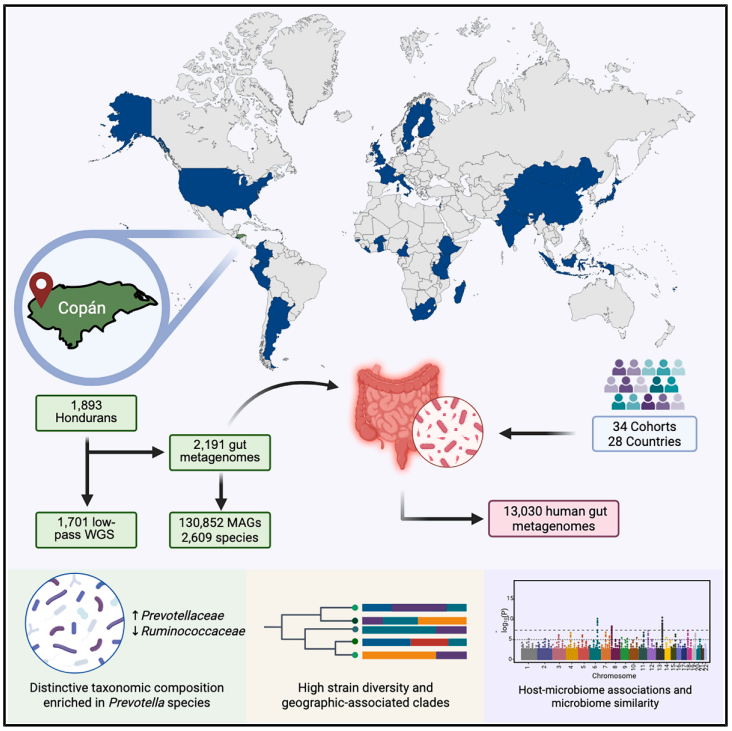
Article

Cell Reports

Characterization of gut microbiomes in rural Honduras reveals uncharacterized species and associations with human genetic variation

Graphical abstract



Authors

Francesco Beghini, Ilana L. Brito, Mark Gerstein, Nicholas A. Christakis

Correspondence

francesco.beghini@yale.edu

In brief

Beghini et al. analyzed 2,191 samples from rural Hondurans, revealing a distinct microbiome compared to other non-western populations, through comprehensive profiling of bacterial, viral, and eukaryotic components and their host associations.

Highlights

- We investigate the composition of 2,191 rural Honduran gut microbiomes
- Honduras microbiome is distinct from other non-western microbiomes
- High strain diversity, uncharacterized species, and high *Blastocystis* prevalence
- Shifts in taxonomic composition associated with SARS-CoV-2 infection

Beghini et al., 2026, Cell Reports *45*, 116724
January 27, 2026 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.celrep.2025.116724

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Characterization of gut microbiomes in rural Honduras reveals uncharacterized species and associations with human genetic variation

Francesco Beghini,[1,9,*] Ilana L. Brito,[2] Mark Gerstein,[3,4,5,6,7] and Nicholas A. Christakis[1,7,8]
[1]Human Nature Lab, Yale University, New Haven, CT 06511, USA
[2]Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY 14853, USA
[3]Program in Computational Biology and Biomedical Informatics, Yale University, New Haven, CT 06511, USA
[4]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA
[5]Department of Computer Science, Yale University, New Haven, CT 06511, USA
[6]Department of Biomedical Informatics and Data Science, Yale University, New Haven, CT 06511, USA
[7]Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA
[8]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511, USA
[9]Lead contact
*Correspondence: francesco.beghini@yale.edu
https://doi.org/10.1016/j.celrep.2025.116724

SUMMARY

The gut microbiome is integral to human health, yet research data to date have emphasized industrialized populations. Here, we performed large-scale shotgun metagenomic sequencing on 1,893 individuals from rural Honduras, providing the most comprehensive microbiome dataset from Central America. We identify a distinct microbial composition enriched in *Prevotella* species. Longitudinal analysis in 301 individuals reveals microbiome instability, with shifts in taxonomic diversity and metabolic potential, including changes associated with severe acute respiratory syndrome coronavirus 2 infection. Additionally, we characterize the gut virome and eukaryotic microbiome, identifying uncharacterized viral taxa and a high prevalence of *Blastocystis* species in individuals with greater microbial diversity. Finally, by integrating host genomic data, we uncover significant host-microbiome associations, highlighting the influence of human genetic variation on microbial composition. These findings expand our understanding of microbiome diversity in non-industrialized populations, underscoring the need for global microbiome research.

INTRODUCTION

Extensive work has investigated the role that the human microbiome plays in health and illness, and how variations in diet, lifestyle, social interactions, and environment are relevant to its composition.[1–3] At present, existing sampled microbiomes are heavily skewed toward western and industrialized countries in which individuals have generous access to antibiotics and other medications and a diet rich in ultra-processed food and low in fiber.[4,5] Yet, more than 80% of the world population lives outside Europe and North America. An exploration of people's microbiome composition from geographically diverse areas is necessary for a more comprehensive understanding of variation in microbiome composition, as well as its association with human health and even with host genetic background. Previous studies have documented that non-western populations harbor a more diverse microbiome, with several taxa that do not occur in other populations[6,7]; in fact, some microbes appear to have even become extinct in western populations.[8]

Modern tools have simplified the process of exploring this variation. Advances in sequencing and computational methods have enabled the reconstruction of more refined metagenomic assembled genomes (MAGs) and the consequent expansion of the set of observed species in the human gut. Thus, MAGs reconstructed from under-represented regions can provide ecological and functional insights into the human diversity of the gut microbiome. Here, we perform shotgun metagenomic sequencing on 1,893 individuals living in the isolated western highlands of Honduras, which is the largest shotgun metagenomic dataset collected in the Central America region and from a single non-western population to date. In addition, using saliva, we performed low-pass sequencing of the human genome to conduct microbiome-genome association analyses. Furthermore, by putting the Honduran microbiome in a more global context, we describe the extent of microbial diversity observed in Honduras and report several uncharacterized species. We also explore associations between host human genotypes and microbiome occurrence.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

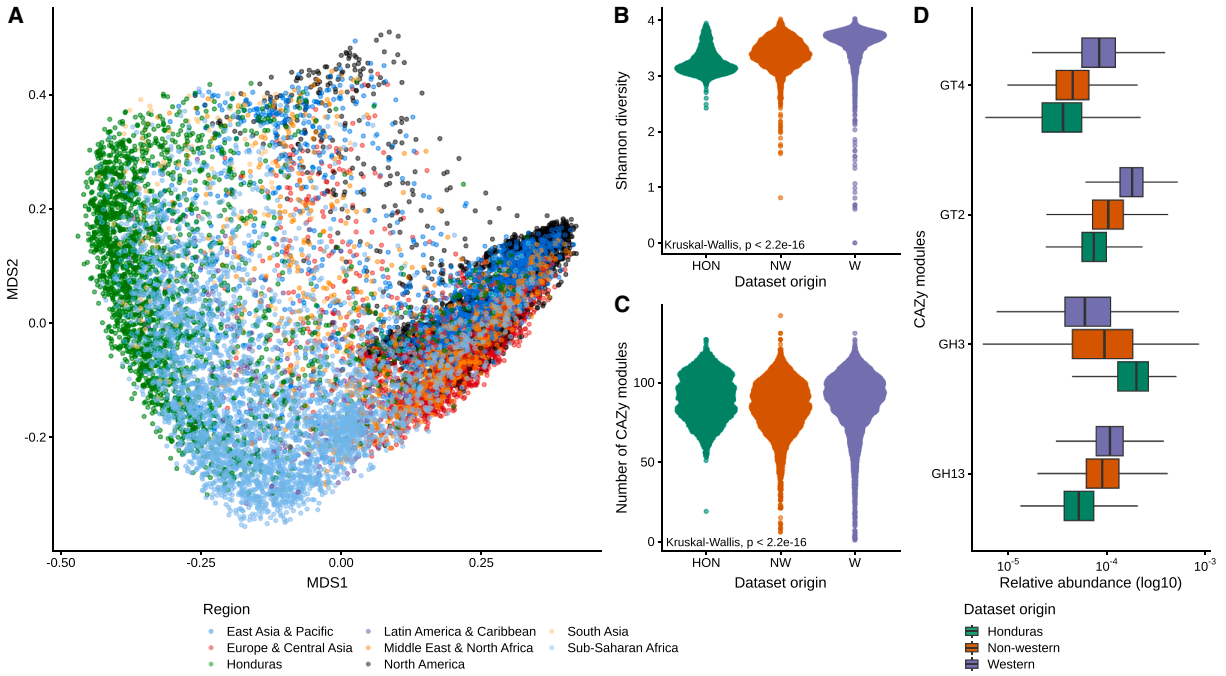


Figure 1. The Honduras gut microbiome in a global context

PCoA (A) of the Bray-Curtis dissimilarity calculated on the SGB relative abundances of 13,030 samples. Honduras taxonomic composition is distinct from other non-western cohorts (PERMANOVA $p = 0.001$). Honduras samples have lower diversity (B) but more richness (C) in terms of CAZy modules, and they also have certain modules present in higher abundance compared to other datasets (D). Boxplot whiskers extend from the hinge to no further than 1.5× interquartile range (IQR) from the hinge.

RESULTS

Gut microbiome genome catalog

We enrolled a total of 1,893 people from 26 geographically isolated villages in rural Honduras. The villages varied in size from 66 to 434 residents. This cohort includes participants from 18 villages that are part of a previous study.[1] The age of the participants ranges from 15 to 94 years (mean 40.73 years, SD = 17). The majority of participants, 63.9% (n = 1,210), are women.

Stool samples were collected starting in 2020, and a subset of 301 individuals were sampled again after two years for a longitudinal analysis. We performed shotgun metagenomics on a total of 2,191 samples, which generated a total of 27.5 Tbp of data (mean 12.53 Gbp, SD 4.4 Gbp). We then proceeded to include our cohort in a larger context by integrating our analyzed samples with human gut metagenomes from 34 different datasets spanning different lifestyles and geographical locations, resulting in a total of 13,030 metagenomes[9–37] (Figure S1; Tables S1 and S2).

SGB-level (species-level genome bins)[21] composition was determined with MetaPhlAn 4.[38] Similar to what has been previously observed in other non-western populations,[8,30,39] the overall species-level composition in the Honduras dataset is dominated by *Prevotella copri* clades, now reclassified as the *Segatella* genus; in particular, 52% of the samples have *P. copri* clade A as the most abundant species, with a decreasing fraction of samples showing *Prevotella hominis* and Spirochaetia SGB3548 as the most abundant species. *Lacrimispora amygdalyna* (SGB4716) was identified as the most prevalent species among our Honduras samples, present in virtually all the samples (99.5%). While we did observe some overlap in composition between individuals in Honduras and African and Middle Eastern cohorts, the Honduras abundance profiles are nevertheless distinguishable (PERMANOVA $p = 0.001$) (Figure 1A) with several Prevotellaceae species found to be more abundant in our cohort than other non-western cohorts (Table S3). The species identified as the strongest distinguishing marker between Honduras and other non-western cohorts was Veillonellaceae SGB5809. Overall, the increased abundance of the whole Prevotellaceae family is the main distinguishing feature that differentiates Honduras from other non-western cohorts (linear mixed model $\beta = 22.88$, SEM = 0.50, $p = 0$). Also, Ruminococcaceae family bacteria were found to be less abundant in Honduras, although to a lesser extent (linear mixed model $\beta = -7.13$, SEM = 0.23, $p = 8.38 \times 10^{-188}$).

We then compared the overall distribution of specific bacterial families to assess differences in composition between our cohort and other western and non-western cohorts. Families Spirochaetaceae, Prevotellaceae, and Succinivibrionaceae are commonly found to be increased in populations with traditional lifestyles (VANISH taxa), while industrialized populations have higher abundances of Verrucomicrobiaceae, Akkermansiaceae, and Bacteroidaceae (BloSSUM taxa)[7,8] (Figure S2). We identified a total of 173 species from the VANISH families and 189

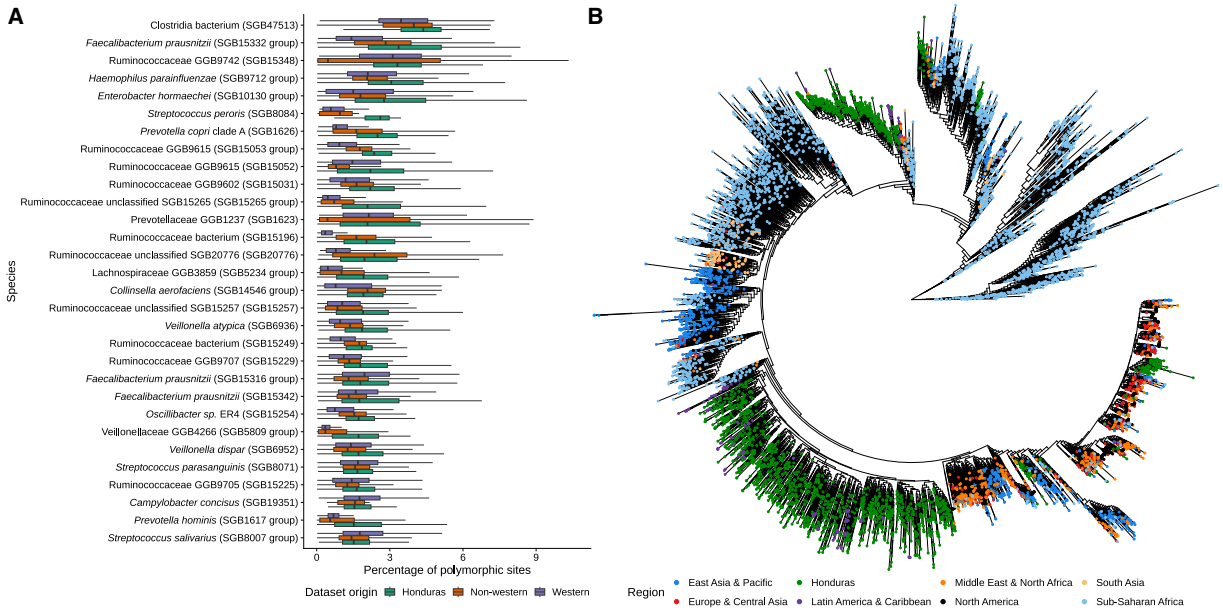


Figure 2. High strain diversity in Honduras metagenomes

(A) Distribution of the percentage of polymorphic sites for 30 species identified by StrainPhlAn 4.

(B) Phylogenetic tree of 7,071 *Prevotella copri* clade A strains reconstructed by StrainPhlAn 4.

Strains reconstructed from Honduras samples are segregated in branches significantly associated with Honduras origin (D statistic: −0.67). Boxplot whiskers extend from the hinge to no further than 1.5× IQR from the hinge.

species from the BloSSUM group. The overall distribution of BloSSUM taxa in the Honduras population resembles the one observed in other non-western cohorts, with slightly higher abundances of Bacteroidaceae taxa (Wilcoxon rank-sum test $p = 2.15 \times 10^{-22}$) and lower abundances of Verrucomicrobiaceae than other non-western cohorts (Wilcoxon test $p = 1.14 \times 10^{-48}$). All VANISH families are present in higher abundances, especially Prevotellaceae (Wilcoxon rank-sum test $p = 1.92 \times 10^{-290}$). Spirochaetaceae and Succinivibrionaceae are found in lower abundances, but higher than other cohorts (Wilcoxon rank-sum test $p = 2.01 \times 10^{-45}$ and $p = 3.27 \times 10^{-1}$, respectively). In addition, our cohort had richer samples than other analyzed countries (Wilcoxon rank-sum test $p < 0.05$) except for datasets from Ethiopia, Fiji, Liberia, Ghana, Indonesia, and Burkina Faso (Table S4; Figure S3).

We did not observe any differences in alpha diversity between people living in different villages in Honduras, but some villages could be distinguished according to differences in the overall composition (PERMANOVA $p = 0.001$). Although most people self-reported their ethnicity as Maya Ch'orti', we did not observe any difference in microbiome composition based on self-reported ethnicity.

Looking at the functional composition, the Honduran microbiome is also distinct (Figure S4). In terms of predicted metabolic pathways, samples from Honduras are separated on the principal coordinate analysis (PCoA) space from other western and non-western cohorts (PERMANOVA $p = 0.001$, Table S5). This is also apparent when looking at the distribution of carbohydrate-active enzymes (CAZy), which digest carbohydrates. We observed that non-western cohorts exhibited lower richness of CAZy modules compared to western cohorts, while Honduras showed lower Shannon diversity, but higher richness than both western and non-western cohorts (Figures 1B and 1C), although the distribution of the Shannon diversity for the Honduras samples had lower variance compared to the other categories, and differences between individual countries could be observed as well, especially in the tail of the distribution (Figure S5). Differential abundance analysis showed that two CAZy classes, GH (glycoside hydrolase) and GT (glycosyl transferases), were present at different levels in our cohort. GH modules are enzymes that break down carbohydrates into simpler sugars, and their presence is associated with diets high in fiber.[40] GT modules are glycosyltransferases, enzymes involved in a wide range of roles from cell membrane formation to antibiotic production. Higher median abundance of GH3 modules and a lower median abundance of GH13, GT2, and GT4 modules were observed in Honduras compared to other cohorts (Figure 1D). Analysis of functional annotations of the reconstructed MAGs showed that those modules were present in higher prevalence in genera like *Prevotella*, *Bacteroides*, *Phocaeicola*, and *Butyribacter*.

Strain-level analysis

To expand the understanding of intra-species diversity in our cohort, we investigated the strain-level diversity distribution by performing an analysis with StrainPhlAn. We found a wide range of strain diversity compared to the same species present in the other datasets (Figure 2A; Table S6). *Haemophilus parainfluenzae* and *Enterobacter hormaechei* were species with

substantially higher variation, with more than 3% of polymorphic sites, and hence the median of polymorphic rates from Honduras strains was higher than that of strains from other cohorts (Table S3). One species in particular, the butyrate-producing 0*Butyrivibrio crossotus,* presented high genetic diversity in our Honduras sample. *B. crossotus* is a component of the healthy gut microbiome; its prevalence is higher in samples from non-industrialized settings, and the species encodes several carbohydrate-active enzymes able to digest fiber.[41,42] We identified *B. crossotus* in nearly 80% of Honduras samples (n = 1,729), with relative abundances as high as 15% in some samples. *P. copri* clade A, the most abundant species in our dataset, showed high intra-species diversity as well, with samples split into two separate and diverging clades, suggesting high geographical adaptation or the presence of another *P. copri* clade[39] not detected by the markers present in the reference database (Figure 2B).

Phylogenetic analysis of the *P. copri* strains retrieved from the Honduras samples revealed the presence of clades enriched in samples from a common origin, also previously reported by other studies.[43,44] Honduras strains are segregated in branches that rarely overlap with samples from other geographical origins, with the exception of very few strains retrieved from El Salvador, a neighboring country; that is, we observed a significant association with Honduras origin (D statistic −0.67). The same close association with geography was observed for SGB1614, now identified as *P. copri* clade H, in which multiple monophyletic groups are composed of strains from identical geographic locations, mostly from non-westernized populations. In addition, we observed the presence of a sub-species branch that is particularly divergent, with all strains retrieved from Honduras (Figure S6). This same pattern of phylogenetic clades enriched in samples retrieved from Honduras was observed in multiple species (not just *P. copri*), revealing a strong geographic association of clades (n = 1,054 species with D statistic < 0).

We further improved the characterization of our cohort by reconstructing MAGs (STAR Methods). After quality control, we recovered a total of 130,852 MAGs meeting the medium-quality thresholds proposed by MIMAG (>50% completeness and <10% contamination), with 63,459 of them being classified as high-quality (>90% completeness and <5% contamination)[45] (Figure S7; Table S7). After deduplication, this yielded a total of 2,609 representative genomes (Figure 3).

The species with the greatest number of genomes reconstructed is *Agathobacter faecis*, previously known as *Roseburia faecis*, with 1,494 genomes reconstructed and a prevalence of 91%. *Prevotella* was the genus with the largest number of reconstructed genomes (n species = 72, n genomes = 17,594), with several species identified as the most prevalent in our dataset (Table S8; Figure S8). Excluding species from the *Prevotella* genus, the second most prevalent species was identified as *Faecalibacterium longum*, with a prevalence of 93%. Genus CAG-269 from Clostridia showed the largest number of species-clusters (n species = 74, n genomes = 1,641). Bacterial genomes spanned 22 different phyla, with most of the species assigned to Bacillota A (n = 1,365), followed by Pseudomonadota (n = 207), Bacteroidota (n = 271), and Bacillota (n = 380).

Archaeal organisms were reconstructed as well, for a total of 9 species from the Thermoplasmatota (n = 6), Methanobacteriota (n = 2), and Thermoproteota (n = 1) phyla. The most prevalent archaeal organism was identified in *Methanobrevibacter smithii*, which was found in 51% of the samples with 803 reconstructed genomes. We identified a second *M. smithii* clade as well, classified by the Genome Taxonomy Database (GTDB) as *Methanobrevibacter smithii* A, now renamed as *Candidatus Methanobrevibacter intestini*.[46] Such species were identified in nearly 30% of the samples, and we observed (in 716 samples) a mutual exclusion between *M. smithii* and *C. M. intestini.*

We then used two approaches to estimate how many species were not reported in other databases. First, we evaluated the taxonomy assignment from GTDB,[47] and then we quantified the overlap between our reconstructed set of genomes and 289,232 MAGs from the Unified Human Gastrointestinal Genome (UHGG)[48] catalog expanded with the genomes from two large datasets.[49,50] We identified 270 species that lacked a representative from UHGG (Figure S9). Among those species, the families most represented were Bacteroidaceae (n = 21), Acutalibacteraceae (n = 18), Lachnospiraceae (n = 18), and Oscillospiraceae (n = 16). The largest number of genomes, 137, was reconstucted for a species assigned under the CAG-302 genus from the Bacilli class. KEGG module over-representation analysis showed an enrichment of genes involved in the quorum-sensing pathway (ko02024) and a complete set of genes involved in the vancomycin resistance pathway. In addition, several antibiotic resistance genes for beta-lactam and cationic antimicrobial peptides were found among the 217 other species. GTDB failed to assign species-level taxonomy for 1,976 genomes, revealing 391 unknown species spanning 3 families and 19 genera. Most of them are from the Clostridia class; in particular, we reconstructed 215 genomes from an uncharacterized species from the RUG11200 genus.

Among the reconstructed Clostridia species, we reconstructed 109 genomes for SGB41716. Identified by MetaPhlAn as *Lacrimispora amygdalina,* and by GTDB as uncharacterized *Alitiscatomonas* sp900066535, it is a species with 77.56% average nucleotide identity with *Lacrimispora amygdalina*. Metabolic pathways reconstruction and metabolite production prediction showed that the species is a short-chain fatty acids (SCFAs) producer, with model-based evidence of production of butyrate and propionate. We analyzed the population structure of the species by performing a pangenome analysis of the reconstructed genomes. PCoA highlighted the difference of *Alitiscatomonas* spp. strains retrieved from Honduras separated in the PCoA space (Figure S10). Metabolic features are distinguishable between strains, with speciating gene family variants retrieved from Honduras and other non-western datasets involved in the pentose-phosphate pathway (K07404), TCA cycle (K01681), the cysteine and methionine metabolism, a LysR-binding domain, and a dicarboxylate carrier protein (Figure S11).

In addition, we performed gene prediction and annotation on the reconstructed contigs. We built a gene catalog starting from 244,932,121 protein sequences that were clustered at 95% amino acid identity, resulting in 9,883,899 gene families. Among those, 413,626 gene families (4.2%) were annotated with a COG category, although the majority were of unknown function (S category, n = 1,514,506) or had no hit at all (n = 3,470,273). This fraction could represent protein fragments and proteins of non-bacterial origin (e.g., viral and eukaryotic),

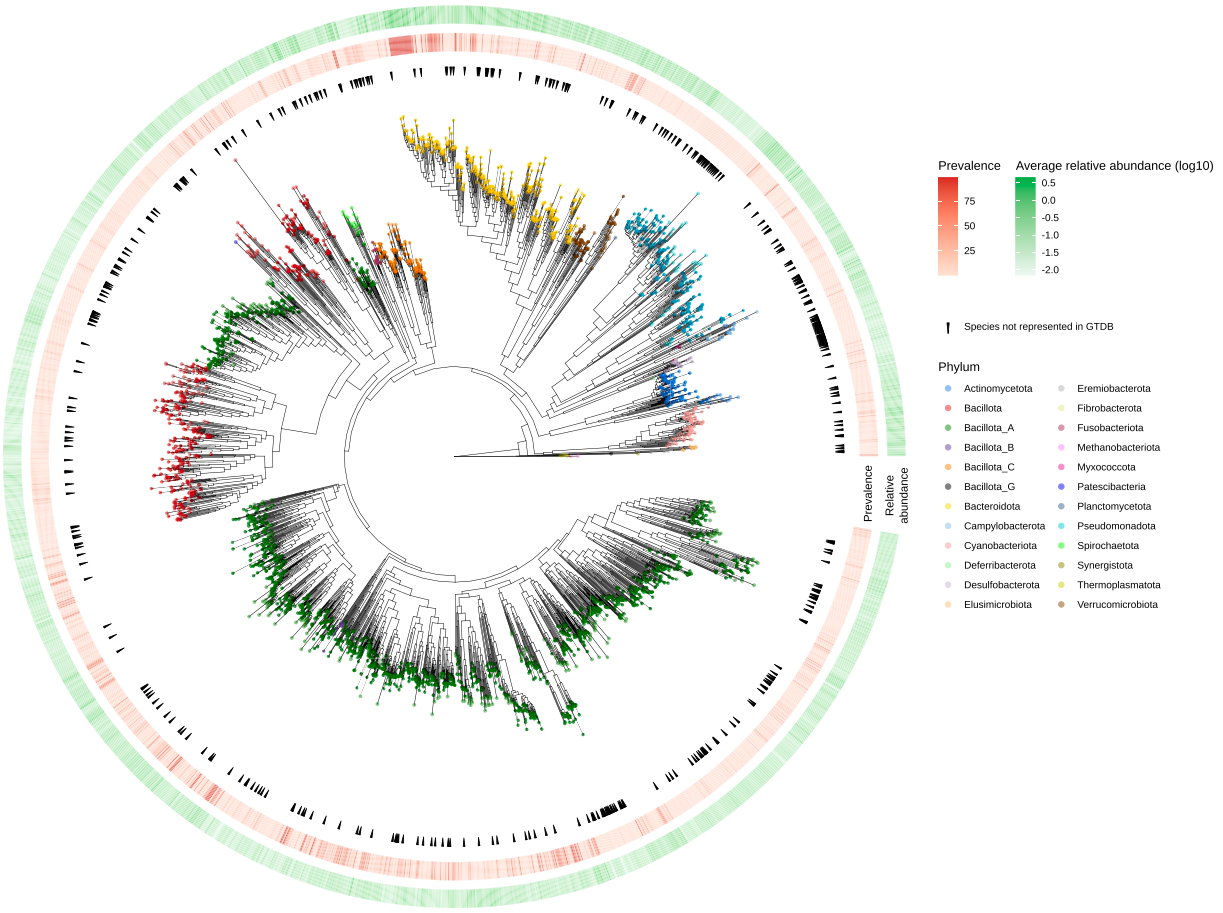


Figure 3. Phylogenetic tree of the 2,609 reconstructed MAG representatives

The tree was built using PhyloPhlAn 3 using the PhyloPhlAn markers. Tips are annotated according to the phylum-level taxonomy predicted by GTDB-tk, and relative abundance and prevalence in the Honduras dataset are reported in the two outermost ring annotations. Species without a representative in GTDB are marked with a triangle.

although 25,535 proteins of eukaryotic origin were detected, mostly from micro-eukaryotes and food.

When combined with gene families reported in the UHGG, we observed that 3,508,259 gene families have no representation within UHGG. Functional annotation with eggNOG showed that 171,380 gene families (4.8%) lacked functional annotation, as extrapolated from COG functions and PFAM annotations. We further investigated the functional nature of those gene families and identified gene families that lacked any type of functional annotation, including broader PFAM annotations. For those with some level of annotation, the majority of the proteins were annotated as part of COG categories M (cell wall/membrane/envelope biogenesis), L (replication, recombination, and repair), and K (transcription).

Non-bacterial components

To test for the presence of eukaryotic microorganisms, 1,148 stool samples were screened for the presence of ova and parasites. The exam showed the colonization by several organisms, such as *Ascaris lumbricoides* (14%, n = 167), *Entamoeba coli* (14%, n = 165), *Entamoeba histolytica* (13%, n = 155), *Blastocystis* spp. (5%, n = 67), *Trichuris* (2%, n = 28), and *Iodamoeba bütschlii* (1.5%, n = 17). Polyparasitism was observed in 151 individuals, mostly occurring with species *Ent*. *coli* and *E*. *histolytica*, followed by *Ent. coli* and *A*. *lumbricoides*. The presence of *Ent*. *coli* and *I*. *butschii* in the human gut has been previously reported as nonpathogenic and usually occurs at low prevalence. The prevalence of *Blastocystis* spp. is surprisingly low in our population compared to previous reports in Central America and other non-western countries.[51–54] High colonization by soil-transmitted helminths like *A*. *lumbricoides* and *Trichiuris* is commonly observed in subtropical countries and is one of the greatest public health concerns in developing countries.[55] Despite higher rates of helminth infections, very few participants reported taking antiparasitic medications prior to our evaluation (n = 2).

We screened for the presence of genomes of eukaryotic origin with EukRep,[56] which identified a total of 129 bins, with 55 bins

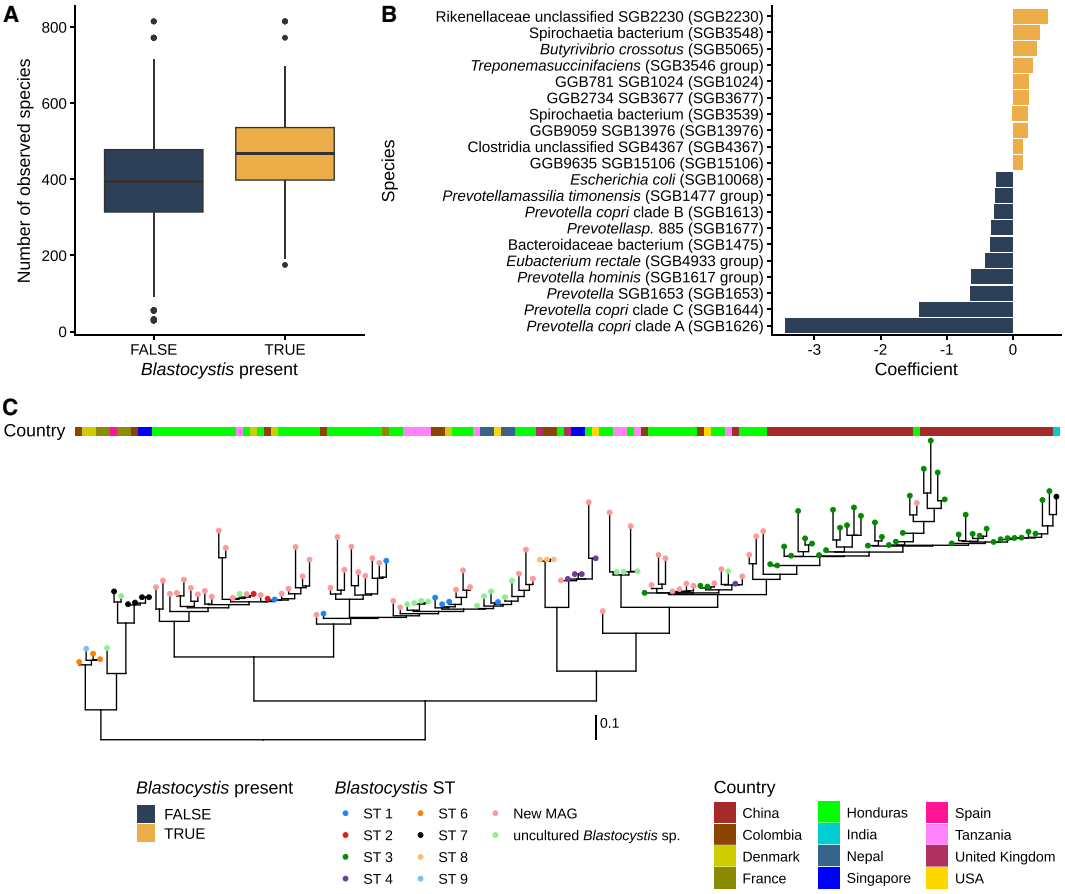


Figure 4. Genetic diversity of *Blastocystis* strains and association with microbial features

Blastocystis carriers have higher alpha diversity compared to non-carriers (A) (Wilcoxon rank-sum test $p = 1.38 \times 10^{-94}$) as well as differentially abundant species (B). (C) Phylogenetic tree of 55 newly reconstructed *Blastocystis* MAGs alongside publicly available isolate genomes. Boxplot whiskers extend from the hinge to no further than 1.5× IQR from the hinge.

assigned to the Stramenopiles lineage (Tables S9 and S10). Further taxonomic assignment determined that all 55 bins were classified in the *Blastocystis* genus, a prevalent inhabitant of the human gut.[52,57,58] Using the 55 newly recovered bins, alongside a catalog of eukaryotic genomes, we proceeded to get estimates of the relative abundances of those organisms (Figure 4C). A total of 13 organisms were identified by mapping the metagenomic reads against the genome catalog using coverM. The most prevalent organism detected was *Blastocystis* spp., present with multiple subtypes (ST1 $n = 530$, ST2 $n = 303$, ST3 $n = 577$, ST4 $n = 3$, ST6 = 1, ST7 $n = 1$, and unknown subtype $n = 79$), followed by *Giardia intestinalis* ($n = 17$) and *Saccharomyces cerevisiae* ($n = 12$). We were not able to find most of the organisms that were discovered through microscopy assessment of the fecal material when using metagenomic mapping. This could be explained by the low number of cysts or eggs in our samples that resulted in a low number of reads mapping to any of those organisms' genomes. In addition, for some of those species, a reference genome is not available, but only ribosomal gene sequences.

Consistent with what has been previously reported, we observed a higher prevalence of individuals carrying at least a *Blastocystis* subtype ($n = 1,010$), and 225 individuals were carriers of multiple subtypes. The overall cohort prevalence is 56%, higher than in other countries where *Blastocystis* prevalence has been investigated.[52,58] Per-village prevalence ranges from 45% to 70%, and the subtype distribution resembles the cohort distribution.

We examined the differences between the microbiome composition of *Blastocystis* carriers and non-carriers and found that carriers have higher alpha diversity than non-carriers (Figure 4A, Wilcoxon rank-sum test observed species $p = 1.38 \times 10^{-94}$; Shannon diversity $p = 1.53 \times 10^{-87}$). We expanded our analysis by looking at associations between *Blastocystis* carriage and microbiome composition (Figures 4B; Table S11). The overall microbiome composition can be differentiated for carriers and non-carriers, with an increased abundance of taxa like *Butyrivibrio crossotus*, *Treponema succinifaciens*, and *Methanobrevibacter smithii* in carriers. We also observed a depletion of

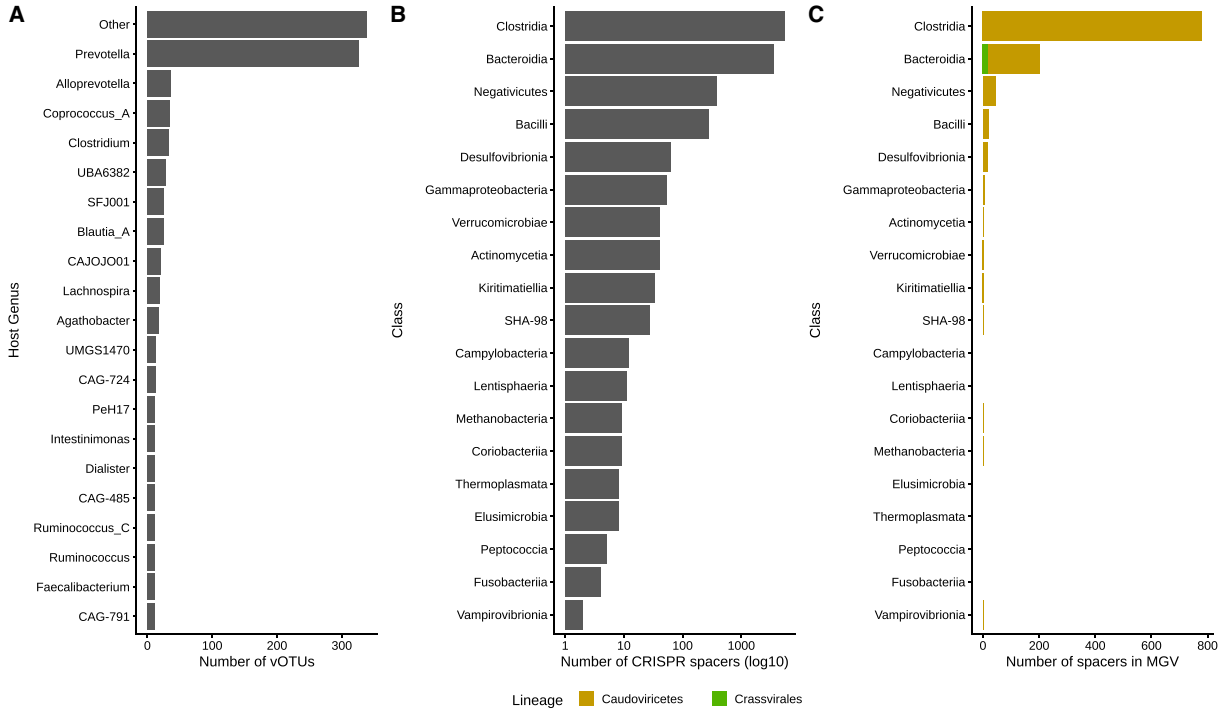


Figure 5. Uncharacterized viral sequences in the Honduras microbiome

Distribution of the predicted host genomes for vOTU representatives (A) and CRISPR spacers among bacterial classes (B). Very few CRISPR spacers were identified in other public databases (C).

several *Prevotella* species and *Eubacterium rectale* in non-carriers, in stark contrast to previous reports, as those species are usually reported to be found with increased abundance in carriers.[52,59] However, this result is similar to one previously observed in a cohort in Colombia.[60] We did not find any changes in the direction of association when looking at individual subtypes, with 28 species associated with one *Blastocystis* subtype and 15 species associated with two or more subtypes (mixed effect linear model $p < 0.05$).

Long-term colonization of *Blastocystis* was observed in our 301 longitudinally sampled individuals, in which 71 individuals were found to be carrying *Blastocystis* after two years. Moreover, 5 individuals carrying multiple subtypes were still carrying the same subtypes after 2 years, with *Blastocystis* ST3 being the most prevalent subtype among long-term carriers ($n = 41$).

Next, to characterize the composition of the gut virome, all bins produced by VAMB were post-processed by the PHAMB pipeline, and, after quality control, we obtained 110,089 viral MAGs that were clustered into 29,958 viral operational taxonomic units (vOTUs) (Table S12). In addition to those MAGs, we obtained 12,136 viral clusters from contigs that were classified as a provirus by CheckV[61] and that were clustered separately. A consecutive clustering with representative sequences from the MGV database[62] highlighted that the great majority of the sequences were of unknown origin, and only 3,353 newly reconstructed vOTUs clustered with an MGV reference sequence and 25,580 newly reconstructed vOTUs with respect to MGV. For almost all vOTUs, geNomad[63] was able to assign them to the Caudoviricetes class, with some of them annotated as families Circoviridae, Crassvirales, and Microviridae.

By using clustered regularly interspaced short palindromic repeat (CRISPR) spacers, we investigated the associations between microbes and viral genomes by predicting CRISPR spacers in all our reconstructed MAGs. This resulted in 104,851 CRISPR spacers, further deduplicated into 4,935 spacers. 48,862 spacers were mapped against a viral MAG, enabling further host-association analysis for 2,417 vOTUs (Figure 5). A total of 403 vOTUs that associated with a bacterial host were associated exclusively with one bacterial species; 541 vOTUs were associated with more than 2 and less than 10 hosts; and 83 vOTUs with 10 or more hosts. We also observed the presence of vOTUs associated with different species across different families, classes, phyla, and some occurrence of *trans*-kingdom association, as 5 associated with both bacterial and archaeal species. The majority of the vOTUs were associated with *Prevotella* species ($n = 326$) as their host, followed by the *Alloprevotella* ($n = 37$), *Coprococcus* ($n = 34$), and *Clostridium* ($n = 33$) (Figure 5A). Among the vOTUs that were represented in MGV, we observed a similar pattern in assigned taxonomies, with vOTUs with *Prevotella* species as the predicted host present in higher abundance and higher prevalence.

vOTU representatives were mapped against the original metagenomes to produce relative abundance profiles. Overall, viral reads accounted, on average, for 4.44% of the original

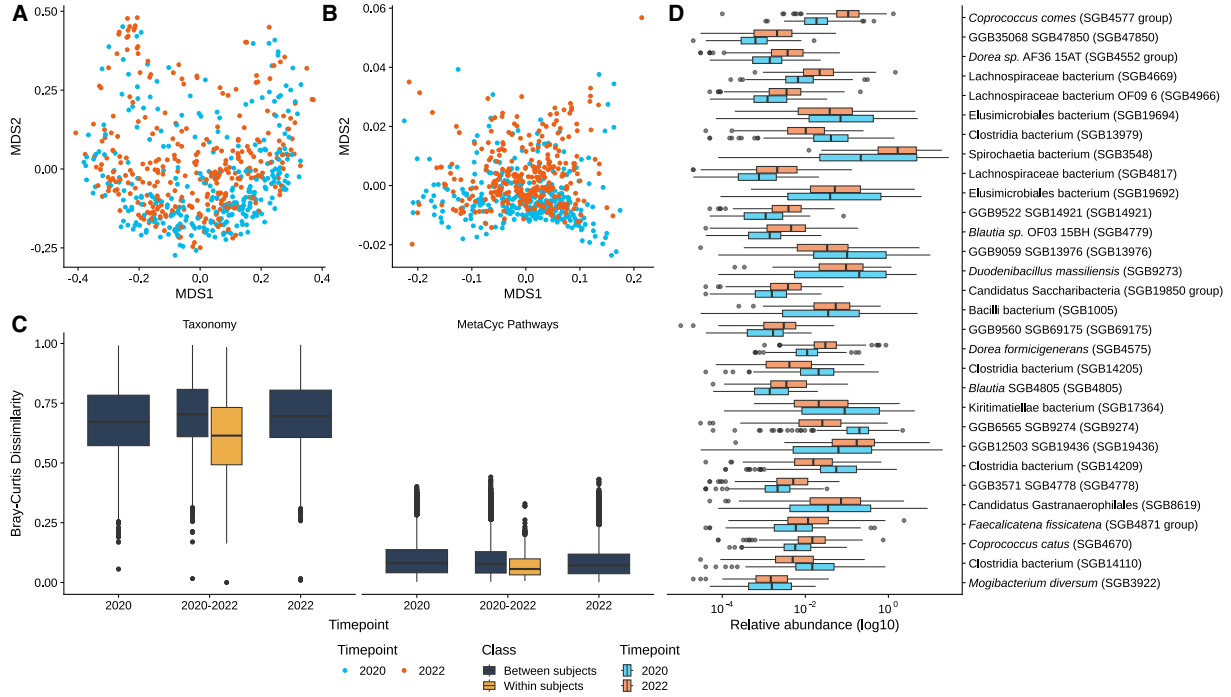


Figure 6. Longitudinal analysis of the Honduras microbiome

PCoA on the Bray-Curtis dissimilarity calculated on species-level (A) and MetaCyc pathways (B) relative abundances. Individuals are more similar to each other after 2 years, but intra-individual beta diversity decreased (C, Wilcoxon rank-sum test: $p < 2.2 \times 10^{-16}$). Linear mixed models identified 286 species differentially abundant between individuals during and after infection with SARS-CoV-2 (D). Boxplot whiskers extend from the hinge to no further than 1.5× IQR from the hinge.

metagenomes (1.68% SD); individuals showed a median value of 130 vOTUs present in each individual (mean 142.4, SD 66.13). Viral diversity was found to be positively correlated with bacterial diversity, as measured by the number of identified species (Spearman correlation ρ = 0.73). We observed as well that many viral OTUs were private, i.e., present exclusively in the individual from which the viral MAG originated.

A total of 255 vOTUs were annotated as Crassvirales, the family of crAss-phage, a globally widespread gut microbiome component, and were found in 1,129 samples, roughly 50% of the total, with an average abundance of 5.5%. Using skani, we then aligned the representative sequences of the identified *Crassvirales* vOTUs to the crass-phage contigs identified and annotated from prior work.[64] The crass-phage catalog contains 673 genomes, and almost half of the vOTUs identified in our dataset were represented by one of the already known crass-phage genomes. Among these, we observed that most individuals have vOTUs assigned to the candidate family Deltacrassvirinae (n = 42), followed by Beta (n = 32) and Epsilon (n = 23).

Longitudinal analysis of the gut microbiome

Prior work has documented the overall stability of the microbiome in the short and long term, with the ability to recover after exposure to microbiome-altering drugs, changes in diet, or colonization by pathogens.[65–67] We investigated changes in the gut microbiome for a subset of individuals (n = 301) who were re-sequenced after 2 years in 2022.

We observed a significant decrease in Shannon diversity and in the number of observed species between the two timepoints (Wilcoxon rank-sum paired test, Shannon diversity p = 0.025, and number of species $p = 1.5 \times 10^{-7}$). We also found an overall change in species composition structure (PERMANOVA p = 0.007, Figure 6A). The variation of the microbiome structure across time points within subjects was less than that observed between individuals (within-subject versus between-subject Bray-Curtis dissimilarity, Wilcoxon rank-sum test: $p < 2.2 \times 10^{-16}$, Figure 6C). We did not observe a significant change in the beta diversity of the functional potential, which was observed to be more stable than the taxonomic composition, suggesting preservation of microbiome functionalities.

Intra-individual bacterial species stability, measured with the intra-class correlation coefficient (ICC) on most prevalent species (prevalence greater than 0.5), showed very low stability, with a median value of 0.09, with 17 species having an ICC value greater than 0.3. We then compared ICC values calculated on a longitudinal subset of samples from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB), including only samples from healthy donors,[16] and showed higher species stability (ICC median 0.25, max 0.94).

A similar pattern of changes in composition was observed when looking at individual strain composition, with a median strain-retention rate of 0.26 (lower than other cohorts[44]). Looking into individual species retention rates, we found that several

Prevotella strains are the most stable, with one out of three individuals retaining the same strain after 2 years.

These findings suggest that the gut microbiome of a traditional lifestyle in the non-western population might be more prone to variation over time, compared to western populations, given the fact that changes in medication usage, diet composition, or major diseases were not observed (although the intervening pandemic of COVID-19 may have played a role). In recent years, there has been growing evidence that infection with COVID-19 is associated with changes in the composition of the intestinal tract. In the follow-up visit, individuals were asked to report whether they had an infection confirmed by a doctor, and most of the individuals reported not having been infected with COVID-19 (n = 296).

We then compared the self-reported assessments with results from a rapid immunoassay test to identify antibodies to SARS-CoV-2. A large fraction of the tested individuals (n = 264) tested positive for the presence of immunoglobulin G (IgG) only, while only 3 individuals were positive for both IgG and IgM. Hence, we compared microbial profiles between the two timepoints using MaAsLin2 (STAR Methods). We identified 286 species differentially abundant between individuals before and after infection with SARS-CoV-2 (Figures 6D; Tables S13 and S14). *Coprococcus comes* and *Dorea* sp. AF36-15AT were the two species most enriched at the second time point among individuals positive for the presence of IgG only, while several Spirochaetia and Elusimicrobiales species were depleted among the same group of individuals. We also observed a major reduction in the relative abundance of several species from the Prevotellaceae family, although they remained the most dominant taxa.

Finally, viral diversity was observed to be increased over the 2-year period (median T1 116 vOTUs, median T2 131 vOTUs, Wilcoxon rank-sum test p = 0.0005).

Human genetic variation and microbiome associations

The genetic background of Honduras has been described as a mix of mostly Spanish, African, and native American ancestries, primarily Mayan.[68] Copán, the region in which our cohort is located, was the center of the Mayan civilization, and most of the people still identify themselves as Maya Ch'orti'.

Using low-pass sequencing and genotype imputation, we performed genomic characterization of 1,889 individuals in our cohort. The average sequencing coverage for the samples was 4X. The final dataset resulted in 70,723,799 SNVs identified in 1,701 samples, after retaining individuals with both human genome sequencing and shotgun metagenomic microbiome sequencing samples.

After quality control, we retained 227,216 SNVs, and we proceeded to compare human genetic variation with microbial variation. The first 10 genetic components from the principal-component analysis explained most of the genetic variance in the dataset, and we proceeded to calculate the microbial variation that could be explained by host genetic variation in species and MetaCyc pathways composition. To reduce the total number of tests performed, we filtered for SGB present in at least 20% of the population, retaining a total of 557 species. Using linear models, we fit each microbial species' relative abundance, predicting it with the first 10 genetic components in order to calculate the percentage of explained variance (Figure S12; Table S15).

The total bacterial species variation explained by the genetic component was 2%, with 125 out of 557 species reaching statistical significance. One species in particular, Lachnospiraceae SGB4777, reached an R^2 of 9.7% (Permutation test $p = 1 \times 10^{-6}$) and was found to be associated with the PC1, PC2, and PC8.

We performed the same analysis on 571 MetaCyc pathway abundances, which showed a smaller amount of total variance explained (0.5%), but a larger number of pathways reaching statistical significance (false discovery rate [FDR] < 0.05, n = 289). The top pathways with higher R^2 values are bacterial pathways involved in the biosynthesis of lipids, amino acids, cofactors, and carbohydrates (Table S16).

We then examined the association between genetic variation and overall microbial diversity, as summarized by the weighted UniFrac distance. The microbiomeGWAS tool[69] identified 15 SNVs with minor-allele frequency (MAF) greater than 0.05 associated with differences in beta diversity (Table S17). Variants were identified close or within the genes *DAB1*, *AP1S3*, *HTT*, *COX7B2*, *GABRA4*, *ARSB*, *RASGRF2*, *CALN1*, *NCKAP1L*, *DGKH*, and *CYB5B*.

Next, to study the association between individual species and host variation, we performed a GWAS on SNVs with MAF greater than 0.05 and genus-level, species-level, and MetaCyc pathways abundances using Matrix eQTL (quantitative trait locus) (Figure S13).

After combining genus- and species-level relative abundances and treating them as microbial traits, we tested 9,399,332 microbiome-QTLs and identified 280 associations with FDR < 0.05 (Table S18). Only 7 microbiome-QTLs reached the study-wide statistical significance threshold ($p = 2.93 \times 10^{-11}$) (Figure S14). Several SGB from the Lachnospiraceae and Ruminococcaceae were found associated with distinct SNVs as well as the uncharacterized Ruminococcaceae genus GGB9729.

The strongest signal was identified on chr22:22208496, located within the *IGL* immunoglobulin lambda locus gene. Abundances of Lachnospiraceae SGB4817 were found to be decreased in individuals with one or two copies of the T allele.

When looking at the SNVs identified as significant by both microbiomeGWAS and microbiome-QTL models, we identified a total of 2 SNVs in common between the two models, rs7679715 and rs17599102, both located on chromosome 4; rs17599102 is located on the *GABRA4* gene and rs7679715 on the *COX7B2* gene, but still proximal to *GABRA4*. Individuals homozygous for both the rs17599102 A/A and rs7679715 C/C genotypes showed decreased median abundances of Ruminococcaceae SGB15234 and Clostridia SGB4372 and smaller beta-diversity.

We tested associations between SNV and pathways as well, with a total of 7,378,167 microbiome-QTLs. Although no microbiome-QTLs reached the study-wide significance threshold, 119 were statistically significant with FDR < 0.05 (Table S19).

Using KING, we calculated the pairwise kinship coefficient for each pair of participants (n = 1,622,078 pairs), and kinship values were grouped in degrees of relation by using pre-calculated thresholds (STAR Methods). Participants in our cohort are, on

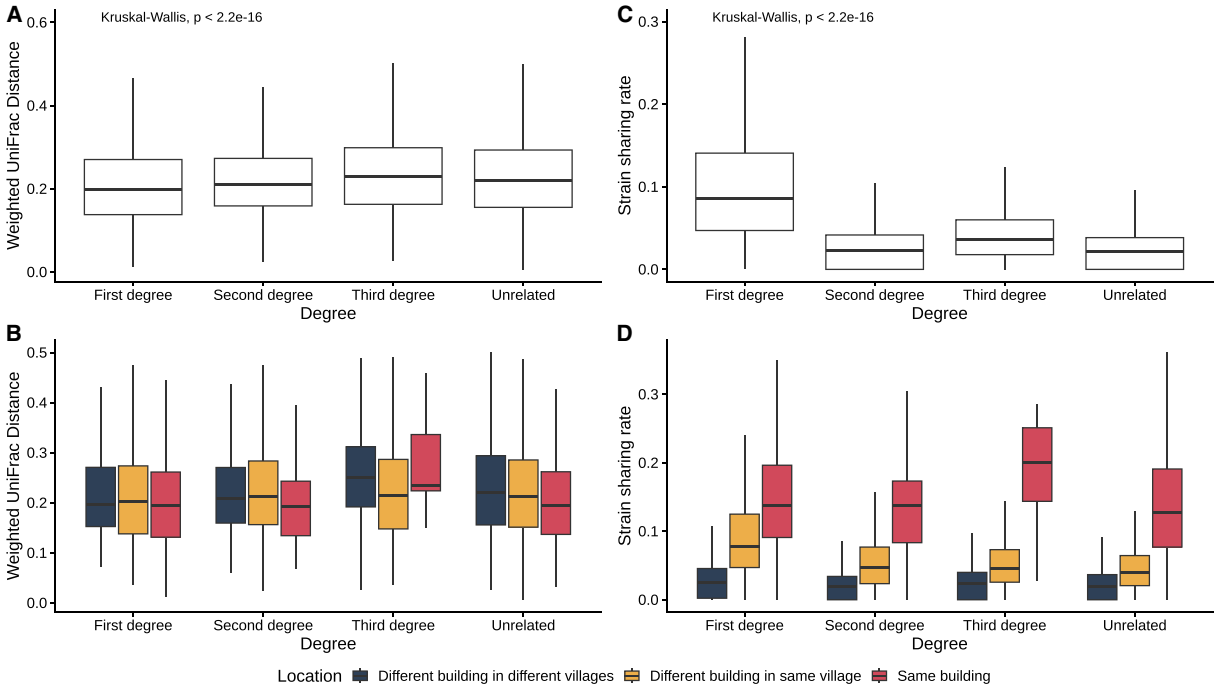


Figure 7. Microbiome similarity and relatedness

Distribution of the weighted UniFrac distance (A–B) and strain sharing rate (C–D) between different relationship degrees. All comparisons between microbiome similarity and location within degrees are statistically significant (Dunn test $p < 0.05$). Boxplot whiskers extend from the hinge to no further than 1.5× IQR from the hinge.

average, related to each other at the second degree ($\phi = 0.1313$), and this was observed when looking at the cohort as a whole and each separate village. Although some villages have both higher and lower kinship coefficients, we did not find any correlation with village size.

Using the weighted UniFrac distance and the strain-sharing rate calculated for all the pairs of individuals, we looked at the similarity of the microbiome of people who are more genetically similar (Figure 7). The distribution of the microbiome similarity among pairs of people varies according to the three levels of relationship degree—first to second degree, second to third degree, and third degree to unrelated—such that more genetically similar people have lower UniFrac distances (regardless of whether such a pair of people actually has a relationship) (Kruskal-Wallis test $p < 2.2 \times 10^{-16}$). However, this analysis does not hold if we restrict the pairs of people to be village co-residents. Because of the overwhelming shared environment, increased genetic similarity within a village is not associated with increased microbiome similarity, as this effect reflects shared environment and lifestyle rather than shared genetics per se.[70,71] No comparison between individuals living in the same building across the relationship degrees reached statistical significance (Dunn Test $p > 0.05$). In all the cases observed, the microbiomes of related pairs resemble unrelated individuals less (albeit with $p = 0.59$).

We observed a similar trend when looking at the number of strains shared between individuals. Regardless of their relationship degree, the number of strains shared across individuals living in different villages is relatively low, but it is higher in individuals living in the same village and in the same building. We observed an increased strain-sharing rate among third-degree-related individuals living in the same building, compared to first- and second-degree pairs. We hypothesized this was due to the inclusion of pairs of partnered individuals, as we previously observed higher relatedness between participants, but the median kinship coefficients among partners ($\phi = 0.1425$) are closer to a second-degree relationship than a third-degree relationship.

When looking at actual relationships between individuals known to be socially connected (since we have detailed social network maps of the villages),[1] and, after correcting for several covariates (such as BMI, age, gender, and diet score), we observed that genetically similar individuals with an existing social relationship (such as a friend) have increased strain-sharing rate (linear mixed model kinship $\beta = 0.11$, $p < 2 \times 10^{-16}$; relationship $\beta = 0.84$, $p < 2 \times 10^{-16}$; interaction between kinship and relationship $\beta = 4.713 \times 10^{-2}$, $p = 1.42 \times 10^{-5}$). In other words, information about the genetic similarity of socially connected people further enhances the ability to predict their microbiome similarity, and this also holds even after accounting for diet, medication usage, and socio-economic and demographic status (which have an effect on the existence of a relationship and not on the increase of shared strains).

DISCUSSION

We present a comprehensive characterization of the gut microbiome of an isolated, rural Honduran population, significantly expanding our understanding of microbial diversity in non-western settings. Through shotgun metagenomic sequencing, we generated one of the most extensive microbiome datasets in Central America, uncovering uncharacterized bacterial species by reconstructing MAGs. These findings expand previous observations that traditional-lifestyle communities harbor unique microbial ecosystems that are distinct from those of industrialized populations, reinforcing the idea that modernization, via dietary changes, antibiotic usage, and other means, has contributed to a substantial reshaping of the human gut microbiome.

By integrating our dataset with global microbiome cohorts, we identified several *Prevotella* species found in higher abundance in the Honduran gut microbiome compared to other non-western populations, and the presence of clades strongly associated with our cohort. Alongside the geographical association, higher genetic diversity within species was observed, suggesting divergent clades or the carriage of multiple strains. We also observed that non-western cohorts have lower richness of CAZy modules compared to the western cohorts, while Honduras showed higher richness than both the western and non-western cohorts. This is in contrast to previous reports,[8] where higher CAZy richness was reported in non-western cohorts, probably due to the fact that previous comparisons were drawn from under-sampled populations. This can be put in the context of the Honduras diet: beans, a source of complex carbohydrates, needs to be broken down by the GH3 enzyme family, which was found in increased abundance, and more prevalent in *Prevotella* and *Butyribacter* species that are able to ferment fiber and produce SCFAs. The observation of a lower abundance of GH13 enzymes seems to be in contradiction to its function of digesting starches, as rice is a predominant part of the everyday diet as well. When cooked and cooled, starches like rice undergo a process called retrogradation, which creates a fiber-like starch, further fueling fermentation in the colon rather than being absorbed as sugar.

We also expanded the UHGG by reconstructing and identifying prokaryotic species and proteins. In addition to prokaryotic diversity, we explored the broader microbial ecosystem by looking at eukaryotic and viral components. We identified and reconstructed 55 genomes of several subtypes of *Blastocystis* spp., a common gut protist. Notably, our findings suggest a higher prevalence of *Blastocystis* in individuals with greater microbiome diversity, reinforcing its potential role as a marker of gut ecosystem richness. We also reconstructed several thousand viral genomes, expanding the viral mappability of our metagenomes by over half, including uncharacterized bacteriophages and members of the Crassvirales family, which are thought to play critical roles in microbiome stability and bacterial population control.

The longitudinal component of our study provided insights into microbiome stability and temporal dynamics. Our sample collection effort began in the early months of 2020, but some villages were not targeted as we had to postpone the collection to 2022 due to restrictions caused by the COVID-19 pandemic. This fortuitous change in the collection process allowed us to observe significant shifts in microbial composition, with notable changes in taxonomic diversity and functional profiles. While the gut microbiome exhibited resilience, the overall microbial diversity declined over time, a pattern that differs from observations in the western cohorts. This could be attributed to several factors, including possibly the COVID-19 pandemic, which caused significant microbiome alterations.

By leveraging low-pass human genome sequencing, we identified several microbial species and metabolic pathways that exhibit significant associations with host genetic components. These findings suggest that specific genetic factors may influence microbial colonization and metabolic potential, reinforcing the concept of microbiome adaptation to the host. Our genome-wide association analysis further identified specific genetic loci linked to microbiome variation, shedding light on potential host determinants of gut microbial diversity.

Overall, this study represents an important step forward in uncovering the vast microbial diversity of under-represented human populations. By characterizing the Honduran gut microbiome, we provide insights into microbial ecology and host-microbiome interactions. Our findings underscore the importance of expanding microbiome research beyond industrialized nations to gain a more comprehensive and global perspective on human microbial diversity.

Limitations of the study

A limitation of this study is that our sample was from a particular setting of isolated villages, and this may capture a founder effect. In addition, our longitudinal sampling targeted only a small fraction of the full-sized cohort, which limited the resolution for a better understanding of the effect of the SARS-CoV-2 pandemic and other factors.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Francesco Beghini (francesco.beghini@yale.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- Metagenomic reads and associated sample metadata have been deposited at the NCBI Sequence Read Archive database and Zenodo, respectively. Data are publicly available as of the date of publication, with accession numbers listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank the NOMIS Foundation, Rothberg Catalyzer, Schmidt Futures, Paul Graham, and Pershing Square Philanthropies for their support. The underlying cohort was supported by the Gates Foundation.

AUTHOR CONTRIBUTIONS

F.B., I.L.B., and N.A.C. conceived and designed the study; N.A.C. supervised the project; F.B. and N.A.C. contributed to the methodology design and

analytical approach; F.B., I.L.B., and N.A.C. assembled the data; F.B. performed the analysis and interpreted the data; and F.B., I.L.B., M.G., and N.A.C. wrote the manuscript.

DECLARATION OF INTERESTS

Authors declare that they have no competing interests.

STAR★METHODS

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

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Enrollment of participants
  - Sample collection
  - Library preparation and sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Read-based taxonomic and potential functional profiling
  - Metagenomic assembly and contig binning
  - Alitiscatomonas sp900066535 pangenome analysis
  - Eukaryotic contig binning and prevalence estimation
  - Viral contig binning and prevalence estimation
  - Human genome sequencing
  - Association analyses

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2025.116724.

Received: April 8, 2025
Revised: August 10, 2025
Accepted: November 21, 2025
Published: December 18, 2025

REFERENCES

1. Beghini, F., Pullman, J., Alexander, M., Shridhar, S.V., Prinster, D., Singh, A., Matute Juárez, R., Airoldi, E.M., Brito, I.L., and Christakis, N.A. (2025). Gut microbiome strain-sharing within isolated village social networks. Nature *637*, 167–175.
2. Shridhar, S.V., Beghini, F., Alexander, M., Singh, A., Juárez, R.M., Brito, I.L., and Christakis, N.A. (2024). Environmental, socioeconomic, and health factors associated with gut microbiome species and strains in isolated Honduras villages. Cell Rep. *43*, 114442.
3. Gacesa, R., Kurilshikov, A., Vich Vila, A., Sinha, T., Klaassen, M.A.Y., Bolte, L.A., Andreu-Sánchez, S., Chen, L., Collij, V., Hu, S., et al. (2022). Environmental factors shaping the gut microbiome in a Dutch population. Nature *604*, 732–739. https://doi.org/10.1038/s41586-022-04567-7.
4. Abdill, R.J., Adamowicz, E.M., and Blekhman, R. (2022). Public human microbiome data are dominated by highly developed countries. PLoS Biol. *20*, e3001536.
5. 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.
6. Segata, N. (2015). Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity. Curr. Biol. *25*, R611–R613.
7. Sonnenburg, J.L., and Sonnenburg, E.D. (2019). Vulnerability of the industrialized microbiota. Science *366*, eaaw9255.
8. Sonnenburg, E.D., and Sonnenburg, J.L. (2019). The ancestral and industrialized gut microbiota and implications for human health. Nat. Rev. Microbiol. *17*, 383–390.
9. Asnicar, F., Berry, S.E., Valdes, A.M., Nguyen, L.H., Piccinno, G., Drew, D.A., Leeming, E., Gibson, R., Le Roy, C., Khatib, H.A., et al. (2021). Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat. Med. *27*, 321–332.
10. Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe *17*, 690–703.
11. Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W., Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. Nature *535*, 435–439.
12. Dhakan, D.B., Maji, A., Sharma, A.K., Saxena, R., Pulikkan, J., Grace, T., Gomez, A., Scaria, J., Amato, K.R., and Sharma, V.K. (2019). The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigaScience *8*, giz004. https://doi.org/10.1093/gigascience/giz004.
13. Hall, A.B., Yassour, M., Sauk, J., Garner, A., Jiang, X., Arthur, T., Lagoudas, G.K., Vatanen, T., Fornelos, N., Wilson, R., et al. (2017). A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. Genome Med. *9*, 103.
14. Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.
15. Schirmer, M., Franzosa, E.A., Lloyd-Price, J., McIver, L.J., Schwager, R., Poon, T.W., Ananthakrishnan, A.N., Andrews, E., Barron, G., Lake, K., et al. (2018). Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nat. Microbiol. *3*, 337–346.
16. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662.
17. Kaur, K., Khatri, I., Akhtar, A., Subramanian, S., and Ramya, T.N.C. (2020). Metagenomics analysis reveals features unique to Indian distal gut microbiota. PLoS One *15*, e0231197.
18. Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xi, X., Liang, Z., Hou, Q., Zhou, B., et al. (2016). Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. Sci. Rep. *6*, 34826.
19. Lokmer, A., Cian, A., Froment, A., Gantois, N., Viscogliosi, E., Chabé, M., and Ségurel, L. (2019). Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. PLoS One *14*, e0211139.
20. Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. *6*, 6505.
21. 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.
22. Pehrsson, E.C., Tsukayama, P., Patel, S., Mejía-Bautista, M., Sosa-Soto, G., Navarrete, K.M., Calderon, M., Cabrera, L., Hoyos-Arango, W., Bertoli, M.T., et al. (2016). Interconnected microbiomes and resistomes in low-income human habitats. Nature *533*, 212–216.
23. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

24. Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. Nature *513*, 59–64.

25. Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., and Candela, M. (2015). Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr. Biol. *25*, 1682–1693.

26. Rosa, B.A., Supali, T., Gankpala, L., Djuardi, Y., Sartono, E., Zhou, Y., Fischer, K., Martin, J., Tyagi, R., Bolay, F.K., et al. (2018). Differential human gut microbiome assemblages during soil-transmitted helminth infections in Indonesia and Liberia. Microbiome *6*, 33.

27. Rubel, M.A., Abbas, A., Taylor, L.J., Connell, A., Tanes, C., Bittinger, K., Ndze, V.N., Fonsah, J.Y., Ngwang, E., Essiane, A., et al. (2020). Lifestyle and the presence of helminths is associated with gut microbiome composition in Cameroonians. Genome Biol. *21*, 122.

28. Smits, S.A., Leach, J., Sonnenburg, E.D., Gonzalez, C.G., Lichtman, J.S., Reid, G., Knight, R., Manjurano, A., Changalucha, J., Elias, J.E., et al. (2017). Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science *357*, 802–806.

29. Tamburini, F.B., Maghini, D., Oduaran, O.H., Brewster, R., Hulley, M.R., Sahibdeen, V., Norris, S.A., Tollman, S., Kahn, K., Wagner, R.G., et al. (2022). Short- and long-read metagenomics of urban and rural South African gut microbiomes reveal a transitional composition and undescribed taxa. Nat. Commun. *13*, 926.

30. Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., et al. (2019). The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. Cell Host Microbe *26*, 666–679.e7.

31. Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat. Med. *25*, 968–976.

32. Yassour, M., Jason, E., Hogstrom, L.J., Arthur, T.D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S.M., et al. (2018). Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe *24*, 146–154.e4.

33. Ye, Z., Zhang, N., Wu, C., Zhang, X., Wang, Q., Huang, X., Du, L., Cao, Q., Tang, J., Zhou, C., et al. (2018). A metagenomic study of the gut microbiome in Behcet's disease. Microbiome *6*, 135.

34. Yu, J., Feng, Q., Wong, S.H., Zhang, D., Liang, Q.Y., Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut *66*, 70–78.

35. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized nutrition by prediction of glycemic responses. Cell *163*, 1079–1094.

36. Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. *10*, 766.

37. Zhu, F., Ju, Y., Wang, W., Wang, Q., Guo, R., Ma, Q., Sun, Q., Fan, Y., Xie, Y., Yang, Z., et al. (2020). Metagenome-wide association of gut microbiome features for schizophrenia. Nat. Commun. *11*, 1612.

38. Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat. Biotechnol. *41*, 1633–1644.

39. Blanco-Míguez, A., Gálvez, E.J.C., Pasolli, E., De Filippis, F., Amend, L., Huang, K.D., Manghi, P., Lesker, T.-R., Riedel, T., Cova, L., et al. (2023). Extension of the Segatella copri complex to 13 species with distinct large extrachromosomal elements and associations with host conditions. Cell Host Microbe *31*, 1804–1819.e9.

40. 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.

41. Agus, A., Clément, K., and Sokol, H. (2021). Gut microbiota-derived metabolites as central regulators in metabolic disorders. Gut *70*, 1174–1182.

42. Wibowo, M.C., Yang, Z., Borry, M., Hübner, A., Huang, K.D., Tierney, B.T., Zimmerman, S., Barajas-Olmos, F., Contreras-Cubas, C., García-Ortiz, H., et al. (2021). Reconstruction of ancient microbial genomes from the human gut. Nature *594*, 234–239.

43. Suzuki, T.A., Fitzstevens, J.L., Schmidt, V.T., Enav, H., Huus, K.E., Mbong Ngwese, M., Grießhammer, A., Pfleiderer, A., Adegbite, B.R., Zinsou, J.F., et al. (2022). Codiversification of gut microbiota with humans. Science *377*, 1328–1332.

44. Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. *27*, 626–638.

45. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731.

46. Chibani, C.M., Mahnert, A., Borrel, G., Almeida, A., Werner, A., Brugère, J.-F., Gribaldo, S., Finn, R.D., Schmitz, R.A., and Moissl-Eichinger, C. (2022). A catalogue of 1,167 genomes from the human gut archaeome. Nat. Microbiol. *7*, 48–61.

47. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics *38*, 5315–5316.

48. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.

49. Maghini, D.G., Oduaran, O.H., Wirbel, J., Olubayo, L.A.I., Smyth, N., Mathema, T., Belger, C.W., Agongo, G., Boua, P.R., Choma, S.S., et al. (2024). Expanding the human gut microbiome atlas of Africa. Preprint at bioRxiv. https://doi.org/10.1101/2024.03.13.584859.

50. Carter, M.M., Olm, M.R., Merrill, B.D., Dahan, D., Tripathi, S., Spencer, S.P., Yu, F.B., Jain, S., Neff, N., Jha, A.R., et al. (2023). Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. Cell *186*, 3111–3124.e13.

51. Perea, M., Vásquez, V., Pineda, V., Samudio, F., Calzada, J.E., and Saldaña, A. (2020). Prevalence and subtype distribution of Blastocystis sp. infecting children from a rural community in Panama. Parasite Epidemiol. Control *9*, e00139.

52. Beghini, F., Pasolli, E., Truong, T.D., Putignani, L., Cacciò, S.M., and Segata, N. (2017). Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. ISME J. *11*, 2848–2863. https://doi.org/10.1038/ismej.2017.139.

53. Rojas-Velázquez, L., Morán, P., Serrano-Vázquez, A., Fernández, L.D., Pérez-Juárez, H., Poot-Hernández, A.C., Portillo, T., González, E., Hernández, E., Partida-Rodríguez, O., et al. (2018). Genetic diversity and distribution of Blastocystis subtype 3 in human populations, with special reference to a rural population in central Mexico. BioMed Res. Int. *2018*, 3916263.

54. Nieves-Ramírez, M.E., Partida-Rodríguez, O., Laforest-Lapointe, I., Reynolds, L.A., Brown, E.M., Valdez-Salazar, A., Morán-Silva, P., Rojas-Velázquez, L., Morien, E., Parfrey, L.W., et al. (2018). Asymptomatic intestinal colonization with protist Blastocystis is strongly associated

with distinct microbiome ecological patterns. mSystems *3*, e00007. https://doi.org/10.1128/mSystems.00007-18.

55. Health Organization, W. (2012). Accelerating work to overcome the global impact of neglected tropical diseases a roadmap for implementation. https://apps.who.int/iris/bitstream/handle/10665/338712/WHO-HTM-NTD-2012.5-eng.pdf.
56. West, P.T., Probst, A.J., Grigoriev, I.V., Thomas, B.C., and Banfield, J.F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome Res. *28*, 569–580.
57. Andersen, L.O., and Stensvold, C.R. (2016). Blastocystis in Health and Disease: Are We Moving from a Clinical to a Public Health Perspective? J. Clin. Microbiol. *54*, 524–528.
58. Piperni, E., Nguyen, L.H., Manghi, P., Kim, H., Pasolli, E., Andreu-Sánchez, S., Arrè, A., Bermingham, K.M., Blanco-Míguez, A., Manara, S., et al. (2024). Intestinal Blastocystis is linked to healthier diets and more favorable cardiometabolic outcomes in 56,989 individuals from 32 countries. Cell *187*, 4554–4570.e18.
59. Deng, L., Wojciech, L., Gascoigne, N.R.J., Peng, G., and Tan, K.S.W. (2021). New insights into the interactions between Blastocystis, the gut microbiota, and host immunity. PLoS Pathog. *17*, e1009253.
60. Castañeda, S., Muñoz, M., Villamizar, X., Hernández, P.C., Vásquez, L.R., Tito, R.Y., and Ramírez, J.D. (2020). Microbiota characterization in Blastocystis-colonized and Blastocystis-free school-age children from Colombia. Parasit. Vectors *13*, 521.
61. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585.
62. Nayfach, S., Páez-Espino, D., Call, L., Low, S.J., Sberro, H., Ivanova, N.N., Proal, A.D., Fischbach, M.A., Bhatt, A.S., Hugenholtz, P., and Kyrpides, N.C. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat. Microbiol. *6*, 960–970.
63. Camargo, A.P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P.S.G., Nayfach, S., and Kyrpides, N.C. (2024). Identification of mobile genetic elements with geNomad. Nat. Biotechnol. *42*, 1303–1312.
64. Yutin, N., Benler, S., Shmakov, S.A., Wolf, Y.I., Tolstoy, I., Rayko, M., Antipov, D., Pevzner, P.A., and Koonin, E.V. (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. Nat. Commun. *12*, 1044.
65. Faith, J.J., Guruge, J.L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A.L., Clemente, J.C., Knight, R., Heath, A.C., Leibel, R.L., et al. (2013). The long-term stability of the human gut microbiota. Science *341*, 1237439.
66. Gilbert, J.A., Blaser, M.J., Caporaso, J.G., Jansson, J.K., Lynch, S.V., and Knight, R. (2018). Current understanding of the human microbiome. Nat. Med. *24*, 392–400.
67. Chen, L., Wang, D., Garmaeva, S., Kurilshikov, A., Vich Vila, A., Gacesa, R., Sinha, T., Segal, E., Weersma, R.K., and Wijmenga, C. (2021). The long-term genetic stability and individual specificity of the human gut microbiome. Cell *184*, 2302–2315.e12.
68. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the population genetic history of the Caribbean. PLoS Genet. *9*, e1003925.
69. Hua, X., Song, L., Yu, G., Vogtmann, E., Goedert, J.J., Abnet, C.C., Landi, M.T., and Shi, J. (2022). MicrobiomeGWAS: A tool for identifying host genetic variants associated with microbiome composition. Genes *13*, 1224.
70. Vilchez-Vargas, R., Skieceviciene, J., Lehr, K., Varkalaite, G., Thon, C., Urba, M., Morkūnas, E., Kucinskas, L., Bauraite, K., Schanze, D., et al. (2022). Gut microbial similarity in twins is driven by shared environment and aging. EBioMedicine *79*, 104011.
71. Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M.A., Knight, R., Ober, C., Spector, T.D., Bell, J.T., Clark, A.G., and Ley, R.E. (2016). Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe *19*, 731–743.
72. Cantu, V.A., Sadural, J., and Edwards, R. (2019). PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. PeerJ. https://doi.org/10.7287/peerj.preprints.27553.
73. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
74. Blanco-Miguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., et al. (2022). Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. Preprint at bioRxiv. https://doi.org/10.1101/2022.08.22.504593.
75. Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife *10*, e65088. https://doi.org/10.7554/eLife.65088.
76. Mallick, H., Rahnavard, A., McIver, L.J., Ma, S., Zhang, Y., Nguyen, L.H., Tickle, T.L., Weingart, G., Ren, B., Schwager, E.H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. PLoS Comput. Biol. *17*, e1009442.
77. Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834.
78. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.
79. Nissen, J.N., Johansen, J., Allesøe, R.L., Sønderby, C.K., Armenteros, J.J.A., Grønbech, C.H., Jensen, L.J., Nielsen, H.B., Petersen, T.N., Winther, O., and Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. Nat. Biotechnol. *39*, 555–560.
80. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.
81. Woodcroft, B.J. (2024). galah: More scalable dereplication for metagenome assembled genomes (Github).
82. Schwengers, O., Jelonek, L., Dieckmann, M.A., Beyvers, S., Blom, J., and Goesmann, A. (2021). Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb. Genom. *7*, 000685. https://doi.org/10.1099/mgen.0.000685.
83. Aroney, S.T.N., Newell, R.J.P., Nissen, J.N., Camargo, A.P., Tyson, G.W., and Woodcroft, B.J. (2025). CoverM: Read alignment statistics for metagenomics. Preprint at arXiv. https://doi.org/10.48550/ARXIV.2501.11217.
84. Zimmermann, J., Kaleta, C., and Waschina, S. (2021). Gapseq: Informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. Genome Biol. *22*, 81.
85. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.
86. Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat. Commun. *11*, 2500.
87. Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J.A., Gladstone, R.A., Lo, S., Beaudoin, C., Floto, R.A., et al. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. *21*, 180.
88. Huerta-Cepas, J., Forslund, K., Szklarczyk, D., Jensen, L.J., von Mering, C., and Bork, P. (2016). Fast Genome-wide Functional Annotation

through Orthology Assignment by eggNOG-Mapper (Cold Spring Harbor Laboratory), pp. 076331. https://doi.org/10.1101/076331.

89. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359.
90. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing genome assembly and annotation completeness. Methods Mol. Biol. *1962*, 227–245.
91. Johansen, J., Plichta, D.R., Nissen, J.N., Jespersen, M.L., Shah, S.A., Deng, L., Stokholm, J., Bisgaard, H., Nielsen, D.S., Sørensen, S.J., and Rasmussen, S. (2022). Genome binning of viral entities from bulk metagenomics data. Nat. Commun. *13*, 965.
92. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
93. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
94. Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat. Genet. *53*, 120–126.
95. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
96. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.
97. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
98. Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353–1358.
99. 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.
100. Airoldi, E.M., and Christakis, N.A. (2024). Induction of social contagion for diverse outcomes in structured experiments in isolated villages. Science *384*, eadi5147. https://doi.org/10.1126/science.adi5147.
101. BMTagger ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/.
102. Valles-Colomer, M., Blanco-Míguez, A., Manghi, P., Asnicar, F., Dubois, L., Golzato, D., Armanini, F., Cumbo, F., Huang, K.D., Manara, S., et al. (2023). The person-to-person transmission landscape of the gut and oral microbiomes. Nature *614*, 1–11.
103. Stražar, M., Temba, G.S., Vlamakis, H., Kullaya, V.I., Lyamuya, F., Mmbaga, B.T., Joosten, L.A.B., van der Ven, A.J.A.M., Netea, M.G., de Mast, Q., and Xavier, R.J. (2021). Gut microbiome-mediated metabolism effects on immunity in rural and urban African populations. Nat. Commun. *12*, 4845.
104. Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. *38*, 1079–1086.
105. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J., and Levy Karin, E. (2021). Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics *37*, 3029–3031.
106. Aroney, S.T.N., Newell, R.J.P., Nissen, J., Camargo, A.P., Tyson, G.W., and Woodcroft, B.J. (2024). CoverM: Read Coverage Calculator for Metagenomics (Zenodo). https://doi.org/10.5281/ZENODO.10531254.
107. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
chemagic Stool gDNA extraction kit	Revvity, Inc.	CMG-1076
KAPA Hyper Library Preparation Kit	KAPA Biosystems	Part#KK8504
Deposited data		
Metagenomic reads	This paper	NCBI SRA BioProject PRJNA999635
Sample metadata	This paper	https://doi.org/10.5281/zenodo.17289392
Software and algorithms		
R (version 4.2.0)	R Core Team	https://www.r-project.org/
Prinseq lite (version 0.20.2)	Cantu et al.[72]	https://github.com/uwb-linux/prinseq
BMTagger	N/A	https://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger
Trimmomatic (version 0.36)	Bolger et al.[73]	https://github.com/timflutre/trimmomatic
MetaPhlAn (version 4.0.1)	Blanco-Miguez et al.[74]	https://github.com/biobakery/MetaPhlAn
HUMAnN (version 3.0)	Beghini et al.[75]	https://github.com/biobakery/HUMAnN
MaAsLin 2	Mallick et al.[76]	https://github.com/biobakery/Maaslin2
metaSPADES (version 3.13.1)	Nurk et al.[77]	https://github.com/ablab/spades/
minimap2 (version 2.24)	Li[78]	https://github.com/lh3/minimap2/
VAMB (version 3.0.9)	Nissen et al.[79]	https://github.com/RasmussenLab/vamb/
checkM (version 2.0)	Parks et al.[80]	https://github.com/chklovski/CheckM2
Galah (version 0.4.0)	Woodcroft[81]	https://github.com/wwood/galah
GTDB-tk (version 2.3.2)	Chaumeil et al.[47]	https://github.com/Ecogenomics/GTDBTk
Bakta (version 4.1)	Schwengers et al.[82]	https://github.com/oschwengers/bakta
CoverM (version 0.6.1)	Aroney et al.[83]	https://github.com/wwood/CoverM
gapseq (version 1.3.1)	Zimmermann et al.[84]	https://github.com/jotech/gapseq
MMseqs2 (Commit abab073)	Steinegger and Söding[85]	https://github.com/soedinglab/mmseqs2
PhyloPhlAn (version 3.1)	Asnicar et al.[86]	https://github.com/biobakery/phylophlan
Panaroo (version 1.4.0)	Tonkin-Hill et al.[87]	https://github.com/gtonkinhill/panaroo/
eggnog-mapper (version 2.1.12)	Huerta-Cepas et al.[88]	https://github.com/eggnogdb/eggnog-mapper/
EukRep (version 0.6.6)	West et al.[56]	https://github.com/patrickwest/EukRep
Metabat2 (version 2.15)	Kang et al.[89]	https://bitbucket.org/berkeleylab/metabat/
BUSCO (version 5)	Seppey et al.[90]	https://gitlab.com/ezlab/busco
PHAMB (version 1.0.1)	Johansen et al.[91]	https://github.com/RasmussenLab/phamb
CheckV (version 1.0.1)	Nayfach et al.[61]	https://bitbucket.org/berkeleylab/checkv/src/master/
Bowtie 2 (version 2.5.1)	Langmead and Saltzberg[92]	https://github.com/BenLangmead/bowtie2
Genomad (version 1.7.4)	Camargo et al.[63]	https://github.com/apcamargo/genomad
BLAST (version 2.5.0)	Altschul et al.[93]	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast
NVIDIA Clara Parabricks (version 4.1.0)	N/A	https://www.nvidia.com/en-us/clara/genomics/
GLIMPSE (version 1.1.1)	Rubinacci et al.[94]	https://github.com/odelaneau/glimpse
bcftools (version 1.90)	Danecek et al.[95]	https://github.com/samtools/bcftools
PLINK (version 1.90b7.1)	Chang et al.[96]	https://github.com/chrchang/plink-ng/
KING (version 2.3.2)	Manichaikul et al.[97]	https://www.kingrelatedness.com/
Matrix eQTL (version 2.3)	Shabalin[98]	https://cran.r-project.org/web/packages/MatrixEQTL/index.html
MicrobiomeGWAS (version 1.0)	Hua et al.[69]	https://github.com/lsncibb/microbiomeGWAS
Other		
curatedMetagenomicData	Pasolli et al.[99]	https://github.com/waldronlab/curatedMetagenomicData
BioRender	N/A	https://BioRender.com/5nc7ivu

METHOD DETAILS

Enrollment of participants

Our study was conducted in the department of Copán, Honduras, in an area of over 200 square miles of mountainous terrain with an estimated total population of 92,000 people. Villages and participants selected for this study were previously enrolled in the Honduras Social Network Targeting RCT project[100] aimed to affect reproductive, maternal, neonatal and child health outcomes by the adoption of practices and knowledge dissemination. Villages were chosen to vary by wealth, population size, elevation, isolation index, and time to the main road.

Village communities in western Honduras are characterized by geographic isolation and an economy reliant on subsistence agriculture and coffee cultivation. Geographical isolation is further exacerbated during the wet season when rainfall and flooding impede traveling. With healthcare facilities not usually located within the same village, residents are required to travel to a nearby city for clinical or hospital care. A high prevalence of daily intake of non-steroidal anti-inflammatory drugs use was reported, while most individuals had not consumed antibiotics in the preceding six months.

The local diet is predominantly composed of staples such as tortillas, rice, and beans, supplemented with eggs and chicken a few days per week.

Sample collection

Samples were collected between January and February 2020, before the SARS-CoV-2 pandemic put on halt the collection effort. Collection was resumed later in the spring of 2022. Participants were instructed on how to self-collect the fecal samples using in-person education and a leave-behind training manual and asked to return samples to the local team or to the designated community liaison within 1hr of producing them. Fecal samples are collected using sterile collection tubes with a scoop. Samples were then stored in liquid nitrogen at the collection site and then moved to a −80 C° in Copán Ruinas, Honduras. Samples were then shipped on dry ice to the USA and stored in −80°C freezers.

Library preparation and sequencing

DNA was extracted using the Chemagic Stool gDNA extraction kit and libraries were prepared using the KAPA Hyper Library Preparation kit. Shotgun metagenomic sequencing was performed on an Illumina NovaSeq 6000 platform at a targeted sequencing depth of 50Gbp. Samples not reaching the desired sequencing depth were resequenced on a separate run.

Raw metagenomic reads were deduplicated using prinseq lite (version 0.20.2,[72]) with default parameters. The resulting reads were screened for human contamination (hg19) with BMTagger[101] and then quality filtered with trimmomatic[73] (version 0.36, parameters ILLUMINACLIP:nextera_truseq_adapters.fasta:2:30:10:8:true SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:50).

QUANTIFICATION AND STATISTICAL ANALYSIS

Read-based taxonomic and potential functional profiling

Quantification of organisms' relative abundance was performed using MetaPhlAn 4[74], which internally mapped the metagenomes against a database of ∼5.1M marker genes describing more than 26∼ species-level genome bins (SGB). For species present in at least 10% of samples strain-level profiling was performed by StrainPhlAn 4[74] using default parameters.

Phylogenetic trees generated by StrainPhlAn were annotated with country of origin or world region as a trait, and phylogenetic signal was calculated by using the phylo.d function as implemented in the caper R package.

Functional potential analysis was performed using HUMAnN 3.0.[75] Gene family's profiles were normalized using relative abundances and collapsed into MetaCyc pathways and KO identifiers. CAZy profiles were obtained by regrouping the HUMAnN gene family profiles using a custom mapping file.

Microbiome species and CAZy richness and diversity was estimated using the Shannon entropy index and the total number of observed species (relative abundance greater than zero). In-silico rarefaction was performed before the calculation of each index using a custom script on the SAM file produced by MetaPhlAn. Multidimensional scaling analysis (vegan cmdscale function) was performed on the Bray-Curtis dissimilarity (vegan vegdist function) calculated on the relative abundances obtained by MetaPhlAn4 or HUMAnN 3.0 MetaCyc pathways.

Differential abundance analysis was performed using mixed-effect linear models as implemented in Maaslin2.[76] Basic covariates (age, gender, BMI) were included in all models as fixed effects as well as village as random effect. Longitudinal models included the individual ID as random effect. Multiple hypothesis testing was performed using the FDR procedure.

A total of 34 additional studies comprising 13,030 human gut metagenomes were downloaded by accessing curatedMetagenomicData[99] and processed with the same pipeline used to process our collected data. Five additional datasets, including samples from South Africa, Ghana, Tanzania, Kenya, Burkina Faso, Guinea Bissau, Argentina, Colombia, and Nepal,[29,49,50,102,103] not present in curatedMetagenomicData to date, were downloaded and processed using the same tools used before. Due to high memory requirement, 38 samples from the CarterM_2023 Tanzanian cohort were not included in the functional analysis.

Metagenomic assembly and contig binning

De-novo metagenomic assembly was performed on all samples using metaSPADES[77] (version 3.13.1, default parameters), and contigs longer than 1000bp were discarded. Metagenomic reads were mapped against the assembled contigs using minimap2[78] (version 2.24, parameters: -ax sr -t 12 -N50). The resulting alignments were used as input for the VAMB pipeline for contig binning[79] (version 3.0.9, parameters: –minfasta 500000). The automatic binning procedure generated a total of 226,039 MAGs that were then subjected to quality control to evaluate completeness and contamination using CheckM2[80] (version 1.0.2). We selected a total of 130,852 MAGs passing the thresholds for medium-quality genomes proposed by another study[45] being at least 50% complete and displaying less than 10% of contamination.

MAGs were dereplicated at 95% ANI using galah[81] and cluster representative, selected using the formula proposed by one study,[104] were taxonomic annotated with GTDB-tk[47] obtaining a total of 2,609 metagenomic species clusters. Raw reads were then mapped against the catalog of representative clusters with bwa-mem and the resulting alignments were processed by CoverM[83] to obtain relative abundances of the species clusters (parameters: min-read-aligned-percent 70, min-read-percent-identity 90, min-covered-fraction 10).

MAGs were annotated with Bakta[82] (version 4.1, default parameters) using the ''Full'' database, and predicted proteins from each MAG were functionally annotated using eggnog-mapper.[88]

Reconstruction of metabolic models was performed by gapseq[84] using the ''doall'' pipeline on each MAG, then species-level pangenomes were generated by ''gapseq pan''. Prediction of metabolic by-products was performed using the R package cobrar.

Proteins predicted by Bakta were iteratively clustered with MMseqs2[85] using the cluster pipeline to build non-redundant catalogs first at 100% sequence identity and 80% sequence coverage and then at 95% sequence identity and 80% coverage.

Species-level representatives identified by galah were used as input to PhyloPhlAn[86] to generate a maximum-likelihood phylogenetic tree (parameters -d phylophlan –diversity high –accurate).

Alitiscatomonas sp900066535 pangenome analysis

GFF annotations and associated genomic sequences for MAGs assigned with *Alitiscatomonas* sp900066535 taxonomy by GTDBtk were used to build the species' pangenome with panaroo[87] (–clean-mode moderate –merge_paralogs). The resulting presence/absence matrix was subsequently filtered for fragmented genes, pseudogenes, and genes with unusual lengths with the panaroo-filter-pa tool. Population structure was analyzed with PCA calculated using the Jaccard distance on the filtered presence/absence matrix. Enrichment of gene families was determined by fitting generalized linear models (GLM) on each pangenome feature using as outcome a binary variable signaling whether the MAG was reconstructed from Honduran samples. All p-values were FDR corrected.

Eukaryotic contig binning and prevalence estimation

Scaffolds generated by metaSPADES were screened for presence and recovery of eukaryotic contigs with EukRep[56] using default parameters. Samples with non-zero sequences were mapped against the original metagenome with minimap2 and then binned with metabat2. Eukaryotic genome completeness was estimated using BUSCO (v5)[90] using parameters '–auto-lineage-euk -m genome' and the 'eukaryota_odb10' lineage. Taxonomic assignment of binned contigs was performed with the MMseqs2 taxonomy workflow.[85,105]

The retrieved eukaryotic MAGs were used to build a minimap2 index and mapped against the original metagenomes using CoverM[106] to estimate their relative abundance using parameters ''—min-covered-fraction 80 –min-read-percent-identity 95 –min-read-aligned-percent 80''.

Viral contig binning and prevalence estimation

Viral OTUs (vOTUs) were obtained by parsing the previously processed VAMB bins with PHAMB[91] and then quality checked with CheckV (workflow 'end_to_end').[61]

CheckV result tables were parsed in order to identify viral contigs. We retained contigs annotated as complete, medium- and high-quality contigs with $\geq$75% average amino acid identity (AAI), >30% alignment fraction of the contig, and less than 15% difference in genome size from the closest CheckV database hit. Low quality contigs with more than 10 genes, 10% of which annotated as host genes, and 40% annotated as viral genes, were included as well to account for potential uncharacterized viruses. Proviral contigs annotated at least as medium quality were included as well.

Viral and proviral contigs were clustered and dereplicated at 95% ANI across 85% of the shorter sequence using the scripts provided by CheckV to generate vOTUs. Taxonomic assignation was performed by Genomad[63] using the end-to-end workflow. Additional taxonomic assignation was performed by clustering the vOTU representative sequencing with the MGV genome database using the same approach described above.

The representative sequences for the vOTUs were indexed with BowTie2[92] and mapped against the original metagenomes using CoverM[83] to estimate their relative abundance using parameters ''—min-covered-fraction 80 –min-read-percent-identity 95 –min-read-aligned-percent 80''.

vOTUs discovered in contigs of bacterial MAGs were host annotated according to the bacterial host. The host taxonomy of other viruses was derived using CRISPR spacers. CRISPR spacers were internally identified by Bakta using PILER-CR. The combined set of CRISPR spacers mined from MAGs were then mapped to all vOTUs using blastn[93] (v2.13). Subsequently, spacer hits were

processed such that only CRISPR-spacer matches with ≥90% sequence identity over ≥90% of spacer length and a maximum of 3 mismatches were retained. Spacer hits were used to calculate the taxonomy at the species level on the predicted host bacteria using a majority rule.

Human genome sequencing

Participants were instructed how to produce at least 5 mL of saliva, and these samples were frozen, stored, and transported using procedures similar to the microbiome specimens. Saliva samples were subjected to low-pass genome sequencing on the Illumina NovaSeq 6000 after performing DNA extraction. Reads were mapped against the human genome (version hg38) using FQ2BAM as implemented in the NVIDIA Clara Parabricks pipelines, and the resulting BAM files were processed with GLIMPSE[94] to perform low pass imputation against the 1000 Genomes reference panel.[107] Resulting VCF files were merged into a single BCF file with bcftools[95] and processed with PLINK[96] using parameters "–hwe 1e-3 –maf 0.05 –indep-pairwise 50 10 0.1". This resulted into 13,651,831 variants across 2,240 individuals. We then applied filters to retain samples with paired metagenome and performed LD pruning, which yielded 1,701 individuals and 227,216 variants. Principal Component Analysis was calculated with PLINK using the "–pca" parameter.

Kinship analysis was performed using KING, and the ranges used for inferring the degree of relation from kinship coefficients were taken from the original publication.

Association analyses

Microbiome association analysis was performed by fitting multiple linear models as implemented in Matrix eQTL[98] to predict microbiome features (species-level and MetaCyc pathways abundances) after including gender, age, BMI, and the first 10 genetic principal components as covariates. We included species present in at least 10% of the participants. SNPs identified as statistically significant were associated with the closest gene (nearest function in SummarizedExperiment R package). microbiomeGWAS[69] was used to perform a GWAS on the Weighted UniFrac distance matrix, including the same covariate used for the microbiome QTL analysis. A GWAS using the beta-diversity distance matrix is built on the idea that if a genetic variant affects the microbiome, individuals who share more alleles at a specific locus (for example, both having AA means they share two alleles, while one having AA and the other GG means they share none) will have lower beta-diversity distances.