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], Ilana L. Brito[3]*, and Nicholas A. Christakis[1,2,4,5]*

[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] Department of Statistics and Data Science, Yale University; New Haven, CT, USA
[5] 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 regions of the world, data linking microbiome features from such settings to diverse phenotypes remains uncommon. Here, using metagenomic data from a community-based cohort of 1,187 people from isolated villages in the Mesoamerican highlands of Western Honduras, we report 7,117 statistically robust associations spanning 788 gut microbial species (including both known and unknown taxa) and 126 phenotypes (including physical and mental health, medication use, diet, animal exposure, and social and economic measures). We report 394 new associations with mental health phenotypes alone, as well as 3,004 associations with diverse socioeconomic phenotypes. Distinctly, we also found 1,210 associations with microbiome metabolic pathways. We also report 302 significant associations after including strain-level phylogenies from 666 microbial species. Including the strain-phylogenetic information changes the overall relationship between gut microbiome and these phenotypes, and strain-level phylogenetic information enhances the observed relationship between microbiome and phenotypes as a whole. Our findings suggest new roles that gut microbiome surveillance can play in understanding broad features of individual and public health.

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 various phenotypes may, reciprocally, influence the microbiome, is becoming increasingly clear.[1] Yet these studies have largely focused on industrialized populations.[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).[3,4,5] 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 information) are still uncommonly applied in non-industrialized settings.

The village communities in the western highlands of Honduras are geographically isolated **(Fig. 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 psychological, socioeconomic, and health phenotypes. Our cohort consists of 1,187 people living in 11 villages which are part of a larger cohort developed for a different original purpose.[7] The adult population size in our 11 villages ranges from 66 to 432 individuals, and the average household size is 4.68. The average age of participants was 39.67 (SD=17.06; range: 15 - 93); 62.4% were women and 37.6% were men; and 26.3% of them were married. Each of our 11 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 (**Fig. 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 (**Fig. 1B-C**). This aspect is also reflected within the social networks of the villagers. Villagers located at the topological center of the network have a more similar microbiome to the rest of the village, unlike those at the social periphery ($\beta = 3.66 \times 10^{-5}$, p-value =0.761 from linear regression model; see Methods for details, and also **Fig. 1A**).

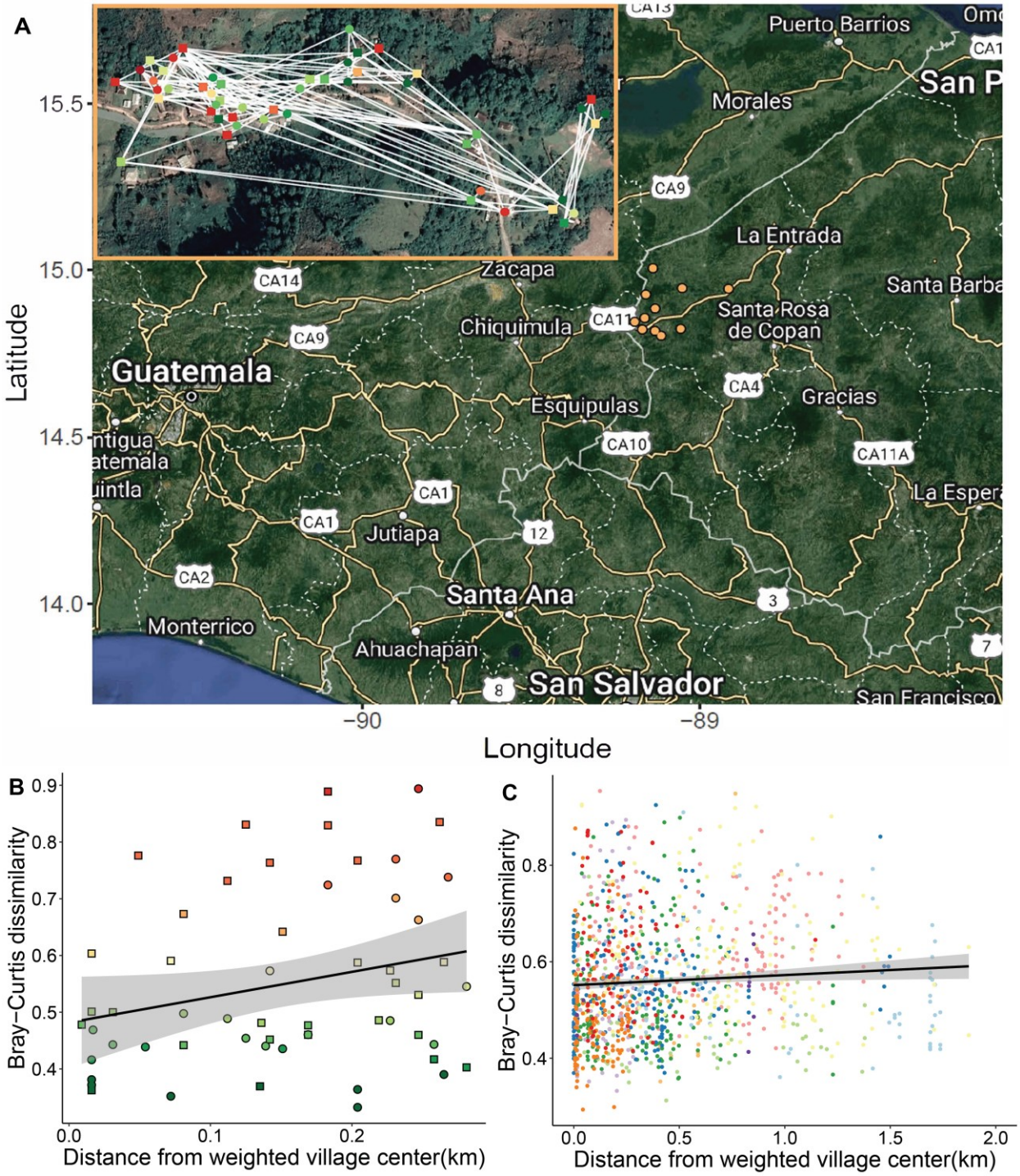


Fig. 1: Geographic overview of the Honduras microbiome project: (A) A satellite view of the Honduran villages (in orange) which constitute the microbiome dataset. On the top left, there is a zoomed-in satellite view of an illustrative village with each inhabitant (n=57) colored with beta diversity (Bray-Curtis dissimilarity) 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 dissimilar. Square nodes indicate male and circle nodes indicate female villagers. (B) Scatter-plot of Bray-Curtis dissimilarity (of the above village) and the distance of households from the

population-weighted village centroid (see methods) shows a positive correlation ($\rho = 0.144$, p-value =0.05) of gut microbiome dissimilarity with distance. Individual dots are colored according to the person's dissimilarity from the village's average microbiome. The light grey areas indicate a 95% confidence interval. (C) Combined plot of all the village Bray-Curtis dissimilarities and distance from village centroids. The black regression line indicates a consistent trend ($\rho = 0.311$, p-value =0.00253) 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.

Overall, we found 7,117 associations when examining 788 microbial species and 126 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 (see Methods) and were corrected for multiple hypothesis testing using a False Discovery Rate (FDR) procedure. Distinctly, we also found 1,210 associations with metabolic pathways (See Supplementary Table 2). The 126 phenotypes exist as continuous, categorical, and discrete variable types (Supplementary Tables 3-5). Of course, 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 refrigerator (Extended Data Fig. 1)). Similarly, 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 (Extended Data Fig. 2).

For the health phenotypes, 722 species were found to be significantly associated with at least one phenotype (Supplementary Table 1 and 3). Among the 722 significant species, 556 of them belonged to *Firmicutes*, making this phylum the one most associated with health phenotypes. Among all the associated species, 28.12% were identified as unknown[8] at several taxonomic levels. uSGB2240 (unknown at the genus level) from the *Rikenellaceae* family was the most frequently associated species, significantly associated with 11 health phenotypes; in particular, it was observed to be depleted in individuals with worse overall health, high BMI, heart disease, intestinal illness, allergies, moderate-severe anxiety, mild depression, and nervousness; and it was enriched in patients with dementia (**Fig. 2A**). Multiple health phenotypes were characterized by similar changes in relative abundances in multiple species. We also observed that different phenotypes shared similar sets of microbial signatures, especially when looking at species reported as depleted. For instance, uSGB2240 is associated with mental health in general (associating with 5 mental health phenotypes). Microbial species from the *Rikenellaceae* family have been previously found to be associated with at least one anxiety disorder.[9]

Furthermore, a total of 167 pathways were associated with at least one health phenotype, totaling 249 pathway associations (for details regarding ascertainment of microbial metabolic pathways, see Methods). Among the 249 associations, physiological variables had 120 associations, followed by 80 associations in chronic illness phenotypes; 20 in medication use; 19 in acute conditions; 8 in personality measures, alcohol, cigarettes, and mental health; and 2 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 over-weight individuals, or diastolic pressure>89 to account for hypertensive individuals) compared to healthy individuals as controls (Extended Data Fig. 3, Supplementary Table 6). High diastolic blood pressure had the strongest effect sizes among the health phenotypes. Three species were associated with four unhealthy phenotypes. *Bacteroides bouchesdurhonensis* was found associated with the anemic range of total hemoglobin, high heart rate, and high blood pressure (Supplmentary Table 6); uSGB6513, an unknown species in the Bacilli family, was associated with high hemoglobin A1c and high blood pressure; and *Clostridia* bacterium (SGB4394) was associated with high hemoglobin A1C (6.5-7), overweight BMI, and high blood pressure (Supplementary Table 6).

We also evaluated whether the diversity of the microbiome was itself associated with various health (and other) phenotypes. The majority of the villagers self-reported themselves as healthy individuals (n=847, 72%) and only 132 villagers (11%) reported having more than one disease. Villagers with reported illnesses (except arthritis) were found to have lower diversity relative to healthy villagers (**Fig.**

2B); in particular, villagers with reported stomach illness had decreased diversity (p-value= 4.9e-5 Wilcoxon Rank Sum test). Villagers taking various medications also had lower diversity (**Fig. 2C**); anti-parasitic drug users showed the lowest diversity (p-value=7.1e-4 Wilcoxon Rank Sum test) followed by anti-diarrheal users (p-value=9e-3 Wilcoxon Rank Sum test) and antibiotic users (p-value=12e-3 Wilcoxon Rank Sum test). We found no material associations with other categories of medications.

Overall, all the health phenotypes put together contribute 7.87% of the total variance explained in microbial species composition (Extended Data Fig. 4 and Supplementary Table 7). Similarly, 15.2% of the variance in pathways composition is relevant to health phenotypes.

We also performed a simple comparison by comparing all chronically ill individuals to those without any chronic conditions. With 848 healthy people and 339 chronically ill people, we use differential abundance analysis. We found that 134 species were significantly differentially abundant (**Fig. 2D**). Among them, *Bacteroides ovatus* and *Bacteroides caccae* have been implicated with IBD.[10,11] Specifically, *B. caccae* has been shown to be linked with an increase in the degradation of mucus and increased inflammation;[12] *Parabacteroides distasonis* has been shown to reduce metabolic dysfunction and obesity;[13] *Alistipes putredinis* has been shown to play a direct role on health through diet[14]; and *Prevotella* sp. 885 has been shown to decrease with advanced Chronic Kidney Disease.[15]

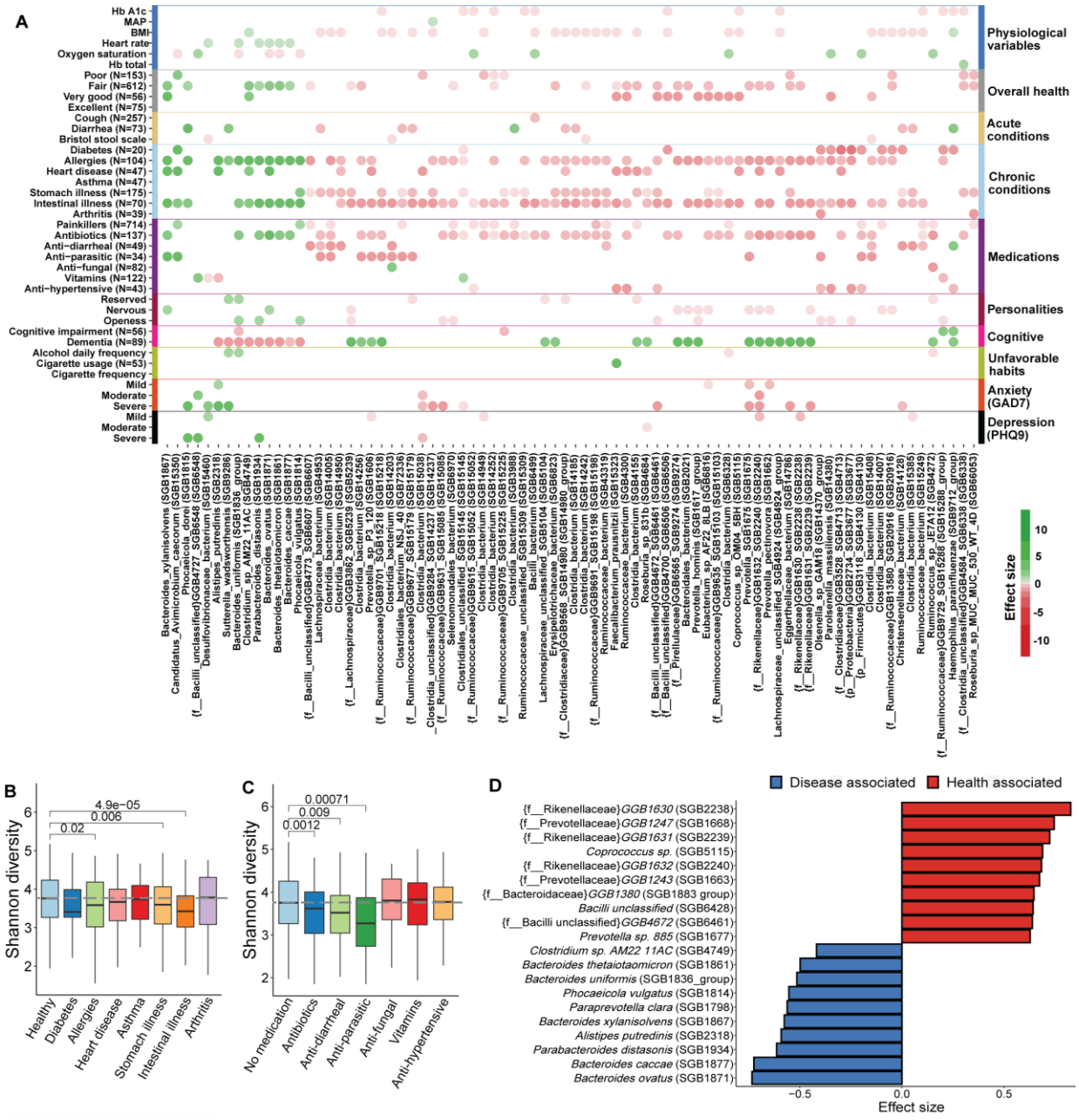


Fig. 2: Microbiome association with health phenotypes: (A) 80 species that best represent gut microbiome associations with health phenotypes (see Supplementary Table 1 for a complete list of associations). The number of individuals (N) manifesting the respective phenotypes is shown in brackets. The presence of color shows p-value significant species for that phenotype (FDR<0.05); the intensity of the color corresponds to the strength of the effect size. Unknown species are indicated with "{}", specifying the taxonomic level at which the species is known. Phenotypes without "N" are recorded for all individuals. (B) Shannon diversity of healthy and chronically ill individuals highlights differences in overall microbiome diversity; healthy individuals are chosen as a reference (gray dashed line). (C) Shannon diversity calculated between different medication use categories; non-medicated individuals (n=793) are chosen as reference (grey dashed line). Pairwise comparisons were performed using the Wilcoxon Rank Sum test and corrected for multiple hypothesis correction. (D) List of the top 10 differentially abundant species between healthy and chronically diseased individuals (those with at least one illness), with the effect size obatined using MaAsLin2 (see Methods).

We also explored possible associations with animal exposure and diet.[1,16,17,18] 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, 632 species were found to be significantly associated with at least one of the phenotypes, resulting in 1,858 associations (**Fig. 3A**, Supplementary Table 4). Among all the associating bacterial species, 26% were unknown. Among the 632 significant species, 471 of them belonged to *Firmicutes*, making this phylum the one most commonly associated with specific animals or food categories. We found 8 pathways associated with exposure to animals (Supplementary Table 2). Animal exposure contributed to 3.48% of the variation in species composition. We found no difference in overall Shannon diversity in individuals exposed to different animal categories (Extended Data Fig. 5).

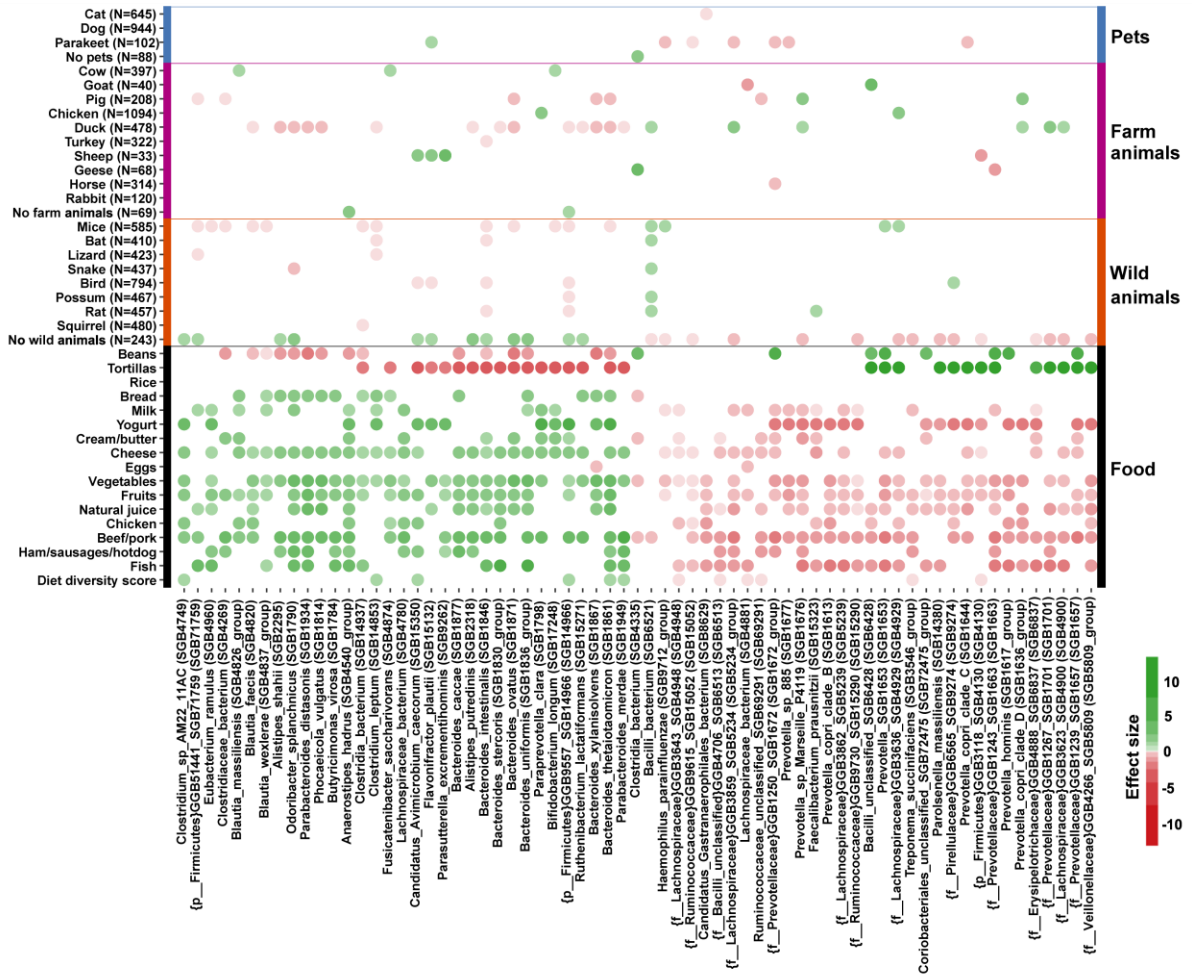


Fig. 3: Microbiome association with animal and food phenotypes: 74 species that best represent gut microbiome associations with animal exposure and food consumption (see Supplementary Table 1 for a complete list of associations). The number of individuals (N) involved in the respective phenotypes is shown in brackets. The presence of color shows p-value significant species for that phenotype (FDR<0.05); the intensity of color corresponds to the strength of the effect size. Unknown species are indicated with "{}", specifying the taxonomic level at which the species is known. The top ribbon indicates the phylum of the associating species. Phenotypes without "N" are recorded for all individuals.

Diet has been extensively studied and shown to have a substantial impact on the gut microbiome.[18,19,20] We assessed associations with microbial features and food frequency consumption (Extended Data Fig. 5) and found 1,319 significant associations with diet (**Fig. 3A**). *Bacteroides intestinalis* (SGB1846) was the species associated with the most food phenotypes. *Bifidobacterium longum* (SGB17248), *Escherichia coli* (SGB10068), and *Klebsiella pneumoniae* (SGB10115) were associated with cream/butter and cheese in our cohort, similar to other studies.[21,22,23] Even though the majority of the individuals were consuming tortillas and beans on a daily basis, we measured diet diversity using the Diet Diversity Score[24] (DDS) (see Methods and Extended Data Fig. 6). We identified a total of 36 significant associations between the DDS and gut microbiome species (**Fig. 3A**); in particular one of the significant species, *Flavonifractor plautii* (SGB15132), was also found in another study with individuals having high DDS.[25] We also found 451 pathway associations with food phenotypes (Supplementary Table 2). Diet was responsible for 3.09% and 4.04% of the variance explained in our sample in species and pathways composition respectively (Extended Data Fig. 4).

For all socioeconomic phenotypes, 718 species were found to be significantly associated with at least one of the phenotypes (**Fig. 4A**, Supplementary Table 1). Among all the 718 associated species, 27% of them were unknown, and 546 of them belong to *Firmicutes*, making it again the most dominant phyla for socioeconomic factors. Moreover, *uSGB4929* of the Lachnospiraceae family is the species with the strongest association, significantly associated with 15 socio-economic phenotypes. We also found 512 associations with pathways, with 3 of them associating with 7 of the socio-economic phenotypes (Supplementary Table 2).

Socioeconomic factors are relevant to many exposures and habits of individuals. Higher monthly expenditures are correlated with a better diet and better household essentials such as having a refrigerator or a paved floor. Indeed, most of the bacteria associated with higher monthly expenditures are the same as the ones associated with better diet quality.[26,27]

Even though all of the subjects in our sample are poor, 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 1 (least wealthy) to 5 (most wealthy)) is 3.21, and the standard deviation was 1.36. In terms of measures of economic status, both monthly expenditure and travel were associated with the microbiome. Of course, wealth was also directly correlated with owning various items (such as a TV or a mobile phone) some of which (such as 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 (Extended Data Fig. 1). This is evident in **Fig. 4A**. The analysis with the wealth index revealed similar patterns of association, where a high index was associated with the same bacterial species as owning expensive items (like glass windows) and conversely. The variance explained by economic factors was 5.03% for species and 5.04% for pathways (Extended Data Fig. 4), indicating the relative importance of economic factors in explaining variation in the gut microbiome.

Moreover, as shown in **Fig. 4B**, even at the level of overall microbial diversity, the subjects from the less well-off households (in the bottom three quintiles of the wealth distribution) had a Shannon index that was higher from that of the subjects from the wealthiest households (in the top quintile).

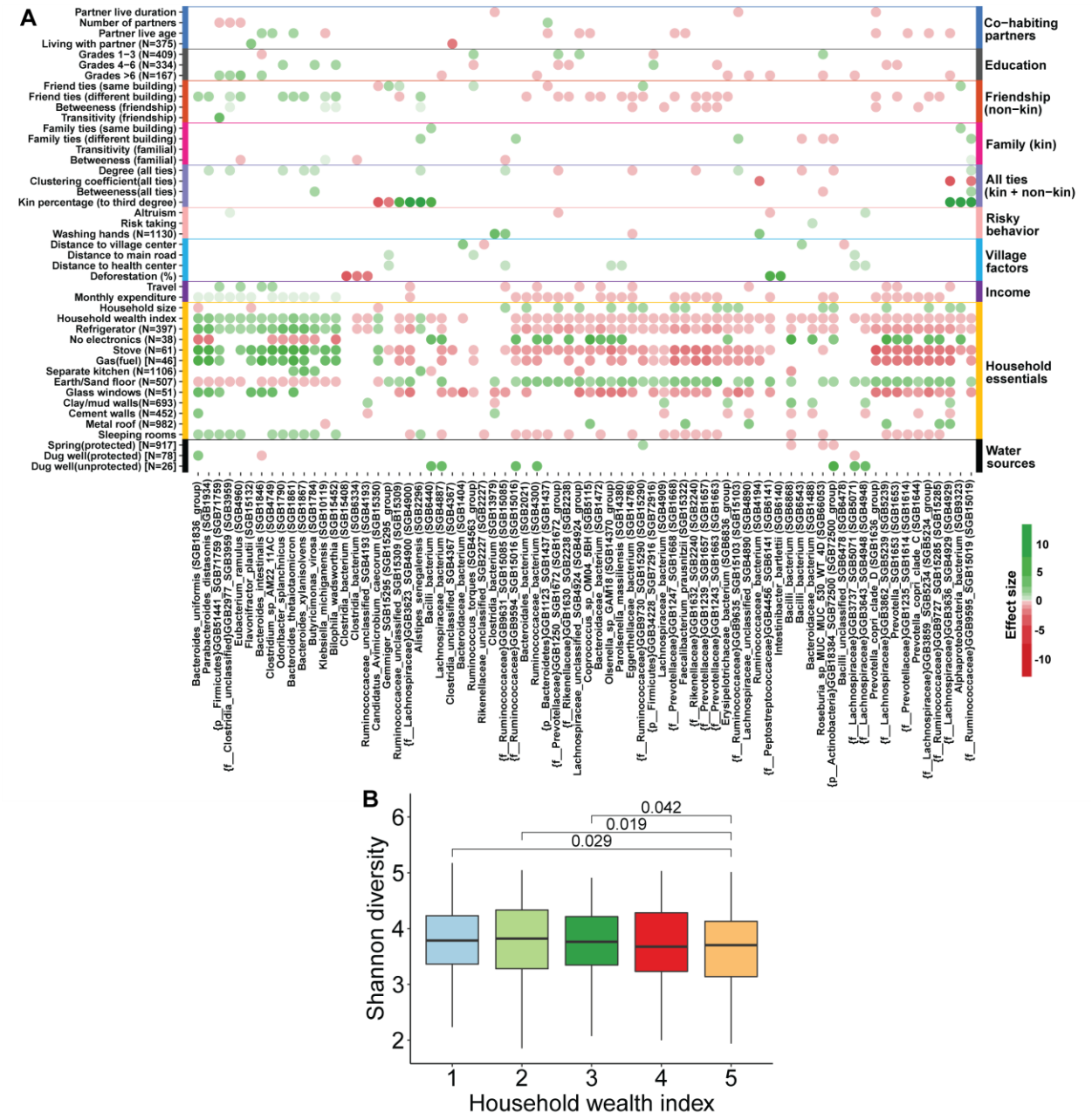


Fig. 4: Microbiome association with socioeconomic and environmental phenotypes: (A) 74 species which best represent gut microbiome associations with socioeconomic, environmental, and social interaction phenotypes (see Supplementary Table 4 for complete list). Social interaction phenotypes included co-habiting partners, friendship (non-kin) networks, familial (kin) network, and "all" (kin+non-kin) networks. The number of individuals (N) involved in the respective phenotypes is shown in brackets. The presence of color shows p-value significant species for that phenotype (FDR<0.05); the depth of color corresponds to the strength of the effect size reported from the mode. Unknown species are indicated with "{}", specifying the taxonomic level at which the species is known. Phenotypes without "N" are recorded for all individuals. (B) 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.

In addition to the above factors, **Fig. 4A** and **Extended Data Fig. 4** show that education (0.27% variance explained in species, 0.38% in pathways) was also related to the gut microbiome and had 127 associations. And environmental exposures in the village, like the extent of nearby deforestation or the distances to the main road, health center, or village center, had 66 significant associations with 0.7% and 2.11% of the variance explained in species and pathways, suggesting such exposures play some (small) role in influencing the overall gut microbiome. In the past, household-level environmental exposures (such as water sources) have also been known to influence both pathogenic and non-pathogenic bacterial species in the human gut microbiome.[28,29,30] Our results reveal 124 unique associations between the household water sources and the gut microbiome, with 0.85% of the species variance and 0.73% of the pathway variance explained. Moreover, Supplementary Table S2 reveals 489 pathway associations with economic variables (including income, risky behavior, village factors, and household variables). The remaining 9 pathway associations were found with water sources. In addition, social network factors had 409 unique associations with the microbiome, with familial factors making up 101 associations, and friendship factors having 159 associations. Overall, social factors made up 1.68% of the variance explained in the case of species and 1.21% of pathways (Extended Data Fig. 4).

Finally, moving beyond species-specific associations with phenotypes, there is meaningful variation between the genetic makeup of the same species across different individuals that is in turn associated with diverse phenotypes. This is shown in **Fig. 5A** with an illustrative example where frequent fruit consumers are clustered together within the strain phylogeny plot of uSGB14230 (in the *Clostridia* family). Other studies have established links between several *Clostridium* species and a good diet.[31,32] This strain-level phylogeny effect reflects the phylogenetic tree structure of this single species (**Fig. 5A**); that is, the fruit-eating phenotype is associated with a particular locations on the strain-level phylogenetic tree of this bacterial species. Therefore, even within the same species, different strains have different phenotype relationships.

Overall, including strain phylogenies in the model of microbiome genotype-phenotype associations (N=83,916) changes the results: by adjusting the model by considering the strain-level phylogeny and then comparing it with the unadjusted model, there is an overall shift of the estimated β-coefficient of 5.12% – across all our species and phenotypes combined (**Fig. 5B**). A total of 45.02% (238) of the associations are present (significant) in both results (with and without strain-phylogenetic effect). But 19.88% (105) associations are present only in the absence of the other level. Among the 238 associations which were present in the models both with and without strain information, the inclusion of strain-phylogenetic effects flips the direction of effect in 28.26% (67) of the associations (Extended Data Fig. 7).

The significant association between species and phenotype after including the strain information into the model can be broken down. After considering all phenotypes and including strain-level phylogenetic information, we found 302 significant associations that can be observed across ten sub-categories of phenotypes across 666 species (Extended Data Fig. 8). The health phenotypes (comprising physiological variables, chronic conditions, medication usage, mental health, etc.), showed 54 associations in total, with 34 species. The food and animal categories had 66 associations with 38 species. As for the socioeconomic factors (comprising economic and social variables, and water sources), they had almost three times as many associations as any other, with 182 associations over 69 species.

Since, as shown in **Fig. 5B**, adding the strain-phylogenetic effect alters the relationship between species and phenotypes, we performed a parallel comparison of the individual effect sizes in the two models. Examining the effects sizes of species (across all phenotypes) revealed 6 species which were significantly altered after adding the strain-phylogenetic effect. The species *Faecalibacterium prausnitzii* (SGB15318 group), *Clostridium sp. NSJ 42* (SGB6174 group), and *Ruminococcaceae bacterium* (SGB15249) undergo the maximum significant change in effect sizes across all phenotypes ($p<0.01$) when including the strain-phylogenetic effect (**Fig. 5C**).

On the other side, 35 phenotypes were significantly different after including the strain-phylogeneitc effect. Among them, the clustering coefficient and transitivity (which are social network properties)

underwent the maximum significant change ($p<0.0001$), and the relationship of antibiotics was also changed substantially (**Fig. 5D**).

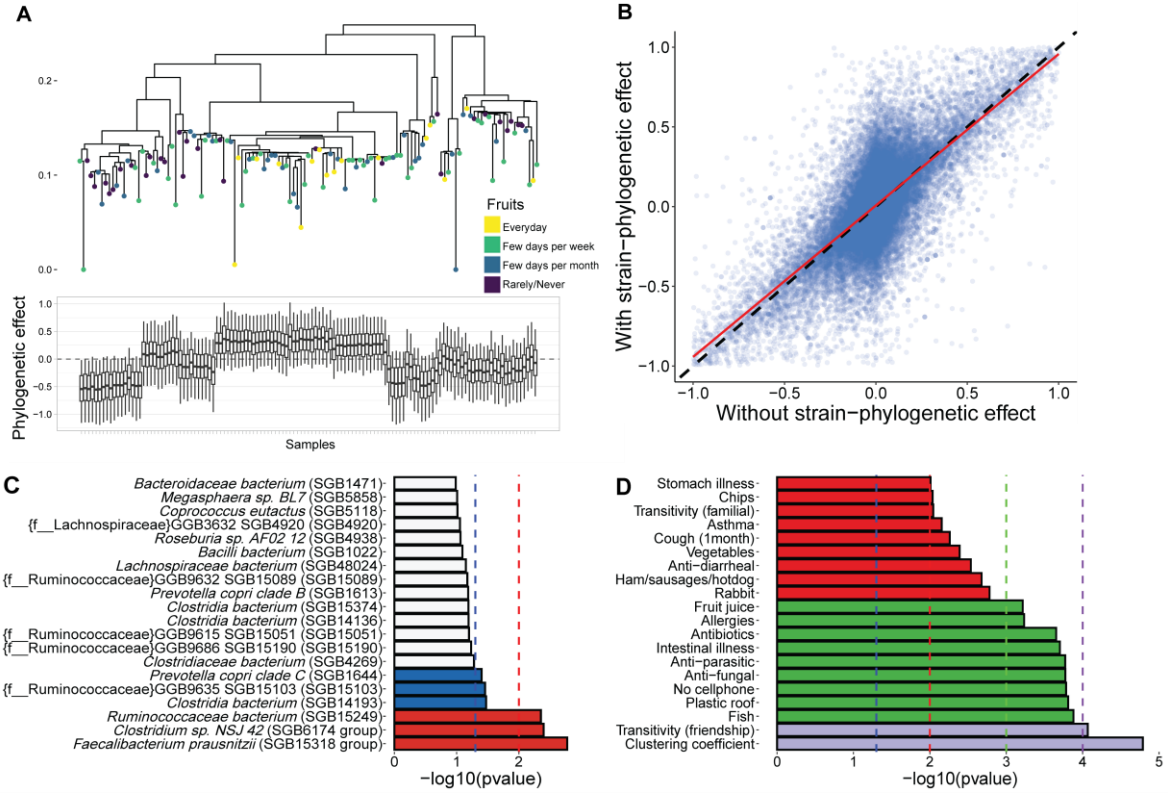


Fig. 5: Microbial strain association with host phenotypes: (A) Strain-level phylogeny of Clostridia uSGB14230 shows leaves that are colored with the frequency of fruit consumption (as an illustrative phenotype) in individual subjects. The strain-phylogenetic effect here is a direct consequence of the overall leaf structure. (B) Comparison of effect sizes in phenotype-species association plot in the presence and absence of strain-phylogenetic information (indicated by blue dots) across 83,916 species-phenotype relationships (see Methods). β-coefficients from the association models with and without including phylogenetic information are positively correlating (Spearman correlation coefficient $\rho=0.57$, $p < 2.2x10^{-16}$), and the red line is the linear fit ($\beta = 0.94882$, intercept $=0.007267$), showing the relationship between the two models. (C) List of species (across all phenotypes) which are most affected after including the strain-phylogenetic effect. Comparisons are generated using Wilcoxon ranked-sum test. Blue and red dashed lines indicate $p<0.05$* and $p<0.01$** respectively. (D) List of phenotypes (across all species) which are most affected after including the strain-phylogenetic effect. Comparisons are generated using Wilcoxon ranked-sum test. Blue, red, green, and purple dashed lines indicate $p<0.05$*, $p<0.01$**, $p<0.001$*** and $p<0.0001$**** respectively.

Integrated, standardized, large-population-based cohorts to study the microbiome are uncommon, but such studies offer the prospect of disentangling 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 developing world setting; by assessing previously uncharacterized taxa; and by using strain-level genomic information, our goal is to advance understanding of the possible relationship of the gut microbiome and a broad range of human phenotypes. Hence, compared to prior research, our work is distinctive in that we

analyze large, non-Westernized rural population, have a very broad range of phenotypes, and report strain-level analyses.

We find that variation in the gut microbiome across individuals living in this traditional way can nevertheless still be at least partly explained by variations in diet, lifestyle, environmental exposures, and health factors. Overall, we found 7,117 unique associations between 788 bacterial species and 126 phenotypes from health, environment, and socioeconomic categories. 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. After including strain-level phylogenies, certain phenotype categories (especially socioeconomic status) had a much higher effect size. Similarly, some species, such as *Faecalibacterium prausnitzii*, were also more affected by including strain-phylogenetic information, acquiring more associations with phenotypes.

Still, despite measuring a large number and variety of phenotypes, only 26.3% of the variation across individuals in microbiome composition was accounted for by these phenotypes, in keeping with prior studies.[1,33,34,35] 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 helps 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 we ascertained in Honduras did combine to account for 26.3% of the species variation (as noted) and 37.4% of the pathway variation; this may be compared to a recent study from the Netherlands were the measured phenotypes accounted for 13% and 16.2% of the variation, respectively,[1] although different methodologies for taxonomic and functional characterization were used, reflecting ongoing methodological advances.

The gut microbiome is known to be related to various health conditions both in humans and in mice,[36] and conditions like cancer, obesity, diabetes, autism, anxiety, and depression can induce shifts in the gut composition (as shown in many mostly Westernized populations).[1,36,37,38,39,40,41,42] Alcohol intake and cigarette use have been linked to gut microbiome dysbiosis, as have medications.[43,44,45,46,47] In keeping with these prior studies, we confirm such findings in this rural non-Western cohort.[3,4,5,48] Indeed, we found 2,255 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 572 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 328 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 36 taxa to be differentially enriched in one of the groups. Among medication users, those taking anti-parasitic medication had the largest drop in overall diversity. Differential abundance analysis revealed 134 significant species, with *uSGB2238* of the Rikenellaceae family showing an increased relative abundance in healthy people, and *Bacteroides ovatus* being more enriched in ill people. Among these groups, 7 of the top 10 differentially abundant species in healthy guts were unknown species in our sample from these isolated Honduras villages.

Another factor which greatly influences the gut microbiome is diet.[36] Our sample population exhibits a consistent diet, with beans and tortillas being consumed by most people on a daily basis. Still, we found 1,471 associations with food categories. A previously studied Dutch cohort found that pets had notable associations with the microbiome,[1] and we also found 617 associations with a broader range of animal exposures. Furthermore, our sample was spread across 11 villages separated in space and elevation, and the overall gut microbiome was observed to vary with relative spatial position within the villages; the dissimilarity score with a village-averaged microbiome increased as subjects lived further

away from the village center. The Dutch cohort also showed that rural residence and greenspace were associated with different microbiome profiles.[1]

Social and economic factors had 3,004 associations, with the bulk of the strong associations coming from unknown species. The gut microbiome had 2,280 unique associations with economic factors alone, making it the second highest associating category of variables we examined, after health. Prior research in Honduras has highlighted the crucial importance of socioeconomic status in addressing health in such communities.[49]

Social interaction is an integral part of Honduran villagers' life. In total, 616 unique associations with social network factors were found. Studies investigating social interactions between mice have shown the evolutionary advantage of having a behavior that enhances social interaction that consequently facilitates microbiome transmission. [36,50,51] In wild mice, social associations are predictive of microbiome composition, and the microbiome is correlated across mice interaction networks.[52] In humans, strain-level similarities have been shown in familial and partner networks within and outside households.[53,54] Whether these interactions translate into exposures that directly contribute to health is an important area for follow-up studies.

Uncharacterized taxa play a vital role in all these associations, as in prior non-Western cohorts.[6] Despite the number of unknown species in Honduran cohort being about a third of total distinguishable species, their relative strength of associations was observed to be higher in all the phenotype categories. Strain-level information is also relevant to the microbiome-phenotype relationship and should optimally be accounted for. For instance, after including this effect, animal exposure and economic phenotypes are the strongest factors associated with the gut microbiome overall.

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.[55,56] 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 can help advance individual and collective health.

Methods

Sample collection, library preparation, and sequencing:

Participants were instructed on how to self-collect the fecal samples using a training module and asked to return 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 Stated of America 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[57]) with default parameters. The resulting reads were screened for human contamination (hg19) with BMTagger[58] and then quality filtered with trimmomatic[59] (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 1,187 samples (with an average 8.68498 x 10^7 reads).

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 2,285 species in our dataset. Among the 2,285 species, 788 species were used for association analysis after filtering for minimum relative abundance values (10^{-3}), and a minimum of 15% prevalence in the population (i.e., >119 people).

We performed strain-level profiling for a subset of species present in at least 10% of the subjects (n=666) 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 Bray-Curtis dissimilarity (vegan vegdist function) calculated on the relative abundances obtained by MetaPhlAn4.

Functional potential analysis was performed using HUMAnN 3.0.[60] Gene family profiles were normalized using relative abundances and collapsed into MetaCyc pathways. In total, we found 1,180 pathways associated with at least one individual.

To understand the amount of variance explained by various factors, we performed a PERMANOVA analysis (adonis function from the vegan package[61]) using the "bray" method; the diversity matrix was calculated on both species-level relative abundances and MetaCYC pathway relative abundances as input, and including 126 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 according to CDC[62] and NBME[63] guidelines (Extended Data Fig. 3).

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).[64] The PHQ9 (Patient Health Questionnaire) score measuring depression was computed in a similar fashion, with the levels being: minimal or none (≤5), mild (6 - 10), moderate (11 - 15), moderately severe (16 - 19), and severe (≥20).[65]

Frequency of intake of various food items were 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)[24] was calculated from 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 consumed on a daily basis, 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 report 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 the least wealthy households to 5 indicating most wealthy households.

We explored associations with several social network features, including 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 comprehend 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.[66]

Model for microbiome-phenotype regression:

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

This mixed model was computed for every species and phenotype pair.

$$\text{Species abundance} \sim \text{age} + \text{sex} + \text{BMI} + \text{batch effect} + \text{bristol stool scale} + \text{DNA concentration} + \text{Sampling date} + 1|\text{village} + \text{phenotype}$$

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

Since technical factors (age, sex, DNA concentration, sequencing batch, sampling date) along with BMI and Bristol stool scale accounted for 2.8% of the species variation and 6.1% of the pathway variation, we used these variables as primary controls in our association models (Extended Data Fig. 4, Extended Data Fig. 9).

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:

For strain-level analysis, we used the ANPAN package. Using the leaf distance in the phylogenetic tree, a linear mixed model – namely Phylogenetic Generalized Linear Mixed Model (ANPAN package v 0.2.0) – was implemented to get associations between phenotypes and strains:

$$\text{Phenotype} \sim \text{age} + \text{sex} + \text{BMI} + \text{batch effect} + \text{bristol stool scale} + \text{DNA concentration} + \text{Sampling date} + \text{species(abundance)} + 1|\text{village} + 1|\text{leaf} + \varepsilon$$

In total, 83,916 associations were explored, coming from 666 species (which met strainphlan marker thresholds) and 126 phenotypes.

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 fixed effects 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 Rigoberto Matute Juarez, Jose Eduardo Gámez, and Eduardo Jose Urrea Carbajal for coordinating the field work; 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.

Author contributions:

Conceptualization: SVS, FB, MA, IB, and NAC; Methodology: SVS, FB, MA, IB, and NAC; Data Collection: SVS, FB, MA, 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.

References:

[1] Gacesa, R., Kurilshikov, A., Vich Vila, A. *et al.* Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732–739 (2022).

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

[3] Mohanan, M. *et al.* The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural India. *JAMA Pediatrics.* **169**, 349–357 (2015)

[4] Young, B. N. *et al.* 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] Hartley, D. Rural health disparities, population health, and rural culture. *American Journal of Public Health* **94**, 1675–1678 (2004)

[6] Pasolli, E. *et al.* 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] Shakya, H. B., Stafford, D., Hughes, D. A. & Keegan, T. 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* (2017)

[8] Blanco-Miguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. *bioRxiv* 2022.08.22.504593 (2022) doi:10.1101/2022.08.22.504593

[9] Domènech, L. *et al.* Changes in the stool and oropharyngeal microbiome in obsessive-compulsive disorder. *Scientific Reports* **12**, 1448 (2022)

[10] Saitoh, S. *et al.* Bacteroides ovatus as the predominant commensal intestinal microbe causing a systemic antibody response in inflammatory bowel disease. *Clinical and Diagnostic Laboratory Immunology* **9**, 54–59 (2002)

[11] Wei, B. *et al.* Molecular cloning of a Bacteroides caccae TonB-linked outer membrane protein identified by an inflammatory bowel disease marker antibody. *Infection and Immunity* **69**, 6044–6054 (2001)

[12] Desai, M. S. *et al.* A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell* **167**, 1339–1353.e21 (2016)

[13] Wang, K. *et al.* Parabacteroides distasonis Alleviates Obesity and Metabolic Dysfunctions via Production of Succinate and Secondary Bile Acids. *Cell Rep.* **26**, 222–235.e5 (2019)

[14] Parker, B. J., Wearsch, P. A., Veloo, A. C. M. & Rodriguez-Palacios, A. The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health. *Frontiers in Immunology* **11**, 906 (2020)

[15] Wu, I.W. *et al.* Integrative metagenomic and metabolomic analyses reveal severity-specific signatures of gut microbiota in chronic kidney disease. *Theranostics* **10**, 5398–5411 (2020)

[16] Balakumar, M. *et al.* Improvement in glucose tolerance and insulin sensitivity by probiotic strains of Indian gut origin in high-fat diet-fed C57BL/6J mice. *European Journal of Nutrition* **57**, 279–295 (2018)

[17] Baxter, N. T. *et al.* Dynamics of Human Gut Microbiota and Short-Chain Fatty Acids in Response to Dietary Interventions with Three Fermentable Fibers. *MBio* **10**, (2019)

[18] Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine* **8**, 343ra82 (2016)

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

[20] Bonaccio, M. *et al.* 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)

[21] Callon, C., Arliguie, C. & Montel, M.-C. Control of Shigatoxin-producing Escherichia coli in cheese by dairy bacterial strains. *Food Microbiology* **53**, 63–70 (2016)

[22] Speranza, B., Liso, A., Russo, V. & Corbo, M. R. Evaluation of the potential of biofilm formation of Bifidobacterium longum subsp. infantis and Lactobacillus reuteri as competitive biocontrol agents against pathogenic and food spoilage bacteria. *Microorganisms* **8**, 177 (2020)

[23] Schiemann, D. A. Occurrence of Klebsiella pneumoniae in dairy products. *J. Milk Food Technol.* **39**, 467–469 (1976)

[24] Krebs-Smith, S. M. *The Effects of Variety in Food Choices on Dietary Quality: A Thesis in Nutrition*. (Pennsylvania State University, 1985).

[25] Xiao, C. *et al.* Associations of dietary diversity with the gut microbiome, fecal metabolites, and host metabolism: results from 2 prospective Chinese cohorts. *The American Journal of Clinical Nutrition* **116**, 1049–1058 (2022)

[26] Darmon, N. & Drewnowski, A. Contribution of food prices and diet cost to socioeconomic disparities in diet quality and health: a systematic review and analysis. *Nutrition Reviews* **73**, 643–660 (2015)

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

[28] McFeters, G. A., Bissonnette, G. K., Jezeski, J. J., Thomson, C. A. & Stuart, D. G. Comparative survival of indicator bacteria and enteric pathogens in well water. *Applied Microbiology* **27**, 823–829 (1974)

[29] Sofi, M. H. *et al.* pH of drinking water influences the composition of gut microbiome and type 1 diabetes incidence. *Diabetes* **63**, 632–644 (2014)

[30] Littleford-Colquhoun, B. L., Weyrich, L. S., Kent, N. & Frere, C. H. City life alters the gut microbiome and stable isotope profiling of the eastern water dragon (Intellagama lesueurii). *Mol. Ecol.* **28**, 4592–4607 (2019)

[31] Guo, P., Zhang, K., Ma, X. & He, P. Clostridium species as probiotics: potentials and challenges. *Journal of Animal Science and Biotechnology* **11**, 24 (2020)

[32] Bader, J., Albin, A. & Stahl, U. Spore-forming bacteria and their utilisation as probiotics. *Beneficial Microbes* **3**, 67–75 (2012)

[33] Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016)

[34] Blacher, E. *et al.* Potential roles of gut microbiome and metabolites in modulating ALS in mice. *Nature* **572**, 474–480 (2019)

[35] Manor, O. *et al.* Health and disease markers correlate with gut microbiome composition across thousands of people. *Nature Communications* **11**, 1–12 (2020)

[36] Cryan, J. F. *et al.* The Microbiota-Gut-Brain Axis. *Physiological Reviews* **99**, 1877–2013 (2019)

[37] Patterson, E. *et al.* Gut microbiota, obesity and diabetes. *Postgraduate Medical Journal* **92**, 286–300 (2016)

[38] Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2008)

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

[40] Murri, M. *et al.* Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study. *BMC Medicine* **11**, 46 (2013)

[41] Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012)

[42] Correction for Hu et al., Microbiota-induced activation of epithelial IL-6 signaling links inflammasome-driven inflammation with transmissible cancer. *Proceedings of the Natural Academy of Sciences* **110**, 12852–12852 (2013)

[43] Sarkola, T., Iles, M. R., Kohlenberg-Mueller, K. & Eriksson, C. J. P. Ethanol, acetaldehyde, acetate, and lactate levels after alcohol intake in white men and women: effect of 4-methylpyrazole. *Alcoholism: Clinical and Experimental Research* **26**, 239–245 (2002)

[44] Allais, L. *et al.* Chronic cigarette smoke exposure induces microbial and inflammatory shifts and mucin changes in the murine gut. *Environmental Microbiology* **18**, 1352–1363 (2016)

[45] Botschuijver, S. *et al.* Intestinal Fungal Dysbiosis Is Associated With Visceral Hypersensitivity in Patients With Irritable Bowel Syndrome and Rats. *Gastroenterology* **153**, 1026–1039 (2017)

[46] Caputi, V. *et al.* Antibiotic-induced dysbiosis of the microbiota impairs gut neuromuscular function in juvenile mice. *British Journal of Pharmacology* **174**, 3623–3639 (2017)

[47] Fröhlich, E. E. *et al.* Cognitive impairment by antibiotic-induced gut dysbiosis: Analysis of gut microbiota-brain communication. *Brain, Behavior, and Immunity* **56**, 140–155 (2016)

[48] Kim, D. A. *et al.* Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *Lancet* **386**, 145–153 (2015)

[49] Arps, S. Socioeconomic status and body size among women in Honduran Miskito communities. *Annals of Human Biology* **38**, 508–519 (2011)

[50] Rosenberg, E., Sharon, G. & Zilber-Rosenberg, I. The hologenome theory of evolution contains Lamarckian aspects within a Darwinian framework. *Environmental Microbiology* **11**, 2959–2962 (2009)

[51] Carlson, S. J. *et al.* 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. *Frontiers in Neuroscience* **13**, 72 (2019)

[52] Raulo, A. *et al.* Social networks strongly predict the gut microbiota of wild mice. *ISME Journal* **15**, 2601–2613 (2021)

[53] Brito, I. L. *et al.* Transmission of human-associated microbiota along family and social networks. *Nature Microbiology* **4**, 964–971 (2019)

[54] Valles-Colomer, M., Blanco-Míguez, A., Manghi, P. *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).

[55] Xiao, L. *et.al*. Large-scale microbiome data integration enables robust biomarker identification. *Nature Computational Science* **2**, 307-316 (2022)

[56] Wynford-Thomas, R., & Robertson, N. P. The economic burden of chronic neurological disease. *Journal of Neurology*. **264**, 2345-2347(2017).

[57] Cantu, V. A., Sadural, J. & Edwards, R. *PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets*. https://peerj.com/preprints/27553/ (2019) doi:10.7287/peerj.preprints.27553v1

[58] BMTagger. ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/.

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

[60] Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, (2021).

[61] Oksanen, J. *et al.* The vegan package. *Community ecology package* **10**, (2008).

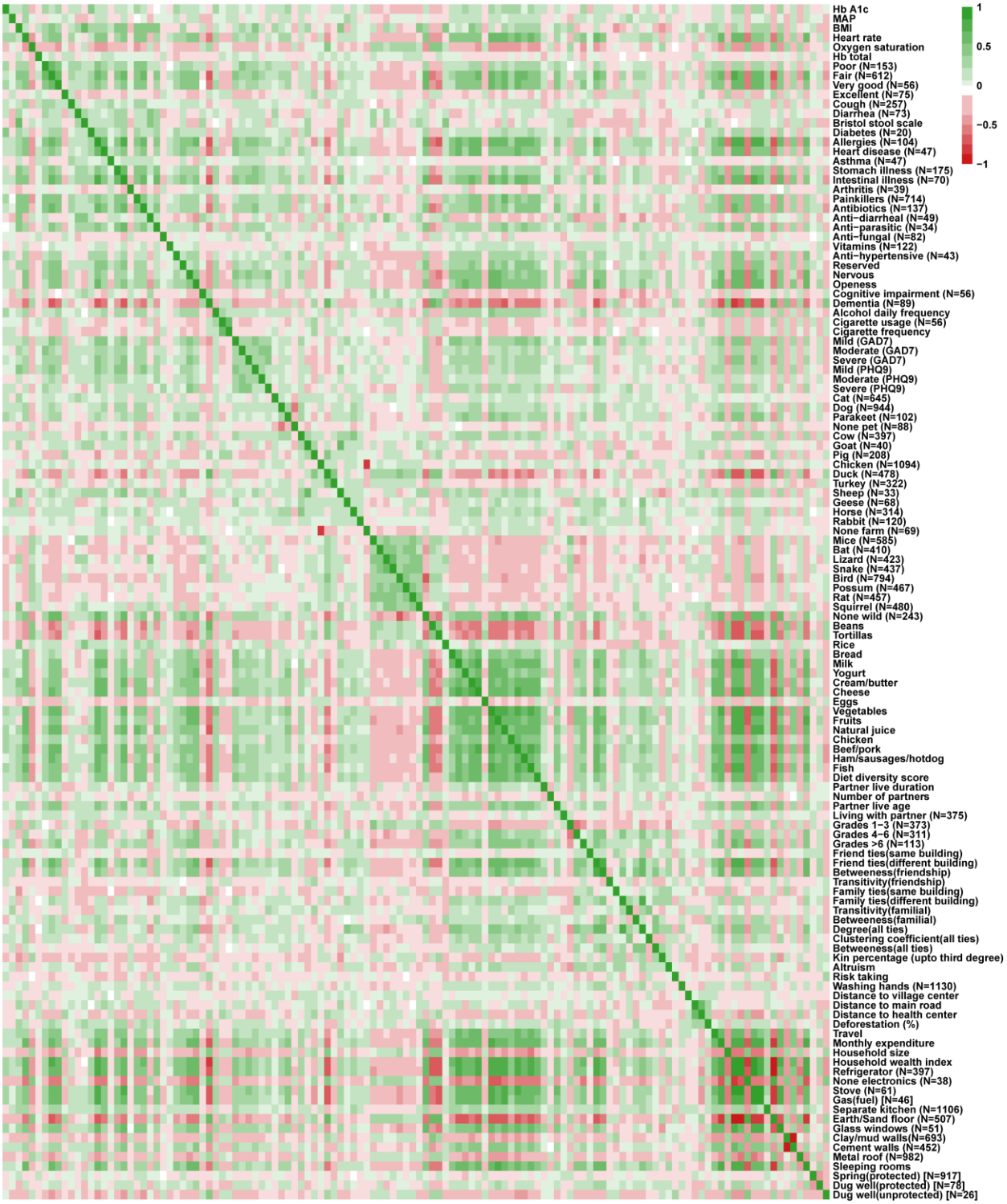
[62] CDC official website: https://www.cdc.gov/

[63] NBME reference values: https://www.nbme.org/sites/default/files/2020-07/Laboratory_Reference_Values.pdf

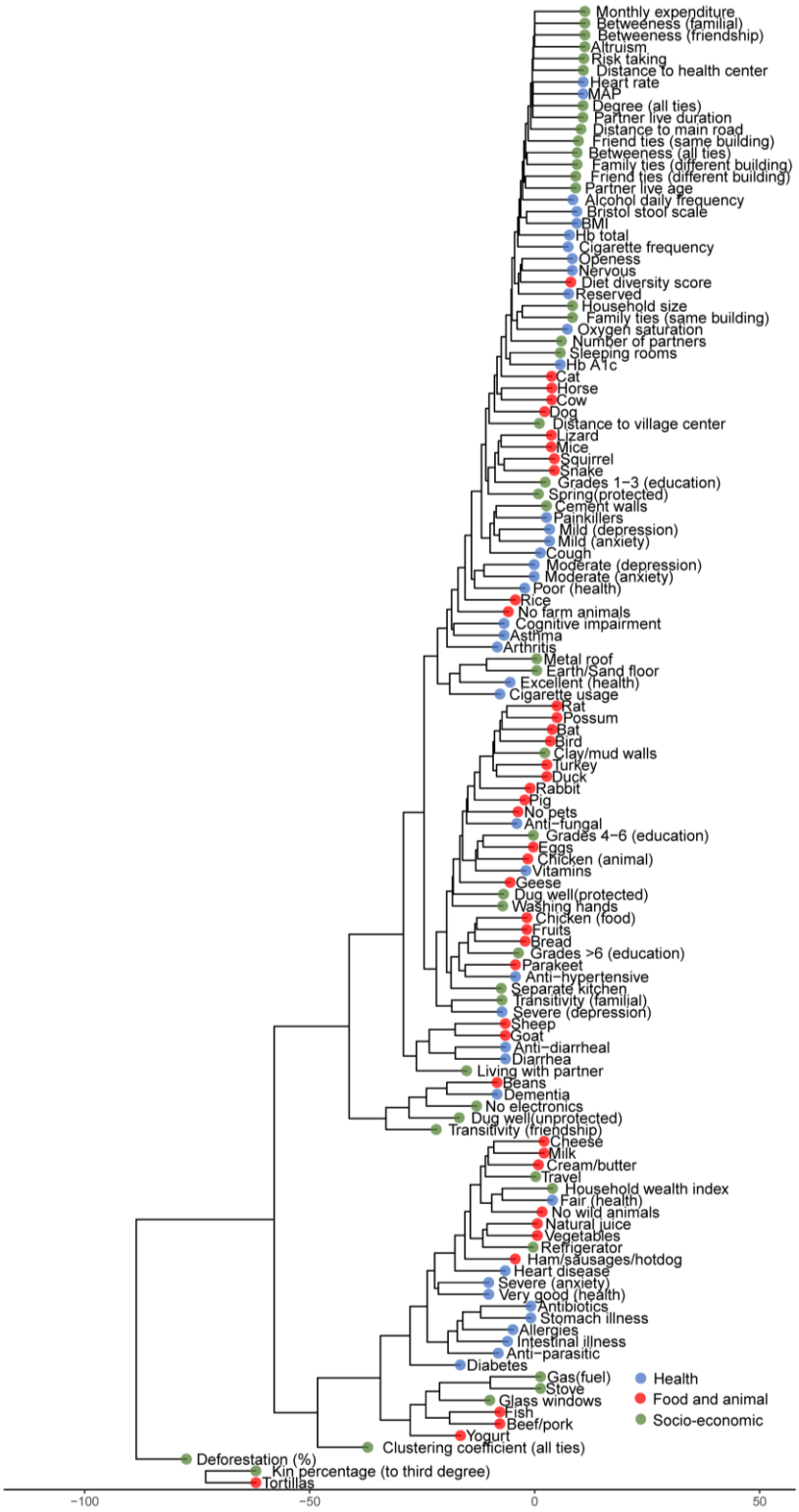
[64] Williams, N. The GAD-7 questionnaire. *Occupational Medicine* **64**, 224–224 (2014)

[65] Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* **16**, 606–613 (2001)

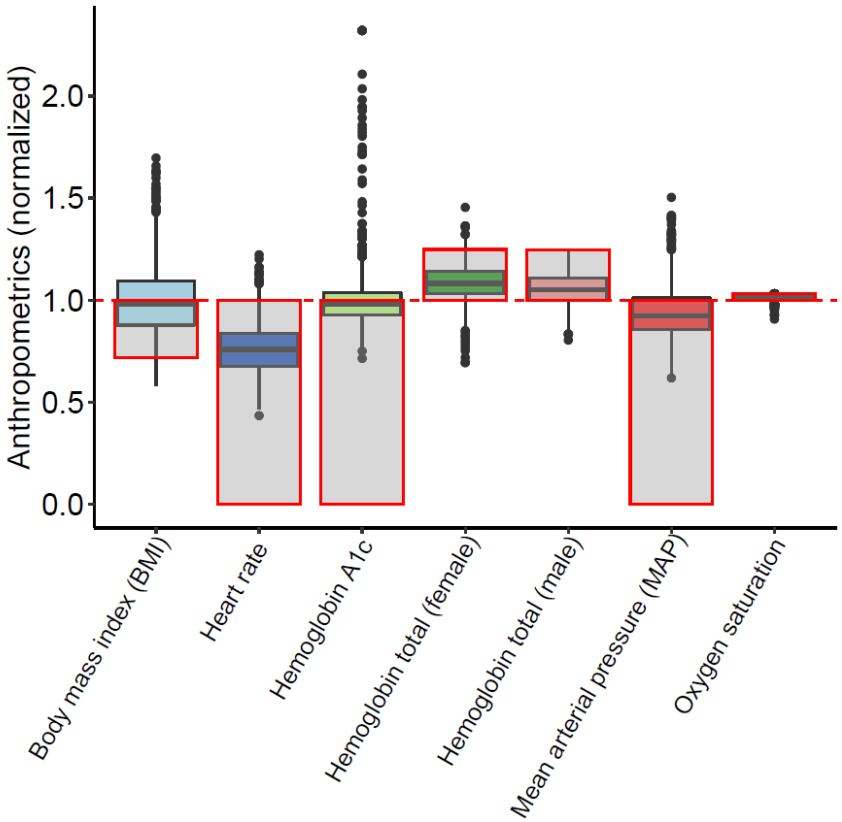
[66] Kahle D, Wickham H (2013). "ggmap: Spatial Visualization with ggplot2." *The R Journal*, **5**(1), 144–161



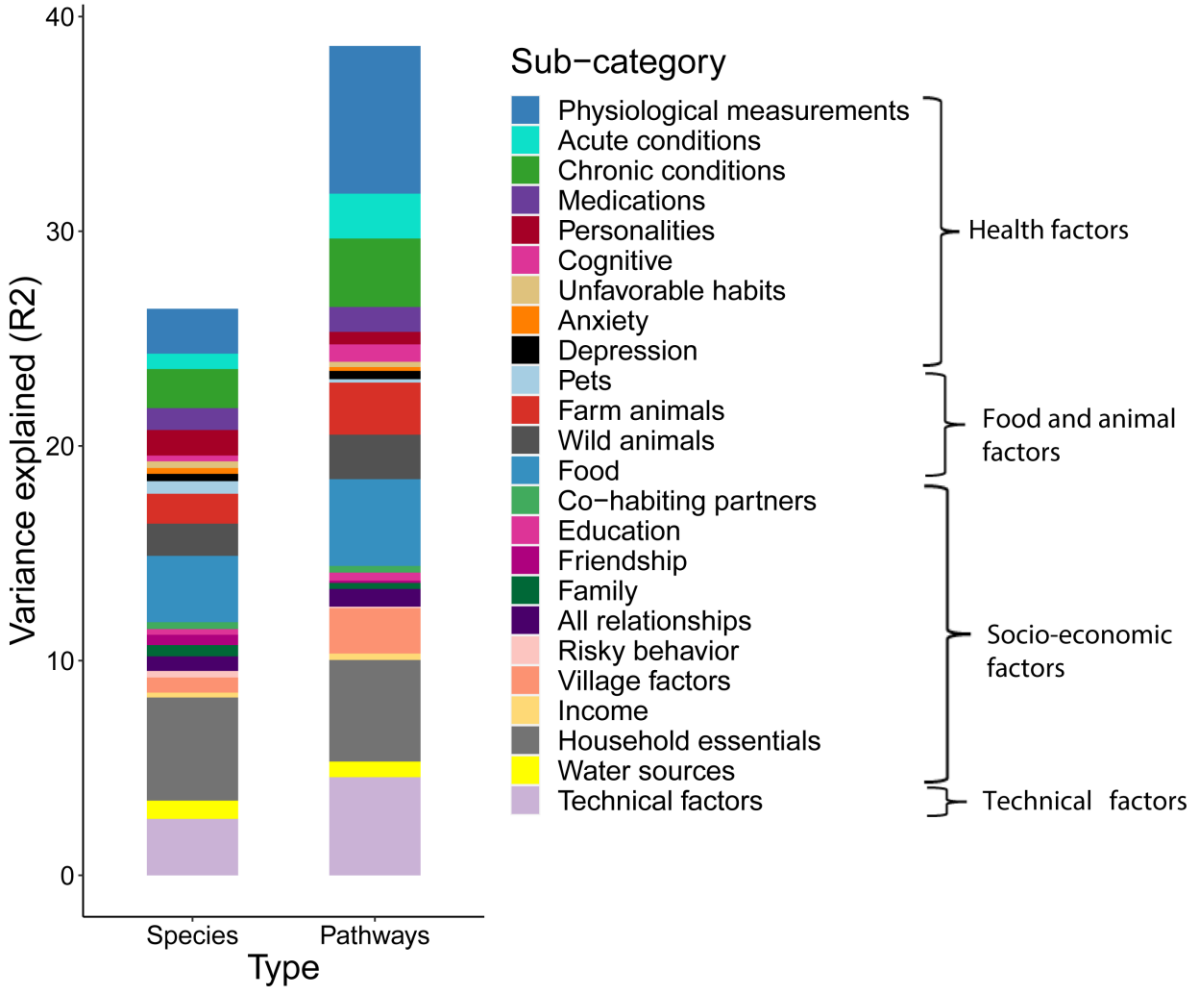
Extended Data Fig. 1 (Phenotype-phenotype correlation): A matrix showing raw correlations between the phenotypes from every category (health, food and animals, socio-economic factors). Column names are same as the rownames indicated on the right side of the matrix. Color ranges from positive (green) to negative (red) correlations.



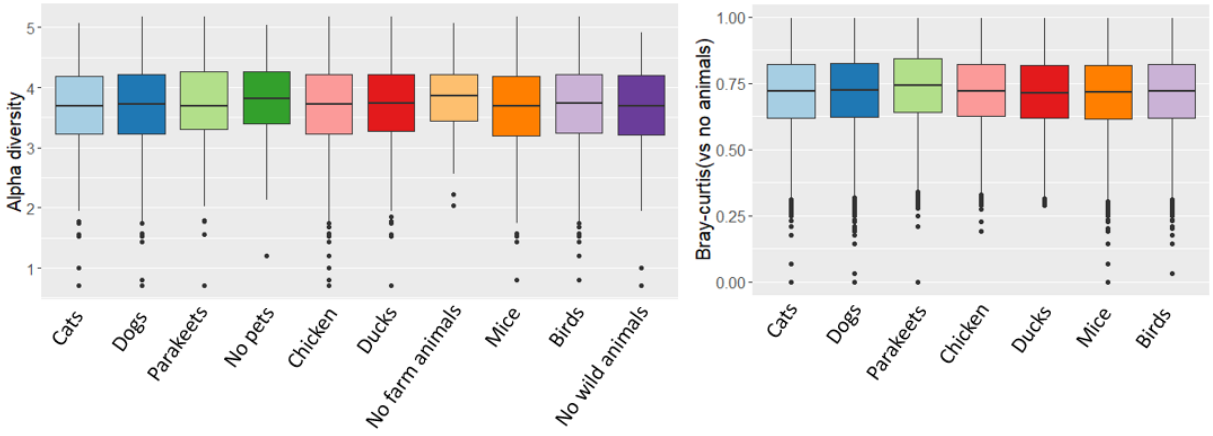
Extended Data Fig. 2 (Phenotype-microbiome association clustering): Effect sizes from associations of all 126 phenotypes with 788 species are hierarchically clustered with respect to phenotypes. This phenotype tree is another representation of how similarly behaving a pair of phenotypes are with respect to how they associate with the gut microbiome overall.



Extended Data Fig. 3 (Health measurements against standardized values): Graphical visualization of physiological measurements (anthropometrics) of all N=1,187 villagers, with the grey box indicating normal values of each respective physiological measurement. The red box indicates the bounding limit of healthy ranges.

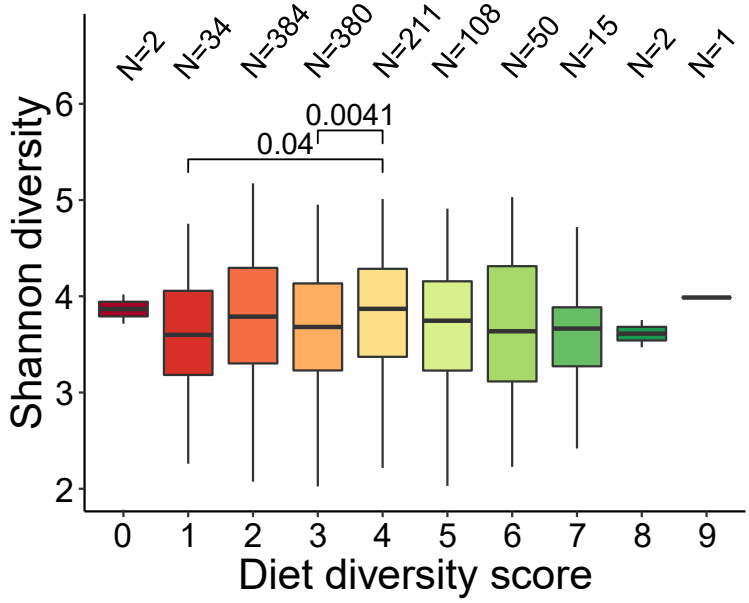


Extended Data Fig. 4 (Variance explained): PERMANOVA analysis (999 permutations, p-value<0.001) computed on all phenotypes shows the variance explained in species and pathway compositions with a breakdown of sub-categories of all phenotypes (health, food and animal, socioeconomic factors). Technical factors include age, sex, DNA concentration, sequencing batch, and sampling date. (See Supplementary Table 7 for complete breakdown of variance explained in each sub-categories)

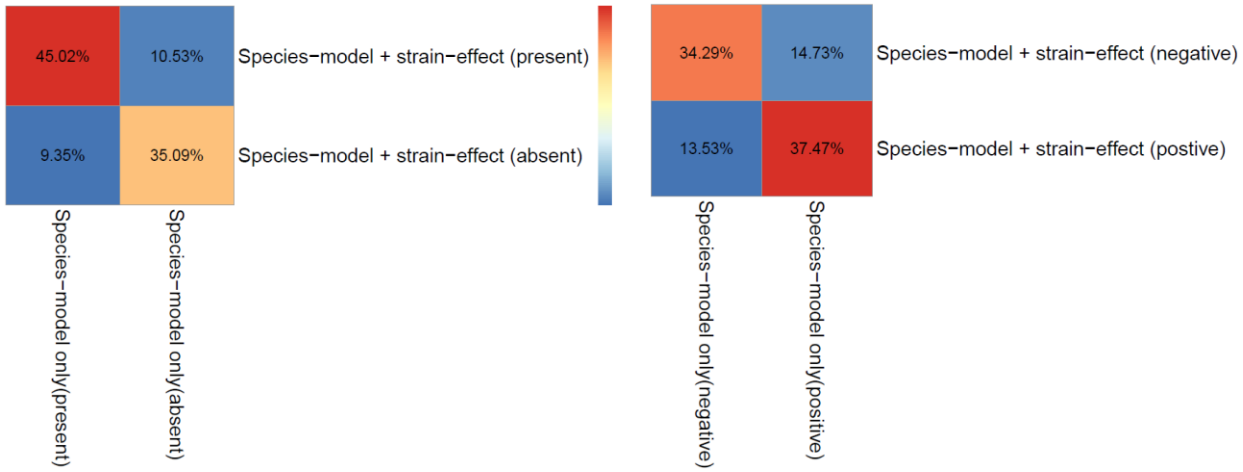


(A) (B)

Extended Data Fig. 5: (A) Alpha diversity among villagers expose to animals. (B) Pair-wise Bray-Curtis dissimilarity between villagers who are exposed to animals compared to unexposed villagers.

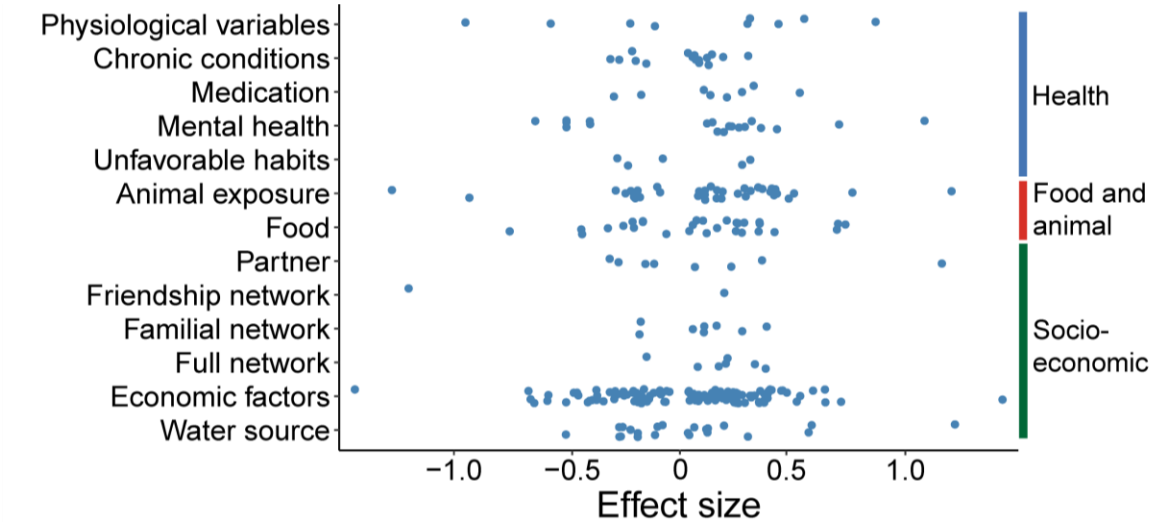


Extended Data Fig. 6 (Diet diversity score): Plot showing the Shannon diversities of individuals with varying diet diversity scores. Individuals belonging to DDS score of 4 are significantly different from those in group 1 and 3. All comparisons are performed using Wilcoxon ranked sum test.

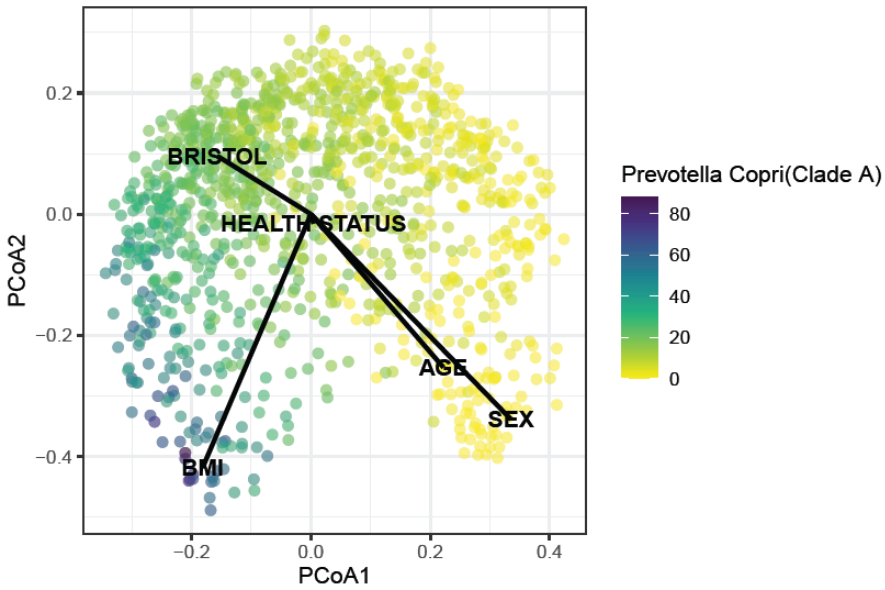


(A) (B)

Extended Data Fig. 7: (A) Side-by-side comparison of significant associations (FDR<0.05) in both models (with and without strain-phylogenies). Each quadrant indicates presence and absence of associations in either model. (B) Figure showing direction/sign flipping of common significant associations in both models.



Extended Data Fig. 8 (Species-phenotype association with strain effect) Effect size distribution of 302 significant associations (FDR<0.05) with all phenotypes (shown as categories and sub-categories) with microbiome species, after incorporating strain-level phylogenetic information. Each dot represents a species with the corresponding phenotype category. In general, it is apparent that socio-economic factors have a lot more associations after including the strain effect.



Extended Data Fig. 9: Principal Coordinates Analysis (PCoA) of the Bray-Curtis dissimilarity computed using the species-level relative abundances (legend) generated by MetaPhlAn4. Health status, age, sex, body mass index (BMI), and Bristol stool scale are shown as arrows along with the direction of influence. Samples are colored with the relative abundances of *Prevotella copri (clade A)*.