Ultra-deep Sequencing of Hadza Hunter-Gatherers Recovers Vanishing Microbes

Authors: Bryan D. Merrill[1*], Matthew M. Carter[1*], Matthew R. Olm[1*], Dylan Dahan[1], Surya Tripathi[1], Sean P. Spencer[1], Brian Yu[2], Sunit Jain[2], Norma Neff[2], Aashish R. Jha[3], Erica D. Sonnenburg[1+], Justin L. Sonnenburg[1,2,4+]

Affiliations:

[1]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA.

[2]Chan Zuckerberg Biohub, San Francisco, CA, USA.

[3]Genetic Heritage Group, Program in Biology, New York University Abu Dhabi, Abu Dhabi, UAE.

[4]Center for Human Microbiome Studies, Stanford University School of Medicine, Stanford, CA, USA.

*These authors contributed equally to this work

[+]Corresponding author. Email: erica.sonnenburg@stanford.edu, or jsonnenburg@stanford.edu

Abstract

The gut microbiome has been identified as a key to immune and metabolic health, especially in industrialized populations[1]. Non-industrialized individuals harbor more diverse microbiomes and distinct bacterial lineages[2], but systemic under-sampling has hindered insight into the extent and functional consequences of these differences[3]. Here, we performed ultra-deep metagenomic sequencing and laboratory strain isolation on fecal samples from the Hadza, hunter-gatherers in Tanzania, and comparative populations in Nepal and California. We recover 94,971 total genomes of bacteria, archaea, bacteriophage, and eukaryotes, and find that 43% are novel upon aggregating with existing unified datasets[4,5]. Analysis of *in situ* growth rates, genetic *pN/pS* signatures, and high-resolution strain tracking reveal dynamics in the hunter-gatherer gut microbiome that are distinct from industrialized populations. Industrialized versus Hadza gut microbes are enriched in genes associated with oxidative stress, possibly a result of microbiome adaptation to inflammatory processes. We use phylogenomics to reveal that global spread of the spirochaete *Treponema succinifaciens* parallels historic human migration prior to its extinction in industrialized populations. When combined with a detailed definition of gut-resident strains that are vanishing in industrialized populations, our data demonstrate extensive perturbation in many facets of the gut microbiome brought on by the industrialized lifestyle.

Recognition of work with indigenous communities

Research involving indigenous communities is needed for a variety of reasons including to ensure that scientific discoveries and understanding appropriately represent all populations and do not only benefit those living in industrialized nations[3,6]. Special considerations must be made to ensure that this research is conducted ethically and in a non-exploitative manner. In this study we performed deep metagenomic sequencing on fecal samples that were collected from Hadza hunter-gatherers in 2013/2014 and were analyzed in previous publications using different methods[2,7]. A material transfer agreement with the National Institute for Medical Research in Tanzania specifies that stool samples collected are used solely for academic purposes, permission for the study was obtained from the National Institute of Medical Research (MR/53i 100/83, NIMR/HQ/R.8a/Vol.IX/1542) and the Tanzania Commission for Science and Technology, and verbal consent was obtained from the Hadza after the study's intent and scope was described with the help of a translator. The publications that first described these samples included several

scientists and Tanzanian and Nepali field-guides as co-authors for the critical roles they played in sample collection, but as no new samples were collected in this study, only scientists who contributed to the analyses described here were included as co-authors in this publication. It is currently not possible for us to travel to Tanzania and present our results to the Hadza people, however we intend to do so once the conditions of the COVID-19 pandemic allow it.

Main

The gut microbiome is increasingly recognized as a critical aspect of human health, but microbiome studies are heavily biased towards western industrialized populations[3]. Microbiota composition varies across lifestyles with those from non-industrialized populations harboring greater diversity and distinct microbes known as VANISH (Volatile and/or Associated Negatively with Industrialized Societies of Humans) taxa[8–14]. Analogously, microbiomes of industrialized populations are enriched for BloSSUM (Bloom or Selected in Societies of Urbanization/Modernization) taxa. The transition to an industrialized microbiome is observed in immigrants to the U.S. supporting a causal role of lifestyle[15]. The presence of VANISH taxa in non-industrialized societies around the world and ancient humans underscores their potential evolutionary importance[16,17]. Human-associated microbial lineages have been passed across hominid generations across evolutionary time[18,19], raising the possibility that human biology has become reliant upon functions and cues that these microbes provide[20]. Current understanding of VANISH taxa is primarily based on 16S rRNA sequencing[21], and therefore lacks phylogenetic resolution and genomic/functional insight. A higher-resolution view, including an understanding of VANISH functional capacity, growth dynamics, and dispersal patterns is needed to understand microbiome change induced by the industrialized lifestyle.

Metagenomic sequencing has transformed our ability to understand microbes without culturing, but most microbiome studies use relatively shallow sequencing **(Fig. 1A)**. Deeper sequencing improves detection of resident microbes[22] (including microbial eukaryotes[23]) and provides insight from recently developed techniques including *in situ* growth rate prediction, high-precision strain-tracking, *de novo* genome recovery, and microdiversity analysis[24,25].

Here we present ultra-deep metagenomic sequencing of the Hadza hunter-gatherer gut microbiome. The Hadza reside near Lake Eyasi in the central Rift Valley of Tanzania, live in bush

camps of approximately 5 to 30 people, move between camps approximately every 4 months, primarily drink from water springs and streams, and eat a diet that includes foraged tubers, berries, honey, and hunted animals[26]. They are among the last remaining populations in Africa that continue a form of the ancestral foraging legacy of our human species.

Recovery of Hadza-associated genomes and isolates

We performed metagenomic sequencing on stool samples collected from 167 Hadza individuals (including 33 infants and 6 mothers[27]) between September 2013 and August 2014[2,7], 56 Nepali individuals[28] and 12 Californians[1] **(x`tary Table 1**). The Nepali samples are from four populations living on a lifestyle gradient: foragers (Chepang), and agrarians (Raute and Raji, recent agrarians; Tharu, longtime agrarians[28]). The Hadza, Nepali, and Californian samples were sequenced to approximately 25 giga base pairs (Gbp), an exceptional depth relative to prior studies (**Fig. 1A; Extended Data Fig. 1**), and includes the most deeply-sequenced human gut metagenome (210 Gbp) to date.

Using multi-domain assembly, binning, and read-mapping, we recovered 48,185 bacteria (**Fig. 1B**), 290 archaea, and 17 eukaryote **(Fig. 1C)**, and 34,552 bacteriophage (**Fig. 1D**) metagenome-assembled genomes (MAGs) from Hadza samples (see methods for details, **Supplementary Table 2**). MAGs recovered from the Hadza expand the Unified Human Gastrointestinal Genome (UHGG, v1) database[4] bacterial and archaeal species count by 25.4% and 14.3%, respectively, and the Metagenomic Gut Virus (MGV) catalog[5] viral species count by 23.7%. Over half (59.7%) of the 6.6 million protein families found in Hadza gut microbes are absent from the UHGP-95 (v1) protein database[4] **(Extended Data Fig. 2)**. 52 bacterial strains (**Supplementary Table 3)** were isolated and sequenced from Hadza stool samples, which comprise of 31 different bacterial species belonging to 4 phyla (including 9 strains of *Bifidobacterium infantis*[27]); 20 of these species have no previously cultured representative from human stool and 9 species are novel relative to UHGG v1 (**Extended Data Fig. 3**).

Of the 21 eukaryotic genomes recovered in this study, 17 are from the Hadza and 4 from the Nepali samples (**Fig. 1C; Supplementary Table 2**). All Nepali and the majority of Hadza genomes are from the genus *Blastocystis* (n=14), a prevalent member of the mammalian gut microbiota[106]. Of the 7 other eukaryotic genomes recovered from the Hadza gut, one is a remarkably large and

complete genome of a stingless bee (232 megabase pairs and 92.3% complete), the honey and larvae of which are known to be consumed by the Hadza[107], and four are novel *Amoebae* (n=2) and *Trepomonas* (n=2) genomes **(Fig. 1C)**. While a comprehensive genome database does not yet exist for eukaryotes known to colonize the human gut, genomes from these species are not present in NCBI GenBank[108] (a repository of genomes sequenced from all environments).

Metagenomic reads generated here were mapped to three custom databases containing full genome sequences of species-level representatives for the bacteria/archaea (n=5,755) bacteriophage (n=16,899), and eukaryote (n=12) genomes (see methods for details). Over 80% of the metagenomic reads from Hadza, Nepali, and Californian samples map to these databases **(Fig. 1E)**. Notably, the Hadza have higher bacterial, bacteriophage, and archaeal diversity than other populations in this study**,** with the exception of Nepali Forager bacteriophage diversity (**Fig. 1F)**. This increased diversity was not due to increased sequencing depth as an *in-silico* rarefaction analysis revealed more total and novel species of bacteria, archaea, and bacteriophage in Hadza samples compared to other populations across a range of sequencing depths (**Fig. 1G; Extended Data Fig. 4)**. Analysis of exceptionally deeply-sequenced samples (≥ 50 Gbp) suggests that the Hadza gut microbiome contains two-to-four times the number of bacterial species when compared to Californians at similar sequencing depths (**Extended Data Fig. 5).**

VANISH microbes abound in the Hadza

To explore the extent to which the Hadza microbiome differs from other populations, we curated a dataset of 1,800 human gut metagenomes from 21 published studies[11,15,29–44](industrial, n=950; transitional, n=583; Hadza hunter-gatherers from this study, n=135; and other hunter-gatherers, n=132; **Extended Data Fig. 6A-B, Supplementary Table 4)**. Analysis of the hunter-gatherer samples demonstrates that much diversity and distinguishing taxa are recovered with deeper sequencing, so subsequent compositional analysis was focused on the deeply sequenced Hadza samples (**Extended Data Fig. 6C-F**). The presence of each species within our bacterial/archaeal genome database was determined for each sample (**Fig. 2A**, **Supplementary Table 5**); and VANISH (n=124) and BloSSUM (n=63) taxa were defined as those that are most significantly enriched in the Hadza and industrial populations, respectively (Fisher's exact test; ≥95th percentile; **Fig. 2B; Extended Data Fig. 7**).

Most VANISH taxa (n=120; 96%) and all BloSSUM taxa (n=63; 100%) are detected in "transitional" samples (taken from populations intermediate between hunter-gatherer/forager and industrialized lifestyles). These taxa are typically found at intermediate prevalence, consistent with the extent of lifestyle change corresponding to the magnitude of microbiome shifts[45] (**Fig. 2C**). Interestingly, BloSSUM taxa have higher *in situ* growth rates than VANISH taxa in transitional samples **(Fig. 2D)** and are anti-associated with the presence of *Blastocystis*, even within individuals from industrialized populations (**Extended Data Fig. 8**). Replication rate differences may indicate a competitive advantage of BloSSUM taxa in the industrialized gut versus the slower replicating VANISH taxa.

We investigated the functional consequences of the trade-off between VANISH and BloSSUM taxa concomitant with lifestyle change. The extraordinary level of novelty present in the Hadza gut **(Extended Data Fig. 7)** precludes the use of most gene annotation pipelines, and we thus focused our functional analysis on protein domains (Pfams), which represent broad, evolutionary conserved functional units[46]. Functional analysis identified 145 and 588 Pfams that are more prevalent in VANISH and BloSSUM taxa, respectively ($p < 0.01$; Fisher's exact test, Benjamini p-value correction; **Fig. 2E**; **Supplementary Table 6**). Pfams most associated with VANISH taxa point to a relatively outsized use of metal ions, peptidases, and RNA methylation. BloSSUM Pfams are associated with antioxidant and redox sensing functionality, perhaps reflecting increased oxygen tension associated with inflammation or an altered epithelial metabolic state in the industrialized gut[21,47]. These differences demonstrate that VANISH and BloSSUM taxa are not functionally redundant.

Treponema succinifaciens **dispersal mirrors human migration**

Several species of the phylum Spirochaetota were identified as VANISH taxa in this study **(Supplemental Table 5)**. Spirochaetota in general, and especially the most well-studied species *Treponema succinifaciens,* are known to be depleted in industrialized microbiomes[8]. Here we leveraged i) deep sequencing we performed on Hadza, Nepali, and Californian samples using consistent methods, and ii) new Spirochaetota genomes recovered in this study (n=1047) to conduct a robust analysis of Spirochaetota prevalence across lifestyles. Our recovered

Spirochaetota MAGs belong to the *Treponemataceae*, *Sphaerochaetaceae*, or *Brachyspiraceae* families and span 26 species (including a sequenced isolate of *Treponema perunse*[48]), 16 of which are novel relative to the UHGG. The relative abundance of Spirochaetota species decreases with increased industrialization and no Spirochaetota genomes are detected within Californians **(Fig. 3A)**. Hadza Spirochaetota genomes fall into three diverse families also found in other populations **(colored boxes, Fig. 3B)** suggesting that Spirochaetota are a core component of the non-industrialized microbiome and highly susceptible to loss upon lifestyle change.

The MAGs recovered here increase the number of publicly available *Treponema succinifaciens* genomes from 125 to 346 (276% increase), enabling a robust phylogenomic analysis of the species (**Fig. 3C**). We identified both Hadza-specific and globally distributed clades of *T. succinifaciens* and observed an association between phylogeny and continent of origin (delta statistic d=7.79, p-value<0.0001)[49]. To model the dispersal of *T. succinifaciens* between human populations, we performed stochastic character mapping on the phylogenetic tree of MAGs in which the population where each MAG was recovered is coded as a trait of the genome and the frequency of "transition events" between each pair of populations is quantified[50] (**Fig. 3D**). The 4 most frequent transition events between populations are from the Hadza to other populations, accounting for 46.7% of all transition events, suggesting that *T. succinifaciens* was carried along the out-of-Africa human dispersal routes[51]. The congruence of *T. succinifaciens* phylogenomics with known patterns of past human migration is consistent with its dispersal being linked to close human contact (e.g., vertical transmission), as has been described for *Helicobacter pylori*[19,52].

Evolution, growth, and dispersal in the Hadza gut

The high sequencing depth and sample number achieved in this study provide an unprecedented opportunity to investigate *in situ* growth rates, microdiversity, and strain sharing within a hunter-gatherer population. The Hadza gut microbiome has been shown to undergo seasonal cycling in carbohydrate-active enzyme (CAZyme) and species composition[2,7]. Here we confirm these findings using deeper sequencing and updated metagenomic methods **(Extended Data Fig. 9)** and for the first time identify seasonal cycling in bacterial replication rates **(Extended Data Fig. 10A).** Analysis of intra-genic *pN/pS* ratios, a measure of bias towards non-synonymous mutations that suggests directional or diversifying selection, reveals a roster of functions with higher *pN/pS* ratios

($p < 0.01$; n=693); many are associated with extracellular or membrane-bound proteins, such as Ig-like folds, pilin motifs, and collagen-binding proteins (**Extended Data Fig. 10B**).

Family relation and cohabitation are among the strongest factors associated with microbial strain sharing in industrial populations[53,54], but it is unknown whether these patterns hold for hunter-gatherer populations like the Hadza. We performed a high-resolution strain-tracking analysis (threshold for same strain = 99.999% popANI) and found that family members share more recently-transmitted strains than unrelated individuals among the Hadza **(Fig. 4A, Supplementary Table 7)**. Interestingly, strain sharing among members of the same bush camp approaches that between members of the same family **(Fig. 4B)**, and this effect is stronger in some bush camps **(Fig. 4B)**. For example, individuals from the Hukamako camp share more strains with one another than family members share on average across all camps. Drinking water source (e.g., spring, stream, riverbed, etc.) and season (late dry, early dry, early wet, or late wet) have been previously linked to gut microbiome similarity[2,28], and here we demonstrate that these factors are also linked with the sharing of identical microbial strains **(Fig. 4A)**. Overall, these results point to the importance of environmental factors, kinship, and bush camp membership (a social structure with no equivalent in the industrialized populations) in driving strain dispersal among hunter-gatherers.

Discussion

Here we elucidated many novel facets of lifestyle differentiation in the gut microbiome using exceptionally deep metagenomic sequencing of non-industrialized populations, particularly the Hadza hunter-gatherers of Tanzania. The discovery of numerous novel clades of bacteria, archaea, bacteriophage, and eukaryotes highlight a leap in understanding of non-industrialized microbiomes and reframe the incompleteness and bias of commonly used genomic reference databases. Functional differences in the gut microbiomes of humans living different lifestyles affirm the ability of our intestinal inhabitants to continually adapt to selective pressures in the gut environment. The VANISH taxa found in present-day Hadza may represent lineages of microbes that shaped human development throughout our species' long history as foragers. Global phylogenomic analysis of the commensal spirochaete, *Treponema succinifaciens*, shows strain relatedness consistent with known human migration patterns prior to industrialization. Extending deep metagenomic sequencing to populations living across additional geographies will enable a

better understanding of which microbes traveled with, were lost, or gained in human populations as we spread around the planet. An important challenge is to characterize the impact of these microbes on human physiology and determine whether the absence or presence of species and functions are detrimental to human health. Overall, our results conclusively show that the differences between industrialized and non-industrialized microbiomes go well beyond simple taxonomic membership and diversity. These findings have substantial implications for how the microbiome may be investigated toward improving the health of both industrialized and non-industrialized populations.

Acknowledgements: We are indebted to the participants in this study. We acknowledge the numerous people and organizations who provided logistical support and conducted sample collection in the USA, Tanzania, and Nepal, including Dorobo Safaris, the Human Food Project, John Changalucha, Alphaxard Manjurano, Maria Gloria Domiguez-Bello, Michelle St. Onge, Allison Weakley, Samuel Smits, Gabriela Fragiadakis, Hannah Wastyk, Yoshina Gautam, Dinesh Bhandari, Sarmila Tandukar, Katharine Ng, Guru Prasad Gautam, Jeevan B. Sherchand, and members of the Gardner lab at Stanford. The sequencing depth and breadth of this study was made possible by the Chan-Zuckerberg Biohub. JLS is a Chan-Zuckerberg Biohub Investigator. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding: National Institutes of Health grant DP1-AT009892 and R01-DK085025 (JLS)
National Institutes of Health grant F32DK128865 (MRO)
NSF Graduate Research Fellowship grant DGE-1656518 (DD)
NSF Graduate Research Fellowship grant DGE-114747 (BDM)
Stanford Graduate Smith Fellowship (DD, MMC)
National Institutes of Health training grant 4 T32 AI007328-30 (MRO).
NYUAD Faculty Research Fund (ARJ)

Author contributions: Conceptualization: BDM, MMC, MRO, DD, ARJ, EDS, JLS
Methodology: BDM, MMC, MRO, DD, SJ, JLS
Software: BDM, MMC, MRO, DD, SJ
Investigation: BDM, MMC, MRO, DD
Resources: BY, NN, ARJ

Writing - Original Draft: BDM, MMC, MRO, DD, EDS, JLS

Writing - Review & Editing: BDM, MMC, MRO, DD, ARJ, EDS, JLS

Visualization: BDM, MMC, MRO

Funding acquisition: JLS, EDS, MRO, DD

Supervision: EDS, JLS

Project administration: JLS

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. Metagenomic reads and *de novo* genomes are being submitted to the short read archive (SRA) and GenBank and this manuscript will be updated with additional accessions when the submission is complete.

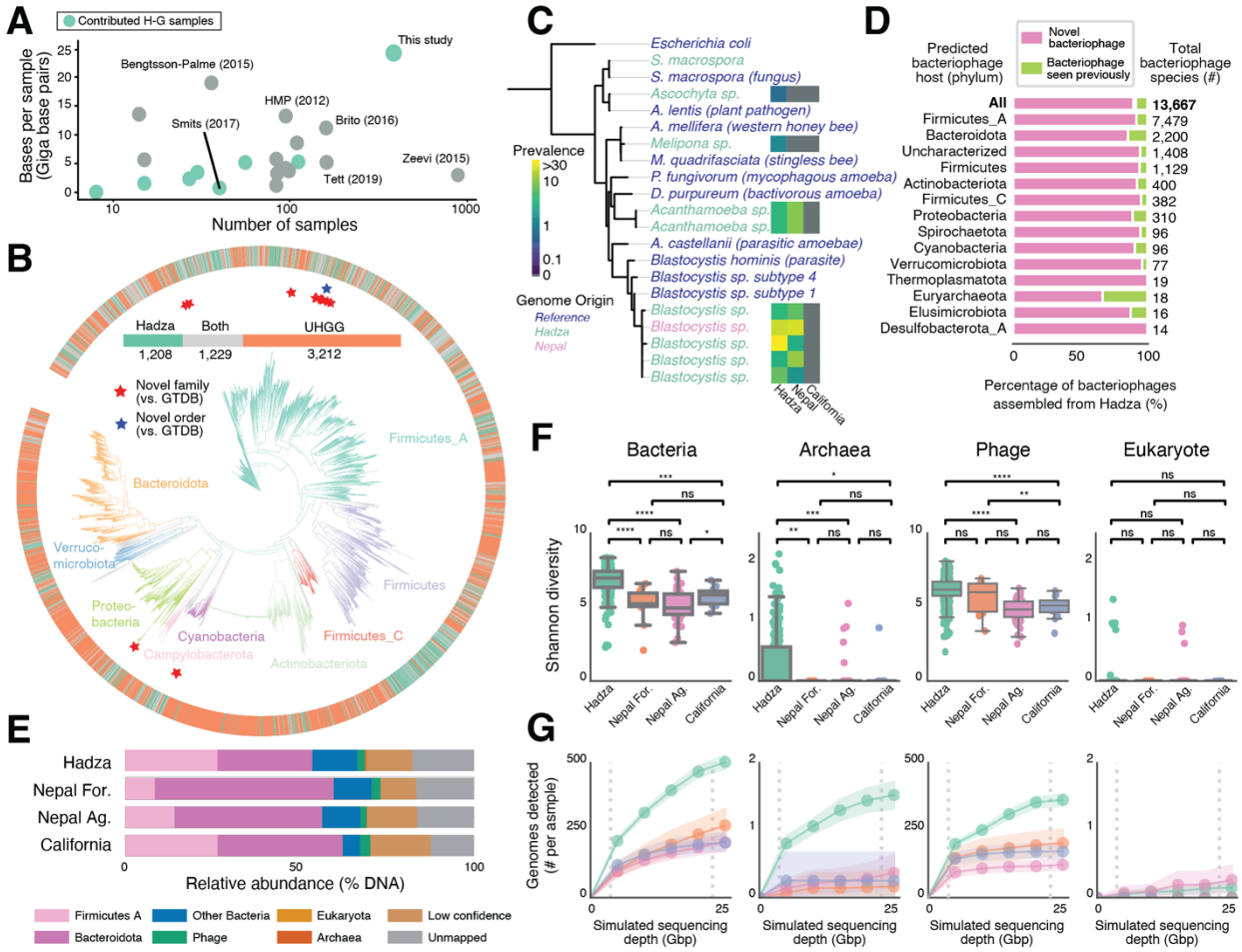


Fig. 1. The Hadza gut microbiota contains substantial multi-domain novelty.

(A) Number of samples versus the number of bases sequenced per sample for 19 previously published metagenomic data sets and the present study. Points are green if they contributed hunter-gatherer (H-G) samples to public databases and gray if they did not.

(**B**) Phylogenetic tree of bacterial species-level representative genomes (SRGs) from Hadza and UHGG based on bacterial single copy gene alignment; branch colors correspond to phyla. SRGs from species-level groups consisting of only genomes assembled from the Hadza or only UHGG are colored green and orange in the outer ring, respectively. The number of SRGs found in the Hadza, UHGG, or both is shown as a horizontal line. Hadza genomes that are novel at the family or order level according to GTDB are annotated with red and blue stars, respectively.

(C) A phylogenetic tree of eukaryotic genomes recovered from the Hadza and Nepali based on universal single copy genes. Public reference genomes are marked with blue text labels. The heatmap shows the prevalence of the individual *Blastocystis* species in the Hadza, Nepali and

Californian cohorts.

(**D**) The percentage of bacteriophage species clusters assembled from the Hadza that are novel at the species level according to MGV, categorized by phylum of the predicted host. Bacteriophages without a host prediction are labeled "Uncharacterized".

(E) The percentage of metagenomic reads mapping to various domains averaged across all metagenomic samples from each population. The phyla "Bacteriodota" and "Firmicutes_A" are shown separated from other bacteria. "Unmapped" depicts the percentage of reads that do not map to any genomes, and "Low confidence" depicts the percentage of reads that map to genomes with less than 50% genome breadth.

(F) The Shannon diversity of bacteria, archaea, bacteriophage, and eukaryote genomes in metagenomes sequenced in this study. P-values from two-sided Mann-Whitney-Wilcoxon test with multiple hypothesis correction; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$, ns: $p \geq 0.05$.

(G) Collectors' curves depicting the average number of genomes detected per sample in each population sequenced in this study after rarefication to various sequencing depths. The vertical dotted lines indicate the average per-sample sequencing depth of this study (~23 Gbp) and the average depth of samples studied in Nayfach et al. (~4 Gbp; ref. [4]). Shaded areas around lines indicate 95% confidence intervals. "Nepal For." includes the Chepang foragers, while "Nepal Ag." includes Raute, Raji, and Tharu agrarians.

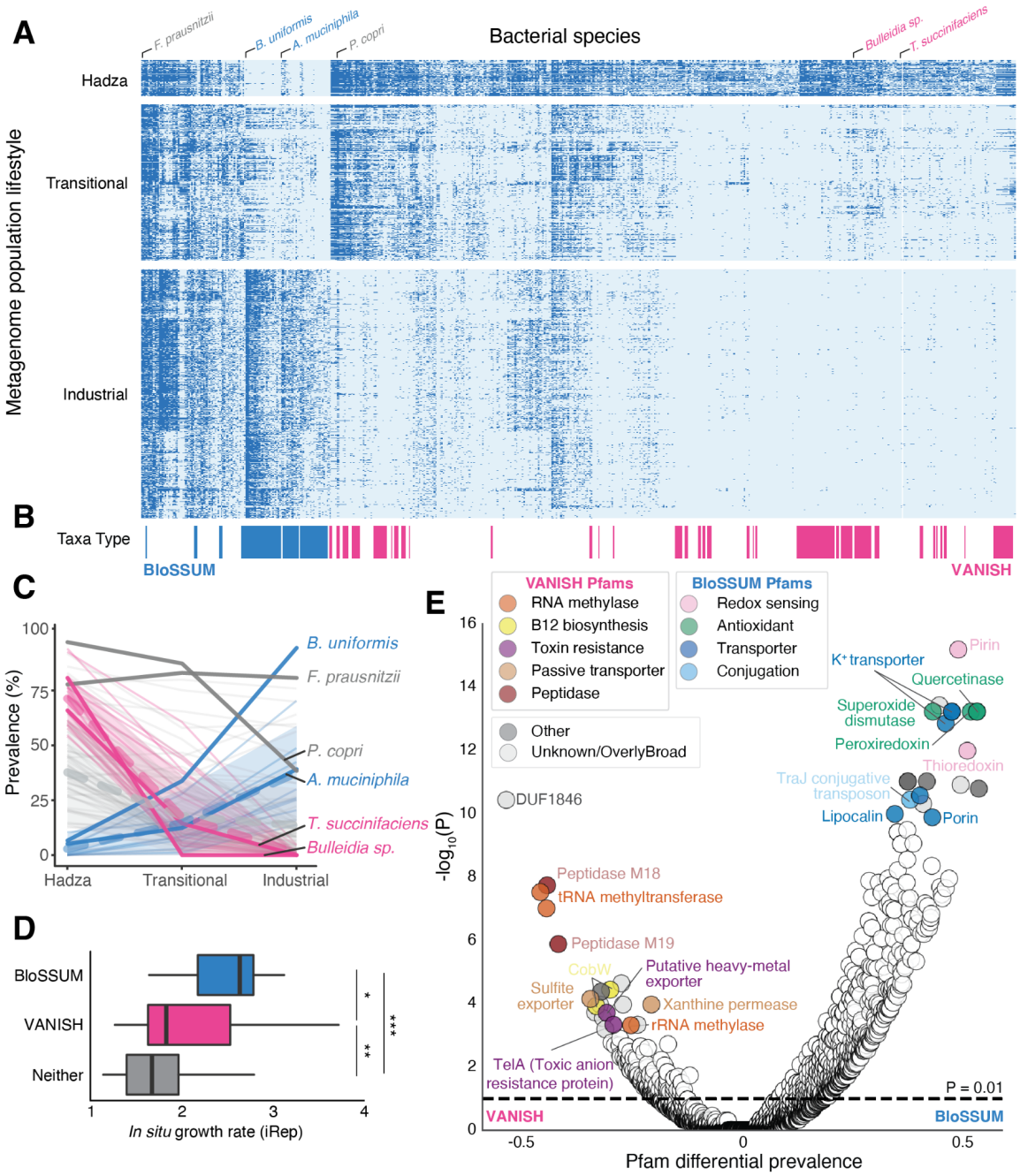


Fig. 2. VANISH and BloSSUM taxa have distinct global prevalence, function, growth rates and covariance with eukaryote detection.

(**A**) A heatmap depicting the presence of 524 SRGs (columns) within metagenomic samples from populations living different lifestyles (rows). Darker blue indicates SRG presence, lighter blue indicates SRG absence. SRGs with >30% prevalence among all samples in any lifestyle category were included.

(B) SRGs were classified as "BloSSUM" or "VANISH" based on their prevalence across

lifestyles (see methods for details). Colored bars correspond to columns in the heatmap.

(C) The prevalence of VANISH (magenta), BloSSUM (blue) and non-enriched taxa (gray) in the Hadza, transitional lifestyle populations and industrial lifestyle populations. Dashed lines connect median prevalence across the taxa in each category surrounded by standard deviation (color shaded regions). Solid lines show the median prevalence for 6 representative taxa in each of these lifestyle groups.

(**D**) The *in situ* growth rate of SRGs in metagenomes from Nepali individuals, stratified by status as "VANISH" (middle), "BloSSUM" (bottom), or neither (top) (* $P \leq 0.05$; ** $P \leq 0.01$; Wilcoxon rank-sum test).

(E) The association of Pfams with VANISH or BloSSUM genomes. The x-axis displays the fraction of BloSSUM genomes a Pfam is detected in minus the fraction of VANISH genomes a Pfam is detected in (Pfam differential prevalence). The y-axis displays the p-value resulting from Fisher's exact test with multiple hypothesis correction.

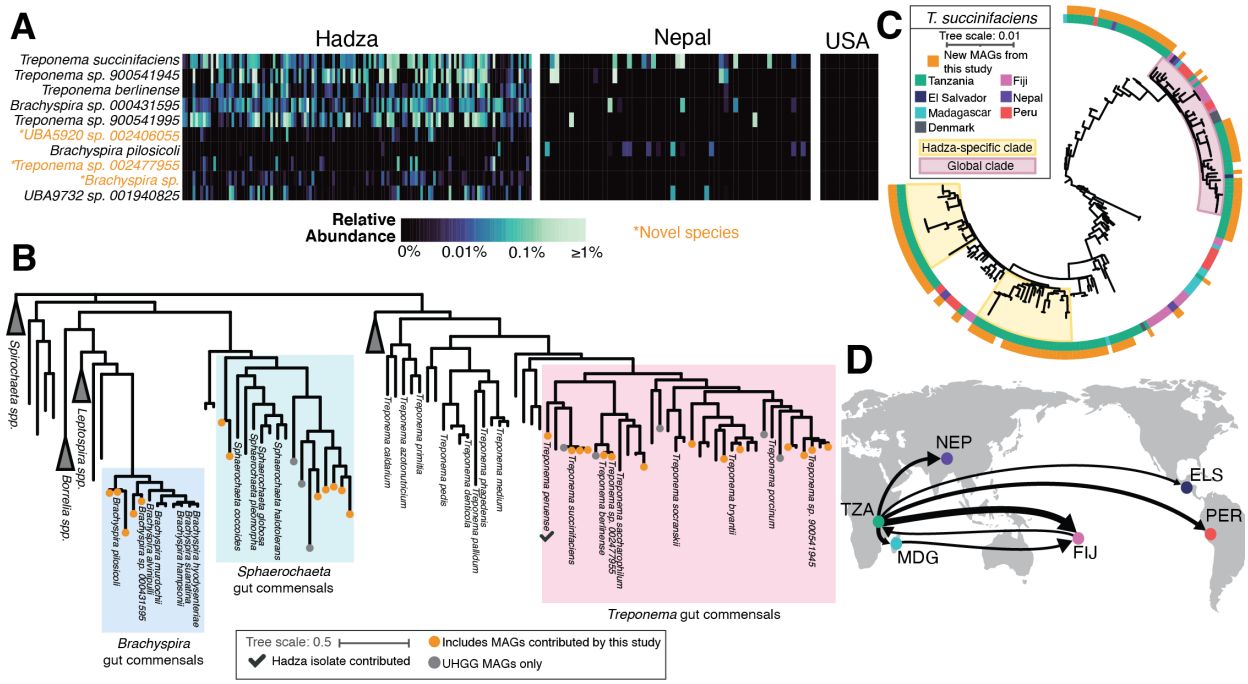


Fig. 3. Spirochaetes that are highly abundant in the Hadza are absent in industrial samples. (**A**) A heatmap showing the relative abundance of the 10 most prevalent spirochaete species in the Hadza, Nepali, and American cohorts. All samples are sequenced to approximately the same sequencing depth.

(**B**) A phylogenetic tree of all spirochaete species using genomes from NCBI, the UHGG and the species-representative genomes added in this study. Clades of commensal organisms representing the genera *Brachyspira*, *Spirochaeta*, and *Treponema* are highlighted.

(**C**) A phylogenetic tree of all *Treponema succinifaciens* MAGs in the UHGG in addition to new MAGs recovered in this study (annotated in outer ring). The inner ring is colored based on the country of origin of the individual contributing the MAG.

(**D**) World map showing locations of populations from which *T. succinifaciens* MAGS were recovered as nodes (TZA = Tanzania, MDG = Madagascar, NEP = Nepal, FIJ = Fiji, PER = Peru, ELS = El Salvador). Arrows indicate the detection of transition events between populations. Thickness of the arrow indicates frequency of the transition event (thickest arrow is Tanzania to Fiji, 17.1%). The top 7 most frequent transition events are shown, accounting for 65.7% of all transitions.

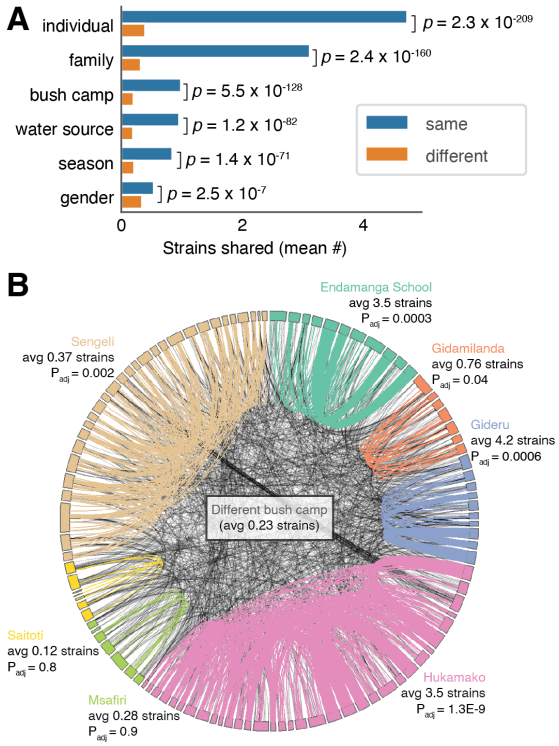


Fig. 4. Microbial strains in the Hadza exhibit are highly shared within bush camps.

(A) The mean number of strains shared between Hadza adults broken down by various types of familial relationships. Exact p-values shown from Wilcoxon rank-sum test.

(B) Rectangles along the circumference represent Hadza individuals and each link drawn between boxes indicates a shared strain. Links between members of the same bush camp are colored based on the bush camp; links between bush camps are colored black. The mean number of strains shared between members of the same bush camp and the p-value comparing strains sharing among members of that bush camp vs members from different bush camps are shown (Wilcoxon rank-sums test).

Materials and Methods

Sample collection

Samples from Tanzania are from 2013-2014 and described previously [2,7]. Permission was obtained from the National Institute of Medical Research and the Tanzania Commission for Science and Technology. For longitudinal samples, one sample from each individual was marked "high_prority" **(Table S1)** and used as noted in statistical analyses that are not robust to multiple samples from the same individual. Nepal samples were obtained previously[28] approved by the Ethical Review Board of the Nepal Health Research Council (NHRC) and the Stanford University Institutional Review Board (IRB). U.S. samples were obtained previously[55] . All human samples were collected after receiving informed consent from participants.

Library preparation and sequencing

Shotgun metagenome sequencing was performed on extracted DNA (MoBio PowerSoil) as described previously[2,7]. Deeper shotgun metagenome sequencing was performed on samples extracted using phenol:chloroform:isoamyl alcohol described previously[2]. 101 Hadza individuals were sampled once and 66 individuals were sampled longitudinally. DNA extraction was performed using mechanical extraction (n=318), phenol chloroform extraction (n=38), or both (n=32).

Libraries were prepared using half-reactions (Nextera Flex), using a minimum of 10 ng of DNA and 6 or 8 PCR cycles to minimize amplification bias using a different 12 base pair unique dual-indexed barcode. Libraries were quantified (Agilent Fragment Analyzer) and size-selected (AMPure XP beads,Beckman), targeting a fragment length of 450bp (insert size 350 bp).

Paired-end sequencing (2x140bp) was performed on a NovaSeq 6000 using S4 flow cells at Chan Zuckerberg Biohub (San Francisco, CA, USA). Samples were randomized across runs and sequenced repeatedly until the target depth was reached. Minimum target depth for each sample was 50 million paired-end reads (~14 Gbp) with a subset of samples sequenced to a minimum target depth of 100 million paired-end reads (~28 Gbp). A total of 8,148 giga base pairs (Gbp) of metagenomic data were generated from 388 Hadza metagenomes (range = 0.7 - 210.3 Gbp, mean = 21.0 Gbp, std dev = 14.5 Gbp), 57 Nepali metagenomes (1,794 Gbp total, range = 14.9 – 84.9

Gbp, mean =31.5 Gbp, std dev = 11 Gbp), and 12 California metagenomes (418 Gbp total, range = 25.2 – 56.8 Gbp, mean = 34.8 Gbp, std dev = 9.2 Gbp) for a total of 10.4 Tbp.

Metagenome quality control and assembly

Raw sequencing reads were demultiplexed and data originating from the same libraries were concatenated prior to analysis. Raw reads were processed using BBtools suite[56]. Exact duplicate reads (subs=0) were marked (clumpify), adapters and low-quality bases were trimmed (bbduk;trimq=16 minlen=55), trimmed reads were mapped (BBmap) against the human genome (hg19) with masks over regions conserved broadly in eukaryotes, and duplicate reads were removed. FastQC[57] was used to ensure read quality. BBMerge was used to merge reads that could be joined unambiguously using the recommended settings (rem k=62 extend2=50 ecct vstrict)[58].

Metagenomes were assembled individually (metaSPAdes[59] ; v3.13) using unmerged forward/reverse and merged reads (-k 21,33,55,77) with error-correction enabled. Assembly size and contig metrics were evaluated (QUAST[60] v5.0) and filtered to contigs >=1500 bp for all subsequent analyses. Gene-calling was performed on all assemblies (Prodigal [61] ;v2.6.3) in metagenome mode.

Strain isolation and genome sequencing

Stool resuspended in PBS was plated on CHG, YCFA (Anaerobe Systems), MRS (Sigma Aldrich), BSM (BBL), Colombia (Anaerobe Systems), BHI (Sigma Alrdich), LKV (Anaerobe Systems), *Treponema* media (DSM Medium 275), and milk-enriched media under anaerobic conditions. Individual colonies were re-streaked and then biotyped on a Bruker MALDI-TOF microflex to determine taxonomy. Colonies were grown in liquid media of the same type as the originating agar plate in anaerobic conditions. For isolating *Treponema*, 0.5% agar was added to the liquid media before making plates. Treponema strains were isolated after removing the top layer of agar to harvest colonies within the agar. Many of these isolated strains are not currently amenable to freezer storage and liquid-culture-based propagation in isolation.

Genomic DNA was extracted (Qiagen DNeasy Blood and Tissue). Libraries were prepared using half-reactions of the Nextera Flex kit, a minimum of 10 ng of DNA as input, 6 or 8 PCR cycles to minimize PCR amplification bias and a different 12 base pair unique dual-indexed barcode. Libraries were quantified (Agilent Fragment Analyzer) and size-selected (AMPure XP beads; Beckman), targeting a fragment length of 450bp (insert size of 350 bp). Paired-end sequencing (2x140bp) was performed on a NovaSeq 6000 using S4 flow cells at Chan Zuckerberg Biohub. Assembly of genomes was performed by trimming using BBduk (trimq=30), normalizing read depth using BBnorm (target=320, min=2), and assembled using SPAdes v3.13.1 (-k 21,33,55,77,99,127)[62]. Genomes were assessed for completeness and contamination using CheckM (1.1.2)[63].

Bacterial and archaeal genome recovery and refinement

A novel "co-mapping" approach was developed to leverage contig depth information from multiple, closely related samples and improve genome bin recovery from single-sample assemblies. MASH sketches (-s 1000000 -k 32 -m 2)[64] were created from reads in each metagenome individually, and sketches were compared in a pairwise manner. For each assembly, reads from that sample and the nine next-closest related samples by MASH distance were mapped (Bowtie2[65];--very-sensitive -X 1000) and genome bins generated using contig depth for all 10 samples (MetaBAT2[66];v2.15, default settings). For California samples, only samples taken from the same individual were co-mapped. Genome bin quality was assessed using CheckM (v1.1.2)[63] and anvi'o[67] (v6.3).

Bins were refined using MAGpurify v2[68] (using weighted mode for gc_content, tetra_freq, and coverage). The database used by Nayfach et al.[68] for conspecific analysis was augmented by adding all bins that were >=95% complete and <=5% contaminated (CheckM and anvi'o). For each species-level group, only the highest-quality genome bin for each individual was included. Flagged contigs were removed. Rarely, a module suggested the removal of >25% of a bin's length, and in such cases that module was turned off. Genomes with ≥50% completeness and <10% contamination according to CheckM were retained, in accordance with MIMAG standards[69].

Evaluation of self- and co-mapping relative to isolate genomes

Isolate genomes from Hadza stool samples were de-replicated (dRep v3.2.2;-s 100000, -sa 0.99). The highest scoring isolate as the representative when multiple isolates from the same secondary cluster were isolated from the same sample. 19 representative isolates were identified from samples that also had metagenome sequencing, assembly, and binning. Representative isolates and bins (>=50% complete, <5% contamination) generated using self-mapping and co-mapping were compared (MASH;-s 100000), selecting most similar bin MASH distance <0.05), with co-mapping and self-mapping recovered 17 and 10 bins representing isolates, respectively, with no significant differences in quality.

Creating bacteria / archaeal species-level genome database

Bacterial and archael genomes sharing ≥ 95% average nucleotide identity (ANI) over 30% of their length were considered the same species[70]. Species-level groups were determined using dRep (v3.0.0[71] ;--S_algorithm fastANI --multiround_primary_clustering --clusterAlg greedy -ms 10000 -pa 0.9 -sa 0.95 -nc 0.30 -cm larger) based on the ANI between all genomes within each species-level group. Each genome was assigned a "centrality" score according to its average ANI to all other genomes in the group. The highest score genome was chosen as representative for each species-level group using the formula: score = (1*completeness) - (5*contamination) + (0.5*log10(ctg_N50)) + (1*log10(contig_bp)) + (2*(centrality-0.95)*100).

Centrality was calculated between all genomes in the UHGG genome database (v1.0) using the species-grouping[4], and species representatives were chosen as above. Representatives from *de novo* genomes generated here and from the UHGG database (v1.0) were compared (dRep;--S_algorithm fastANI --multiround_primary_clustering --clusterAlg greedy -ms 10000 -pa 0.9 -sa 0.95 -nc 0.30 -cm larger). Representatives for each species-level group were chosen using the formula: score = (1*completeness) - (5*contamination) + (0.5*log10(ctg_N50)) + (1*log10(contig_bp)). Representatives were compared using the same dRep command, and winners were chosen using the same scoring criteria. Species-level group membership was back-propagated to the original bins.

Annotating bacteria / archaeal genomes and assessing genomic novelty

Taxonomy was determined for all species-level representative genomes using GTDB (v95)[72]. Novelty against UHGG were determined based on the species-level clustering described above. Only genomes that pass both the MIMAG genomic standards used in this study (≥50% completeness and <10% contamination) and the standard used during UHGG creation (completeness - (5*contamination) > 50) were considered in comparisons against UHGG. Species groups containing only genomes recovered from the Hadza were considered novel relative to UHGG.

A phylogenetic tree was made (GtoTree (v1.5.36)[73] with bacterial gene sets (-H Bacteria). All other settings were default. The tree was visualized using iTol[74] with taxonomy provided by GTDB.

Eukaryotic genome recovery and analysis

EukRep (v0.6.6)[75] was employed on all assemblies (default settings) and if a genome bin was both >5 mega base pairs and >80% eukaryotic according to EukRep, it was called eukaryotic. EukCC (v1.1)[76] was run on eukaryotic bigs using database eukcc_db_20191023_1

Proteins identified via EukCC were compared against UniRef100[77] (downloaded 3/5/2020) using DIAMOND[78] with a maximum e-value of 0.0001 (blastp -f 6 -e 0.0001 -k 1). The resulting taxonomy was parsed with tRep (https://github.com/MrOlm/tRep/tree/master/bin)[79]. Eukaryotic genomes with the same species-level taxonomy that originated from the same metagenomic sample were presumed to be from the same organism, were merged into a single file and re-analyzed using EukCC and tRep.

Phylogenetic tree was created (GToTree; v1.5.36)[73] "GToTree -H Universal_Hug_et_al -j 4 -B -c 1 -t") with a custom set of public reference genomes. Tree was visualized using iTol[74].

Creating eukaryotic species-level genome database

To identify eukaryotic species that may be present in the metagenomics sequenced in this study and which did not have genomes recovered using the pipeline described above, we ran the program EukDetect[80] on all metagenomes sequenced in this study. Five species were detected in

at least two samples with "perecent_observed_markers" ≥ 50, and reference genomes for these five species were included in the eukaryotic species-level genome database. In addition to these five genomes, the highest quality representative genome from each of the seven species of eukaryotes recovered in this study was included in the eukaryotic species-level genome database.

Metagenome reads were mapped onto the eukaryotic species-level genome database (Bowtie 2[65]) and the resulting mappings were processed (inStrain quick_profile; v1.2.14[24] and CoverM v0.4.0 (https://github.com/wwood/CoverM)). A species was "present" if the breadth of coverage according to inStrain exceeded 0.1.

Viral genome recovery

CheckV[81] (version 0.8.1, end-to-end mode, database v1.0) was run on all assembled contigs >=1500bp. Contigs predicted to contain one or more proviruses were run iteratively through CheckV (up to 5 rounds) until CheckV assumed the remaining region was viral. For provirus iterations only yielding an HMM-based completeness estimates, the most complete fragment was selected and excised from the parent contig. For provirus iterations with AAI (Average Amino acid Identity)-based completeness predictions, the fragment with the length closest to expected length was selected and excised from the parent contig. Viral contigs were passed through the MGV viral detection pipeline[5] and Bacphlip (v0.9.6) was run to assign a lytic and temperate score[82].

Creating bacteriophage species-level genome database

The 40,171 viruses recovered in this study were clustered into species-level groups as described previously[5] (blastn --min_ani 95 --min_qcov 0 --min_tcov 85, https://github.com/snayfach/MGV/tree/master/ani_cluster), and the longest viral contig in each cluster was selected as the representative. To measure novelty versus MGV, the 16,899 species-level representatives were subsequently clustered with the 54,118 MGV cluster representatives into species-level groups using the same method, and clusters without an MGV genome were considered novel.

Viral host prediction

Host prediction was performed on the 40,171 viruses as described previously [5]. Briefly, CRISPR spacers were identified (PILER-CR[83] and CRT[84]) And BLASTN[85] used to search viruses for CRISPR spacers identified from bins reported here and UHGG v1 (-dust no -word_size 18). CRISPR spacer hits were retained if there was a maximum of one mismatch or gap over >=95% of the spacer length. Additionally, hs-blastn[86] was used to identify >=1kb and >=96% DNA identity hits between all UHGG and newly-recovered genomes and viruses reported here. All viral connections to host genomes were aggregated, and host taxonomy was assigned based on the lowest host taxonomic rank that had >70% agreement across CRISPR or BLASTN.

Characterizing diversity

Reads from all metagenomes generated here were mapped to the bacterial/archaeal, bacteriophage, and eukaryote species-level genome databases (Bowtie 2[65]). Resulting mappings were processed (inStrain quick_profile; v1.2.14[24] and CoverM v0.4.0 (https://github.com/wwood/CoverM)). Prokaryotes where the representative genome was detected at ≥ 0.5 breadth (i.e. at least half of bases were covered by at least 1 read) were considered present. Bacteriophages and eukaryotes breadth thresholds were 0.75 and 0.1, respectively.

Relative abundance (% DNA) was calculated as (# reads mapping a genome / total # reads in metagenome). Shannon diversity was calculated based on relative abundance (% DNA) values (scikit-bio (http://scikit-bio.org)) .

Rarefaction analysis

In silico rarefaction was performed on samples sequenced to ≥ 50 Gbp using the InStrain auxiliary script "rarefaction_curve.py" (v0.3.0) (https://github.com/MrOlm/inStrain/blob/master/auxiliary_scripts/rarefaction_curve.py) on a .bam file of reads mapped with Bowtie 2[65]. For other rarefaction curves **(Fig. 1G, Extended Data Fig. 4C)** an alternative *in silico* rarefaction technique was used. Genomes with < 50% breadth were removed from the analysis, and for each rarefaction level 1) a scaling threshold was established based on the total sequencing depth (scaling factor = rarefaction depth / total sequencing depth), 2) scaled genome coverage was calculated by each genome by multiplying

un-rarefied coverage by this scaling factor, and 3) genomes with scaled coverage ≥ 1 were considered detected.

Collating previously published human gut metagenomic samples

Prevalence of microbial species across lifestyle was characterized using a curated collection of 2122 metagenomes including samples from industrial[29–32,34,40,87,88], transitional[15,17,35–38,40,41,89], and hunter-gatherer populations[36,38,42–44]. Samples were binned using the U.N. Human Development Index (HDI)[90,91]. Samples from individuals < 3 years old were excluded. For longitudinal samples, a single sample was randomly selected resulting in 137 Hadza samples. Reads were processed as described above. Samples with fewer than 60 genomes detected were excluded.

Hadza sample ERR7803603, sequenced to a depth of 210 Gbp, was determined to be the deepest human gut metagenome sequenced as of 28 Feb 2022 by downloading all summary metadata from NCBI SRA with the search term "(txid408170[Organism:noexp]) AND WGS[Strategy]" and sorting by decreasing base pairs sequenced.

Species prevalence analysis

All reads generated here and publicly available were mapped to the bacterial/archaeal species-level genome database (Bowtie2[65]), and resulting mappings were processed using inStrain quick_profile (v1.2.14)[24] and CoverM v0.4.0 (https://github.com/wwood/CoverM)). Species detected at ≥ 0.5 breadth were considered present and prevalence was calculated as the percentage of metagenomes in which the species was present.

Genomes were assigned to VANISH or BloSSUM using p-values resulting from Fisher's exact test on the following contingency table: [[(# Hadza samples where genome is found, # industrial samples where genome is found), (# Hadza samples genome is not found, # industrial samples where genome is not found)]]. All p-values were ranked and a percentile score was assigned. Genomes in the 95th percentile or greater where Hadza prevalence was higher were "VANISH" taxa. Those in the 95th percentile or greater where industrial prevalence was higher were "BloSSUM".

Heatmaps displaying species prevalence data were created using the R package "pheatmap" (v1.0.12). Principal coordinate analysis was performed on the species prevalence data using the vegdist function in the package "vegan" [92] (v2.5-6) and the function cmdscale from the package "stats" (v4.0.4).

Growth rate analyses

InStrain profile (v1.2.14) (26) was run on all .bam files created as described in the "species prevalence analysis" section. All iRep values for genomes with ≥ 50% genome breadth and with values < 5 were considered valid. Seasonality of iRep values was plotted using seaborn v0.11.1[93] "lineplot" with the default estimator (mean) and 95% confidence interval for error bars.

Blastocystis analysis

Presence or absence of each *Blastocystis* MAG was determined as described above. The top two most prevalent *Blastocystis* MAGs were most closely related to *Blastocystis ST1* and *Blastocystis ST4*, respectively(tRep; https://github.com/MrOlm/tRep/tree/master/bin)[79]. Wilcoxon rank sum test was used to determine if presence of a *Blastocystis* genome was correlated with total relative abundance of VANISH taxa and BloSSUM taxa separately. Linear discriminant analysis was performed using the "lda" function from the package MASS (v7.3) to determine the effect size of each association.

Seasonality analysis

Principal coordinate analysis was performed on the Bray-Curtis distance between all Hadza samples in our study. Relative abundance was aggregated at the taxonomic level of family to mirror initial analysis done in Smits, et al.[2]. The adonis function in the R package "vegan" was used to test significance by season. Subject ID was used as a sub-stratum. A Wilcoxon rank-sum test was used to determine whether samples varied in composition along the major axis of variation, aggregated by season.

The average relative abundance of each species-level group in our bacterial/archaeal species-level genome database was calculated for each sub-season. Taxa that observed cyclical

abundance over the course of a year was determined (Kruskal-Wallis test; p-values were Bonferroni-adjusted to control for multiple hypothesis testing).

CAZyme annotation was performed using dbCAN_v9 HMMs[94] (http://bcb.unl.edu/dbCAN2/download/Databases/V9/dbCAN-HMMdb-V9.txt). Proteins were searched against the HMM collection using hmmscan[95] and filtered using the "hmmscan-parser.sh" script provided with dbCAN2. Seasonal CAZyme analysis was performed using previously described seasonal delineations[2].

Protein clustering and novelty assessment

Predicted proteins were clustered at 95% identity (MMseqs2[96] ; v12.113e3; easy-linclust --cov-mode 1 -c 0.8 --kmer-per-seq 80 --min-seq-id 0.95 --compressed 1). Novelty relative to UHGP-95 (v1.0)[4] was determined by clustering together UHGP-95 with our *de novo* representative proteins (MMseqs2) and back-propagating to the initial *de novo* clustering to calculate the number of protein clusters assembled from each lifestyle **(Extended Data Figure 2)**. Representative proteins were also compared against UniRef100 using DIAMOND[78]. Novel proteins were defined when the representative protein was not related to any protein in the UniRef100 database with ≥ 95% amino acid identity.

Protein annotation

Proteins were annotated (Pfams (v32)[97] ;hmmsearch [95]), filtered (hmmsearch --cut_ga --domtblout), and protein domain overlap was resolved (cath-resolve-hits.ubuntu14.04[98]; --input-format hmmer_domtblout --hits-text-to-file).

Pfam enrichment analysis

For each Pfam, the number of VANISH and BloSSUM genomes with at least one gene containing a Pfam was recorded as "c1" and "c2", respectively. Pfams found more often in one genome set or the other were detected using a Fisher's exact test on the following contingency table: [[c1, (# VANISH genomes) - c1], [c2, (# BloSSUM genomes) - c2]]. Multiple hypothesis correction was performed using the FDR method[99]). Pfam differential prevalence was calculated as (c2 / (# BloSSUM genomes)) - (c1 / (# VANISH genomes)).

Spirochaetes analysis

Spirochaete genomes from the bacterial/archaeal species-level genome database and NCBI were de-replicated (dRep; --S_algorithm fastANI -ms 10000 -pa 0.9 -sa 0.95), and a phylogenetic tree was generated (GtoTree;v1.5.36)[61,73,95,100–102]) from bacterial (-H Bacteria) gene sets. All other settings were default. The tree was visualized using iTol[74] and colored by taxonomy provided by GTDB.

A tree of *Treponema succinifaciens* in the bacterial/archaeal species-level genome database was generated using GtoTree; v1.5.36)[61,73,95,100–102] with IQ-Tree[103] from bacterial (-H Bacteria) gene sets (completeness threshold 75% with "-G 0.75"). We used country-of-origin information (re-coded as continent-of-origin) as a trait of each genome to measure the degree of phylogenetic signal in the geographic spread of the MAGs ("delta" function from Borges, et al.[49]). *P*-value of the delta statistic was performed using 100 calculations with randomly permuted tree tip labels.

Stochastic character mapping of *Treponema succinifaciens*

Stochastic character mapping was performed using SIMMAP via the "make.simmap" function ("phytools" R package[104]). We applied the character mapping on the marker-based tree of *T. succinifaciens* GToTree generated MAGs (described above). "Country of origin" of each MAG served as a trait and inferred ancestral character states on phylogeny (equal rates model, repeated 100 times to calculate average # of character changes and direction of host transfer events).

Pfam pN/pS analysis

The *pN/pS* was calculated using inStrain (v1.2.14) (inStrain profile --database_mode)[24] on mappings to the bacterial/archaeal species-level genome database, using the predicted genes. All genomes detected with < 80% breadth were excluded from analysis. For remaining genomes, genes with "SNV_count" < 5 were excluded. If <10 genes in a genome fit this criteria, the genome was excluded. Genes with ≥ 5 "SNV_count" and a blank "pNpS_variants" value were assigned a "pNpS_variants" of 100. Genes were sorted according to "pNpS_variants", and genes in the top and bottom 10% of "pNpS_variants" were recorded. How many times each Pfam was detected on any genes that passed the above filters ("trial_count") and how many times the Pfam

was in genes in the top and bottom 10% of genes based on “pNpS_variants” (“top_success_count”, “bottom_success_count”) was noted.

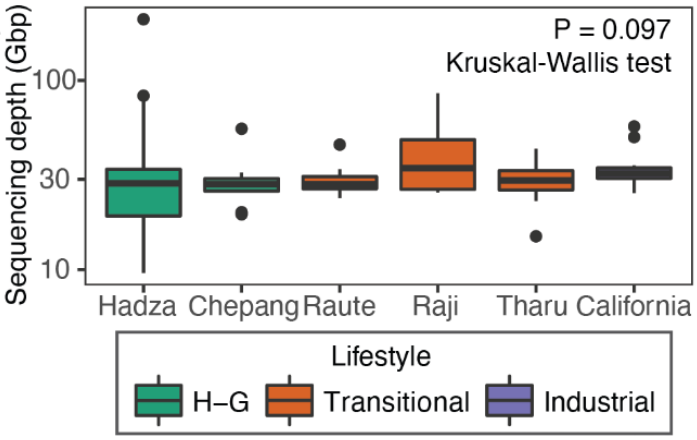
To determine Pfams in the top or bottom 10% of “pNpS_variants” more often than expected by chance, genes detected in less than 5 samples were excluded, the number of times a gene was in the top 10%, bottom 10%, and seen total was scaled (“trial_count”/5), and the scaled “top_success_count”, “bottom_success_count”, and “trail_count” values were summed together. Probability that the “top_success_count” or bottom_success_count” was due to random chance was calculated using binomial statistics (Python Scipy[105]). P-values reported as 0 were set to 1E-300 and multiple hypothesis correction was performed (FDR [99]). Mean Pfam pN/pS was calculated as the average “pNpS_variants" of all genes on genomes with ≥ 80% breadth and a non-blank “pNpS_variants” value.

The procedure described above was repeated using “coverage” instead of “pNpS_variants” to detect Pfams associated with genes with higher or lower coverage than others. To avoid mis-mapping (recruiting genes from other populations), all Pfams with uncorrected p-values < 0.01 were excluded from the “pNpS_variants” analysis.

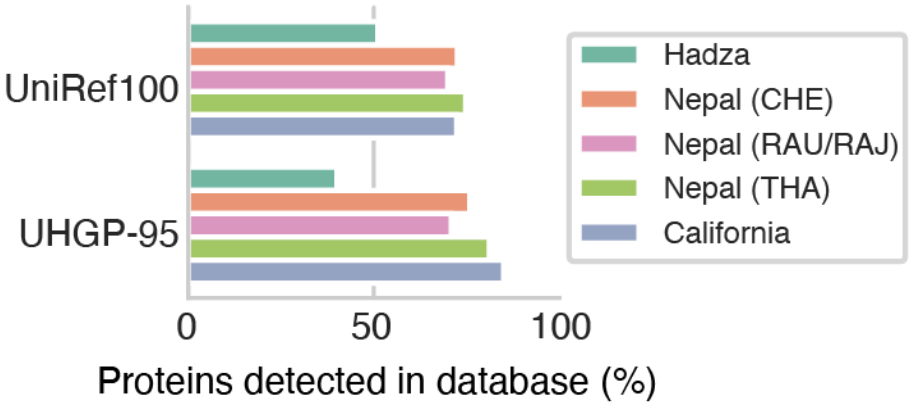
Strain sharing analysis

Genome detection was defined as minCov breadth ≥0.5 (i.e. at least half of bases were covered by at least 5 reads) as measured using “inStrain profile”. Each species detected in more than one individual was compared using inStrain compare. Where a genome was detected in more than 120 samples, samples were divided into groups of equal size such that no group had more than 120 samples, and “inStrain compare” was then run on each group. A distance matrix was created for each species based on resulting popANI values and used to cluster each species into individual strains using “average” hierarchical clustering with a threshold of 99.999% popANI (Scipy cluster). Strains shared between sample pairs were calculated based on this strain definition, and P-values were calculated only considering pairs of samples in which both samples were from Hadza adults.

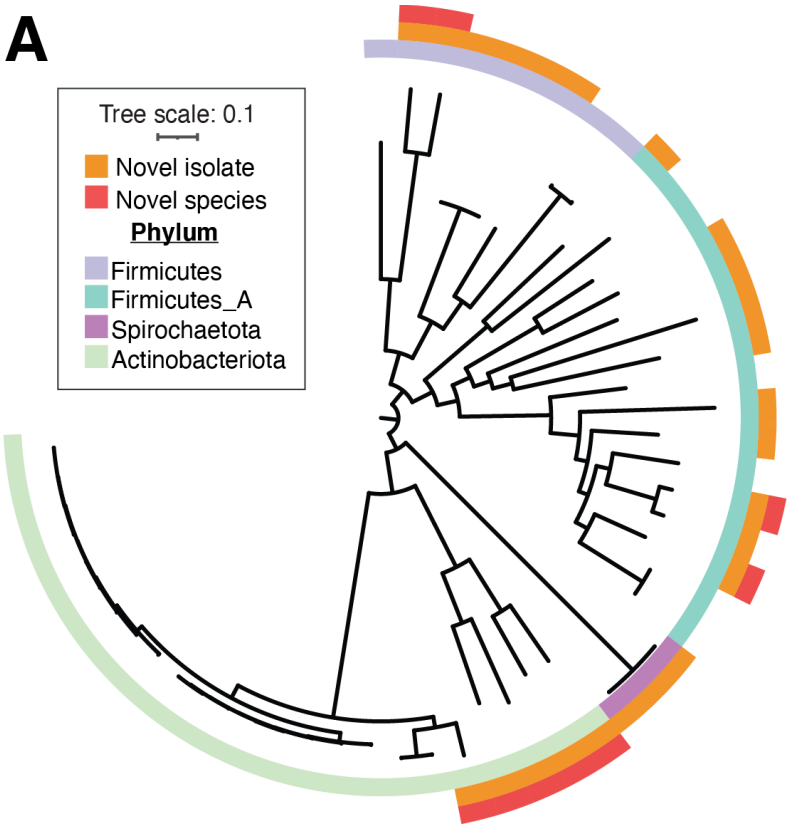
Extended Data Figures



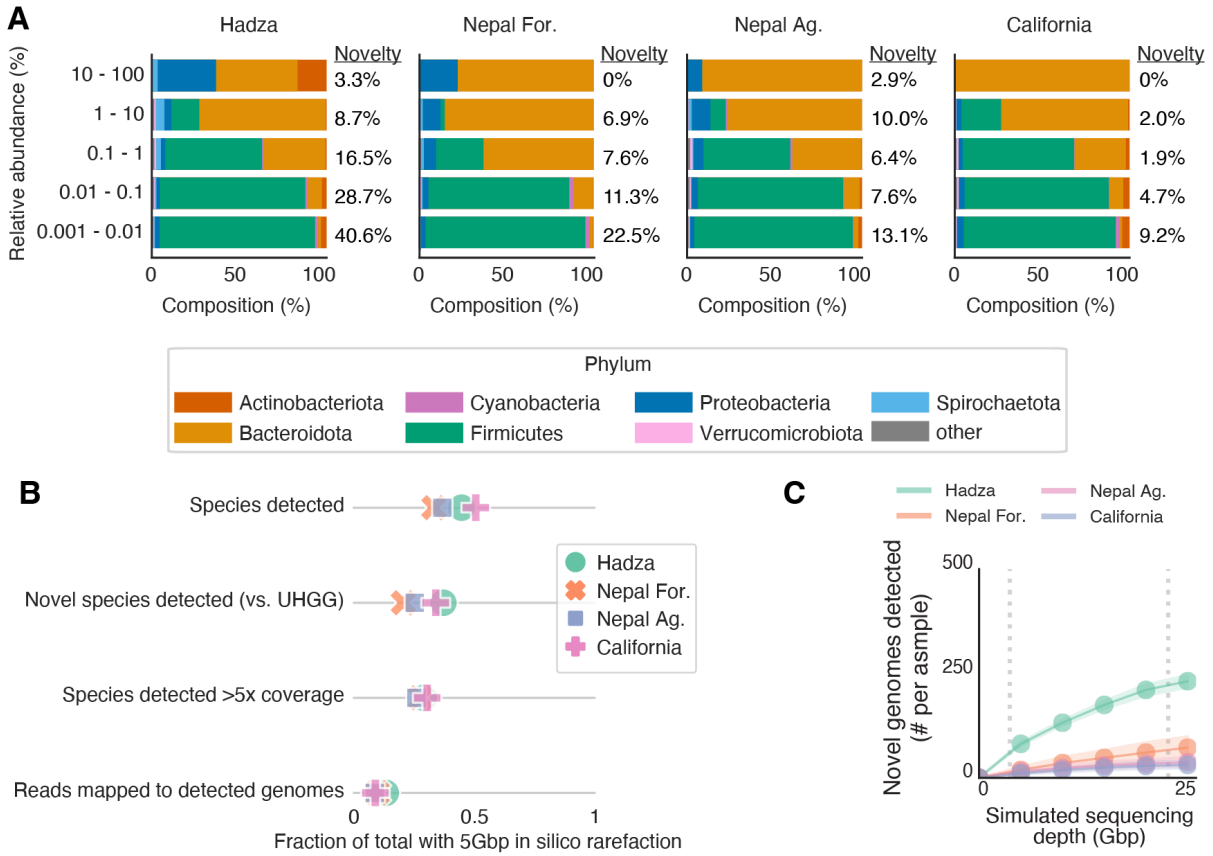
Extended Data Figure 1. The median metagenomic sequencing depths of populations sequenced in this study.
A box plot showing the distribution of sequencing depth, in giga base pairs (Gbp) for each of the populations sequenced in this study. The Chepang foragers and Raute, Raji, and Tharu agrarians are the Nepali populations. The populations do not differ significantly by sequencing depth (P = 0.097, Kruskal-Wallis test).



Extended Data Figure 2. The Hadza gut microbiome has extensive functional novelty. For each population in this study, the percentage of predicted proteins from recovered genomes that are present in the UniRef100 and UHGP-95 protein databases

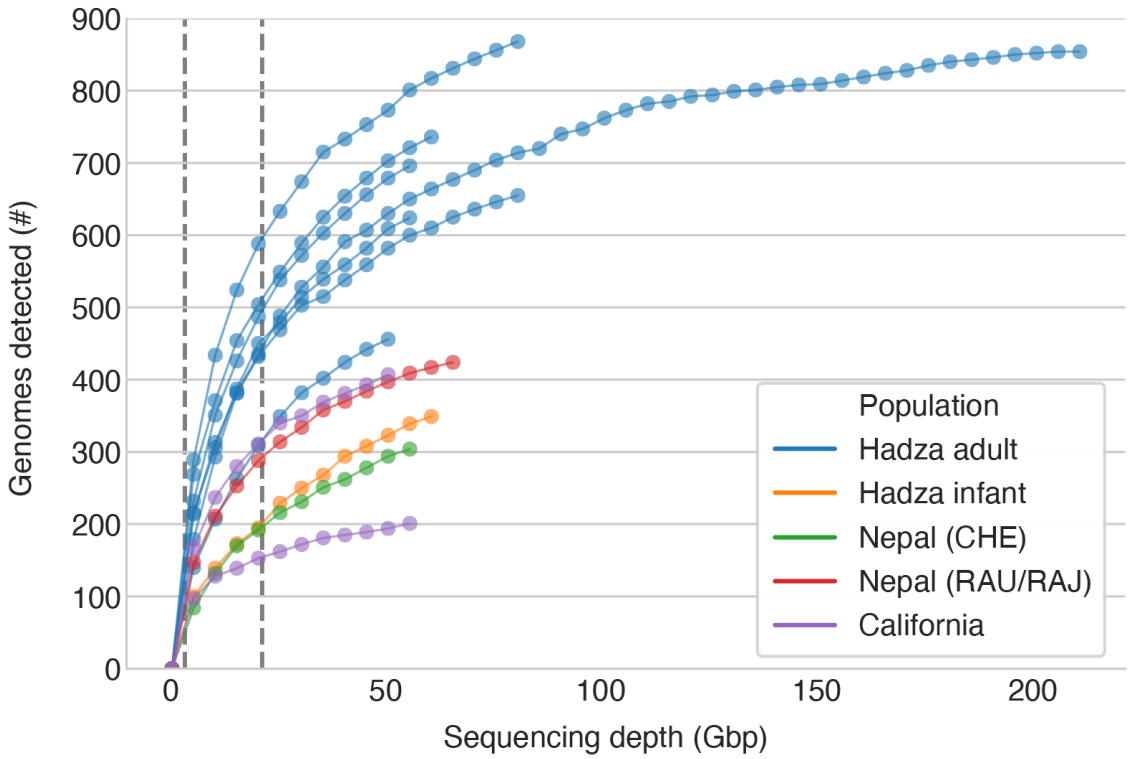


Extended Data Figure 3. Phylogenies of strains isolated from Hadza stool samples.
(A) A phylogenetic tree of all isolate genomes sequenced in this study. The tree is decorated with phylum of each species (inner ring), whether the species is newly isolated for the first time (middle ring) and whether the species is novel relative to UHGG (outer ring).

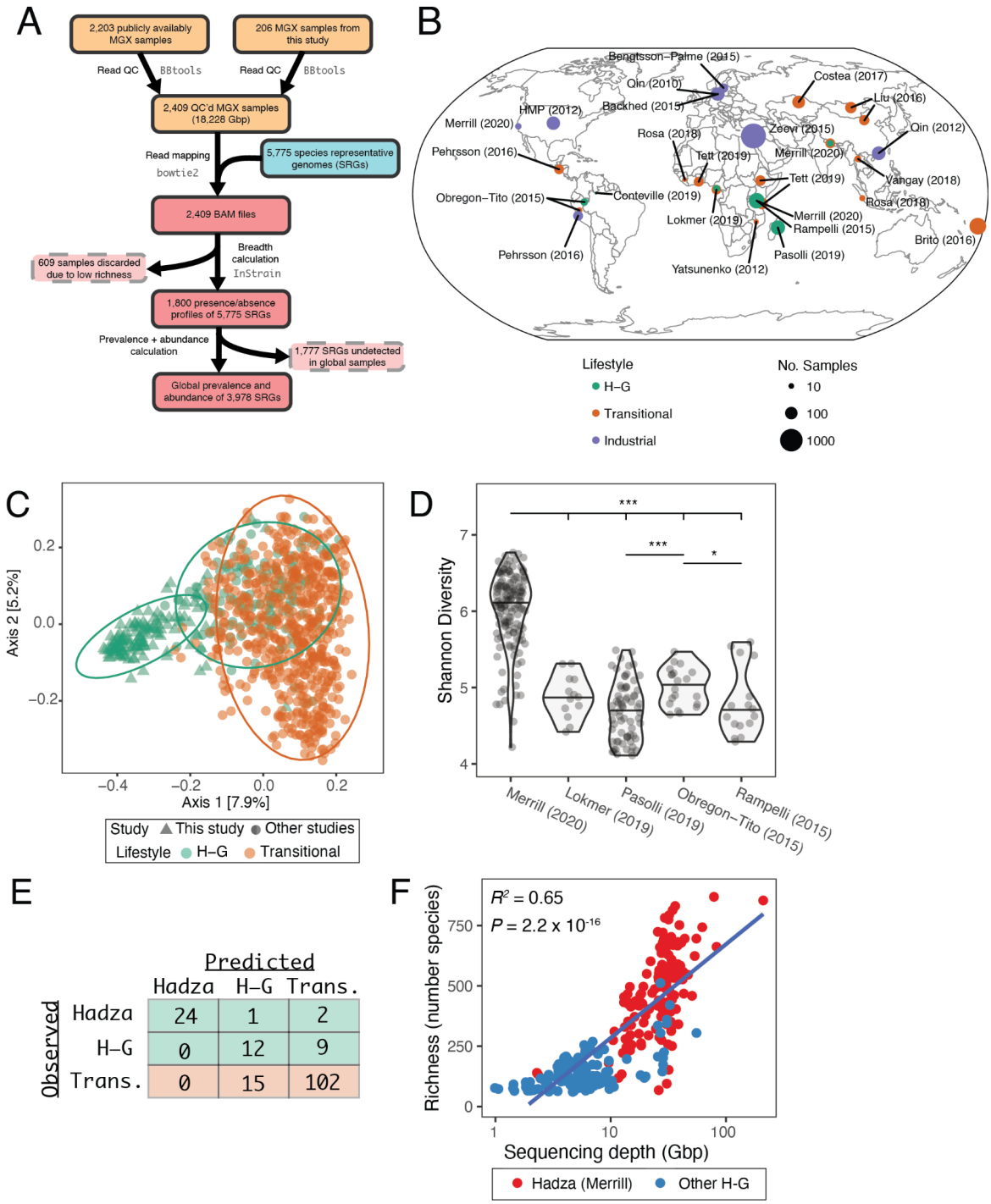


Extended Data Figure 4. Increased sequencing depth results in the detection of novel and phylogenetically distinct taxa.

(**A**) Taxonomic distribution of organisms present at different ranges of relative abundance levels (horizontal stacked bar plots) and the percentage of species that are novel according to GTDB (text percentages right of horizontal bars). Organisms detected at low relative abundance levels are more likely to be novel than those that are more abundant. **(B)** A depiction of how metrics would be different if 5 Gbp (approximately the average depth sequenced in previous large-scale genome binning studies) had been sequenced in the study rather than ~25 Gbp. 'Nepal For.' includes the Chepang foragers, and the 'Nepal Ag.' group includes Raute, Raji, and Tharu agrarians. **(C)** Collectors curves show the number of novel bacterial genomes detected in the four populations we sequenced by limiting sequencing depth of all samples to 5, 10, 15, 20, and 25 Gbp. The vertical dotted lines indicate the average per-sample sequencing depth of this study (~23 Gbp) and the average per-sample sequencing depth of Nayfach et al. (~4 Gbp; ref. [4]). Shaded areas around lines indicate 95% confidence intervals. "Nepal For." includes the Chepang foragers, while "Nepal Ag." includes Raute, Raji, and Tharu agrarians.



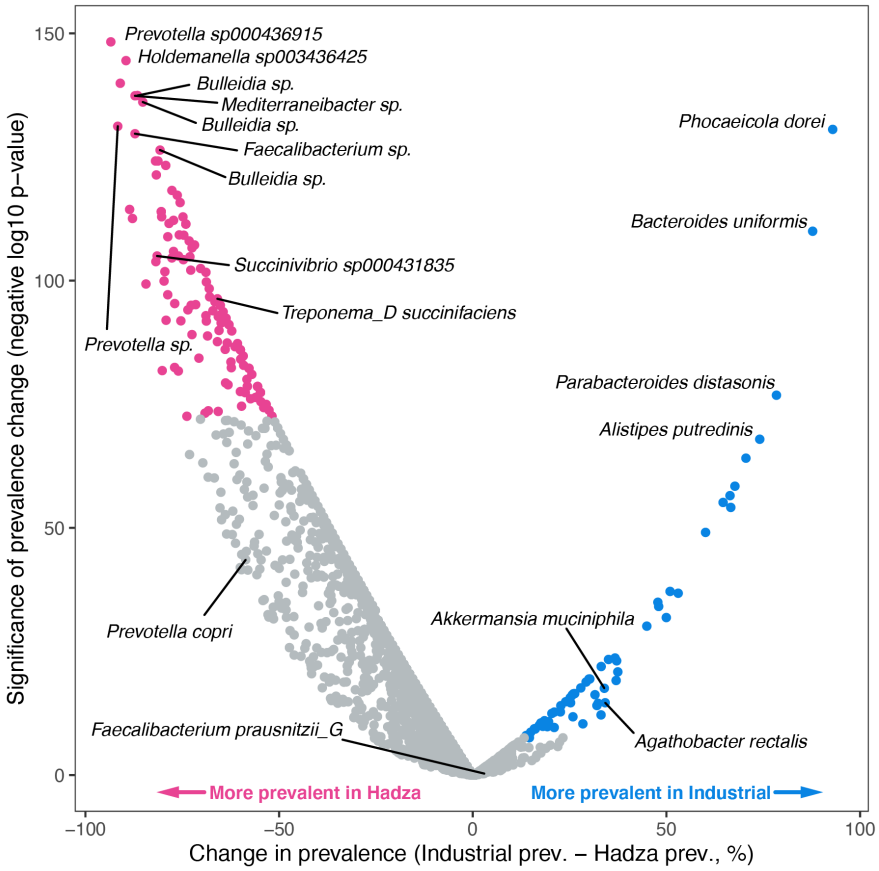
Extended Data Figure 5. Rarefaction analysis of deeply sequenced samples.
The number of genomes detected (breadth > 0.5) in individual samples sequenced in this study by limiting sequencing depth to 5 Gbp increments. Each line represents an individual sample from which ≥ 50Gbp of trimmed, filtered reads were generated. Lines are colored by the population of the individual that gave the sample. The vertical dotted lines indicate the average per-sample sequencing depth of this study (~23 Gbp) and the average per-sample sequencing depth of samples used in Nayfach et al. (~4 Gbp; ref. [4]).



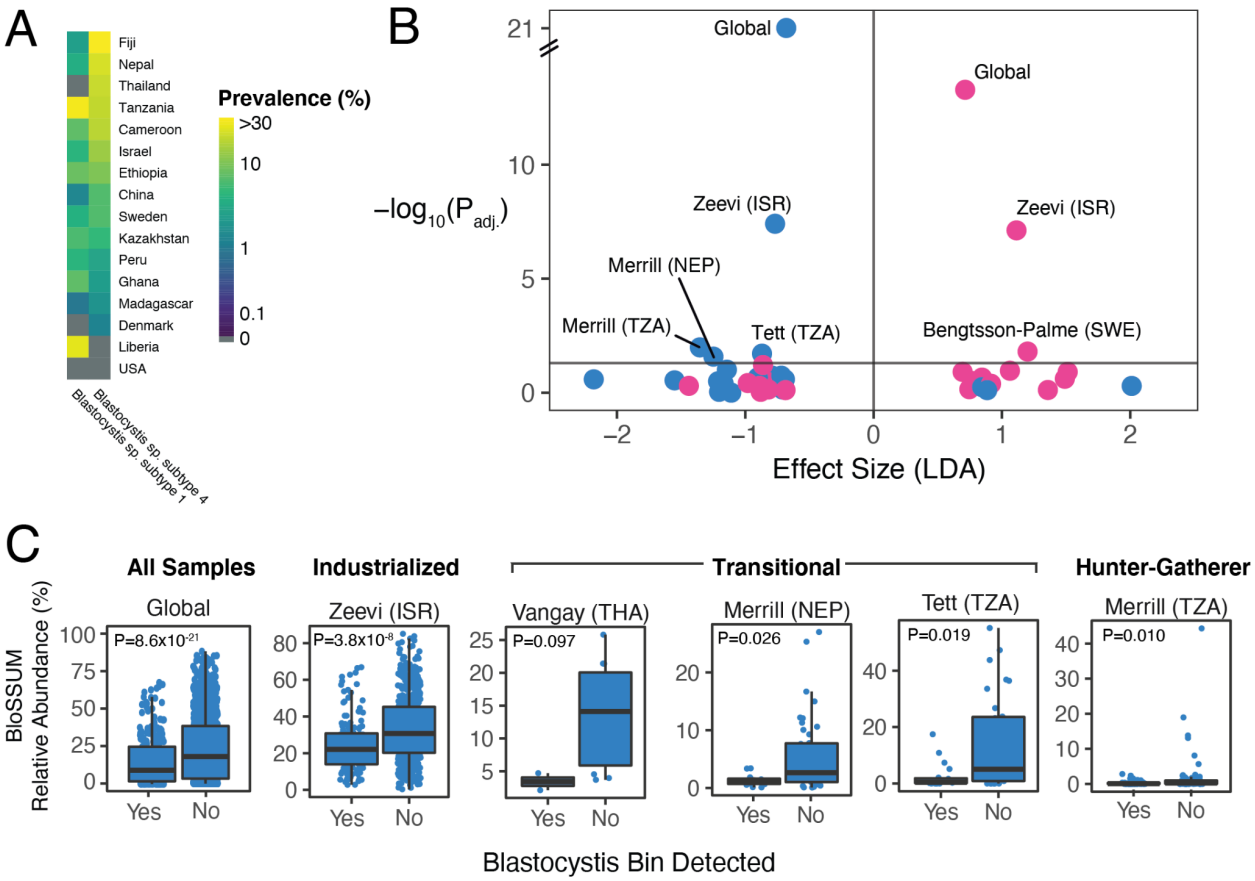
Extended Data Figure 6. Global metagenomics data set and analysis of publicly available hunter-gatherer samples.

(A) A flowchart showing the computational pipeline used to analyze global metagenomics samples. **(B)** A world map showing the geographic locations of global metagenomics samples. Dots are colored based on the lifestyle of the study population and the size of the dots indicate the number of samples contributed by that population. 'H-G', Hunter-Gatherer. (**C**) PCoA plot of H-G (green) and transitional (orange) samples in our global metagenomics data set. Triangles are

samples sequenced in this study (Hadza and Nepali samples). Circles are samples from other studies. Distance matrix was generated with Jaccard similarity between samples. (**D**) Shannon diversity of H-G samples in our global metagenomics data set. Significance between groups was calculated using Wilcoxon rank-sum test (*** $P < 0.001$, * $P < 0.05$). (**E**) Confusion matrix from a random forest classifier built to predict the lifestyle of Hadza samples from this study and H-G and transitional lifestyle samples from publicly available studies. 100% of Hadza samples were classified as Hadza, H-G samples were correctly classified 53% of the time and transitional samples were classified correctly 91% of the time. (**F**) Scatter plot showing sequencing depth versus richness (number of observed species). Linear regression model of richness against sequencing depth reveals a highly significant association ($P = 2.2 \times 10^{-16}$). 'H-G', Hunter-Gatherer.



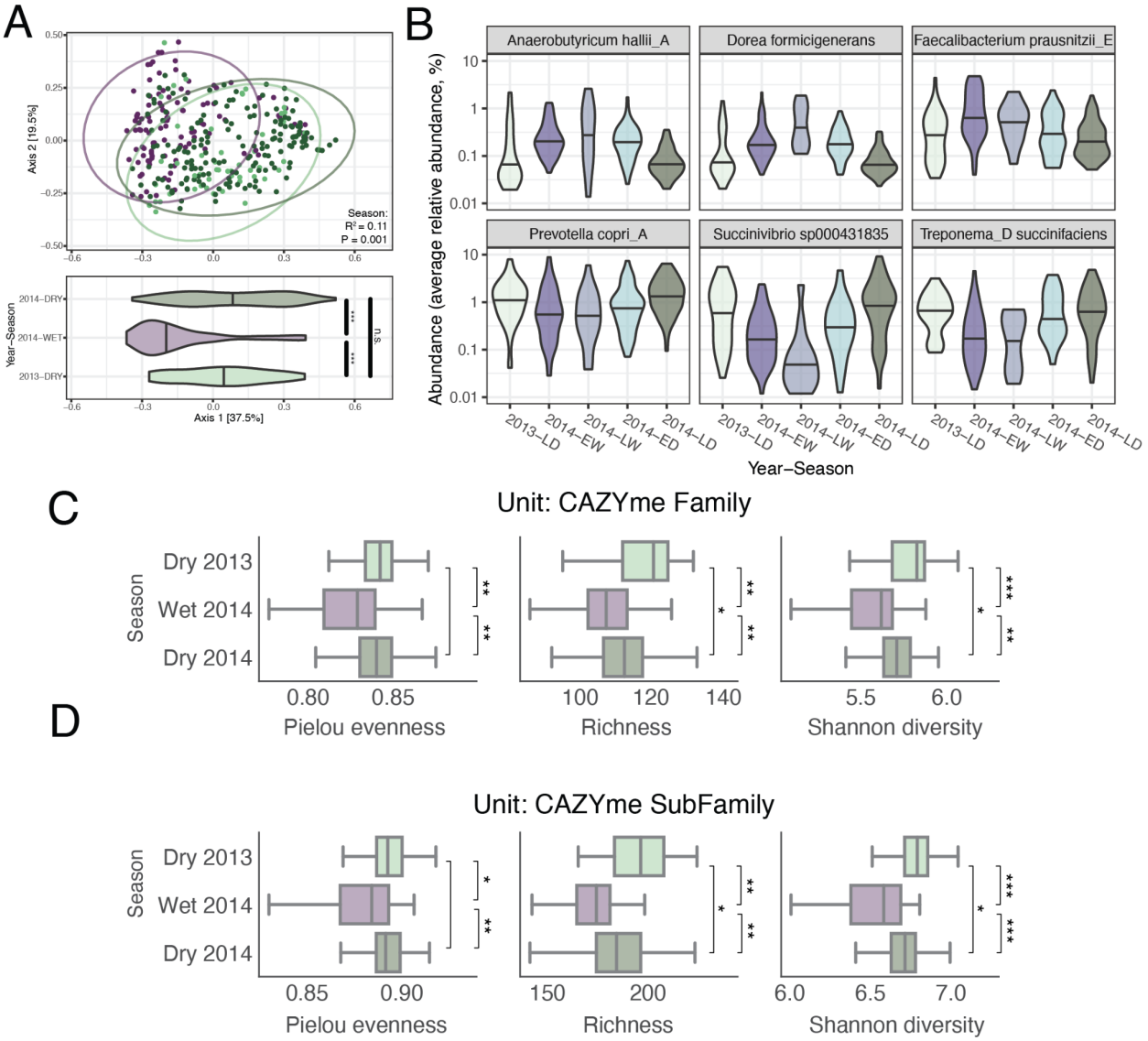
Extended Data Figure 7. Lifestyle-specific enrichment of bacterial and archaeal taxa. Volcano plot showing enrichment of each species in either Hadza or industrial samples in our global metagenomics data set. Dots colored magenta are in the 95th percentile most enriched in the Hadza and are deemed VANISH taxa (124 total). Dots that are colored blue are in the 95th percentile most enriched in industrial samples and are deemed BloSSUM (63 total).



Extended Data Figure 8. **Global prevalence of two most prevalent *Blastocystis* MAGs and their association with VANISH and BloSSUM taxa.**

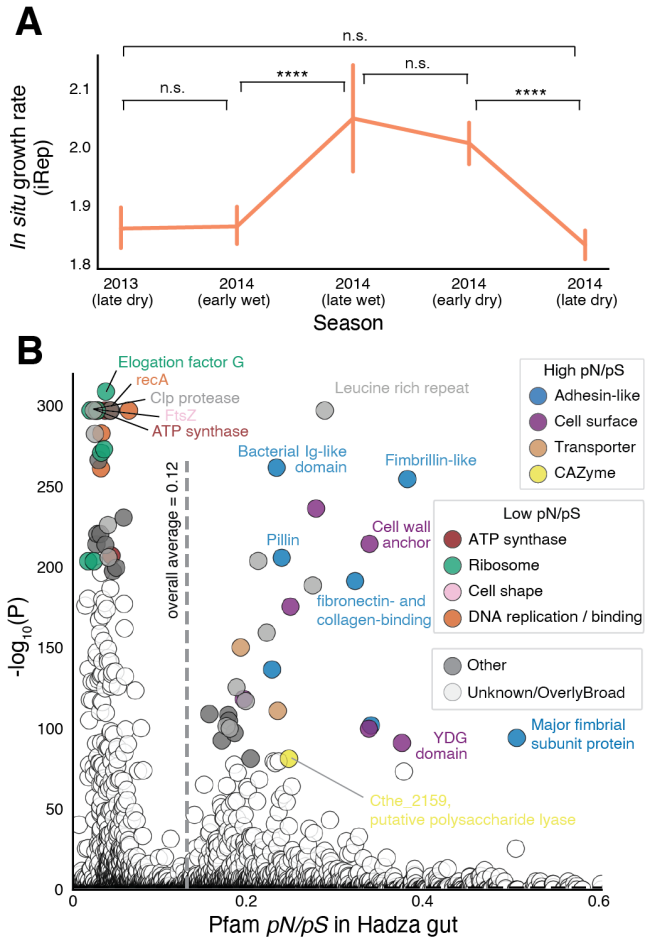
(**A**) A heatmap showing the prevalence of the two most prevalent *Blastocystis* MAGs (subtype 1 and subtype 4) in 16 different countries in our global metagenomics database. (**B**) A volcano plot showing associations between the presence or absence of either *Blastocystis* genome and the relative abundance of VANISH (magenta) or BloSSUM (blue) taxa. *P*-values were determined with a Wilcoxon rank-sum test and then adjusted with the Benjamini-Hochberg method to correct for multiple hypothesis testing. Threshold for significance of the adjusted *p-value*s is *P*=0.05 (or -log10(*P*)=1.3). Effect size was determined by linear discriminant analysis. The data points labeled "Global" are the associations for all samples in our global metagenomics data set. Other data points are for individual studies within the global metagenomics data set (annotated by first author of study and country of origin of the metagenomes). Across all studies we found that Blastocystis presence was positively and negatively associated with the total abundance of VANISH (P=5.1x10^{-14}) and BloSSUM (P=8.6x10^{-21}) taxa, respectively. (**C**) Boxplots showing the summed relative abundance of BloSSUM taxa per sample and whether *Blastocystis* was detected in that sample. Associations shown are for the entire global metagenomics data set and 5 additional populations (NEP = Nepal, TZA = Tanzania, THA = Thailand, ISR = Israel) from

three lifestyles labeled above the plots. P-values shown are the results of Wilcoxon rank-sum tests.



Extended Data Figure 9. Seasonality in Hadza gut microbiome.
(A) A principal coordinate analysis of Hadza samples where the Bray-Curtis distance matrix was generated by calculating the relative abundance of each taxonomic family in our bacterial/archaeal species-level genome database using InStrain (top panel). Samples are colored by season. Season explains a significant amount of the variation in the data ($P = 0.001$, $R^2 = 0.09$; ADONIS, using Subject ID as a strata). Sub-season also explains a significant amount of variation in the data ($P = 0.001$, $R^2 = 0.14$; ADONIS, using Subject ID as a strata). The bottom panel shows a violin plot of each sample's PCo1 position, grouped by season. Samples collected in the dry season are significantly different from the wet season ($P = 1.2 \times 10^{-10}$ and $P = 2 \times 10^{-16}$ for 2013-DRY:2014-WET and 2014-WET:2014-DRY comparisons, respectively; Wilcoxon test). The samples collected in each dry season do not differ significantly from each other ($P=0.34$; Wilcoxon test). **(B)** The violin plots depict distribution of relative abundance for 6 SRGs that vary significantly over the 5 sub-seasons. The top three sub-panels depict species-level groups that have higher abundance in the wet seasons. The bottom three sub-panels depict

species-level groups that have higher abundance in the dry seasons. (*Bulleidia sp.*, P-adjusted = 7.5×10^{-20}; *Dorea formicigenerans*, P-adjusted = 1.2×10^{-16}; *Holdemanella sp003436425*, P-adjusted = 6.5×10^{-16}; *Prevotella copri*_A, P-adjusted = 0.0054; *Succinivibrio sp000431835*, P-adjusted = 4.7×10^{-7}; *Treponema_D succinifaciens*, P-adjusted = 0.012; Kruksal-Wallis test). 2013-LD (Late Dry); 2014-EW (Early Wet); 2014-LW (Late Wet); 2014-ED (Early Dry); 2014-LD (Late Dry). (**C**) For Hadza gut metagenomes sequenced in this study, genes present (≥ 80% breadth of coverage) on detected genomes (≥ 50% breadth of coverage) were annotated against the CAZyme database. CAZyme Pielou evenness (left), total richness (middle), and Shannon diversity (right) were calculated using the summed relative abundance of genomes containing each GH and PL CAZyme Family (for example, 'GH16') or (**D**) SubFamily (for example, 'GH16.7'). Values for samples collected from Hadza individuals in different seasons were compared using a two-sided Wilcoxon rank-sum test (* $P < 0.01$; ** $P < 1 \times 10^{-5}$; *** $P < 1 \times 10^{-10}$).



Extended Data Figure 10. Microdiversity and growth rates of Hadza gut bacteria. **(A)** *In situ* growth rate measurements of all taxa detected in Hadza adult metagenomes across seasons. Error bars indicate 95% confidence intervals. (n.s. $P > 0.05$; **** $P \leq 0.0001$; Wilcoxon rank-sum test). **(B)** Pfams with high or low *pN/pS* values in Hadza fecal metagenomes. The x-axis displays the mean *pN/pS* value of all genes annotated with each Pfam within Hadza fecal metagenomes. The y-axis displays the probability that the number of times genes annotated as each Pfam were in the top 10% or bottom 10% of all genes on detected genomes was due to random chance (binomial test with multiple hypothesis correction). The 30 Pfams with the lowest p-values for low and high *pN/pS* were manually annotated with broad functional categories.

SI Guide

Supplementary Tables

Supplementary Table 1	Description of Hadza, Nepali, and Californian cohorts
Supplementary Table 2	Comprehensive genome information info (including representative genomes and other genomes)
Supplementary Table 3	Roster of strains isolated from Hadza stool (including culturing information)
Supplementary Table 4	Global metagenomics data set broken down by sample
Supplementary Table 5	Prevalence/abundance data for each species-level representative genome in our bacterial/archaeal species-level genome database
Supplementary Table 6	Pfam info (lifestyle-enrichment and pN/pS data)
Supplementary Table 7	Strain sharing data between Hadza adult samples

References

1. Wastyk, H. C. *et al.* Gut-microbiota-targeted diets modulate human immune status. *Cell* **184**, 4137-4153.e14 (2021).
2. Smits, S. A. *et al.* Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802–806 (2017).
3. Abdill, R. J., Adamowicz, E. M. & Blekhman, R. Public human microbiome data are dominated by highly developed countries. *PLoS Biol.* **20**, e3001536 (2022).
4. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0603-3.
5. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* **6**, 960–970 (2021).
6. Green, E. D. *et al.* Strategic vision for improving human health at The Forefront of Genomics. *Nature* **586**, 683–692 (2020).
7. Fragiadakis, G. K. *et al.* Links between environment, diet, and the hunter-gatherer microbiome. *Gut Microbes* **10**, 216–227 (2019).
8. Sonnenburg, E. D. & Sonnenburg, J. L. The ancestral and industrialized gut microbiota and implications for human health. *Nat. Rev. Microbiol.* **17**, 383–390 (2019).
9. Martínez, I. *et al.* The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. *Cell Rep.* **11**, 527–538 (2015).
10. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* (2012) doi:10.1038/nature11053.
11. Clemente, J. C. *et al.* The microbiome of uncontacted Amerindians. *Sci Adv* **1**, (2015).
12. Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z. & Dominguez-Bello, M. G. The infant microbiome development: mom matters. *Trends Mol. Med.* **21**, 109–117 (2015).

13. Modi, S. R., Collins, J. J. & Relman, D. A. Antibiotics and the gut microbiota. *J. Clin. Invest.* **124**, 4212–4218 (2014).
14. Blaser, M. J. The theory of disappearing microbiota and the epidemics of chronic diseases. *Nat. Rev. Immunol.* **17**, 461–463 (2017).
15. Vangay, P. *et al.* US Immigration Westernizes the Human Gut Microbiome. *Cell* **175**, 962-972.e10 (2018).
16. Wibowo, M. C. *et al.* Reconstruction of ancient microbial genomes from the human gut. *Nature* **594**, 234–239 (2021).
17. Tett, A. *et al.* The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* **26**, 666-679.e7 (2019).
18. Moeller, A. H. *et al.* Rapid changes in the gut microbiome during human evolution. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16431–16435 (2014).
19. Linz, B. *et al.* An African origin for the intimate association between humans and Helicobacter pylori. *Nature* **445**, 915–918 (2007).
20. Blaser, M. J. & Falkow, S. What are the consequences of the disappearing human microbiota? *Nat. Rev. Microbiol.* **7**, 887–894 (2009).
21. Sonnenburg, J. L. & Sonnenburg, E. D. Vulnerability of the industrialized microbiota. *Science* **366**, (2019).
22. Jin, H. *et al.* Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome. *Gut Microbes* **14**, 2021790 (2022).
23. Olm, M. R. *et al.* Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* **7**, 26 (2019).

24. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-020-00797-0.

25. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* **34**, 1256–1263 (2016).

26. Marlowe, F. *The Hadza: Hunter-gatherers of Tanzania*. (University of California Press, 2010).

27. Matthew R. Olm, Dylan Dahan, Matthew M. Carter, Bryan D. Merrill, Brian Yu, Sunit Jain, Xian Dong Meng, Surya Tripathi, Hannah Wastyk, Norma Neff, Susan Holmes, Erica D. Sonnenburg, Aashish R. Jha, Justin L. Sonnenburg. Robust Variation in Infant Gut Microbiome Assembly Across a Spectrum of Lifestyles. *Manuscript under review* (2022).

28. Jha, A. R. *et al.* Gut microbiome transition across a lifestyle gradient in Himalaya. *PLoS Biol.* **16**, e2005396 (2018).

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

30. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

31. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).

32. Bengtsson-Palme, J. *et al.* The Human Gut Microbiome as a Transporter of Antibiotic Resistance Genes between Continents. *Antimicrob. Agents Chemother.* **59**, 6551–6560 (2015).

33. Lloyd-Price, J. *et al.* Erratum: Strains, functions and dynamics in the expanded Human

Microbiome Project. *Nature* **551**, 256 (2017).

34. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
35. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).
36. Obregon-Tito, A. J. *et al.* Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* **6**, 6505 (2015).
37. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* (2016) doi:10.1038/nature18927.
38. Lokmer, A. *et al.* Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. *PLoS One* **14**, e0211139 (2019).
39. Liu, W. *et al.* Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Sci. Rep.* **6**, 34826 (2016).
40. Pehrsson, E. C. *et al.* Interconnected microbiomes and resistomes in low-income human habitats. *Nature* **533**, 212–216 (2016).
41. Rosa, B. A. *et al.* Differential human gut microbiome assemblages during soil-transmitted helminth infections in Indonesia and Liberia. *Microbiome* **6**, 33 (2018).
42. 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).
43. Conteville, L. C., Oliveira-Ferreira, J. & Vicente, A. C. P. Gut Microbiome Biomarkers and Functional Diversity Within an Amazonian Semi-Nomadic Hunter–Gatherer Group. *Front.*

Microbiol. **10**, (2019).

44. Rampelli, S. *et al.* Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr. Biol.* **25**, 1682–1693 (2015).
45. Tamburini, F. B. *et al.* Short- and long-read metagenomics of urban and rural South African gut microbiomes reveal a transitional composition and undescribed taxa. *Nat. Commun.* **13**, 926 (2022).
46. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
47. Litvak, Y., Byndloss, M. X. & Bäumler, A. J. Colonocyte metabolism shapes the gut microbiota. *Science* **362**, (2018).
48. Belkhou, C. *et al.* Treponema peruense sp. nov., a commensal spirochaete isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **71**, (2021).
49. Borges, R., Machado, J. P., Gomes, C., Rocha, A. P. & Antunes, A. Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics* **35**, 1862–1869 (2019).
50. Suzuki, T. A. *et al.* Codiversification of gut microbiota with humans. *bioRxiv* 2021.10.12.462973 (2021) doi:10.1101/2021.10.12.462973.
51. Liu, H., Prugnolle, F., Manica, A. & Balloux, F. A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* **79**, 230–237 (2006).
52. Falush, D. *et al.* Traces of human migrations in Helicobacter pylori populations. *Science* **299**, 1582–1585 (2003).
53. Faith, J. J. *et al.* The Long-Term Stability of the Human Gut Microbiota. *Science* **341**, 1237439–1237439 (2013).
54. Valles-Colomer, M. *et al.* Variation and transmission of the human gut microbiota across

multiple familial generations. *Nat Microbiol* **7**, 87–96 (2022).

55. Wastyk, H. C. *et al.* Gut Microbiota-Targeted Diets Modulate Human Immune Status. *Cold Spring Harbor Laboratory* 2020.09.30.321448 (2020) doi:10.1101/2020.09.30.321448.
56. Bushnell, B. BBTools software package. *URL http://sourceforge. net/projects/bbmap* **578**, 579 (2014).
57. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. (2010).
58. Bushnell, B., Rood, J. & Singer, E. BBMerge--accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).
59. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
60. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
61. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
62. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes DE Novo Assembler. *Curr. Protoc. Bioinformatics* **70**, e102 (2020).
63. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
64. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, (2016).
65. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,

357–359 (2012).

66. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
67. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
68. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* (2019) doi:10.1038/s41586-019-1058-x.
69. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
70. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5**, (2020).
71. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
72. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz848.
73. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**, 4162–4164 (2019).
74. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).

75. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* **28**, 569–580 (2018).

76. Saary, P., Mitchell, A. L. & Finn, R. D. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol.* **21**, 244 (2020).

77. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

78. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

79. Olm, M. R. *et al.* Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria. *Science Advances* **5**, eaax5727 (2019).

80. Lind, A. L. & Pollard, K. S. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome* **9**, 58 (2021).

81. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).

82. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396 (2021).

83. Edgar, R. C. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**, 18 (2007).

84. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).

85. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

86. Chen, Y., Ye, W., Zhang, Y. & Xu, Y. High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Res.* **43**, 7762–7768 (2015).

87. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* (2017) doi:10.1038/nature23889.

88. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690–703 (2015).

89. Liu, W. *et al.* Corrigendum: Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Sci. Rep.* **7**, 39576 (2017).

90. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053-2067.e18 (2021).

91. Human Development Data Center. http://www.hdr.undp.org/en/data.

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

93. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

94. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).

95. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

96. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

97. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky995.

98. Lewis, T. E., Sillitoe, I. & Lees, J. G. cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. *Bioinformatics* **35**, 1766–1767 (2019).

99. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* vol. 57 61 (Austin, TX, 2010).

100. Edgar, R. C. MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. *bioRxiv* 2021.06.20.449169 (2021) doi:10.1101/2021.06.20.449169.

101. Gutierrez, S. C., Martinez, J. M. S. & Gabaldón, T. TrimAl: a Tool for automatic alignment trimming. *Bioinformatics* **25**, 1972–1973 (2009).

102. Tange, O. *GNU Parallel 2018*. (Lulu.com, 2018).

103. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

104. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

105. Jones, E., Oliphant, T. & Peterson, P. SciPy: Open source scientific tools for Python. *URL http://scipy. org* (2001).

106. Clark, C. G., van der Giezen, M., Alfellani, M. A. & Stensvold, C. R. Recent developments in Blastocystis research. *Adv. Parasitol.* **82**, 1–32 (2013).

107. Marlowe, F. W. *et al.* Honey, Hadza, hunter-gatherers, and human evolution. *J. Hum. Evol.* **71**, 119–128 (2014).

108. NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **45**, D12–D17 (2017).