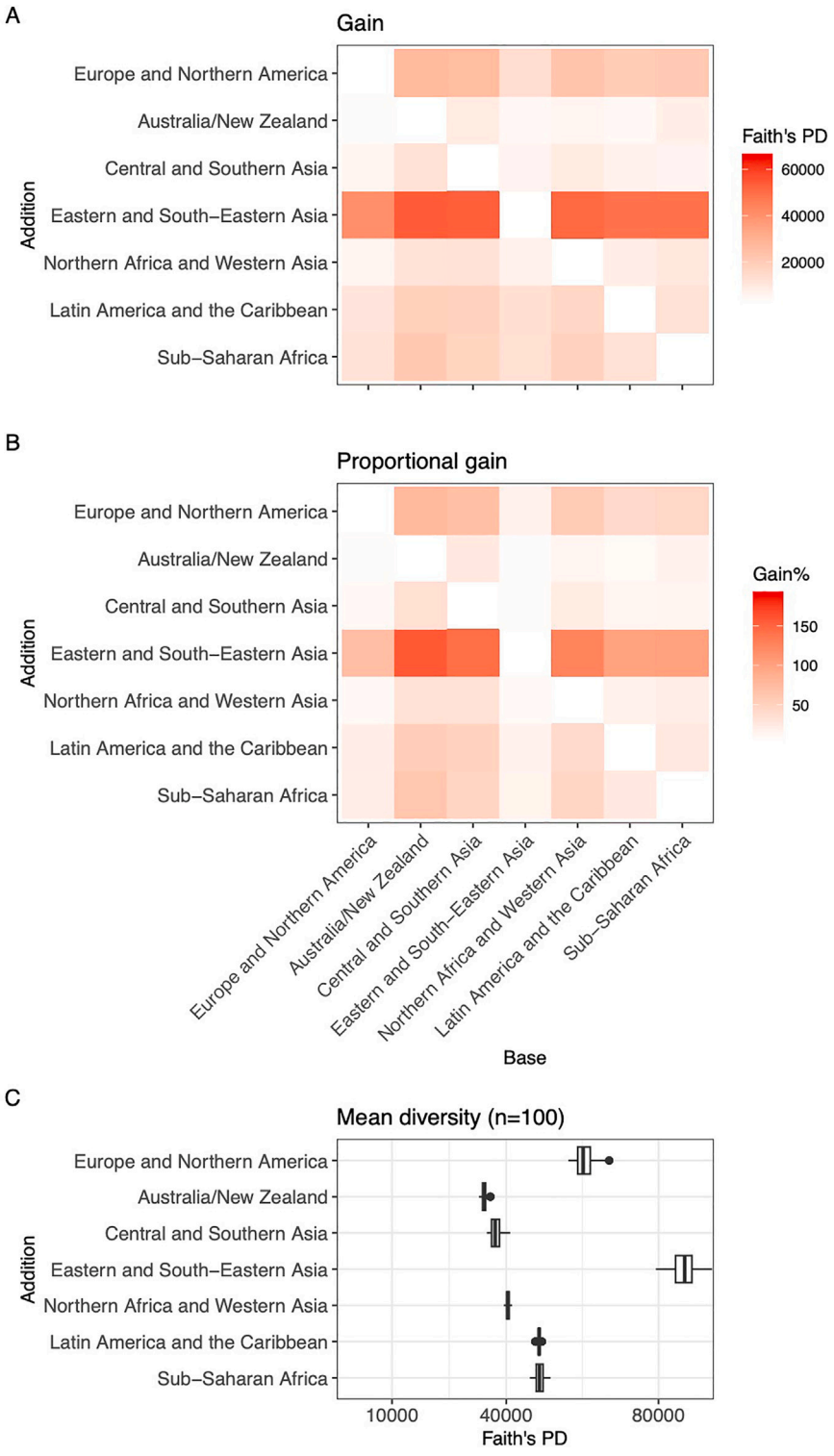
Figure S1. Commonly observed taxa, related to Figures 1, 2, and 4

(A and B) Prevalence at the family and genus levels. The most prevalent taxa observed in the compendium. This figure extends the information in Figures 1C–1E, which lists higher taxonomic orders. The reads in each sample are assigned the most specific taxonomic name possible, down to the genus level. Each panel illustrates results when these assignments are consolidated at the family level (top) and genus level (bottom), and in each, the y axis lists the 10 most prevalent taxa at that level, and the x axis indicates the number of samples in which that taxon was observed. The five most prevalent taxa in the compendium are each assigned a color, which are used in the panels to indicate the phylum of each taxon.

(legend continued on next page)

(C) Relative abundance across samples. Another version of the data illustrated in Figure 4D. A stacked bar plot illustrating the relative abundance of 5,000 randomly selected samples from the compendium. Each vertical bar represents a single sample, and the colored sections each represent the relative abundance of a single phylum in that sample. Samples are ordered by the relative abundance of Firmicutes, the most prevalent phylum

(D) Rarefaction diversity estimates. The x axis indicates Shannon diversity, and the y axis lists all regions evaluated in our rarefaction analysis. Each point indicates the mean alpha diversity calculated over 1,000 iterations (see STAR Methods). The error bars indicate the 95% confidence interval, calculated using the R "t.test" function. The means here are also illustrated as the points in Figure 2C.



(legend on next page)

Figure S2. Phylogenetic gain, related to Figure 2

(A) A heatmap indicating overall gain when the phylogenetic diversity of a "base" region (x axis) is compared with the same measure after its combination with an "addition" region (y axis). The color scale indicates the increase in the Faith's PD score of the base region. The numbers illustrated here indicate the mean gain observed over 100 random subsamplings of 1,000 samples from each region.

(B) This heatmap plots the same data, but gain is scaled to the base region's score, and color indicates the percentage increase in Faith's PD for the base region.

(C) A boxplot illustrating the Faith's PD score for each region over 100 subsampling iterations of 1,000 samples each. The x axis indicates the Faith's PD score, and the y axis lists each region measured.

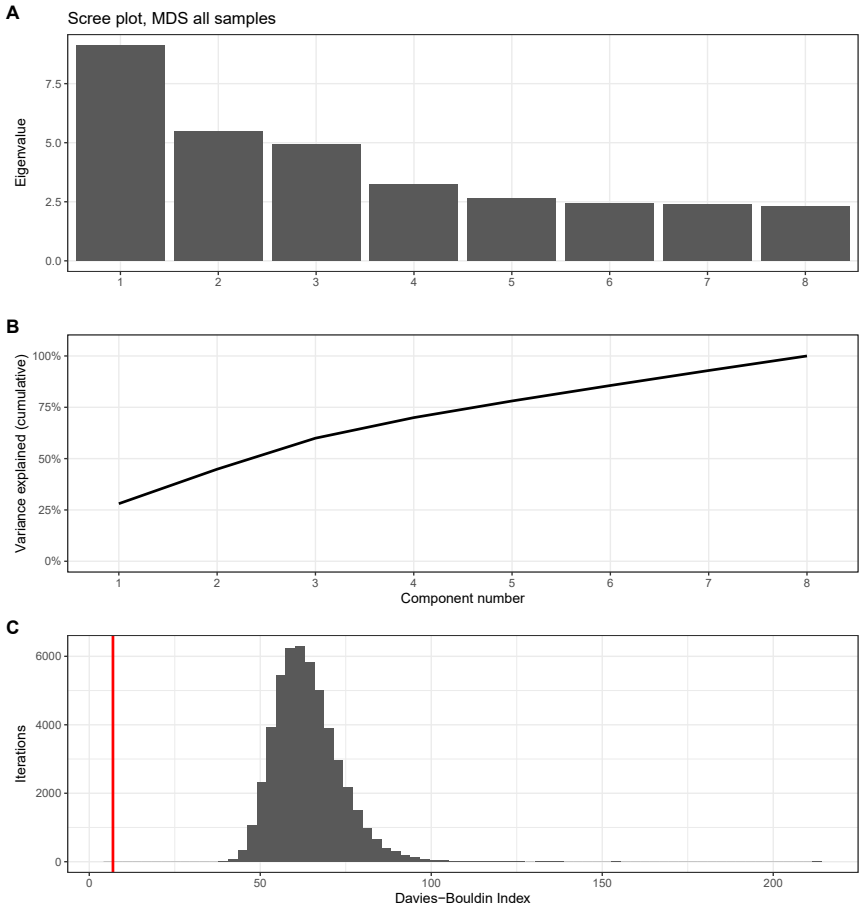


Figure S3. Scree plot for ordination analysis, related to Figure 2

(A) This plot illustrates the importance of all eight axes calculated during the ordination analysis illustrated in Figures 2E and 2F. The x axis of both panels indicates the extracted components. In the top, the y axis indicates the eigenvalue assigned to each component. These values are used to calculate the variance explained by each component individually, and their cumulative total is illustrated in the bottom.

(B) Cluster strength analysis. A histogram illustrating the results of a bootstrap analysis evaluating the strength of the regional clusters formed in the ordination space. Each iteration (total 250,000) shuffled the region labels attached to all samples and generated a single score (Davies-Bouldin index) for the iteration. The x axis indicates the score, and the y axis indicates how many iterations had a score in that bin. The red line indicates the score (6.93) for clusters defined by the real regions.

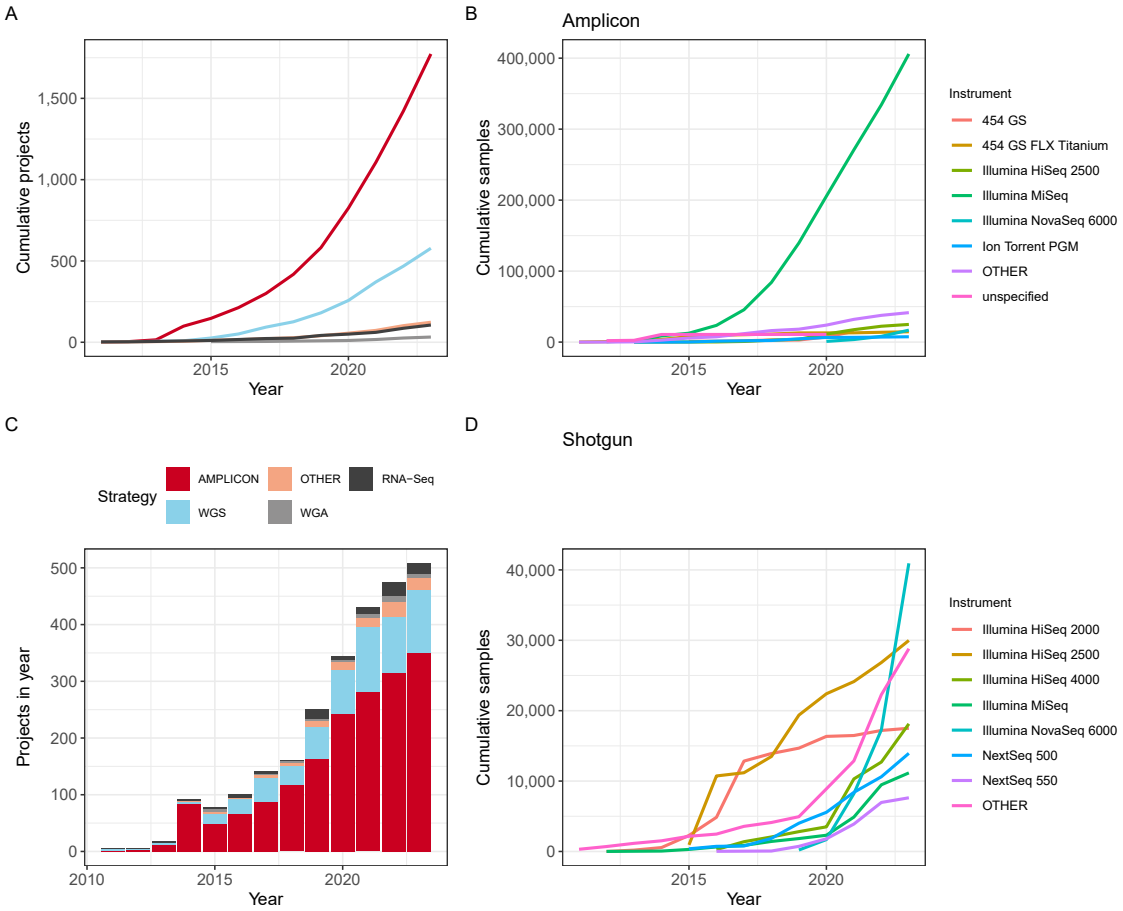


Figure S4. Technical factors over time, related to Figure 3

(A) Reported sequencing library strategies. This extends the data illustrated in Figure 3A by plotting projects, rather than samples. The x axis indicates the year of publication, the y axis indicates the cumulative number of projects, and each line indicates the most common reported values for library strategy.

(B) Sequencing platform over time, 16S rRNA gene amplicon sequencing. The x axis indicates year, the y axis indicates the cumulative samples deposited, and each colored line indicates the most frequently reported sequencing platforms.

(C) This plots the same data as (A) but without cumulative numbers. The x axis indicates year of publication, the y axis indicates total projects deposited in that year, and each colored bar segment indicates the reported library strategy.

(D) Sequencing platform over time, shotgun metagenomic sequencing. The x axis indicates year, the y axis indicates the cumulative samples deposited, and each colored line indicates the most frequently reported sequencing platforms.

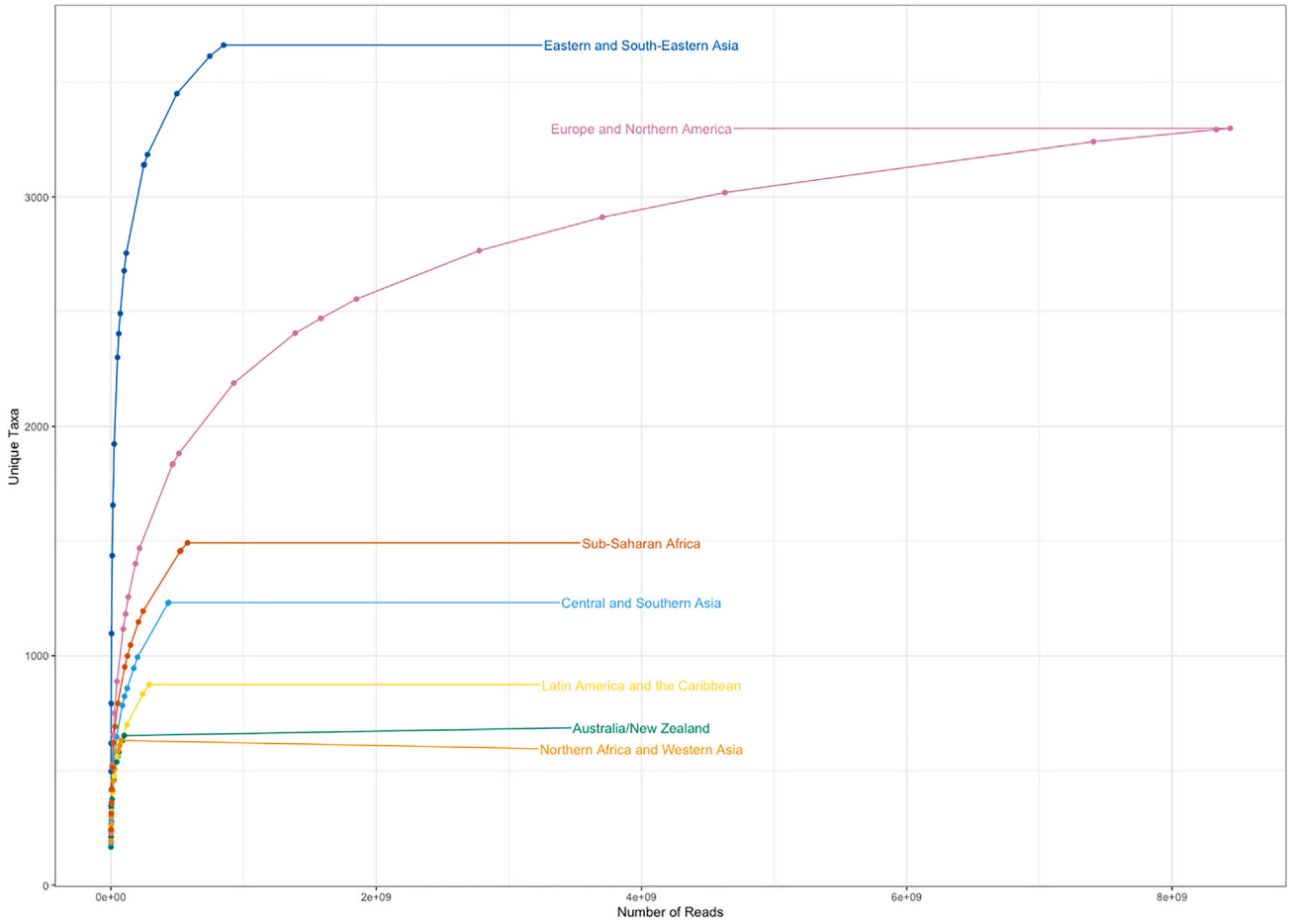
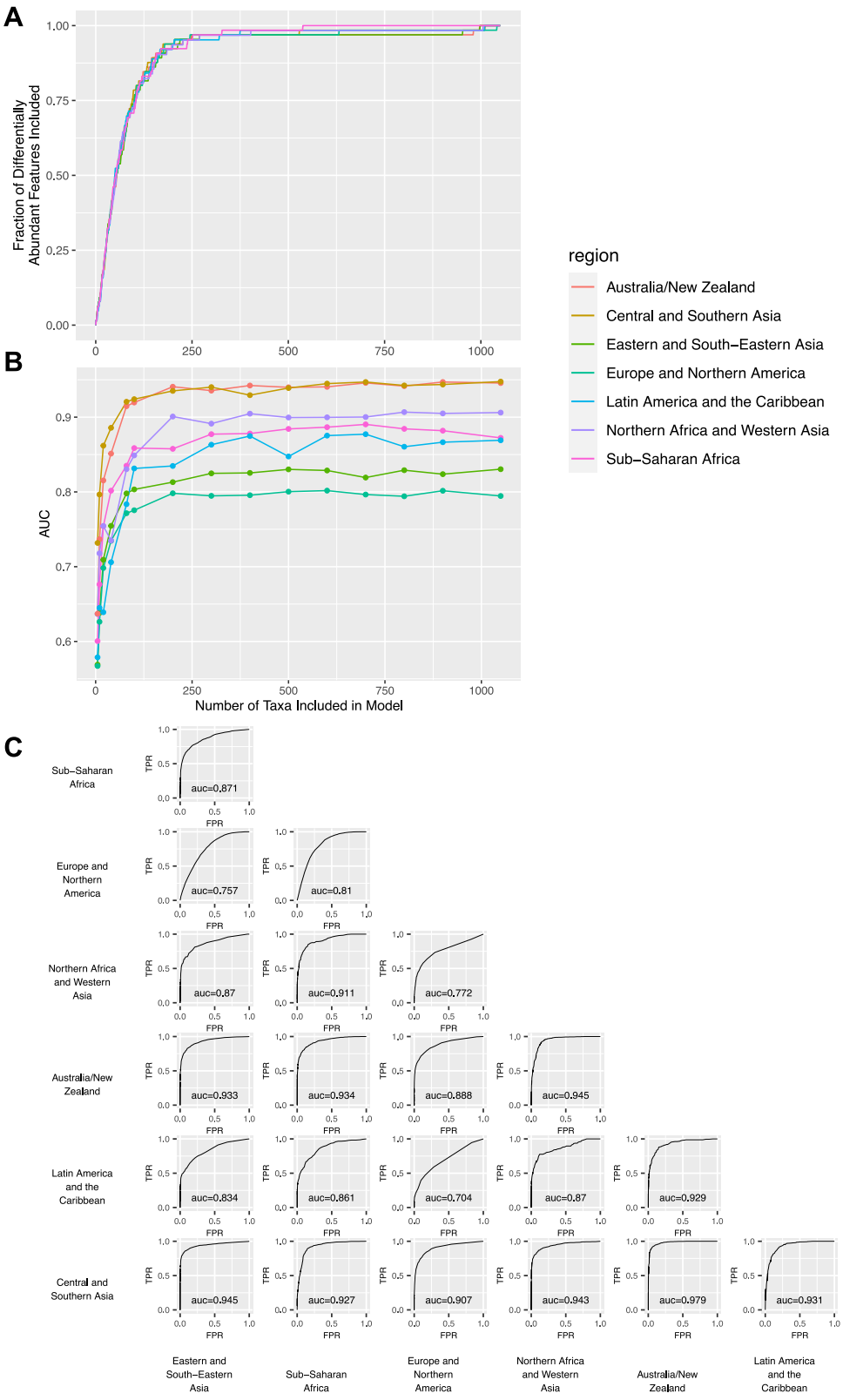


Figure S5. Region-level discovery as read count increases, related to Figure 4

Using the same sampling scheme as Figure 4A, we sampled increasing numbers of microbiomes from each geographic region. For each sample size, we calculated the mean number of reads in the selected samples (indicated on the x axis) and the mean number of unique taxa observed in the selected samples (indicated on the y axis).



(legend on next page)

Figure S6. Classifier results, related to Figures 5 and 6

(A) An illustration of the most important variables in the one-versus-all classifiers described in Figures 6C and 6D. For each region-level classifier, which infers whether a given sample is from a single region, we ranked the importance of all taxa in the model (see STAR Methods). Each line illustrates a single region's model. The y axis illustrates the proportion of differentially expressed taxa in that region (Figure 5B) that are included in the top X most important taxa. The x axis illustrates the rank of the taxa, with 1 being the most important.

(B) An illustration of the effectiveness of models trained only on the X most important taxa, as determined by the final model. The x axis describes the number of taxa included in a regional one-versus-all model, and the y axis indicates the AUC of the model.

(C) One-versus-one classifiers. Each panel illustrates the ROC curve for a one-versus-one classifier tasked with differentiating samples from two regions. The x axis of each panel illustrates the false-positive rate, and the y axis indicates the true-positive rate.