medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Environmental, Socioeconomic, and Health Factors Associated with Gut Microbiome Species and Strains in Isolated Honduras Villages

Shivkumar Vishnempet Shridhar[1,2†], Francesco Beghini[1†], Marcus Alexander[1], Adarsh Singh[3], Rigoberto Matute Juárez[4], Ilana L. Brito[3]*, and Nicholas A. Christakis[1,2,5,6]*

[1] Yale Institute for Network Science, Yale University; New Haven, CT, USA.
[2] Department of Biomedical Engineering, Yale University; New Haven, CT, USA.
[3] Meinig School of Biomedical Engineering, Cornell University; Ithaca, NY, USA.
[4] Soluciones para Estudios de la Salud; Copán, Honduras
[5] Department of Statistics and Data Science, Yale University; New Haven, CT, USA
[6] Department of Medicine, Yale School of Medicine; New Haven, CT, USA

†Co-first authors
*Corresponding and co-senior authors. Email: nicholas.christakis@yale.edu & ibrito@cornell.edu

ABSTRACT

Despite a growing interest in the gut microbiome of non-industrialized countries, data linking microbiome features from such settings to diverse phenotypes remains uncommon. Using metagenomic data from a community-based cohort of 1,871 people from isolated villages in the Mesoamerican highlands of western Honduras, we report novel associations between bacterial species and phenotypes. We also find an uncharacterized *Lachnospiraceae* species associated with 22 different phenotypes, and little overlap with a prior Dutch Microbiome Project. Furthermore, including strain-phylogenetic information modifies the overall relationship between the gut microbiome and the phenotypes, especially in some phenotypes like household wealth. Coincidentally, wealthier individuals also have a higher number of polymorphic sites. Our analysis suggests new roles that gut microbiome surveillance can play in understanding broad features of individual and public health.

INTRODUCTION

Thanks to long-run investments in gut microbiome research in industrialized countries, the pervasive role that the human microbiome plays in influencing health-related and other phenotypes, or how, reciprocally, various phenotypes may influence the microbiome, is becoming increasingly clear (*1*,*2*). However, the majority of the human population lives outside of North America and Europe, and nearly half of the human population lives outside urban areas. Non-industrialized populations often experience problems with access to healthcare resources, have distinctive patterns of social interactions (e.g., low population density, fewer contacts with strangers), and have other distinctive exposures (e.g., animals and diet) (*3*–*5*). Prior studies of non-industrialized populations have documented the presence of rich uncharacterized taxa that are often absent in industrialized cohorts (*6*). And advances in genomics (such as strain-level methods) are still uncommonly applied to cohorts from non-industrialized settings. Here, we elucidate the relationship between the microbiome, characterized at the species and strain levels, on the one hand, with a very broad array of biological and social phenotypes, on the other. And we explore the role that strain-level information may play in these relationships.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

RESULTS

Isolated Setting in Western Honduras

The village communities in the western highlands of Honduras are geographically remote **(Figure 1A)**, consisting in a large proportion of the descendants of Mayan peoples who today still form traditional face-to-face social networks and who depend on subsistence agriculture and coffee cultivation. We collected population-level data in these small communities, including deep sequencing data and a comprehensive set of both individual and community-level characteristics regarding diverse socioeconomic, psychological, and health phenotypes. Our cohort consists of 1,871 people living in 19 villages which are part of a larger cohort developed for a different original purpose (*7*). The adult population in our 19 villages ranges from 66 to 432 individuals. The average age of participants was 41 (SD=17; range: 15 - 93); 63.7% were women; and 41.8% were married. Each of our 19 villages has its own intricately connected social networks with minimal inter-village contact, and they are not only separated by distance, but also by elevation (**Figure 1A**).

Variations in microbiome composition can be appreciated even within the same village. For instance, there is a pattern of decreasing similarity as individuals live farther away from the village center, in the geographic periphery (Pearson correlation coefficient $\rho = 0.311$, $P = 0.0022$, **Figure 1B-C**). In contrast, villagers located at the topological center of the social network within each village have a more similar microbiome to the rest of the village, unlike those at the social periphery (Linear regression $\beta = 3.66 \times 10^{-05}$, $P = 0.761$); (see Methods for details, and also the inset in **Figure 1A**).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

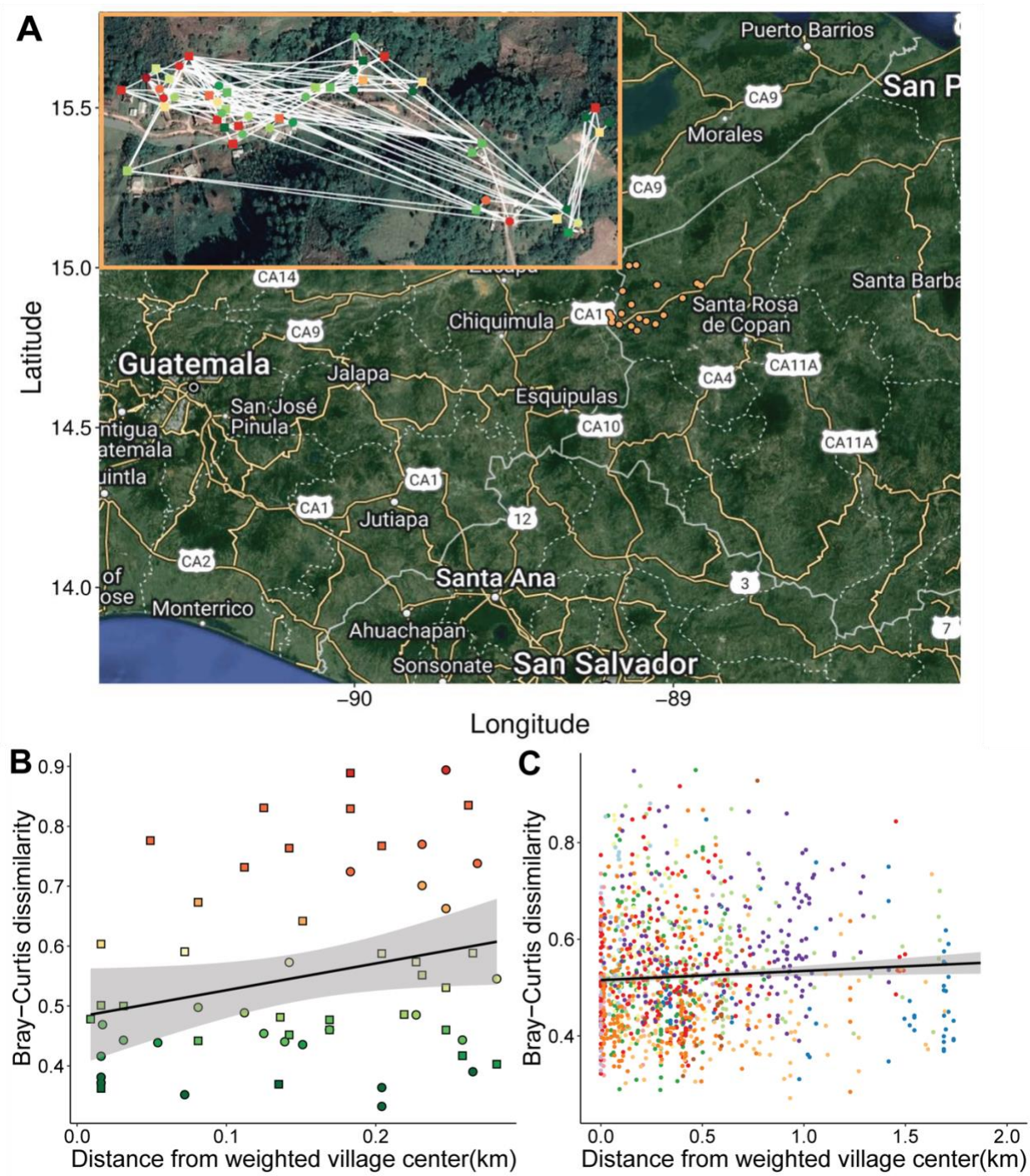


Figure 1: Geographic overview of the Honduras microbiome project: (A) A satellite view of the Honduran villages (in orange) that constitute the microbiome dataset. In the inset, there is a zoomed-in satellite view of an illustrative village with each inhabitant (n=57) colored with the respective Bray-Curtis dissimilarity value relative to the average microbiome composition of the rest of the village, and connected by white edges which represent social interactions between individuals. Green nodes are indicative of being very similar in microbiome composition to the rest of the village, whereas red nodes are more dissimilar. Square nodes indicate males and circle nodes indicate females. (B) Scatterplot of Bray-Curtis dissimilarity (of the village shown above) and the distance of households from the population-weighted village centroid (see Methods) shows a positive correlation (Pearson correlation coefficient $\rho = 0.144$, $P = 0.05$) between gut microbiome dissimilarity and distance from the village center. Individual dots are colored according to the person's dissimilarity from the village's average microbiome.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

The light grey areas indicate a 95% confidence interval. (C) Combined plot of all the Bray-Curtis dissimilarities and distances from village centroids in all the villages. The black regression line indicates a consistent trend (Pearson correlation coefficient $\rho = 0.311$, $P = 2.2 \times 10^{-03}$) of increasing microbiome dissimilarity with regard to the distance from the village centroid. The individual dots are colored according to the village they belong to. The light grey areas indicate a 95% confidence interval.

Species and phenotypes

Overall, we found 2,148 associations when examining 639 microbial species and 123 phenotypes (including physical and mental health, medication use, diet, animal exposure, and social and economic measures) (**See Supplementary Table 1**). All comparisons involved appropriate statistical controls (including for batch effects; see Methods) and were corrected for multiple hypothesis testing using a False Discovery Rate (FDR) procedure. Distinctly, we also found 988 associations with metabolic pathways (**See Supplementary Table 2**). The 123 phenotypes are variously measured as continuous, categorical, and discrete variables (**Supplementary Tables 3-5**). As expected, several of the phenotype variables were correlated (for example, individuals with high Hemoglobin A1c strongly correlated with reporting a diagnosis of diabetes, and the household wealth index correlated with owning a TV (**Figure S1**)). Similarly, the clustering of phenotypes based on species effect sizes (obtained from the species-phenotype association models) showed that multiple phenotypes within different categories have similar microbial signatures (**Figure S2**).

Health phenotypes

We found a total of 402 species to be significantly significant associated with at least one health phenotype (**Supplementary Table 1 and 3**). Among the 402 significant species, 302 of them belonged to the phylum Firmicutes, making it the most associated with health phenotypes. Among all the associated species, 34.58% were identified as unknown (*8*) at several taxonomic levels. uSGB2239 (unknown at the genus level) from the *Rikenellaceae* family and *Parolsenella massiliensis* were the most frequently associated species, significantly associated with 5 health phenotypes; in particular, both were commonly associated with Body Mass Index (BMI), allergies, and intestinal illness (**Figure 2A**). uSGB2239 was also associated with antibiotics and dementia, while *Parolsenella massiliensis* was also associated with anti-hypertensive medication and openness (see Methods). Microbial species from the *Rikenellaceae* family have been previously found to be associated with at least one mental health disorder (Obsessive Compulsive Disorder) (*9*) and to be enriched in type-2 diabetics in a Pakistani cohort (*10*). In another study, *Rikenellaceae* was found to be significantly associated with Indian men having high blood sugar (*11*). Coincidentally, we found that BMI was significantly associated with uSGB2239 of the *Rikenellaceae* family.

Furthermore, a total of 136 pathways were associated with at least one health phenotype, totaling 157 pathway associations (for details regarding the ascertainment of microbial metabolic pathways, see Methods). Among the 157 associations, physiological variables had 85 associations, followed by 24 associations in chronic illness phenotypes; 26 in medication use; 2 in acute conditions; 19 in personality measures, alcohol, cigarettes, and mental health; and 1 in overall health (**Supplementary Table 2**).

We performed association analysis for the subset of individuals falling in unhealthy ranges of various phenotypes (i.e., BMI<18 and BMI>25 to account for underweight and overweight individuals, or diastolic pressure>89 to account for hypertensive individuals) compared to healthy individuals (**Figure S3**, **Supplementary Table 6**). A total of 73 species were associated with multiple unhealthy phenotypes, of which, uSGB14313 of the Clostridia family was associated with 3 phenotypes in unhealthy ranges (Hemoglobin A1c (5.7-6.4), BMI (25-30), and BMI (30-35)) (**Figure S3A**, **Supplementary Table 6**).

We also evaluated whether the diversity of an individual's microbiome (measured with Shannon diversity) was itself associated with various health (and other) phenotypes. The majority of the villagers

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

self-reported themselves as healthy individuals (n=1,407, 75.20%) and only 162 villagers (8.65%) reported having more than one disease. Villagers with reported illnesses (except arthritis and diabetes) were found to have lower diversity relative to healthy villagers (**Figure 2B**); in particular, villagers with reported stomach (Wilcoxon Rank Sum test $P = 1.48x10^{-04}$) and intestinal illnesses (Wilcoxon Rank Sum test $p = 9.31x10^{-05}$) had decreased diversity. Villagers taking various medications also had lower diversity (**Figure 2C**); anti-parasitic drug users showed the lowest diversity (Wilcoxon Rank Sum test P = 0.018) followed by anti-diarrheal users (Wilcoxon Rank Sum test P = 0.032) and antibiotic users (Wilcoxon Rank Sum test $P = 2.26x10^{-04}$). We found no material associations of microbiome diversity with other categories of medications.

We also performed a simple contrast analysis by comparing all chronically ill individuals to those without any chronic conditions. We used differential abundance analysis with 1,407 healthy people and 464 chronically ill people and found that 6 species were significantly differentially abundant (**Figure S3B**). *Lachnospiraceae bacterium* is the only significant species differentially enriched in healthy individuals. On the other hand, uSGB1663 and uSGB27424 of the Prevotellaceae family, *Spirochaetia bacterium*, *Coprococcus*, and uSGB6369 of the Clostridia family were found enriched in diseased individuals.

Overall, all the health phenotypes put together contribute 5.7% of the total variance explained in microbial species composition (**Figure S4, Supplementary Table 7**). Similarly, 11.6% of the variance in pathway composition is relevant to health phenotypes.

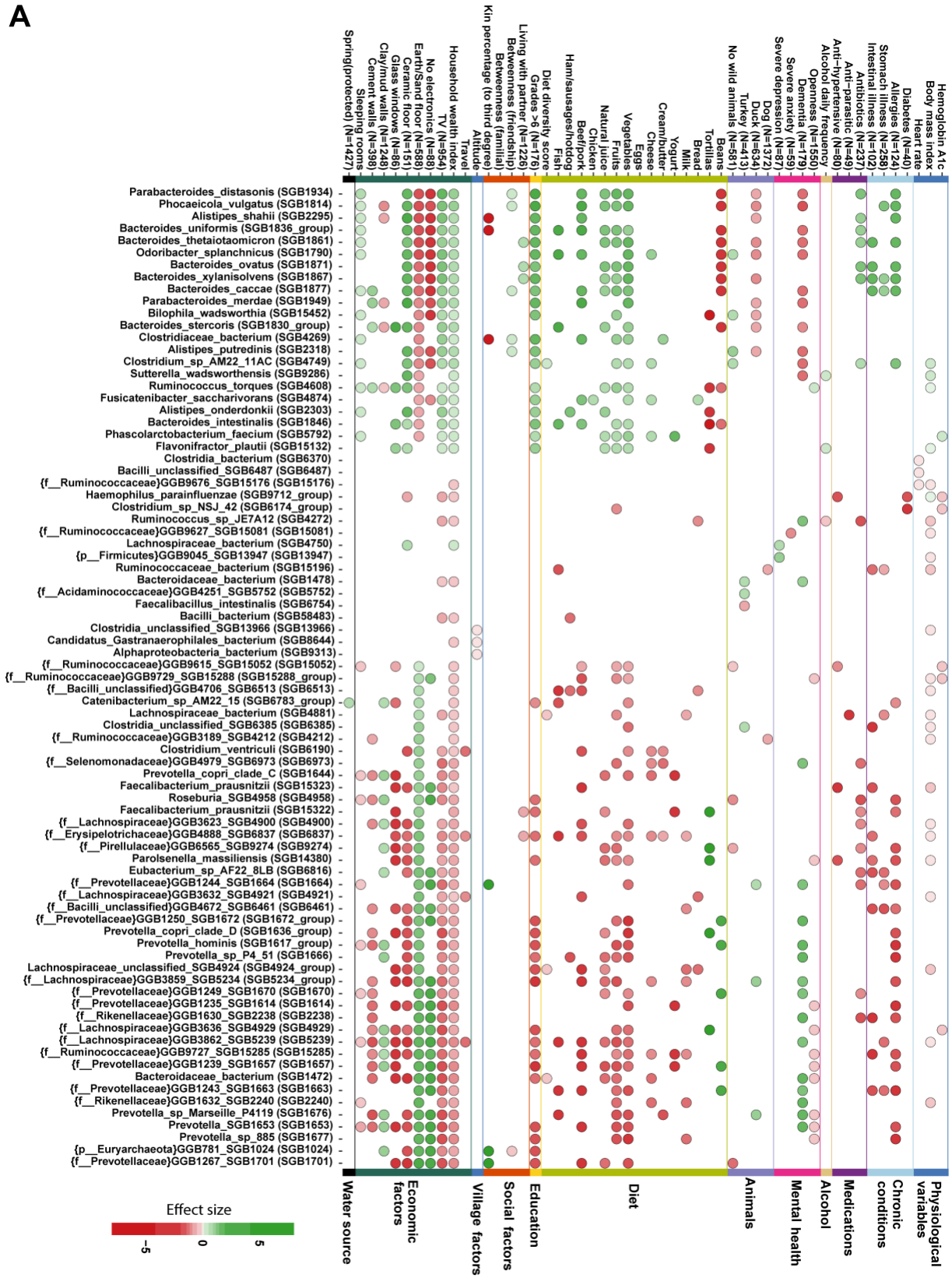
Animal Exposure and Diet Phenotypes

We explored possible associations with animal exposure and diet (*1*, *12–14*). An unusual feature of our setting is that more than 90% of villagers reported having exposure to different types of animals, including wild animals, farm animals, and pets, affording possible zoonotic transmission. Overall, for all food and animal phenotypes, 205 species were found to be significantly associated with at least one of the phenotypes, resulting in 437 associations (**Figure 2A, Supplementary Table 4**). Among all the associating bacterial species, 27% were unknown. Among the 205 significantly associated species, 122 of them belonged to Firmicutes, making this phylum the most commonly associated with specific animals or food categories. We found 10 pathways associated with exposure to animals as well (**Supplementary Table 2**). Animal exposure contributed to 2.3% of the variation in species composition. We found no difference in overall Shannon diversity in individuals exposed to different animal categories (**Figure S5**).

Diet has been extensively studied and shown to have a substantial relationship with the gut microbiome (*14–16*). We assessed associations with microbial features and food frequency consumption and found 360 significant associations with diet (**Figure 2A**). *Bacteroides intestinalis* was the most associating species with food phenotypes, associating with 8 different food types. In the past, *B. intestinalis* has been implicated in the context of dietary fiber contributing to an increase of xylan utilization in the gut (*17*). Even though most of the individuals' daily diet consist of tortillas and beans, we measured diet diversity using the Diet Diversity Score (DDS)(*18*) (see Methods and **Figure S6**). We identified a total of 108 significant associations between the DDS and gut microbiome species (**Figure 2A**). We also found 235 pathway associations with food phenotypes (**Supplementary Table 2**). Diet was responsible for 1.85% and 2.14% of the variance explained in our sample in species and pathways composition, respectively (**Figure S4**).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .



medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

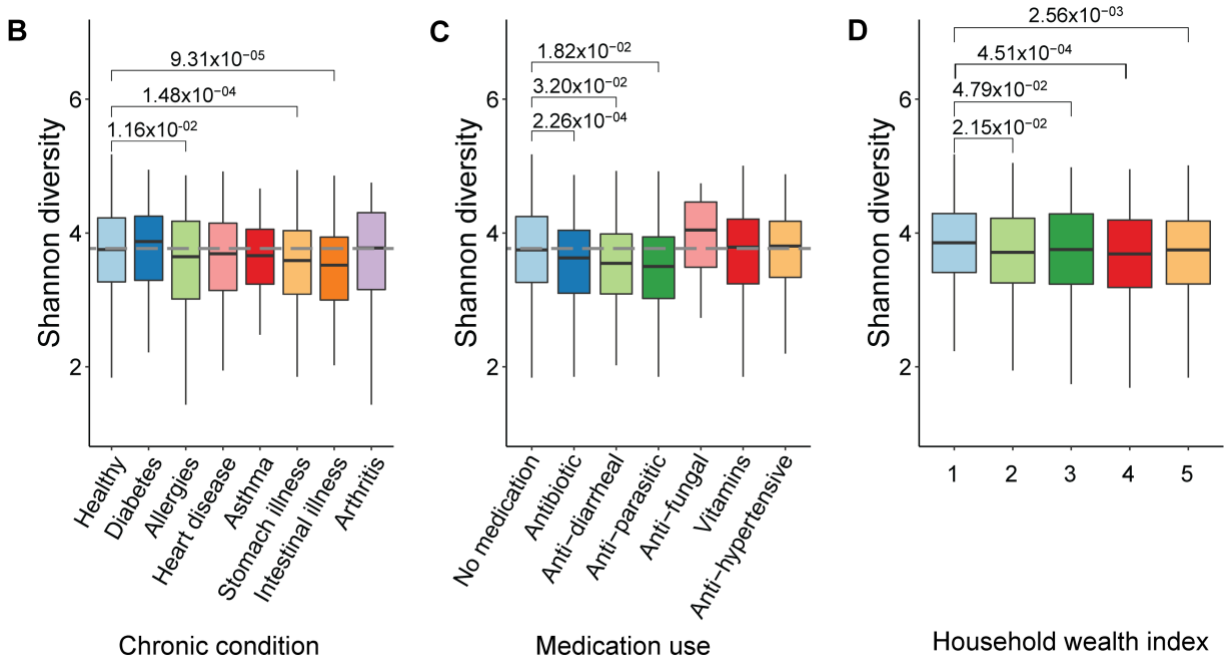


Figure 2: Microbiome association with phenotypes: (A) 81 species that best represent gut microbiome associations with 52 phenotypes (chosen from health, food and animal, and socioeconomic categories, see **Supplementary Table 1** for a complete list of associations). The number of individuals manifesting the respective phenotype is shown in brackets. The presence of color shows significant associations for that phenotype-species pair (FDR<0.05); the intensity of the color corresponds to the strength of the effect size. Negative associations are indicated by red color and positive by green. Unknown species are indicated with "{}", specifying the taxonomic level at which the species is known. Phenotypes without sample size are reported for all individuals. (B) Shannon diversity of healthy and chronically ill individuals highlights differences in overall microbiome diversity; healthy individuals (n=1,407) are chosen as a reference (grey dashed line). (C) Shannon diversity is calculated between different medication use categories; non-medicated individuals (n=1,246) are chosen as reference (grey dashed line). (D) Shannon diversity of villagers belonging to households classified by household wealth index, ranging from 1 (least wealthy) to 5 (most wealthy). All comparisons were performed using the Wilcoxon Rank Sum test and corrected for multiple hypothesis testing.

Socioeconomic Phenotypes

Overall, we found 1,105 significant associations with socioeconomic phenotypes. For all socioeconomic phenotypes, 319 species were found to be significantly associated with at least one of the phenotypes (**Figure 2A**, **Supplementary Table 1**). Among all the 319 associated species, 28.8% of them were unknown and 185 of them belong to Firmicutes, making it again the most associated phylum for socioeconomic factors. Moreover, uSGB5239 of the Lachnospiraceae family is the most associated species, statistically significantly associated with 14 socioeconomic phenotypes. We also found 586 associations with pathways, with one of them associating with 9 socioeconomic phenotypes (**Supplementary Table 2**).

Socioeconomic factors are relevant to many exposures and personal habits. Higher monthly expenditures are correlated with a better diet and better household essentials such as refrigerator or a paved floor. We observed that most of the bacteria associated with higher monthly expenditures are the same as the ones associated with better diet quality (*19*, *20*).

Although all the participants in our study are considered to be living in poverty, economic status still varied among them and was associated with possessions and diets potentially relevant to the microbiome. Overall, the average household wealth index score (ranging from least wealthy (1) to most

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

wealthy (5)) is 3.26 (standard deviation 1.33). In terms of measures of economic status, both monthly expenditure and travel were associated with the microbiome. Total wealth was also correlated with owning various items (such as a TV or a mobile phone) some of which (e.g., a refrigerator or a stove) might affect food consumption and others of which (such as having glass windows, cement walls, more sleeping rooms, an earthen floor, or a metal roof) might affect microbiome exposures via other routes (**Figure 2A, Figure S1**). We observed similar patterns of association where a high wealth index was associated with the same bacterial species associated with owning expensive items (like glass windows), and conversely. The variance explained by economic factors was 4.13% for species and 3.70% for pathways (**Figure S4**), indicating the relative importance of economic factors in explaining variation in the gut microbiome.

With respect to overall microbial diversity, the subjects from the least well-off households had a Shannon index that was increasingly higher than that of the subjects from the wealthier households (in the top 4 quintiles) (**Figure 2D)**.

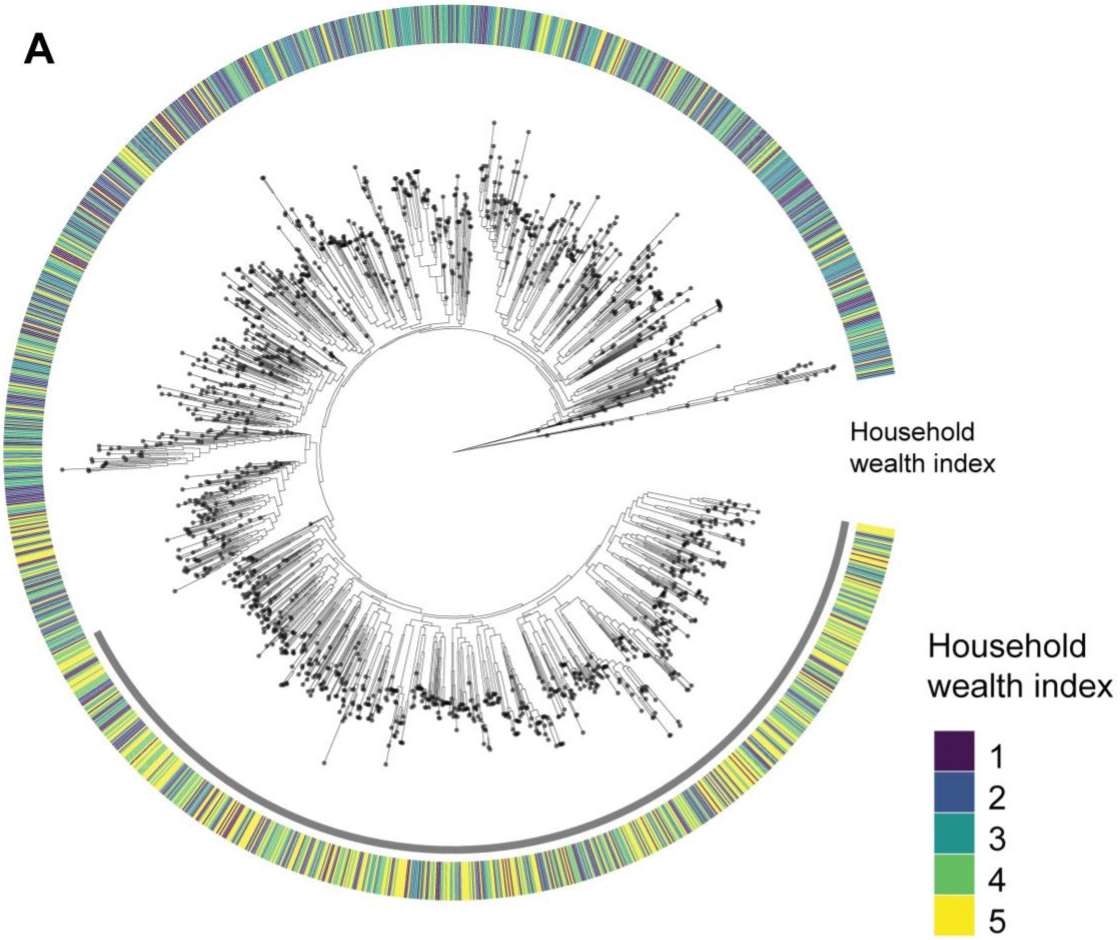
Overall relationship between species and phenotypes

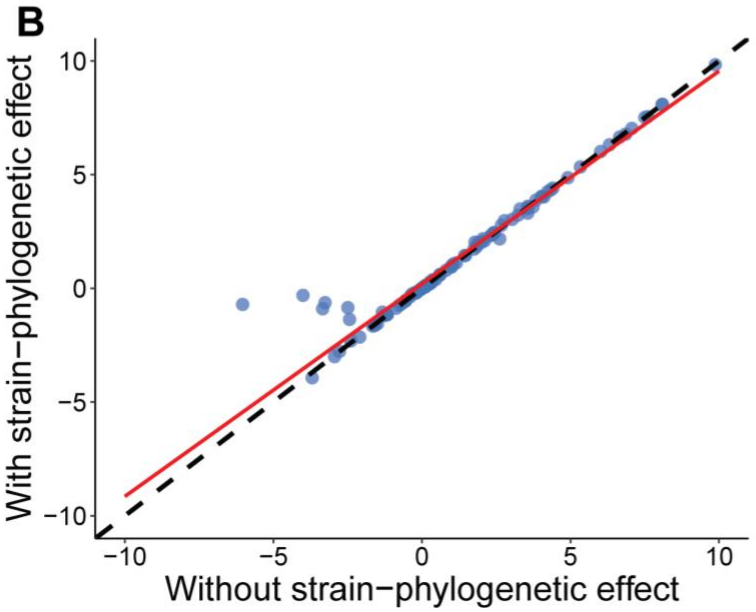
From the clustering of associations (**Figure 2A**) and the dendrogram (**Figure S2**), it can be observed that different phenotypes can be linked together. This link can be visualized through the relationship between the gut microbiome and phenotypes. For example, species that are enriched in socioeconomic phenotypes (such as TV ownership and household wealth index), show a similar pattern in vegetables, fruits, and meat consumption. These phenotypes also form their own collection within the dendrogram (**Figure 2A**, **Figure S2**); previous studies have found diet diversity to be correlated with food security and wealth in rural settings (*21*, *22*). Overall, health, food, animal, and socioeconomic factors are clustered together, which is also visibly demonstrated through the microbiome-phenotype lens. Multiple phenotypes from different categories can also be tied together by a singles species. For instance, uSGB5239 (of the *Lachnospiraceae* family) is associated with 22 different species from the health category (BMI, Stomach illness, Dementia); the food and animal category (Vegetables, Fruits, Natural juice, Beef/pork, Fish); and the socioeconomic category (Grades>6, Travel, Household wealth index, TV, No electronics, Earth/sand floor, Ceramic floor, Glass windows, Clay/mud walls, Cement walls, and Sleeping rooms) (*23*).

Furthermore, we compared our associations with the ones found by the Dutch Microbiome Project. We found 13 species-phenotype associations in common with the Dutch study, 8 of those were with BMI (Body Mass Index), and *Butyrivibrio crossotus*, *Roseburia inulinivorans*, *Faecalibacerium prausnitzii*, *Methanobrevibacter smithii* , *Eubacterium siraeum* *Haemophilus parainfluenzae*, *Mitsuokella multacia*, and *Flavonifractor plautii* were found significantly associated in both datasets for BMI. *Haemophilus parainfluenzae* was also significant for Hemoglobin A1c in both datasets. *Ruminococcus torques* was significant in both datasets for antibiotic use. Finally, monthly income/expenditure had 3 significant species in common: *Alistipes shahii*, *Barnesiella intestinihominis*, and *Flavonifractor plautii*. The presence of very few significant species is largely due to difference in measurements between the cohorts and the relatively low number of common species across the two cohorts (see **Supplementary Table 8 and Figure S9** for a list of possible comparisons).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .





medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Figure 3: Microbial strain association with host phenotypes: (A) In this strain-level phylogeny of *Eubacterium rectale* (SGB 4933 group) (as an illustrative microbe) in 1,610 individuals, leaves are annotated with the household wealth index (as an illustrative phenotype). A cluster of individuals (annotated in grey) situated on a different strain of *Eubacterium rectale* (separated branch) are more likely to have higher wealth compared to rest of the individuals (Fisher's exact test P=0.0004998). (B) Comparison of significant effect sizes obtained from a linear mixed model with and without adding strain-level phylogeny information across significant species-phenotype relationships overall. β-coefficients from the association models with and without including phylogenetic information are positively correlating (Spearman correlation coefficient $\rho=0.989$, $p < 2.2\text{x}10^{-16}$), and the red line is the linear fit ($\beta = 0.9351$, intercept = 0.2016, $p < 2.2\text{x}10^{-16}$), showing the relationship between the two models. The deviation of the red fitted line from the dashed line shows the important effect of adding the strain-level phylogeny in the species-phenotype association model (**Supplementary table 9**).

Relevance of Microbial Strains

Finally, moving beyond species-specific associations with phenotypes, we observed a meaningful variation between the genetic makeup of the same species across different individuals that is in turn associated with diverse phenotypes (**Figure 3A**). For instance, individuals with higher household wealth index are likely to have a different strain of *Eubacterium rectale* compared to less wealthy individuals in a set of 1,610 individuals (p=0.0004998, (Fisher's exact test)) (**Figure 3A**).

Moreover, adding strain-phylogenetic information in the model alters the relationship between species and phenotypes overall (**Figure 3B**) by inducing a small shift. Among all the effect sizes, 0.2% of them switch direction when adding the phylogenetic effect (**Figure S7, Supplementary Table 9**).

Looking deeper into the strain diversity in individuals, we looked at the variation of the percentage of polymorphic sites across individuals and phenotypes. As an illustration, we observed that individuals who are wealthier ($\beta = 0.08345$, $P=1.26\text{x}10^{-07}$) or those consuming a higher number of eggs ($\beta = 0.626$, P=0.1146) had a higher percentage of polymorphic sites (**Figure 4A-B, Supplementary Table 10**). This comports with findings in another study where percentages of polymorphic sites from just *Prevotella copri* strains were found to be different between recent South-Asian Canadian immigrants and first-generation South-Asian Canadians (*24*).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

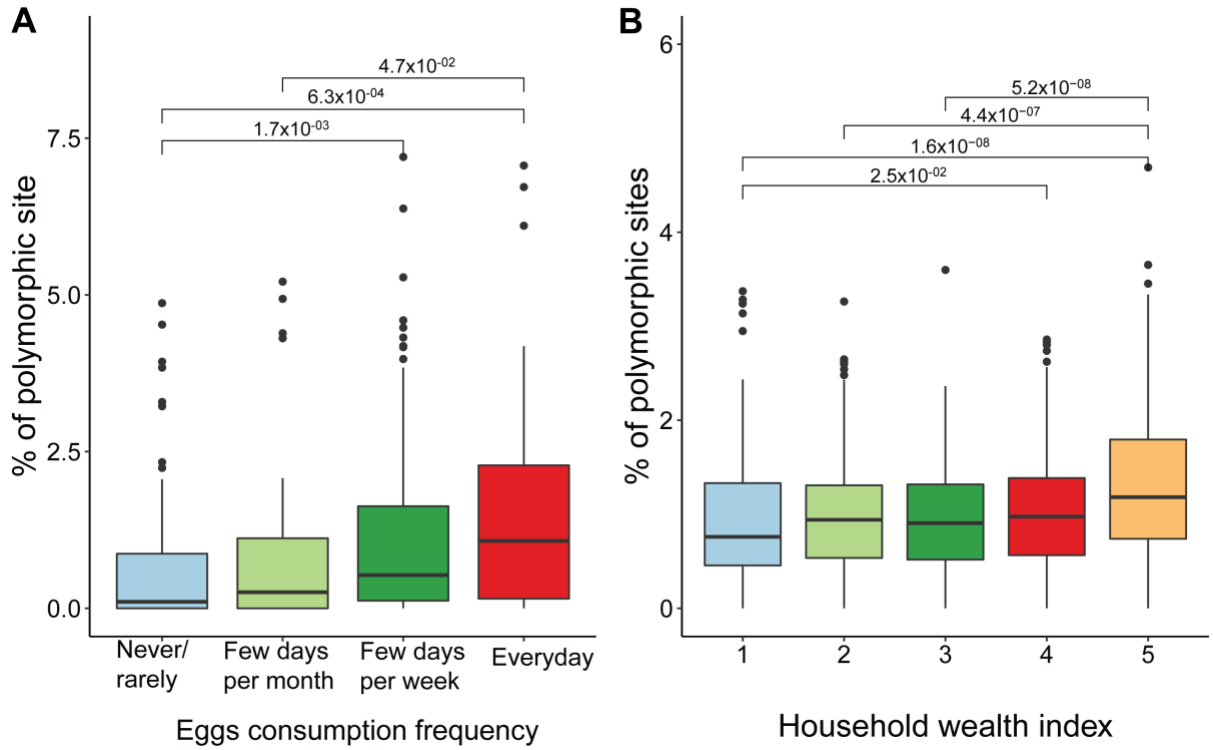


Figure 4: Relationship of variability in polymorphic sites with two host phenotypes: (A) Polymorphic site variability of uSGB4905 shows a gradual increase in the percentage of polymorphic sites in individuals consuming more eggs in their regular diet (N=312). (B) Another demonstration of how variability in polymorphic sites changes with phenotype. Here, progressively wealthier individuals have a higher percentage of polymorphic sites in *Prevotella copri Clade C* (SGB1644) (N=1458).

DISCUSSION

Integrated, standardized, large, population-based cohorts to study the microbiome are uncommon, but such studies offer the prospect of identifying factors shaping the gut microbiome or being shaped by it. By extending our knowledge of the human gut microbiome to a novel population in a lower-and-middle-income (LMIC) setting; by assessing previously uncharacterized taxa; by having a very broad range of phenotypes; and by using strain-level genomic information, our goal is to advance understanding of the possible relationship of the gut microbiome with human phenotypes.

We find that variation in the gut microbiome across individuals living in a traditional way can still be at least partly explained by variations in diet, lifestyle, environment, and health factors. Overall, we found 2,148 unique associations between 639 bacterial species and 123 phenotypes. These associations included many uncharacterized species which in many cases were shown to have a stronger effect than known species. Phenotype associations were also identified after including strain-level phylogenies, which often had a profound effect on the extent of the association between microbiome species and the human phenotypes under consideration.

Still, despite measuring a large number and variety of phenotypes, only 19.2 % of the variation across individuals in microbiome composition was accounted for by these phenotypes, in keeping with prior studies (*1*, *25–27*). This suggests that microbiome composition in individuals may be quite idiosyncratic or may depend on details of social interactions or unmeasured environmental exposures. Rare species may also help account for this variation. The current understanding of how individual and population-level microbiomes come to be shaped is thus still incomplete. Nevertheless, the phenotypes

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

we ascertained in Honduras did combine to account for 19.2% of the species variation (as noted) and 33.4% of the pathway variation; this may be compared to a recent study from the Netherlands where the measured phenotypes accounted for 13% and 16.2% of the variation, respectively, (*1*) although different methodologies for taxonomic and functional characterization were used here, reflecting ongoing methodological advances.

The gut microbiome is known to be related to various health conditions both in humans and in mice (*28*), and conditions like cancer, obesity, diabetes, autism, anxiety, and depression can induce shifts in the gut composition (as shown in many mostly Westernized populations) (*1*, *28–34*). Alcohol intake and cigarette use have been linked to gut microbiome dysbiosis, as have medications (*35–39*). In keeping with these prior studies, we confirm such findings in this rural LMIC cohort (*3–5*, *40*). Indeed, we found 606 associations between the microbiome and health-related phenotypes. Chronic illnesses and medication use were the most strongly associated. Among chronic illnesses, intestinal illnesses show the greatest differences. We uncovered 297 total associations between gut microbiome species and physiological measurement ranges that may be linked to underlying chronic conditions such as obesity, diabetes, and hypertension. Moreover, we find 64 associations with mental health phenotypes alone, a relatively understudied area.

Looking at the overall microbial composition among healthy and chronically ill subjects, the Shannon diversity was generally lower in most of the chronically ill people, especially those with allergies and gastrointestinal illnesses. Moreover, comparing healthy individuals to those who are chronically ill, we found 6 taxa to be differentially enriched in one of the groups. *Lachnospiraceae bacterium* is the only significant species differentially enriched in healthy individuals. On the other hand, 2 unknown and 3 known species: uSGB1663 and uSGB27424 of the Prevotellaceae family, *Spirochaetia* bacterium, *Coprococcus*, and uSGB6369 of the Clostridia family, are differentially enriched in chronically diseased individuals. Among medication users, those taking anti-parasitic medication had the largest drop in overall diversity.

Another factor that greatly influences the gut microbiome is diet. Our sample population exhibits a consistent diet, with beans and tortillas consumed by most people daily. Still, we found 360 associations with food categories. A previously studied Dutch cohort found that pets had notable associations with the microbiome (*1*), and we also found 77 associations with a broader range of animal exposures. Furthermore, our sample was spread across 19 villages separated in space and elevation, and the overall gut microbiome was observed to vary with the relative spatial position within the villages; the dissimilarity score with a village-averaged microbiome increased as subjects lived further away from the village center. Relatedly, we found 3 significant associations with elevation. We also found 6 associations with other village-related factors, such as distance to the health center (which is historically correlated with overall health in middle/low-income areas) (*3*), distance to main road, and number of churches.

Social and economic factors had 1,105 associations, with the bulk of the strong associations coming from unknown species. The gut microbiome had 947 unique associations with economic factors alone, making it the second highest associating category of variables we examined, after health. Wealth differences in individuals can also manifest in the form of more diverse strains in some species being present in wealthier individuals. Prior research in Honduras has highlighted the crucial importance of socioeconomic status in addressing health in such communities (*41*), and the microbiome varies in important ways along this axis, even in this poor setting.

Social interaction is an integral part of Honduran villagers' life. In total, 123 unique associations with various social network factors were found. Studies investigating social interactions between mice have shown the evolutionary advantage of having behavior that enhances social interaction that consequently facilitates microbiome transmission (*28*, *42*, *43*). In wild mice, social associations are predictive of microbiome composition, and the microbiome is correlated across mice interaction networks (*44*). In humans, strain-level similarities have been shown in familial and partner networks within and outside households (*45–47*). Whether these interactions translate into exposures that directly contribute to health is an important area for further work.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Uncharacterized taxa play a vital role in all these associations, as in prior LMIC cohorts (*6*). Despite the number of unknown species in the Honduran cohort being about a third of total species, their relative strength of associations was observed to be higher in all the phenotype categories. Distinctly, strain-level information is also relevant to the microbiome-phenotype relationship and should optimally be accounted for.

By expanding our knowledge of the human microbiome to a novel non-Westernized cohort, it is possible to further our understanding of the role of the gut microbiome in chronic illness and, at the same time, open up opportunities to use such findings to develop inexpensive biomarkers to aid diagnostics in rural settings (*48*, *49*). To the extent that a healthy microbiome is driven by modifiable social and environmental factors (such as diet, smoking, living arrangements, lifestyle, and so on), understanding which factors to target or what possible microbiome-modifying interventions to implement could help advance individual and collective health in diverse settings.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

METHODS

Sample collection, library preparation, and sequencing:

Participants were instructed on how to self-collect the fecal samples using a training module and promptly returned samples to a local team which then stored them in liquid nitrogen at the collection site and then moved them to a -80 C° freezer in Copan Ruinas, Honduras. Samples were then shipped on dry ice to the United States and stored in -80 C° freezers.

Stool material was homogenized using TissueLyzer from Quigen and the resulting lysate was prepared for extraction with the Chemagic Stool gDNA extraction kit (Perkin Elmer) and extracted on the Chemagic 360 Instrument (Perkin Elmer) following the manufacturer's protocol. Sequencing libraries were prepared using the KAPA Hyper Library Preparation kit (KAPA Biosystems). Shotgun metagenomic sequencing was carried out on Illumina NovaSeq 6000. Samples not reaching the desired sequencing depth of 50Gbp were resequenced on a separate run.

Raw metagenomic reads were deduplicated using prinseq lite (version 0.20.2 (*50*)) with default parameters. The resulting reads were screened for human contamination (hg19) with BMTagger (*51*) and then quality filtered with trimmomatic (*52*)(version 0.36, parameters "ILLUMINACLIP:nextera_truseq_adapters.fasta:2:30:10:8:true SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:50").

This resulted in a total of 1,871 samples (with an average of 86.84 x 10^6 reads per sample).

Local involvement in the research

In keeping with proper standards for such research, we worked closely with the local population of Copan, sought feedback and approval from officials at the Ministry of Health (MOH) of Honduras, and endeavored to provide practical benefits to the local community. Here, we briefly summarize this history and outline some of our principles and actions in this regard.

When we began designing this cohort project in 2013 (for the whole cohort of 176 villages and 24,702 people in the parent RCT), the Bill and Melinda Gates Foundation (BMGF) introduced us to the Inter-American Development Bank (IDB), which has been supporting and doing work throughout Latin America, and IDB in turn introduced us to the Honduras MOH. After the cohort was established, we obtained additional funding from other sources as well. Because of this pathway to getting the project launched, we worked with local and regional public health organizations and with local leaders rather than with local academic institutions.

From the outset when the original underlying cohort for this study was impaneled (in 2013), we sought extensive local involvement, beginning with a needs assessment where local village residents told us about topics of concern to them in a series of meetings in villages throughout the Copan region (in the western highlands of Honduras, near the Guatemala border). In addition to extensive community input, we sought input from the MOH. We periodically briefed both the communities and the MOH about our findings.

Copan is a very isolated area, with villages within Copan being further isolated due to varying altitudes (850.13±154.77 m). Over the years, as we built our data collection team in Copan, we developed deep ties to the local community, to local village leaders, to the few local health clinics, and to local transportation and infrastructure providers. Because of these ties and our commitment to the local community, we presented our results directly to these constituencies regularly. We also held two annual joint implementation science conferences with our Honduras and Yale teams (involving many dozens of attendees).

We provided other material benefits to the local community, beyond simply providing them with information. When we tested people for parasites as part of our study, we gave them the results of their tests and arranged for them to be treated. When we tested people for vision, we provided corrective glasses. We solicited ideas from the local community about what infrastructure improvements we could make, and we repaired many local playgrounds and clinics as a result. We arranged for an American company to provide free portable handheld ultrasound devices to the local health clinics, which was much

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

appreciated by local providers. In terms of capacity building, we hired and trained over 100 local people and built capacity in the region; and many of our former data collectors have gone on to work for other public health and development entities. Lastly, we offered a talented young person from Copan a slot as a PhD student in the USA.

Throughout our work in Honduras, and given the extent of local involvement at the regional and MOH levels, we endeavored to act with integrity, curiosity, and respect in all relationships.

Finally, we note that this research would not have been prohibited in the USA. This work is not likely to result in stigmatization, incrimination, or discrimination for the participants, and we have carefully safeguarded all data from threats to the privacy or security of our participants, which has constrained the individual-level data released here.

Taxonomic profiling and diversity analysis:

Quantification of organisms' relative abundance was performed using MetaPhlAn 4 (*8*), which internally mapped the metagenomes against a database of ~5.1M marker genes describing more than 27k~ species-level genome bins (SGB).

We identified a total of 2,508 species in our dataset. Among the 2,508 species, 639 species were used for association analysis after filtering for minimum relative abundance values (10^{-2}), and a minimum of 10% prevalence in the population (n=187).

We performed strain-level profiling for these species with StrainPhlAn4 (*8*)(parameters: "-phylophlan_mode accurate")

Microbiome species richness was estimated using the Shannon entropy index and the total number of observed species (i.e., those with relative abundance simply greater than zero). Multidimensional scaling analysis (vegan cmdscale function) was performed on the Bray-Curtis dissimilarity index (vegan vegdist function) calculated on the relative abundances obtained by MetaPhlAn4.

Functional potential analysis was performed using HUMAnN 3.0. (*53*). Gene family profiles were normalized using relative abundances and collapsed into MetaCyc pathways.

To understand the amount of variance explained by various factors, we performed a PERMANOVA analysis (adonis function from the vegan package (*54*)) using the "bray" method; the diversity matrix was calculated on both species-level relative abundances and MetaCYC pathway relative abundances as input, including the 123 phenotypes variables into the model. All the comparisons were run with 999 permutations.

Phenotype characterization:

We measured a broad range of phenotypes using standard measures (*7*). Description and statistics on all phenotypes can be found in **Supplementary Tables 2-4**. Physiological measurements were deemed within normal limits in accordance with CDC (*55*) and NBME (*56*) guidelines (**Figure S3**).

We used self-reported information to discern whether people were healthy or were diagnosed with various conditions. General anxiety disorder is derived from a set of 7 questions from a self-reported survey-based questionnaire, which assigns a score of 0 to "Not at all", 1 to "Several days", 2 to "More than half the days", and 3 to "Nearly every day". The scores are added up (maximum of 21) and partitioned as: Minimal or none (≤5), Mild (6 - 10), Moderate (11 - 15), and Severe (≥16) (*57*). The PHQ9 (Patient Health Questionnaire) score measuring depression was computed similarly, with the levels being: Minimal or none (≤5), Mild (6 - 10), Moderate (11 - 15), Moderately severe (16 - 19), and Severe (≥20) (*58*). Personality traits like Openness or Nervousness were also based on self-reported questions, where the participants were asked to rate themselves between strongly disagree to strongly agree for each of the personality questions.

The Frequency of intake of various food items was self-reported, ranging from: "Never/rarely" to "Every day". These frequencies were used as input in the diet-microbiome association model. The diet diversity score (DDS) (*18*) was calculated by classifying individual food types into one of the following categories: cereals, roots/tubers, vegetables, fruits, meat/poultry/offal, eggs, fish/seafood, pulses/legumes/nuts, milk and milk products, oils/fats, or sugar/honey. If any of these food items were

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

consumed daily, the respective categories would get 1 for that individual. The sum across these categories would define the DDS score of this individual. The maximum possible DDS score would be 11 and the minimum would be 0.

Numerical values were reported for alcohol frequency and cigarette frequency. The daily alcohol intake ranged from "1 or 2" to "10 or more" drinks. Cigarette usage was reported as a "Yes" or "No".

The household wealth index is computed using Multiple Correspondence Analysis (MCA) based on all the household items. The index ranged from 1, indicating low wealth, to 5, indicating high wealth.

We explored associations with several social network features, including the degree, transitivity, and betweenness centrality of each individual. To uncouple the effects of kin and non-kin social connections, we investigated microbiome associations in familial networks, friendship networks, and combined networks. In the combined network, we computed the amount of kin in a person's first three degrees of social connections (i.e., among a person's friends, friends of friends, and friends of friends of friends) to assess the relative effect of having kin close to a person within the social network. In addition to kin and non-kin relationships, we also explored the microbiome's association with cohabiting partners.

Population-weighted village centroid:

We collected the GPS coordinates (latitude and longitudes) of all the building in the village. Since multiple individuals can reside in a building, the population-weighted centroid was chosen as the reference center of the village, which was then used to compute every individual's distance from this village center. Satellite plots were created using "ggmap" package in R (*59*).

Model for microbiome-phenotype regression:

For the association model with species-level microbiome and phenotypes, a linear mixed model (lmerTest Rpackage (v 3.1)) was used to explore the relationship of the variability in phenotype and the variability in the microbiome.

For every species and phenotype pair, we computed the following model

$$\text{Species abundance} \sim \text{Phenotype of interest} + \text{Age} + \text{Sex} + \text{BMI} + \text{Batch Effect} + \text{Bristol Stool Scale} + \text{DNA concentration} + \text{Sampling date} + (1 \mid \text{Village ID})$$

Species' relative abundances were transformed using the CLR (Centered-Log Ratio) and used as input.

Since basic demographic attributes (age, sex), technical factors (DNA concentration, sequencing batch, sampling date), and BMI and Bristol stool scale accounted for 1.65% of the species variation and 5.86% of the pathway variation, we used these variables as primary controls in our association models (**Figure S4**, **Figure S8**).

Furthermore, all associations were corrected for both microbiome species and phenotype using multiple hypothesis testing (Benjamini-Hochberg correction) and all significant associations are corrected for an FDR (False Discovery Rate) <0.05.

Strain-phenotype analysis and phylogenetic signal:

For strain-level analysis, we used the Almer function from the "evolvability" R package (v 2.0.0). Almer incorporates phylogenetic trees in mixed linear models as a correlated random effects structure.

$$\text{Species abundance} \sim \text{Phenotype of interest} + \text{Age} + \text{Sex} + \text{BMI} + \text{Batch effect} + \text{Bristol Stool Scale} + \text{DNA concentration} + \text{Sampling date} + (1 \mid \text{Village ID}) + (1 \mid \text{phyl})$$

where, "phyl" is the variance-covariance matrix calculated from the species' phylogenetic tree. To evaluate the strain-phylogenetic effect, we compared beta coefficients from this model and the same model without the random effect on the variance-covariance matrix.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

The phylogenetic signal was estimated using the "phylosig" function in "phytools" R package (v 1.9-23) using the 'lambda' method. Overall, among the 78,597 species-phenotype pairs (639 species and 123 phenotypes), 52,864 pairs were chosen after filtering for phylogenetic signal. The phylogenetic signal was estimated for the phylogenetic tree of each species vs phenotype of interest.

Polymorphic sites analysis:

For polymorphic sites, files suffixed with ".polymorphic" in StrainPhlAn4 output were used after discarding 0's in the "percentage of polymorphic sites" column to discard subjects without the species of interest. Wilcoxon rank-sum tests were performed across categories within phenotypes to check for significant changes in polymorphic sites. In addition, linear regression was also performed to investigate the relationship between polymorphic site percentage and individual host phenotypes (see **Supplementary Table 10**).

Differential abundance analysis:

We used the MaAsLin 2 (v 1.0.0) package in R to determine the association between species and disease status (healthy or unhealthy) of individuals and to estimate the effect sizes and adjusted p-values (FDR corrected). A list of significantly positive and negatively associating species was recorded. Species-level relative abundances were normalized (in MaAsLin2) and used as input for MaAsLin2 in which age, sex, BMI, DNA concentration, sampling date, and Bristol stool scale were used as controls and village as a random effect. All the resulting p-values obtained by the MaAsLin2 models were corrected for multiple hypothesis testing using FDR.

Acknowledgments:

We thank all the study participants in Honduras. We thank Jose Eduardo Gámez and Eduardo Jose Urrea Carbajal for coordinating the fieldwork; Rennie Negron, Liza Nicoll, and Thomas Keegan for their support on field operations, data collection, and administrative support; YCGA (Yale Center for Genomic Analysis) for sequencing the metagenomic libraries; and Qiaojuan Shi for processing the specimens and handling the extractions. We thank Michael Baym and Mark Gerstein for helpful comments on the manuscript.

This work was supported by the NOMIS Foundation, with additional support from Schmidt Futures, the Pershing Square Foundation, and the Rothberg Catalyzer Fund. Empanelment of the underlying cohort was supported by the Bill and Melinda Gates Foundation.

Author contributions:

Conceptualization: SVS, FB, MA, IB, and NAC; Methodology: SVS, FB, MA, AS, IB, and NAC; Data Collection: SVS, FB, MA, RMJ, IB, and NAC; Statistical Analysis: SVS, FB, MA; Funding acquisition: NC; Supervision: IB, NAC; Writing: SVS, FB, MA, IB, and NAC.

Competing interests:

The authors declare that they have no competing interests.

Data and materials availability:

The code for replicating the analysis is available at https://github.com/human-nature-lab/Phenotype-paper. Metagenomic sequences for the study participants are deposited in NBCI SRA and available under accession number PRJNA999635.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

REFERENCES

1. R. Gacesa, A. Kurilshikov, A. Vich Vila, T. Sinha, M. A. Y. Klaassen, L. A. Bolte, S. Andreu-Sánchez, L. Chen, V. Collij, S. Hu, J. A. M. Dekens, V. C. Lenters, J. R. Björk, J. C. Swarte, M. A. Swertz, B. H. Jansen, J. Gelderloos-Arends, S. Jankipersadsing, M. Hofker, R. C. H. Vermeulen, S. Sanna, H. J. M. Harmsen, C. Wijmenga, J. Fu, A. Zhernakova, R. K. Weersma, Environmental factors shaping the gut microbiome in a Dutch population. *Nature*. **604**, 732–739 (2022).

2. R. J. Abdill, E. M. Adamowicz, R. Blekhman, Public human microbiome data are dominated by highly developed countries. *PLoS Biol.* **20**, e3001536 (2022).

3. M. Mohanan, M. Vera-Hernández, V. Das, S. Giardili, J. D. Goldhaber-Fiebert, T. L. Rabin, S. S. Raj, J. I. Schwartz, A. Seth, The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural India. *JAMA Pediatr.* **169**, 349–357 (2015).

4. B. N. Young, M. L. Clark, S. Rajkumar, M. L. Benka-Coker, A. Bachand, R. D. Brook, T. L. Nelson, J. Volckens, S. J. Reynolds, C. L'Orange, N. Good, K. Koehler, S. Africano, A. B. Osorto Pinel, J. L. Peel, Exposure to household air pollution from biomass cookstoves and blood pressure among women in rural Honduras: A cross-sectional study. *Indoor Air*. **29**, 130–142 (2019).

5. D. Hartley, Rural health disparities, population health, and rural culture. *Am. J. Public Health*. **94**, 1675–1678 (2004).

6. E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower, N. Segata, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*. **176**, 649-662.e20 (2019).

7. H. B. Shakya, D. Stafford, D. A. Hughes, T. Keegan, R. Negron, J. Broome, M. McKnight, L. Nicoll, J. Nelson, E. Iriarte, M. Ordonez, E. Airoldi, J. H. Fowler, N. A. Christakis, Exploiting social influence to magnify population-level behaviour change in maternal and child health: study protocol for a randomised controlled trial of network targeting algorithms in rural Honduras. *BMJ Open*. **7**, e012996 (2017).

8. A. Blanco-Míguez, F. Beghini, F. Cumbo, L. J. McIver, K. N. Thompson, M. Zolfo, P. Manghi, L. Dubois, K. D. Huang, A. M. Thomas, W. A. Nickols, G. Piccinno, E. Piperni, M. Punčochář, M. Valles-Colomer, A. Tett, F. Giordano, R. Davies, J. Wolf, S. E. Berry, T. D. Spector, E. A. Franzosa, E. Pasolli, F. Asnicar, C. Huttenhower, N. Segata, Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* (2023), doi:10.1038/s41587-023-01688-w.

9. L. Domènech, J. Willis, M. Alemany-Navarro, M. Morell, E. Real, G. Escaramís, S. Bertolín, D. Sánchez Chinchilla, S. Balcells, C. Segalàs, X. Estivill, J. M. Menchón, T. Gabaldón, P. Alonso, R. Rabionet, Changes in the stool and oropharyngeal microbiome in obsessive-compulsive disorder. *Sci. Rep.* **12**, 1448 (2022).

10. A. Saleem, A. Ikram, E. Dikareva, E. Lahtinen, D. Matharu, A.-M. Pajari, W. M. de Vos, F. Hasan, A. Salonen, C. Jian, Unique Pakistani gut microbiota highlights population-specific microbiota signatures of type 2 diabetes mellitus. *Gut Microbes*. **14**, 2142009 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

11. K. S. Singh, D. Paul, A. Gupta, D. Dhotre, F. Klawonn, Y. Shouche, Indian sewage microbiome has unique community characteristics and potential for population-level disease predictions. *Sci. Total Environ.* **858**, 160178 (2023).

12. M. Balakumar, D. Prabhu, C. Sathishkumar, P. Prabu, N. Rokana, R. Kumar, S. Raghavan, A. Soundarajan, S. Grover, V. K. Batish, V. Mohan, M. Balasubramanyam, Improvement in glucose tolerance and insulin sensitivity by probiotic strains of Indian gut origin in high-fat diet-fed C57BL/6J mice. *Eur. J. Nutr.* **57**, 279–295 (2018).

13. N. T. Baxter, A. W. Schmidt, A. Venkataraman, K. S. Kim, C. Waldron, T. M. Schmidt, Dynamics of Human Gut Microbiota and Short-Chain Fatty Acids in Response to Dietary Interventions with Three Fermentable Fibers. *MBio*. **10** (2019), doi:10.1128/mBio.02566-18.

14. N. A. Bokulich, J. Chung, T. Battaglia, N. Henderson, M. Jay, H. Li, A. D Lieber, F. Wu, G. I. Perez-Perez, Y. Chen, W. Schweizer, X. Zheng, M. Contreras, M. G. Dominguez-Bello, M. J. Blaser, Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).

15. A. M. Baumann-Dudenhoeffer, A. W. D'Souza, P. I. Tarr, B. B. Warner, G. Dantas, Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. *Nat. Med.* **24**, 1822–1829 (2018).

16. M. Bonaccio, A. Di Castelnuovo, A. Bonanni, S. Costanzo, F. De Lucia, G. Pounis, F. Zito, M. B. Donati, G. de Gaetano, L. Iacoviello, Moli-sani project Investigators*, Adherence to a Mediterranean diet is associated with a better health-related quality of life: a possible role of high dietary antioxidant content. *BMJ Open*. **3** (2013), doi:10.1136/bmjopen-2013-003003.

17. P.-Y. Hong, M. Iakiviak, D. Dodd, M. Zhang, R. I. Mackie, I. Cann, Two new xylanases with different substrate specificities from the human gut bacterium Bacteroides intestinalis DSM 17393. *Appl. Environ. Microbiol.* **80**, 2084–2093 (2014).

18. S. Kim, P. S. Haines, A. M. Siega-Riz, B. M. Popkin, The Diet Quality Index-International (DQI-I) provides an effective tool for cross-national comparison of diet quality as illustrated by China and the United States. *J. Nutr.* **133**, 3476–3484 (2003).

19. N. Darmon, A. Drewnowski, Contribution of food prices and diet cost to socioeconomic disparities in diet quality and health: a systematic review and analysis. *Nutr. Rev.* **73**, 643–660 (2015).

20. N. Darmon, E. L. Ferguson, A. Briend, A cost constraint alone has adverse effects on food selection and nutrient density: an analysis of human diets by linear programming. *J. Nutr.* **132**, 3764–3771 (2002).

21. A. Rammohan, B. Pritchard, M. Dibley, Home gardens as a predictor of enhanced dietary diversity and food security in rural Myanmar. *BMC Public Health*. **19**, 1145 (2019).

22. N. B. Ali, T. Tahsina, D. M. E. Hoque, M. M. Hasan, A. Iqbal, T. M. Huda, S. El Arifeen, Association of food security and other socio-economic factors with dietary diversity and nutritional statuses of children aged 6-59 months in rural Bangladesh. *PLoS One*. **14**, e0221929 (2019).

23. M. Tang, N. E. Weaver, D. N. Frank, D. Ir, C. E. Robertson, J. F. Kemp, J. Westcott, K. Shankar, A. L. Garces, L. Figueroa, A. K. Tshefu, A. L. Lokangaka, S. S. Goudar, M. Somannavar, S. Aziz, S. Saleem, E. M. McClure, K. M. Hambidge, A. E. Hendricks, N. F. Krebs, Longitudinal reduction in

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

diversity of maternal gut Microbiota during pregnancy is observed in multiple low-resource settings: Results from the Women First trial. *Front. Microbiol.* **13**, 823757 (2022).

24. J. K. Copeland, G. Chao, S. Vanderhout, E. Acton, P. W. Wang, E. I. Benchimol, A. El Sohami, K. Croitoru, J. L. Gommerman, D. S. Guttman, GEMINI Research Team, The impact of migration on the gut metagenome of south Asian Canadians. *Gut Microbes*. **13**, 1–29 (2021).

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

26. E. Blacher, S. Bashiardes, H. Shapiro, D. Rothschild, U. Mor, M. Dori-Bachash, C. Kleimeyer, C. Moresi, Y. Harnik, M. Zur, M. Zabari, R. B.-Z. Brik, D. Kviatcovsky, N. Zmora, Y. Cohen, N. Bar, I. Levi, N. Amar, T. Mehlman, A. Brandis, I. Biton, Y. Kuperman, M. Tsoory, L. Alfahel, A. Harmelin, M. Schwartz, A. Israelson, L. Arike, M. E. V. Johansson, G. C. Hansson, M. Gotkine, E. Segal, E. Elinav, Potential roles of gut microbiome and metabolites in modulating ALS in mice. *Nature*. **572**, 474–480 (2019).

27. O. Manor, C. L. Dai, S. A. Kornilov, B. Smith, N. D. Price, J. C. Lovejoy, S. M. Gibbons, A. T. Magis, Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **11**, 1–12 (2020).

28. J. F. Cryan, K. J. O'Riordan, C. S. M. Cowan, K. V. Sandhu, T. F. S. Bastiaanssen, M. Boehme, M. G. Codagnone, S. Cussotto, C. Fulling, A. V. Golubeva, K. E. Guzzetta, M. Jaggar, C. M. Long-Smith, J. M. Lyte, J. A. Martin, A. Molinero-Perez, G. Moloney, E. Morelli, E. Morillas, R. O'Connor, J. S. Cruz-Pereira, V. L. Peterson, K. Rea, N. L. Ritz, E. Sherwin, S. Spichak, E. M. Teichman, M. van de Wouw, A. P. Ventura-Silva, S. E. Wallace-Fitzsimons, N. Hyland, G. Clarke, T. G. Dinan, The Microbiota-Gut-Brain Axis. *Physiol. Rev.* **99**, 1877–2013 (2019).

29. E. Patterson, P. M. Ryan, J. F. Cryan, T. G. Dinan, R. P. Ross, G. F. Fitzgerald, C. Stanton, Gut microbiota, obesity and diabetes. *Postgrad. Med. J.* **92**, 286–300 (2016).

30. P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, J. I. Gordon, A core gut microbiome in obese and lean twins. *Nature*. **457**, 480–484 (2008).

31. R. E. Ley, P. J. Turnbaugh, S. Klein, J. I. Gordon, Microbial ecology: human gut microbes associated with obesity. *Nature*. **444**, 1022–1023 (2006).

32. M. Murri, I. Leiva, J. M. Gomez-Zumaquero, F. J. Tinahones, F. Cardona, F. Soriguer, M. I. Queipo-Ortuño, Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study. *BMC Med.* **11**, 46 (2013).

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

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

34. Correction for Hu et al., Microbiota-induced activation of epithelial IL-6 signaling links inflammasome-driven inflammation with transmissible cancer. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12852–12852 (2013).

35. T. Sarkola, M. R. Iles, K. Kohlenberg-Mueller, C. J. P. Eriksson, Ethanol, acetaldehyde, acetate, and lactate levels after alcohol intake in white men and women: effect of 4-methylpyrazole. *Alcohol. Clin. Exp. Res.* **26**, 239–245 (2002).

36. L. Allais, F.-M. Kerckhof, S. Verschuere, K. R. Bracke, R. De Smet, D. Laukens, P. Van den Abbeele, M. De Vos, N. Boon, G. G. Brusselle, C. A. Cuvelier, T. Van de Wiele, Chronic cigarette smoke exposure induces microbial and inflammatory shifts and mucin changes in the murine gut. *Environ. Microbiol.* **18**, 1352–1363 (2016).

37. S. Botschuijver, G. Roeselers, E. Levin, D. M. Jonkers, O. Welting, S. E. M. Heinsbroek, H. H. de Weerd, T. Boekhout, M. Fornai, A. A. Masclee, F. H. J. Schuren, W. J. de Jonge, J. Seppen, R. M. van den Wijngaard, Intestinal Fungal Dysbiosis Is Associated With Visceral Hypersensitivity in Patients With Irritable Bowel Syndrome and Rats. *Gastroenterology*. **153**, 1026–1039 (2017).

38. V. Caputi, I. Marsilio, V. Filpa, S. Cerantola, G. Orso, M. Bistoletti, N. Paccagnella, S. De Martin, M. Montopoli, S. Dall'Acqua, F. Crema, I.-M. Di Gangi, F. Galuppini, I. Lante, S. Bogialli, M. Rugge, P. Debetto, C. Giaroni, M. C. Giron, Antibiotic-induced dysbiosis of the microbiota impairs gut neuromuscular function in juvenile mice. *Br. J. Pharmacol.* **174**, 3623–3639 (2017).

39. E. E. Fröhlich, A. Farzi, R. Mayerhofer, F. Reichmann, A. Jačan, B. Wagner, E. Zinser, N. Bordag, C. Magnes, E. Fröhlich, K. Kashofer, G. Gorkiewicz, P. Holzer, Cognitive impairment by antibiotic-induced gut dysbiosis: Analysis of gut microbiota-brain communication. *Brain Behav. Immun.* **56**, 140–155 (2016).

40. D. A. Kim, A. R. Hwong, D. Stafford, D. A. Hughes, A. J. O'Malley, J. H. Fowler, N. A. Christakis, Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *Lancet*. **386**, 145–153 (2015).

41. S. Arps, Socioeconomic status and body size among women in Honduran Miskito communities. *Ann. Hum. Biol.* **38**, 508–519 (2011).

42. E. Rosenberg, G. Sharon, I. Zilber-Rosenberg, The hologenome theory of evolution contains Lamarckian aspects within a Darwinian framework. *Environ. Microbiol.* **11**, 2959–2962 (2009).

43. S. J. Carlson, A. A. O'Loughlin, L. Anez-Bustillos, M. A. Baker, N. A. Andrews, G. Gunner, D. T. Dao, A. Pan, P. Nandivada, M. Chang, E. Cowan, P. D. Mitchell, K. M. Gura, M. Fagiolini, M. Puder, A Diet With Docosahexaenoic and Arachidonic Acids as the Sole Source of Polyunsaturated Fatty Acids Is Sufficient to Support Visual, Cognitive, Motor, and Social Development in Mice. *Front. Neurosci.* **13**, 72 (2019).

44. A. Raulo, B. E. Allen, T. Troitsky, A. Husby, J. A. Firth, T. Coulson, S. C. L. Knowles, Social networks strongly predict the gut microbiota of wild mice. *ISME J.* **15**, 2601–2613 (2021).

45. I. L. Brito, T. Gurry, S. Zhao, K. Huang, S. K. Young, T. P. Shea, W. Naisilisili, A. P. Jenkins, S. D. Jupiter, D. Gevers, E. J. Alm, Transmission of human-associated microbiota along family and social networks. *Nat Microbiol*. **4**, 964–971 (2019).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.11.23288404; this version posted December 12, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

46. M. Valles-Colomer, A. Blanco-Míguez, P. Manghi, F. Asnicar, L. Dubois, D. Golzato, F. Armanini, F. Cumbo, K. D. Huang, S. Manara, G. Masetti, F. Pinto, E. Piperni, M. Punčochář, L. Ricci, M. Zolfo, O. Farrant, A. Goncalves, M. Selma-Royo, A. G. Binetti, J. E. Becerra, B. Han, J. Lusingu, J. Amuasi, L. Amoroso, A. Visconti, C. M. Steves, M. Falchi, M. Filosi, A. Tett, A. Last, Q. Xu, N. Qin, H. Qin, J. May, D. Eibach, M. V. Corrias, M. Ponzoni, E. Pasolli, T. D. Spector, E. Domenici, M. C. Collado, N. Segata, The person-to-person transmission landscape of the gut and oral microbiomes. *Nature*. **614**, 125–135 (2023).

47. J. Pullman, F. Beghini, M. Alexander, S. Vishnempet Shridhar, D. Prinster, I. L. Brito, N. A. Christakis, Detailed social network interactions and gut microbiome strain-sharing within isolated Honduras villages. *bioRxiv* (2023), , doi:10.1101/2023.04.06.535875.

48. L. Xiao, F. Zhang, F. Zhao, Large-scale microbiome data integration enables robust biomarker identification. *Nat. Comput. Sci.* **2**, 307–316 (2022).

49. R. Wynford-Thomas, N. P. Robertson, The economic burden of chronic neurological disease. *J. Neurol.* **264**, 2345–2347 (2017).

50. V. A. Cantu, J. Sadural, R. Edwards, "PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets" (e27553v1, PeerJ Preprints, 2019), , doi:10.7287/peerj.preprints.27553v1.

51. BMTagger. *ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/*.

52. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

53. F. Beghini, L. J. McIver, A. Blanco-Míguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, P. Manghi, M. Scholz, A. M. Thomas, M. Valles-Colomer, G. Weingart, Y. Zhang, M. Zolfo, C. Huttenhower, E. A. Franzosa, N. Segata, Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*. **10** (2021), doi:10.7554/eLife.65088.

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

55. CDC, Centers for disease control and prevention. *Centers for Disease Control and Prevention* (2023), (available at https://www.cdc.gov/).

56. G. Chemistry:, Reference range SI reference intervals SERUM, (available at https://www.nbme.org/sites/default/files/2020-07/Laboratory_Reference_Values.pdf).

57. N. Williams, The GAD-7 questionnaire. *Occup. Med.* . **64**, 224–224 (2014).

58. K. Kroenke, R. L. Spitzer, J. B. Williams, The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).

59. D. Kahle, H. Wickham, Ggmap: Spatial visualization with ggplot2. *R J.* **5**, 144 (2013).