Variation in rural African gut microbiomes is strongly shaped by parasitism and diet

Elise Morton¹, Joshua Lynch¹, Alain Froment², Sophie Lafosse², Evelyne Heyer², Molly Przeworski³, Ran Blekhman#¹ and Laure Ségurel#²

# Co-supervised the work

1) Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN, 55455
2) Eco-anthropology and ethnobiology, UMR 7206, CNRS-MNHN-University Paris 7 Diderot
3) Department of Biological Sciences, Columbia University, New York, NY 10027

Correspondence and requests for materials should be addressed to E.M. (email: emorton@umn.edu), R.B. (email: blekhman@umn.edu), or L.S. (email: lsegurel@mnhn.fr).

Abstract

The human gut microbiome is influenced by its host’s nutrition and health status, and represents an interesting adaptive phenotype under the influence of metabolic and immune constraints. Previous studies contrasting rural populations in developing countries to urban industrialized ones have shown that geography is an important factor associated with the gut microbiome; however, studies have yet to disentangle the effects of factors such as climate, diet, host genetics, hygiene and parasitism. Here, we focus on fine-scale comparisons of African rural populations in order to (i) contrast the gut microbiomes of populations that inhabit similar environments but have different traditional subsistence modes and (ii) evaluate the effect of parasitism on microbiome composition and structure. We sampled rural Pygmy hunter-gatherers

|
as well as Bantu individuals from both farming and fishing populations in Southwest Cameroon and found that the presence of *Entamoeba* is strongly correlated with microbial composition and diversity. Using a random forest classifier model, we showed that an individual’s infection status can be predicted with 79% accuracy based on his/her gut microbiome composition. We identified multiple taxa that differ significantly in frequency between infected and uninfected individuals, and found that alpha diversity is significantly higher in infected individuals, while beta-diversity is reduced. Another factor associated with microbial composition in our data is subsistence mode, notably with some taxa previously shown to differ between Hadza Eastern African hunter-gatherers and Italians also discriminating Pygmy hunter-gatherers from Cameroon from neighboring farming or fishing populations. In conclusion, our results stress the importance of taking into account an individual’s parasitism status in studies of the microbiome, and highlight how sensitive the microbial ecosystem is to subtle changes in host nutrition. Our fine-scale analysis allowed us to identify microbial features that are specific to hunter-gatherers versus ones shared by all rural African populations, increasing our understanding of the influence of subsistence mode and lifestyle on gut microbiome composition.

Introduction

The human gut microbiota, the community of microorganisms inhabiting our gastrointestinal tract, is involved in a number of metabolic and immune functions and is now considered to be an essential determinant of human health\(^1\)\(^-\)\(^3\). Notably, these microbes enable an individual to recover key nutrients from food, and can prevent host infection by opportunistic pathogens\(^4\). Despite considerable plasticity in the structure and composition of an individual’s gut microbiota\(^5\), there are significant correlations between characteristics of the microbiome, dietary composition, host genotype, and patterns of disease\(^3\)\(^,\)\(^6\)\(^,\)\(^7\). These relationships suggest that composition of the microbiome represents a potentially adaptive phenotype that can be targeted by natural selection, especially in regard to adaptation to diet and exposure to pathogens.
Since the Neolithic revolution about 10,000 years ago, human populations have started to diversify their dietary regimes, resulting in the contrasted subsistence modes known today. This major cultural transition has also created new pathogenic pressures due to the proximity of livestock and the increased density of populations. Cultural and environmental differences among populations have resulted in physiological adaptations that can be detected in our genome. Additionally, these factors have likely affected the community dynamics of our gut microbial ecosystem. Dietary changes have been shown to facilitate rapid changes in gut microbiota; however, the roles of habituation (over a lifetime) versus host adaptation (across generations) in these broader patterns are unclear. Understanding the long-term interaction that took place between the dietary specialization of populations and their gut microbiomes is therefore of great interest, notably to understand and predict the effect of recent and rapid changes in lifestyle and food on human health.

Nevertheless, to date, microbiome studies have mostly focused on industrialized populations. The few studies that have included a more diverse array of populations contrasted urban populations in highly industrialized countries to populations in developing countries and found that, outside of age, the factor that most strongly influences the gut microbiome is geography. Such designs do not allow the respective influences of factors such as diet, climate, hygiene, parasitism and host genetics on population differences in the microbiome to be disentangled. While some specific changes in microbial communities have clearly been linked to components of human dietary regimes and the use of antibiotics, the effect of other environmental or host-related factors is not clear. Notably, we do not know the extent to which the observed loss of microbial diversity of the human gut microbiome in urban industrialized populations is attributable to their dietary specialization, changes in sanitation/hygiene practices, or other
factors. This loss of microbial biodiversity is a public health concern, as it may reflect a perturbed ecosystem associated with multiple diseases\textsuperscript{19,20}.

In addition to the loss of microbial diversity, developed countries nearly ubiquitously present a marked decrease in the prevalence of human gut parasites\textsuperscript{21}. Although it is estimated that 3.5 billion people worldwide are infected with some parasite (protozoan or helminth)\textsuperscript{22}, little is known about their role in shaping the gut microbiome. Yet, throughout evolution, gut microbes and gut-dwelling parasites have co-inhabited the human gastrointestinal tract\textsuperscript{23}, and community dynamics are likely determined by current and past interactions (both during an individual's lifespan and throughout evolutionary history) between microbiota, protozoa, helminths, and the host immune response\textsuperscript{24}. For example, it has been shown that direct competition by commensal microbes can provide protection from invading pathogens, and a disturbance to the natural microflora can effectively result in increased susceptibility to pathogens and/or parasites\textsuperscript{25,26}. There is also substantial evidence that these interactions are essential for the development of a healthy immune system, and that the underlying cause of the increased incidence of autoimmune disorders in industrialized countries is the absence of exposure to pathogens and parasites early in life (the "hygiene hypothesis")\textsuperscript{44}. In this context, it is important to evaluate the potential role of parasitism in shaping gut microbiome composition and structure.

Here, we focus on fine-scale comparisons of African rural populations with contrasting modes of subsistence and unequal access to medicine. Our objective is to understand the influence of current and ancestral diet, host genetics, and parasitism on the composition of the human gut microbiome. We focus on populations from Cameroon for which a large diversity of subsistence modes coexist in a restricted geographical area. We include individuals from Pygmy foraging (hunter-gatherer) populations, Bantu farming populations and Bantu fishing populations, all living in a similar rural environment. These populations are almost entirely self-sufficient in food;
their primary source of energy comes from cassava (*Manihot esculenta*), and fish or meat provides the main source of protein. Animal food production for these populations has been estimated to be high compared to elsewhere in Cameroon or Africa\(^2^7\). To account for recent changes in diet, we evaluated current dietary regimes using dietary surveys. We also assessed parasitism status by direct observations of fecal samples under the microscope. The focus on populations living in the tropical rainforest is complementary to previous African populations sampled: the Hadza hunter-gatherers and a population from Burkina Faso, living in the East and West African tropical savanna, respectively\(^1^1,1^4\); and a population from Malawi living in a relatively dry subtropical area of East Africa\(^1^2\). To the best of our knowledge, this study represents the first comparison of the gut microbiome of human populations with limited geographic separation and contrasting subsistence modes, as well as the first characterization of the relationship between microbial communities and intestinal parasites.

### Results

**Description of the samples: host genetics, diet and parasites**

**Host genetics**

We analyzed 64 individuals in seven different villages in Southwest Cameroon (average age 50 years, ranging from 26 to 78 years) corresponding to 20 hunter-gatherers, 24 farmers and 20 individuals from a fishing population (see Fig. 1 and Supplementary Table 1). The Pygmy hunter-gatherers diverged from the other Bantu populations about 60,000 years ago\(^2^8,2^9\) and the farming subsistence mode likely started over the last 5,000 years\(^3^0\). The sampled populations therefore not only have contrasted subsistence modes, but also have different genetic backgrounds.
Diet

We chose these populations because previous work done in 1984-1985, based on nutritional questionnaires and isotopes analyses, showed they had distinct diets. We performed new nutritional frequency surveys to assess how diet had changed during the past 30 years (see Supplementary Table 1). Interestingly, the amount of meat in the hunter-gatherers’ diet has substantially decreased, reflecting the lower abundance of wild game in the forest reserve and the hunting ban applied for some species. In contrast, the consumption of fish has increased in inland populations (especially in farmers), due to the construction of new roads connecting the coastal and inland populations. Similar to the results from 1984-1985, the farmers eat less starchy foods (cassava) than hunter-gatherers and individuals from the fishing population (Wilcox Rank Sum test: p = 0.005 and 0.017, respectively). A principal component analysis on all dietary components revealed roughly three clusters corresponding to the three dietary regimes, with the first axis distinguishing hunter-gatherers from the others, and the second axis separating the farming and fishing populations (see Fig. 1b). The one exception to this pattern concerns farmers from the North (living along the same road as the hunter-gatherers), which cluster with the hunter-gatherers. In the following analyses, we therefore consider this population separately from the farmers living in the South.

BMI

In addition to dietary questionnaires, we assessed the nutritional status of individuals by measuring their BMI (Body Mass Index) (see Supplementary Table 1). Twenty percent of the Pygmy hunter-gatherers were underweight (BMI<18) whereas 12%, 0%, and 4% of the South farmers, North farmers, and individuals from the fishing population were, respectively. Conversely, 0% of hunter-gatherers were overweight (BMI>25) while 12%, 14% and 26% of individuals in the other groups were, respectively. This likely reflects the difference in socio-economical status and access to medicine between these populations.
Parasitism

We assessed the intestinal parasitism of individuals by direct observation of their fecal samples under the microscope. In more than one individual, we detected the presence of *Entamoeba*, as well as eggs of *Ascaris*, *Trichuris*, and *Ancylostoma* (see Supplementary Fig. 1 and Supplementary Table 1). Overall, 89% of hunter-gatherers, 76% of farmers from the South, 100% of farmers from the North, and 58% of individuals from the fishing population were infected by at least one of these parasites. Regarding *Entamoeba*, 37%, 41%, 57% and 16% of individuals were infected in each population, respectively. The reduced rate of parasitism in the fishing population most likely reflects their higher level of hygiene and increased access to medicine.

Characterization of microbiome composition

The fecal microbiota of 69 samples (including 5 biological replicates) were characterized by sequencing of the V5-V6 region of the bacterial 16S ribosomal RNA with the Illumina MiSeq technology. We obtained a total of 12.65 million high-quality reads, resulting in an average of 175,784 reads per sample (+/- 72,822). The average percent of mapped reads per individual was 83% (SD = 7.5%) and did not vary significantly between populations (Welch’s t-test, p > 0.2). The dataset was then rarefied to 50,000 reads/sample (see Supplementary Fig. 2), and reads were clustered into 5039 operational taxonomic units (OTU) at 97% identity.

The five biological replicates (sampling of the same individual few days apart, see Supplementary Table 1) allowed us to compare the microbial differences within individuals to those between individuals. We calculated the UniFrac distance, a phylogenetic based distance metric, which when weighted, accounts for relative abundance of taxa. Because both weighted and unweighted metrics capture different aspects of microbial diversity, we included both...
types of analyses in the manuscript. We found that the average UniFrac distance between replicates of the same individual was lower than between individuals (although only statistically significant for the unweighted distances, one-sided Wilcoxon Rank Sum test: p = 0.003, see Supplementary Fig. 3).

To test which variables were significantly associated with overall microbiome composition, we performed a PERMANOVA analysis on the microbial abundance data. Intergroup comparisons for various variables revealed that infection by the amoeba parasite *Entamoeba*, location, subsistence mode and ancestry (the latter three being nested) were significantly associated with variation in microbiome composition (p = 0.0001, 0.01, 0.003 and 0.01, respectively; Supplementary Table 2). To further characterize patterns of variation that account for phylogenetic relationships of community taxa, we also performed a PERMANOVA analysis on both weighted and unweighted UniFrac distance matrices. Congruent with our previous analysis, we found that infection by *Entamoeba* was the most significant variable for both weighted and unweighted UniFrac distances (p = 0.007 and 0.0001, respectively; Supplementary Table 2). *Entamoeba* infection also provided clear separation along the primary axis of variation of the multidimensional scaling plots (Fig. 2a and Supplemental Fig. 4a). Subsistence and location were both determined to be highly significant based on unweighted UniFrac distances (p = 0.0003 and p = 0.002, respectively), but not weighted (p = 0.14 and p = 0.29, respectively). Because unweighted UniFrac distances assign increased value to rare taxa, this suggests that less abundant taxa are more important in describing differences between the microbiomes across subsistence modes and locations. Furthermore, subsistence provided only weak visual separation along the first two axes of variation for both metrics (Supplementary Fig. 4b-c).
Influence of parasitism on the microbiome

Because of the significant relationship between *Entamoeba* infection status and patterns of variation in the gut microbial communities found in all populations, we further investigated the relationship between infection by this parasite and composition of the microbiome (Fig. 2). As it is difficult to distinguish between the opportunistic pathogenic species (*E. histolytica*) and the strict commensal (*E. dispar*) by microscopy alone, we were unable to characterize this parasite at the species level. However, fewer than 10% of infected individuals were suffering from diarrhea, suggesting that they were not experiencing symptomatic amebiasis.

At the phylum level, we found that 8 of the 13 phyla represented are significantly different between *Entamoeba* infected (*Ent+*) and uninfected (*Ent−*) individuals, with most phyla (except Bacteroidetes and Lentisphaerae) occurring at a higher relative abundance in *Ent+* individuals (see Table 1). When looking at individual taxa, based on a linear regression model, we also identified a number of notable differences between *Ent+* and *Ent−* individuals (Fig. 2b-c, Supplementary Table 3-4), and we found that eighteen of the 93 most abundant taxa (present at ≥ 0.1% in at least 4 individuals) differed significantly in their relative abundance between *Ent+* and *Ent−* individuals (q < 0.05).

These taxonomic signatures for *Entamoeba* infection are so strong that an individual's infection status can be predicted with 79% accuracy using a Random Forests Classifier (RFC) model based on gut microbiome composition (p < 0.001; See Supplementary Fig. 5). Of the ten taxa identified as being the most important in their predictive power, all but *Prevotella stercorea* were significant in our linear regression model (of which all are in higher abundance in *Ent+* individuals except *Prevotella copri*). The reason for the association between *Entamoeba* and these microbes have yet to be identified, but it is noteworthy that the two most important taxa identified in the RFC model, Elusimicrobiaceae unclassified (uncl) and Ruminococcaceae uncl,
include established endosymbionts of protists and common inhabitants of the termite gut\(^{34}\). Spirochaetaceae *Treponema*, the third most important taxon, include species that are established human pathogens and others that have been reported to inhabit the cow rumen, the pig gastrointestinal tract, and the guts of termites\(^{35}\). *Christensenellaceae*, the fourth most important taxon, was recently identified as being the most heritable taxon in an analysis of twins from the UK\(^{3}\). Two taxa in the order Bacteroidales, *Prevotella stercorea* and *Prevotella copri*, the seventh and eighth most important taxa, are the only ones occurring at significantly reduced abundance in infected individuals; *Prevotella* is an important genus of gut bacteria and is underrepresented in Western versus African microbiomes\(^{11}\). While members of the Clostridia and Gammaproteobacteria are more abundant in infected individuals, the pattern for Bacteroidales is the opposite (see Fig. 2b. *Oscillospira uncl* and *Parabacteroides uncl*, the ninth and tenth most important taxa, are associated with the rumen and human intestine, respectively.

Furthermore, when looking at the microbial diversity of *Ent+* versus *Ent-* individuals, we found that the presence of *Entamoeba* is associated with a significant increase in alpha (intra-host) diversity using the Phylogenetic Distance Whole Tree metric (Welch’s t-test: \( p < 0.0001 \), Fig. 3a), as well as using the Shannon and Simpson indices (Welch’s t-test: \( p = 0.001 \) and \( p = 0.025 \), respectively; Supplementary Fig. 6). Interestingly, although the alpha (intra-host) diversity of *Ent+* individuals is significantly higher than *Ent-* individuals, the beta (inter-host) diversity (as estimated by both UniFrac distance metrics) reveals that gut communities across *Ent+* individuals are more similar than across *Ent-* individuals (Welch’s t-test: weighted and unweighted, \( p < 0.0001 \); Fig. 3b and Supplementary Fig. 7). This could suggest that, as alpha diversity increases, there are fewer potential stable states for individual gut communities, or that infection by *Entamoeba* drives changes in the microbiome that are dominant over other factors.
Relationship between specific taxa and microbial community diversity

Because of the striking relationship between *Entamoeba* infection status and alpha diversity, we sought to identify any phyla for which abundance was significantly correlated with community diversity. To account for the effect of *Entamoeba* infection, we added infection status as a binary covariate to our linear model and identified 11 phyla that are significantly correlated with alpha diversity (q < 0.05; see Supplementary Fig. 8). Although, as expected, the majority of these taxa increase in abundance with higher diversity, Bacteroidetes and Proteobacteria exhibit a decrease in relative abundance as alpha diversity increases. This negative relationship suggests that these taxa might be more competitive than others and drive down diversity.

Predicted metagenome

We used the KEGG (Kyoto Encyclopedia of Genes and Genomes) database\(^\text{36}\) and the PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) pipeline\(^\text{37}\) to predict abundances of pathways across individuals (see Supplementary Fig. 9). Considering the 220 most abundant KEGG pathways (comprising ≥ 0.01% of all assigned reads in at least 4 individuals), we identified 19 pathways with significant differences in abundance between *Ent*+ and *Ent*- individuals (linear regression: q < 0.05, see Supplementary Table 6 and Fig. 2d). Of these 19, of particular interest are an increase in amoebiasis (q = 0.001), biosynthesis of the antibiotic tetracycline (q = 0.03), and yeast MAPK signaling pathways (q = 0.01) in *Ent*+ individuals. These changes are largely attributed to Clostridiales and Ruminococcaceae, which occur at significantly greater abundance in *Ent*+ individuals (6.53% vs. 4.53%, q = 0.044; and 29.58% vs. 16.34%, q < 0.0001, respectively, Fig. 2d). Interestingly, the Cellular Antigens pathway, potentially involved in host-microbe and microbe-microbe interactions, is more represented in the predicted metagenomes of *Ent*- individuals (linear regression: q = 0.01). This pathway is predominantly attributed to members of the Enterobacteriaceae family, which was found to be twice as abundant in individuals lacking the parasite.
Finally, we found that there was no statistically significant relationship between infection status for this parasite and most of the covariates tested (including sex, BMI, subsistence, location, or dietary components). The only significant covariate was adult age, with Ent+ individuals being significantly older than Ent- individuals (See Supplementary Fig. 10; Welch’s t-test: p = 0.018). Despite this relationship, age is not a significant predictor of the gut microbial patterns across populations (PERMANOVA for weighted and unweighted UniFrac distance matrices, p > 0.06; Supplementary Table 2). Outside of an association between Ancylostoma and Bacteroidales uncl (q = 0.019), none of the other parasites tested (Ascaris and Trichuris) exhibited a significant association with any taxon, whether individually or as the number of all non-Entamoeba parasite types present. However, the overall composition seems to shift with the number of parasites (see Supplementary Fig. 11), and there is a significant increase in alpha diversity when three parasite species are present (Supplementary Fig. 12).

**Influence of subsistence on the microbiome**

*Microbial community patterns across subsistence*

To investigate the effect of subsistence on the composition of the gut microbiome, we summarized microbial taxonomic composition across the four subsistence groups and their geographic locations (Fig 4a, Supplementary Fig. 13). At the phylum level, we found a significant difference in the relative contribution of Proteobacteria across subsistence (linear regression: p = 0.003, Table 1), with hunter-gatherers having a higher frequency than the fishing population, farmers from the South and the North (23% versus 12.4%, 7.2% and 8.3%, respectively), mirroring the higher frequency in the Hadza hunter-gatherers compared to Italians. Based on a linear regression, we also found that 8 of the most abundant taxa differed significantly across subsistence modes (see Fig. 4b, Supplementary Tables 3-5). Of particular interest is the genus Bifidobacterium, both B. uncl and B. adolescentis, which were found at
higher abundance in the fishing population (means 0.30% and 0.51%, respectively) relative to all other populations (≤ to 0.11% and 0.07%, respectively; linear regression: q = 0.0003 and q = 0.008). This genus is associated with a higher consumption of dairy products, a pattern already observed in a comparison of Italians to Hadza hunter-gatherers\textsuperscript{14} and consistent with the occasional consumption of yogurt in the fishing population. We also found Bacteroidales uncl to occur at significantly lower relative abundance in the fishing population relative to the other three populations (0.7% vs. ≥ 2.4%; linear regression: q = 0.003), a order of bacteria also identified as being less abundant in the Italians versus the Hadza\textsuperscript{14}. In contrast with other Firmicutes genera that tend to be in lower frequency in hunter-gatherers, we found the genus \textit{Sarcina}, a synthesizer of microbial cellulose, to be only present in the hunter-gatherers (means of 0.69% compared to ≤ 0.07% in the other subsistence groups, linear regression: q = 0.007).

Finally, we found three members of the Lachnospiraceae family to be significantly different among populations, with \textit{Ruminococcus uncl} and \textit{Ruminococcus gnavus} being in lower frequency in hunter-gatherers (0.34% and 0.19%, respectively) compared to other populations (0.46-0.86% and 0.41-0.99%, respectively; linear regression: q = 0.030 and 0.006). This family has been linked to obesity\textsuperscript{39} in addition to protection from colon cancer attributable to their production of butyric acid\textsuperscript{40}.

A random forest classifier (RFC) model for microbiome composition revealed an overall accuracy of 59% (p < 0.001) for predicting the four subsistence groups but varied widely across populations (see Supplementary Table 7 and Supplementary Fig. 14a). The hunter-gatherer population was the most distinguishable such that the correct subsistence group was accurately predicted 85% of the time (versus 31% if predictions had been made by chance alone). Individuals of the fishing and South farming populations were predicted with 65% and 47% accuracy, respectively (versus 31% and 27%, by chance), and the North farming population was never predicted correctly (versus 11%, by chance). Furthermore, incorrect assignments for
individuals of the hunter-gatherer, farmers from the South and fishing populations were distributed evenly across all other subsistence groups, with the exception of farmers from the North, to which no individual was predicted to belong. Only five of the top ten taxa identified in the RFC model were determined to be significant in the linear regression (see Supplementary Fig. 14b). This suggests that rather than an individual signature taxon, it is the pattern of abundances of multiple taxa that is important for predicting subsistence. In agreement with our linear regression, the taxon identified as being the most important in distinguishing subsistence groups was *Bifidobacterium uncl* (see Supplementary Fig. 14b and Fig. 4b), occurring at significantly higher frequency in the fishing population (*q* = 0.0003, Supplementary Table 5). *Ruminococcus bromii*, important for degradation of resistant starch\(^\text{38}\), was the second most important taxon, occurring at 0.01%, 0.01%, 0.15%, and 0.12% in the fishing population, farmers from the North, the South, and hunter-gatherers, respectively (*q* < 0.0001) (see Supplementary Fig. 14c). The third, fourth, fifth and eighth most important taxa include members of the Lachnospiraceae family, two of which were found to be significant in the linear model (see above). When grouped together, taxa in this family are less abundant in the hunter-gatherers relative to other subsistence groups (11.3% vs. 15.6-19.6%, respectively), a difference significant only when comparing hunter-gatherers to both farmer populations. Finally, two species of the family Succinivibrionaceae family, *Succinivibrio sp.* and *Ruminobacter sp.*, were also identified as being important taxa in the model, both of which were more abundant in the hunter-gatherers at 9.7% and 3.7%, respectively, vs. less than 5.7% and less than 0.1% for the other three subsistence modes (*q* = 0.068 and 0.057, respectively; see Supplementary Fig. 14c). Both of these taxa, associated with the bovine rumen, were also found in higher frequency in the Hadza hunter-gatherers\(^{14}\). Finally, only five of the top ten taxa identified in the random forest classifier model were determined to be significant in the linear model (see Supplementary Fig. 14b). This suggests that rather than an individual signature taxon, it is the pattern of abundances of multiple taxa that is important for predicting subsistence.
Diet and gut microbial diversity

We found the alpha (intra-host) diversity to be significantly lower in the fishing population than in farmers from the South and the North for the phylogenetic distance whole tree metric (Welch’s t-test: p = 0.021 and p = 0.008, respectively) and only compared to farmers from the North for the Shannon and Simpson metrics (Welch’s t-test: p = 0.017, and 0.021, respectively) (see Fig. 5a, Supplementary Fig. 15-17). Interestingly, the pattern of beta diversity across subsistence modes using both unweighted and weighted Unifrac distance metrics also distinguishes the fishing population from both farmers, such that the within-group variation is significantly higher in the fishing and hunter-gatherer populations compared to both farmers (Welch’s t-test: p < 0.001 for all relevant pairwise comparisons) (see Fig. 5b and Supplementary Fig. 17a-b).

Community differences between subsistence groups based on weighted and unweighted Unifrac distance metrics are the greatest between the fishing population and the other three subsistence groups (Fig. 5c, Supplementary Fig. 17c). Beta diversity is the highest between the fishing population and hunter-gatherers, the two groups for which there is the largest number of significantly differentially abundant taxa (94% of significant taxa overall; Supplementary Table 5). This finding might be expected, given that these populations differ not only in terms of diet, but also in their genetic ancestry and access to medicine. Both farmer populations are slightly more different from the fishing population than from the hunter-gatherers (Fig. 5c, Supplementary Fig. 17c), and present a higher number of significantly differentially abundant taxa relative to the other populations (60% and 50% of the significant OTUs differ between the fishing population and the farmers from the South and the North, respectively, versus 44% and 38% that differ from the hunter-gatherers; Supplementary Table 5). This could suggest that differences in access to medicine, or the occasional consumption of processed food in the
fishing population, has considerable influence on their gut microbiomes, but this requires further investigation.

According to the taxonomy-based predicted metagenome for each subject's gut microbiota, we found that only one pathway, bacterial invasion of epithelial cells, differed significantly across all subsistence types; represented at the highest relative abundance in the hunter-gatherers and lowest in the farmers (linear regression: $q = 0.03$, Supplementary Fig. 18 and Supplementary Table 6). This pathway includes proteins expressed by pathogenic bacteria that are important for entry into host cells. The importance of this difference is unclear, but could be indicative of an increased abundance of pathogens in the microbiomes of hunter-gatherers.

### Discussion

We have conducted the first investigation of the relationship between intestinal parasitism and the human gut microbiome, and found that infection by the amoeboid parasite, *Entamoeba*, outcompetes diet, geographic location and ancestry in predicting composition and structure of the gut microbiome. Furthermore, we have conducted the first analysis assessing the role of subsistence, location and genetic ancestry in shaping the gut microbiota at a local scale. We showed that there is striking variation amongst different rural African populations, indicating that there are multiple signatures of rural, unindustrialized microbiomes.

### Influence of parasitism

The importance of gastrointestinal parasites in human disease is well established, both as infectious agents and in shaping immunity$^{22,41}$, and infection by helminths has notably been found to be a major force underlying the evolution of interleukin genes in humans$^{23}$. It has also been demonstrated that loss of helminth exposure removes the enhanced T helper cell 2 (Th2) and regulatory immune response imparted by these organisms, which is correlated with the
development of an array of immune-mediated diseases$^{42,43}$. This relationship is the basis of the hygiene hypothesis, which proposes that the underlying cause of the high incidence of autoimmune diseases, unique to industrialized countries, is the absence of childhood exposure to infectious agents$^{44}$. Recent research supporting this hypothesis shows that mild and controlled infection by internal parasites can activate an immune response and reduce symptoms of a range of autoimmune diseases$^{45}$. Likewise, the relationship between gastrointestinal microbiota and host immune response has been well established$^{46,47}$. Despite these clear host-parasite and host-microbiome interactions, and the fact that gut parasites and microbes share the same environment, there have been no studies assessing the relationship between these organisms. In this context, it is unclear what the observed differences in gut microbiome composition and structure between Entamoeba positive and negative individuals mean in terms of host health, but it could be that Entamoeba infection indirectly triggers a host immune response that differentially affects the success of different microbes. Alternatively, the effects of Entamoeba infection could be direct, whereby, for example, amoeba feed on certain species of bacteria, allowing others to proliferate. For instance, members of the phylum Bacteroidetes are negatively correlated with alpha diversity. The decrease in these taxa in Ent+ individuals could therefore result in the observed increase in other taxa and overall increase in alpha diversity. Interestingly, Clostridiales Ruminococcaceae, the second most important taxon in the RFC model, significantly more abundant in Ent+ individuals, has been found to be underrepresented in individuals suffering from Crohn’s Disease and Ulcerative Colitis$^{19}$. Likewise, a decreased prevalence of Prevotella copri and Fusobacteria, as observed in Ent+ individuals, was recently shown to be negatively correlated with Rheumatoid arthritis$^{48}$ and incidence of colorectal cancer$^{49}$, respectively, two conditions for which immunity also plays an important role. This suggests that the link between parasite removal and inflammation related human disease could in part be mediated by the gastrointestinal microbiome.
Influence of subsistence and genetic ancestry

In addition to identifying *Entamoeba* infection as an important predictor of gut microbiome composition and structure across populations, we were also able to examine the relative influence of other factors. First, we compared the gut microbiome composition of individuals from the same subsistence mode and genetic ancestry, but coming from different villages. Within the hunter-gatherers, we saw clear differences in composition as well as in diversity between individuals living in Bandevouri versus those living in Makouré and Bidou, although this difference was only significant using the Shannon Index for alpha diversity (p < 0.05; Welch’s t-test). Based on the data we have available, we found that these groups do not differ in terms of diet or parasitism, suggesting a role for other unexplored environmental factors (e.g., water source).

In terms of genetic ancestry, we found that, despite a genetic divergence as old as 60,000 years, the gut microbiome of the Pygmy hunter-gatherers is not strikingly different from that of the Bantu populations. The UniFrac distances and the number of significant taxa are indeed lower between hunter-gatherers and farmers than between the farmers and fishing population, two Bantu groups that share the same genetic ancestry. In agreement with previous studies showing that effects of host genetics are outweighed by diet and only affect specific taxa\(^3,5^0\), this suggests that host genetics alone is not a dominant factor in determining overall shifts in gut microbiome composition.

Nonetheless, in our study, we found key differences distinguishing the microbiota of hunter-gatherers from those of the farming and fishing populations, likely reflecting the influence of their long-term diet. The hunter-gatherers were correctly assigned to their subsistence mode with higher accuracy (85%) relative to the other populations. Furthermore, some of our findings mirror patterns previously observed in the Hadza from Tanzania\(^1^4\) (see Table 1b), suggesting
this ancestral subsistence mode might carry a specific microbial signature. Notably, we found a higher frequency of Proteobacteria in hunter-gatherers compared to the other Cameroonian populations, similar to the relationship between the Hadza and Italians\textsuperscript{14}. Lachnospiraceae uncl, identified as the third important in the RFC model with the tendency to be lower in the Pygmy hunter-gatherers (5.7% versus 7.7-11.3% in other populations, $q = 0.075$), was also found to be in lower frequency in the Hadza compared to Italians\textsuperscript{14}. Finally, \textit{Succinivibrio} and \textit{Ruminobacter} species, found enriched in Hadza, were also identified as important taxa in the RFC model, and occur at higher frequencies in the Pygmy hunter-gatherers (see Table 1b). Thus, all these taxa seem to be a specificity of hunter-gatherer populations, rather than reflecting a difference between industrialized European and rural African populations. \textit{Succinivibrio} is considered to be an opportunistic pathogen, which could mean that hunter-gatherer populations have more opportunistic pathogens than other populations, as proposed by Schnorr et al\textsuperscript{14}. However, while the opportunistic \textit{Treponema} was also found enriched in the Hadza, we found it at a very low frequency in all the populations studied here ($< 3.5\%$, Table 1b), and not differing among subsistence. When looking at other opportunistic genera in the Enterobacteriaceae family, we found that \textit{Shigella} and \textit{Escherichia}, both previously found only in Italian children and not in children from Burkina Faso\textsuperscript{11}, occur at extremely low abundances in all four subsistence groups ($< 0.1\%$, see Table 1b). As for \textit{Klebsiella} and \textit{Salmonella}, neither taxon differed significantly amongst our groups (Table 1b). Thus, there does not seem to be any clear trend for opportunistic pathogens in hunter-gatherer populations compared to others.

\textbf{Influence of geography and industrialized lifestyle}

Amongst the four populations included in this study, the fishing population is the most urbanized due to increased consumption of processed food and access to medicine (resulting in a higher BMI and lower prevalence of intestinal parasites). As such, the characteristics distinguishing the gut microbiomes of the fishing population from the farmers and hunter-gatherers that also differ
between rural populations in developing countries and urban populations in industrialized countries\textsuperscript{12,14} might correspond to signature patterns of a more industrialized lifestyle. In particular, within the phylum Bacteroidetes, we found a lower overall abundance of Bacteroidales uncl in the fishing population relative to the other three populations (Table 1), an order also depleted in Italians compared to Hadza\textsuperscript{14}. High abundance of \textit{Prevotella} and \textit{Bacteroides} have also been shown to represent signatures of the microbiomes for people in developing and industrialized countries, respectively\textsuperscript{11,12,14}. Higher abundances of \textit{Prevotella} are indeed often correlated with increased consumption of carbohydrates and simple sugars, whereas an elevated proportion of \textit{Bacteroides} is associated with a diet richer in protein and fat. Although differences between the populations studied here was not statistically significant, the fishing population harbored the highest abundance of \textit{Prevotella} sp. (30.8%), while the farmers from the North and the hunter-gatherers harbored the lowest (19.2% and 20.2%, respectively). This abundance of \textit{Prevotella} sp. is high relative to other genera in this order across all populations (Table 1b). Species of \textit{Prevotella} were the most reduced in individuals infected with \textit{Entamoeba}, a decrease that correlates significantly with higher alpha diversity.

In conclusion, our study suggests an important role for eukaryotic gut inhabitants and the potential for feedbacks between helminths, protozoa, microbes, and the host immune response, one that has been largely overlooked in studies of the microbiome. Prior analyses of the African gut microbiome have found an enrichment of \textit{Treponema}, Bacteriodetes and \textit{Prevotella} as compared to European populations. Although this enrichment was proposed to be related to diet, our observations indicate that some of these trends could be related to infection by \textit{Entamoeba} (or other parasites). In addition, our results highlight the substantial variability in gut microbiome composition among relatively closely related populations. Using a single population as a representative of a diet or geographical region may therefore be overlooking important fine-
scale patterns in microbiome diversity. Hence, comparative population studies of the human microbiome stand to benefit tremendously from considering variation within a geographic region.

MATERIALS AND METHODS

Sample collection. We sampled 80 volunteers (34 women and 46 men) in seven rural villages (Bidou, Makouré, Bandevouri, Ndtoua, Afan Essokié, Akak and Ebodié) in Southern Cameroon, after obtaining their informed consent for this research project. For each participant, we collected information about their age, gender, anthropometric traits, health status, ethno-linguistic and quantitative nutritional questionnaires. We also collected saliva and fecal samples. The fecal sample was self-collected in the morning and stored in a plastic bag at most 3-4h before further handling. It was then split in two separate samples; one was used to perform the parasitological analysis at a local hospital (fresh or covered with formol) and the other was stored to run the sequencing analyses. This latter sample was handled following previous methods\textsuperscript{51}: the sample was first submerged with pure ethanol for about 24h at room temperature, then the ethanol was poured out of the container and the sample was wrapped in a sterile gauze and deposited on silica gel. The silica gel was then replaced by new gel when it changed colors from orange to yellow, i.e. when it could not absorb further humidity. The samples were then transported back to France and stored at -80°C until they were shipped to Minnesota, USA, on dry ice, and stored there at -80°C until further use. For five individuals, we were able to collect replicate fecal samples at two different time points: four individuals 7 days apart, and one individual 1 day apart.

DNA extraction and 16S rDNA amplification from the fecal sample. Total DNA was extracted directly from approximately 50 mg of each fecal sample using the MO BIO PowerFecal\textsuperscript{TM} DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA, USA) according to the manufacturer's protocol. DNA isolated from fecal samples was quantified using a NanoDrop
(ThermoScientific), and the V5-V6 regions of the 16S rRNA gene were PCR amplified with the addition of barcodes for multiplexing. The forward and reverse primers were the V5F and V6R sets. The barcoded amplicons were pooled and Illumina adapters were ligated to the reads. A single lane on an Illumina MiSeq instrument was used (250 cycles, paired-end) to generate 16S rRNA gene sequences yielding 175,784 Pass Filter (PF) reads per fecal sample (SD = 72,822) and ~12.65 million total PF reads (4.9Gb of data).

**16S rDNA analysis.** Raw Illumina sequences were demultiplexed and filtered using cutadapt 1.7.1 (Martin, 2011) to remove adaptor sequences (Read 1:CTGTCTCTTATACATCTCCGAGCCCACGAGAC, Read 2:CTCTCTCTTATACATCTGCCGCTGCCGACGA), sequences containing ambiguous bases and low quality reads (Phred quality scores < 20). Read pairs were resynced using RISS-UTIL and matching paired-end sequences were merged using FLASH. Merged sequences over 250 bp in length (the maximum length of the V5-V6 region) were removed. The remaining merged sequences were analyzed using the open-source software package QIIME 1.7.0 (Quantitative Insights Into Microbial Ecology). We performed closed-reference Operational Taxonomic Unit (OTU) picking at 97% identity against the May 2013 Greengenes database such that OTUs were assigned based on taxonomic assignment of the reference sequence with a 97% similarity cutoff. Reads which failed to hit a reference sequence with at least 97% identity were discarded. The average number of reads that were assigned to an OTU per population was 82%. All summaries of the taxonomic distributions ranging from phylum to species were generated from the non-rarefied OTU table generated from this analysis.

**Diversity analyses**

To characterize diversity across individuals, rarefaction plots were generated for each sample using the phylogenetic distance metric for diversity. Samples were rarefied to 50,000 reads,
the maximum depth permitted to retain all samples in the dataset. All diversity analyses were conducted on the rarefied OTU table containing 50,000 sequences per sample. Measurements are based on the mean values calculated from 100 iterations using a rarefaction of 10,000 sequences per sample (20% of the total 50,000). Alpha-diversity was calculated for each sample based on phylogenetic diversity, Shannon’s index and the Simpson index. Beta-diversity was assessed based on both unweighted and weighted UniFrac distance metrics using the Greengenes phylogenetic tree. Principal Coordinate Analysis (PCoA) was carried out on the distance matrices. P-values were calculated using the Welch’s t-test. To determine if the UniFrac distances were on average significantly different for groups of samples, we conducted Principal Coordinates Analysis (PCA) to reduce raw gastrointestinal microbial community data into axes of variation. We assessed the significance of each covariate by performing a permutational multivariate analysis of variance (PERMANOVA), a non-parametric test, on both weighted and unweighted UniFrac distance matrices using the “adonis” function from the vegan package in R. This test compares the intragroup and intergroup distances using a permutation scheme to calculates a p-value. For all PERMANOVA tests we used 10,000 randomizations.

**Multivariate analysis of composition data**

Intergroup differences in microbiome composition for subsistence, location, population, BMI, sex, language, age, dietary factors, and parasitism were assessed by PERMANOVA implemented using the “adonis” function of the vegan package in R. Then, to identify taxa significantly associated with each covariate of interest, we used a generalized linear model, FDR corrected for the number of OTUs. For the linear model, OTUs with identical taxonomic identifiers were merged. In parallel, we also restricted the merging only to OTUs names defined at the family, genus or species level. Both results are reported in Supplementary Material ("merged OTUs" versus "partially merged OTUs"). For analyses of both merged and partially
merged OTUs, the resulting taxa were filtered to include only those that occurred at least 0.1% in at least 4 individuals.

**Random forest classifier model**

A random forest classifier with 2000 decision trees was trained on the taxa abundance table consisting of 93 OTUs with 5-fold cross-validation using scikit-learn\textsuperscript{61}. Mean accuracy (the ration of the number of correct predictions relative to the total number of predictions) over the 5 folds was 0.79 (standard deviation 0.09) with p < 0.001 (estimated using 1000 permutation tests with 5-fold cross-validation). The most discriminating taxa were identified by random forest importance values (in scikit-learn random forest importance values are calculated as mean decrease in node impurity). We report the top ten median importance values and 95% confidence intervals from 1000 random forests.

**Metagenomic predictions**

We used PICRUSt v1.0.0 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) to generate taxonomy-based predicted metagenomes for each sample\textsuperscript{37}. Counts from the rarefied OTU Table (50,000 OTUs per sample) were normalized by the predicted 16S rRNA gene abundances and functional predictions of Kyoto Encyclopedia of Genes and Genomes (KEGG)\textsuperscript{36} pathways were determined using pre-computed files for the May 2013 Greengenes database\textsuperscript{65}. Relative abundances of the functional predictions were calculated. We also compared the predicted metagenomes of individuals to determine which functions were enriched or depleted across covariates (subsistence, population, BMI, age, sex, dietary components, and parasitism phenotypes) for abundant (\(\geq 0.1\%\) in at least 4 individuals) and rare (<0.1% in at least 4 individuals) pathways. A linear mixed model was used to determine which predicted pathways were significant (q < 0.05) for all each covariate.
Acknowledgements

The authors thank Karen Tang, Andres Gomez, Michael Burns, and Peter Zee for helpful discussions. This work was carried out using computing resources at the Minnesota Supercomputing Institute.

References


Fig. 1. (a) Map showing the geographic locations of the villages sampled in Southwest Cameroon, the number of samples (N) collected for each subsistence group (the fishing population, farmers from the South, farmers from the North, and hunter-gatherers), and their genetic ancestry (Bantu or Pygmy). (b) Principle Components Analysis based on dietary questionnaires for all 64 individuals. The first two principal components (PC1 and PC2) are shown, with the amount of variation explained reported for each axis.
Fig. 2. Relationship between *Entamoeba* infection (*Ent-* or *Ent+*) and fecal microbiome composition. (a) Multidimensional Scaling plot of unweighted UniFrac distances colored by *Entamoeba* infection. The first two principal components (PC1 and PC2) are shown. (b) Summary of the relative abundance of taxa (≥ 0.1% in at least 4 individuals) for *Ent-* and *Ent+* individuals color coded by phylum (Acinobacteria (Act.) = red, Bacteroidetes (Bact.) = green, Cyanobacteria (Cyan.) = black, Elusimicrobia (Elus.) = gold, Firmicutes (Firm.) = blue, Fusobacteria (Fus.) = pink, Lentisphaerae (Lent.) = yellow, Proteobacteria (Prot.) = purple, Spirochaetes (Spir.) = orange, and Tenericutes (Ten.) = gray). The number of individuals (N) in each population is indicated below the bars. (c) Normalized relative abundance of four taxa significantly associated with *Entamoeba* infection status in the linear regression as well as in the Random Forest Classifier model (q < 0.05). (d) Normalized relative abundance of four KEGG metabolic pathways significantly associated with *Entamoeba* infection status in the linear regression (q<0.05 using the most abundant (≥ 0.4% in at least one group) KEGG (Level 3)).
pathways) (left panel) and the relative contributions of each taxon for each pathway (right panel).
Fig. 3. (a) Comparison of alpha diversity for *Entamoeba* negative (*Ent-*) and positive (*Ent+*) individuals using the phylogenetic distance whole tree metric. (b) Comparison of beta diversity within *Ent-*, within *Ent+*, and between *Ent-* and *Ent+* individuals based on unweighted UniFrac distances. P-values are based on Welch’s t-test.
Fig. 4. Relationship between subsistence modes and fecal microbiome composition. (a) Summary of the relative abundance of taxa (occurring at >= 0.1% in at least 4 individuals) for individuals across subsistence. Taxa are colored by phylum (Actinobacteria (Act.) = red, Bacteroidetes (Bact.) = green, Cyanobacteria (Cyan.) = black, Elusimicrobia (Elus.) = gold, Firmicutes (Firm.) = blue, Fusobacteria (Fus.) = pink, Lentisphaerae (Lent.) = yellow, Proteobacteria (Prot.) = purple, Spirochaetes (Spir.) = orange, and Tenericutes (Ten.) = gray). The number of individuals (N) in each population is indicated below the bars. (b) Relative abundance of four taxa significantly associated with subsistence based on a linear regression model, q < 0.05. Fis = Fishing population; Far(S) = Farmers from the South; Far(N) = Farmers from the North; HG = Hunter-gatherers.
Fig. 5. Comparison of the diversity of gut microbiomes of individuals across subsistence (a) Alpha diversity based on the phylogenetic metric, phylogenetic distance (PD) whole tree. (b) Beta diversity within each subsistence group based on unweighted UniFrac distances. (c) Beta diversity for pairs of subsistence groups based on unweighted UniFrac distances. For interpair comparisons, all are significant ($P < 0.05$) unless specified (n.s.). All $p$-values are based on Welch’s t-tests. Fis = Fishing population; Far(S) = Farmers from the South; Far(N) = Farmers from the North; HG = Hunter-gatherers.
Table 1

<table>
<thead>
<tr>
<th>Phylum (&gt;0.1% in at least 4 ind)</th>
<th>Effect of <em>Entamoeba</em></th>
<th>Effect of subsistence</th>
<th>Previous results from the literature$^{11,14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Ent</em> · <em>Ent</em> · p-val</td>
<td>Fis · Far(S) · Far(N) · HG · p-val</td>
<td></td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>1.4 1.0 0.634</td>
<td>2.1 0.9 0.9 0.8 0.056</td>
<td>I &gt; H</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>35.3 18.6 0.003</td>
<td>33.4 31.7 26.7 26.0 0.708</td>
<td>H &gt; I</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>0.17 0.24 0.006</td>
<td>0.10 0.26 0.18 0.24 0.089</td>
<td></td>
</tr>
<tr>
<td>Elusimicrobia</td>
<td>0.03 0.10 1E-07</td>
<td>0.01 0.08 0.23 0.05 0.206</td>
<td></td>
</tr>
<tr>
<td>Euryarchaeota</td>
<td>0.03 0.09 4E-06</td>
<td>0.02 0.04 0.08 0.07 0.247</td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>47.0 60.8 0.008</td>
<td>49.5 55.3 58.0 47.8 0.518</td>
<td>I &gt; H</td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>0.15 0.02 0.169</td>
<td>0.13 0.21 0.01 0.01 0.376</td>
<td></td>
</tr>
<tr>
<td>Lentisphaerae</td>
<td>0.19 0.15 0.012</td>
<td>0.09 0.26 0.29 0.14 0.189</td>
<td></td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>13.8 14.1 0.915</td>
<td>12.4 7.2 8.3 23.0 0.003</td>
<td>H &gt; I</td>
</tr>
<tr>
<td>Spirochaetes</td>
<td>0.9 2.9 1E-04</td>
<td>1.1 2.6 3.2 0.7 0.574</td>
<td>H &gt; I</td>
</tr>
<tr>
<td>Tenericutes</td>
<td>1.0 1.0 0.001</td>
<td>1.1 1.5 1.9 1.0 0.238</td>
<td></td>
</tr>
<tr>
<td>Verrucomicrobia</td>
<td>0.04 0.16 0.074</td>
<td>0.02 0.04 0.23 0.12 0.078</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.** (a) Frequency (in %) of phyla for *Entamoeba* negative (*Ent*-) and positive (*Ent*+) individuals and for the four subsistence groups (Fis = Fishing population; Far(S) = Farmers from the South; Far(N) = Farmers from the North; HG = Hunter-gatherers). (b) Frequency (in %) of specific taxa of interest previously associated with geography in the four subsistence groups. P-values are based on a linear regression model. The last column indicates whether previous studies$^{11,14}$ found an enrichment of each phylum in Hadza adults (H) versus Italians adults (I), or in Burkina Faso children (BF) versus Italians children (I). The first letters in parenthesis indicate to which phylum each taxa belongs (Act. = Actinobacteria, Bact. = Bacteroidetes, Firm. = Firmicutes, Prot. = Proteobacteria, and Spir. = Spirochaetes).