

Prehistoric Global Migration of Vanishing Gut Microbes With Humans

Corresponding Author: Professor Justin Sonnenburg

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)

This fascinating, well-conducted study investigates the evolutionary divergence between gut microbiota from two human populations (Tsimane and Hadza) that have been isolated from one another for tens of thousands of years. Through a series of sophisticated modeling analyses based on robust population genetic theory, the authors conclude that these two human populations have inherited gut bacteria from a common source (likely their common ancestors in Africa), that these gut bacteria have co-migrated with the populations as they spread from Africa to the Americas, that gene flow between geographically separated co-migrating lineages has been minimal or nonexistent, and that many of these co-migrating lineages have been lost from industrialized populations. The approach—assessing gut bacterial genome divergence between two indigenous human populations through deep metagenomic sequencing—is clever and novel, and it expands our understanding of the evolution of the human gut microbiota. A major novel finding of the study pertains to the absence of gene flow between related bacteria from these human populations. Whereas past work on co-migration / parallel diversification of gut bacteria and hosts (eg in humans and in humans and non-human primate species) has focused on core genes (only a subset of the most conserved loci in bacterial genomes), in the present study the authors present strong evidence of genome-wide genetic isolation between bacteria from these divergent human populations. Thus, the present study shows that the tight relationships between gut symbionts and their host lineages has led to genome-wide divergence of bacterial strains mediated by reduced gene flow following host migration. The authors further employ rigorous molecular clock analyses that corroborate the timing of co-divergence between gut bacteria and hosts.

An additional finding of the paper is that many of the co-migrating taxa have been lost from industrialized human populations. This finding adds evolutionary context to the authors prior work on these VANISH taxa, as well as the work of other groups showing that industrialized human populations harbor reduced gut bacterial diversity, including reduced diversity of gut bacterial lineages inferred to be ancestral to humans. The loss of ancestral taxa may have implications for human health, since the symbionts that have co-evolved with humans may be those whose loss is most costly to hosts.

The paper is well written and fun to read. The figures are beautiful, clear, and compelling. Overall, the methods are sound and easy to follow.

I had questions / comments about the framing, methods, and results that could be addressed.

Specific Comments:

Framing: One critique of the manuscript as written is that it could better highlight the novel aspects of the present study, given what was known previously from other comparative studies. In the introduction and in several places in the text, the authors state or imply that it was not known whether gut bacteria in extant humans were recently acquired from environmental sources or inherited from ancestral host populations. There is a large body of comparative studies that have provided strong evidence that gut bacteria have existed within their host lineages for millions of years. These prior studies include recent genome-resolved metagenomics studies but also date back to > 15 years of comparative microbiome studies. Major novel contributions of the current study appear to be 1) the focus on indigenous, geographically isolated populations

of known divergence time (a major strength) and 2) the sophisticated analyses of gene flow and molecular dating showing genome-wide evidence of genetic isolation whose timing lines up with human migration events. Because this study is not the first to establish that gut bacterial associations with humans are ancestral, I would suggest that the authors tone down some of the claims about the novelty of the results re: establishing that extant gut bacteria represent ancestral associations. The introduction could be easily rewritten to better incorporate knowledge from past work yet still highlight the novel and exciting aspects of the present study.

Contiguosness of MAGs: The approach based on L99 employed to assess gene flow is conceptually sound. However, it was not clear to me how contiguosness of MAGs (or lack thereof) may affect these analyses. Are there systematic differences in N50 between MAGs from different populations that could bias these results? How was variation in contiguosness incorporated?

L99 calculations: Similarly, I wondered how the authors controlled for relatedness of hosts when conducting the within-population L99 measures. For example, were comparisons of hosts from the same family (i.e., close kin) excluded? Given that close kin are known to share strains, this could confound these analyses, which are aimed at quantifying population-level divergence.

Line 381-389: Were any steps taken to account for strain heterogeneity of MAGs? If not, I wondered whether high levels of strain heterogeneity within individual hosts could affect measures of recombination and divergence times.

Molecular Clock analyses: The authors use a classic population genetic equation that models the genetic divergence between lineages as a function of the divergence time, mutation rate, and rate of gene flow. This equation is well suited to estimate divergence times. However, I wondered whether the bacterial lineage divergence time is really the important quantity to establish / test for co-migration, given that related strains from Tsimane and Hadza may coalesce before the divergence time of the host populations. Consider, for example, a case in which the ancestral host population harbored multiple strains of a species, and when the host population diverged (eg leading to the Tsimane and the Hadza), the two descendant populations each inherited a different strain (or different sets of strains) (in a manner analogous to incomplete lineage sorting). In this case, the MRCA for strains in Tsimane and Hadza would predate the timing of the ancestral host population, and estimating the timing of the ancestral host population using bacterial genetic divergence would require some estimate of the ancestral bacterial effective population size (N_e). Given that ancestral N_e when the bacterial lineages split is likely unknown, I do not think the authors need to redo or extend any analyses (what they have done is very reasonable). However, this consideration does affect the interpretation of results and could be discussed somewhere. For example, if the ancestral host population contained strain diversity, then the divergence times of bacterial strains between Tsimane and Hadza may predate the divergence of the host populations, even if strains co-migrated with hosts.

Figure S9: It would be ideal to show the underlying data points in this figure.

Absence of species from industrialized populations: It was not clear to me how absence was assessed. Is this based on lack of MAGs of these species from industrialized populations, or is it based on mapping analyses? Mapping analyses may be more sensitive to detect low abundance bacteria that do not assemble well into MAGs and therefore represent a more conservative approach for testing for absence.

Abstract: "Understanding which microbes have been long-term residents of the human gut, and may have co-evolved with humans, has implications for the importance of microbial biodiversity loss for health. However, the genetic complexities of microbial evolution and the plasticity of gut microbiome composition have made it challenging to define these long-term associations."

I agree with this statement, but several prior studies demonstrated parallel phylogenetic trees between host lineages and gut bacterial lineages/communities, including studies focused on human populations (such as the Suzuki/Ley paper cited) as well as studies focused on human and non-human primate species (the earliest of these go back 15 years, eg, Ochman et al., 2010 PLOS Biology: <https://doi.org/10.1371/journal.pbio.1000546>.) From these studies, it has been established that lineages of gut bacteria (including lineages identified in the present study) have been 'long-term residents of the human gut'. Many of these microbes not only predate split of human populations from one another but also the split of *Homo sapiens* from other primate species.

Introduction: "However, the comparatively recent ages of existing paleofeces (~1000-200060 years old), combined with the challenges of analyzing ancient DNA due to its degradation over time, makes it difficult to determine whether these microbes were also present in more ancestral human populations (~10-100kya) or whether they reflect a more recent acquisition from a common environmental source."

Even without very old ancient DNA samples, comparative studies have established that many lineages of gut bacteria were not recently acquired from common environmental sources. It would not be correct to imply that DNA material from ancient specimens is required to understand ancestral states (most macroevolutionary and molecular evolutionary studies do not have very old DNA material). For example, it is known from cophylogenetic studies that *Buchnera* first formed endosymbiosis with aphid hosts hundreds of millions of years ago, even though we have no ancient DNA samples dating back 100+ million years. Similarly, the results of Suzuki and Ley's paper are not consistent with recent symbiont acquisition, nor are the results of comparative studies of nonhuman primates.

Introduction: “potentially consistent with a past history of co-migration”

I don't think the adverb 'potentially' is needed, unless there are specific points that the authors can make to call into question these past results.

Introduction: However, interpreting these correlations remains challenging due to the complex ways that microbial genomes evolve. Traditional phylogenetic methods break down for many species of gut bacteria, where horizontal gene transfer and homologous recombination produce mosaic genomes that cannot be represented by a single phylogenetic tree. Moreover, existing strain-level analyses have been biased toward species in industrialized populations, which are rarely observed in ancient paleofeces. These sampling biases have hampered our ability to infer the evolutionary history of microbial taxa that are the strongest potential candidates for ancestral co-migration.

I do not fully follow this line of reasoning. While it is true that core-gene phylogenies do not fully capture HGT/recombination events, this does not mean that core-gene phylogenies cannot be used to understand how core genomes of bacteria diversify or co-migrate with hosts. The past phylogenetic analyses from which evidence of co-migration was gleaned were based on single-copy core gene alignments, which contain genes that are generally not horizontally transmitted between distantly related bacteria. Past work (eg Suzuki and Ley et al) has shown that the lineage trees of the bacteria are associated with the lineage trees of the host – that result supports long-term associations between the bacteria and the host, even if it does not fully capture the history of recombination and HGT that those bacteria have experienced since. If HGT of these core genes was rampant, we would not expect to see signals of parallel diversification between hosts and symbionts.

Line 120: “therefore requires that they must have shared a common ancestor within the last ~100-200k year”

Given the uncertainty around bacterial molecular clocks, this statement may be too strong. As the authors note in the supplemental material, the estimate could be off by an order of magnitude.

Line 131: “The broad overlap between these distributions indicates that most of the diversity within the Tsimane population is much older than the onset of geographic isolation”

This could be edited to make clear that you mean diversity “within bacterial species”. Most gut bacterial diversity within the Tsimane must be older than the onset of their geographic isolation from the Hadza, given that the microbiota contains multiple phyla. The result reported in this section is consistent with the possibility that bacterial species contain multiple ancestral strains, each of which co-migrated with hosts.

Lines 172-177:

This conclusion seems self evident, ie, if two host populations share a species it does not mean they share the same strain. If there is a broader point here, it could be made more clearly.

(Remarks on code availability)

The documentation of the code is comprehensive. The analyses are reproducible with the code provided in the github repo.

Referee #2

(Remarks to the Author)

This is an interesting paper from Carter and colleagues. The paper is well written and attempts to answer an exciting question in microbiome research: how connected are specific gut microbial species with human evolutionary history, and to what extent have industrialization and lifestyle changes affected this association? The paper uses deep metagenomic sequencing and population genetic analyses to show that gut microbial species (some of which absent from industrialized populations) have likely co-migrated with humans. I believe this is an important paper with a proof of concept that deep, strain-resolved metagenomic sequencing combined with population genetic analyses is a powerful approach for reconstructing the evolutionary history of the human gut microbiome across diverse populations. In addition, the biological results, although maybe not surprising, are exciting. At the same time, there are a few issues:

1. Justification of parameter choice:

one big question is regarding the parameter choices for MAG generation and subsequent analyses. The analysis relies on established thresholds for MAG quality (e.g. 50% completeness and 10% contamination), which are in line with current standards. However, they remain arbitrary to some degree. This is important because variations in these thresholds can affect the inclusion or exclusion of certain MAGs, potentially affecting the observed diversity and the number of shared microbial species or strains. Similarly, the use of a 10% identical gene cutoff to define recent clonal relationships is somewhat subjective. The age boundary linked to this threshold varies across species and is calibrated only through estimates of bacterial mutation rates, which themselves have broad confidence intervals. Small changes to this cutoff might shift the inferred rates of recent strain sharing, especially in comparisons across populations. Other steps (filtering criteria for core gene identification, clustering thresholds, and minimum gene alignment lengths) can also introduce parameter-dependent variation. Together, these choices could change the resolution and sensitivity of downstream analyses. It is unclear if conclusions in the paper change under alternative parameter settings. This can potentially be addressed by testing different parameters in the MAG assembly and strain-sharing analyses, performing sensitivity analyses to determine how varying these cutoffs affect the qualitative results, and reporting the degree of stability or variability in key findings as parameters are adjusted.

Along the same lines, one specific area is the section on microbial biodiversity loss, which relies heavily on the parameters chosen to define the presence or absence of a taxon in industrialized populations. The sample size required to confidently call a taxon as “absent” is not clearly specified, and increasing the number of samples could alter detection rates. This is

especially true for rare taxa. In addition, details on the sequencing depth and the computational strategy for taxon identification (whether through MAG assembly, read mapping, reference-based profiling) can influence which species are detected and classified as “vanishing” or persistently absent. All of these methodological choices can have a big effect on the observed patterns and conclusions. Similar to the above, an evaluation across a range of parameters and approaches would help establish the robustness of the findings. Sensitivity analyses and modeling the probability of taxon detection as a function of sample size, using data from other rare taxa, would be especially helpful.

2. Statistical and computational issues:

There are a few questions regarding the statistical and computational analysis that I outline below.

Multiple testing: The manuscript makes hundreds of comparisons across different microbial species, but in most places I could not find a mention of correction for multiple hypothesis testing (e.g., FDR). This can potentially cause inflation of false positive rates in the presented results. This is relevant throughout the paper, but especially in the large-scale pairwise ANI calculation, strain sharing rates and percentage of clonally related pairs (Fig 2), gene flow metrics (Fig 3).

Potential confounders and batch effect: It was not clear how potential confounders (such as sample collection approach, sample preservation and handling, sequencing platform, sampling year, sequencing coverage, QC values, sequencing batch of course, etc) were controlled for in the statistical analysis. This could lead to systematic errors between cohorts, so it would be helpful to incorporate this across analyses throughout the paper. This issue appears throughout the Results (Fig 1, 2, 3) when comparing different populations (Tsimane vs. Hadza vs. Industrialized).

Uncertainty estimation: many population-level estimates (e.g. prevalence, strain sharing rates, split time estimates) are presented as single numbers, and lack estimation of uncertainty (confidence or credible intervals) making it challenging to evaluate the reliability of these estimates. This is also relevant in the context of point 1 above.

3. Data sharing:

One of the major strengths of the paper is the deep metagenomic data generated. As such, a significant limitation of this study is the lack of immediate public availability of the data. This includes both the raw sequencing data and the processed data tables underlying the analysis. Open access to raw data is critical for independent validation, reproducibility, and reanalysis as new computational tools and reference databases emerge. Equally important are the processed data tables, which include quality-controlled MAGs, strain-sharing matrices, species presence/absence profiles, and other summary statistics that support key results and figures in the paper. Without access to these, the scientific community is unable to evaluate the analytical pipelines, test alternative parameter choices, or benchmark findings against other datasets. Most importantly, making both raw and processed data available enables meta-analyses that put findings in a broader context.

Although it is stated that the raw sequencing data will be made available upon publication through the data access procedures established by the Tsimane Health and Life History Project, this approach still presents challenges. Restricted or conditional access can impede timely and widespread data reuse, and may slow down validation or meta-analyses. In addition, sharing protected data can be done through standard databases (e.g. NIH dbGap) - as is, it is not clear where and how the data will be available, what is the process for getting accesses, what are the criteria for determining whether access will be granted, etc. Following the link in the paper (<https://tsimane.anth.ucsb.edu/data.html>) and clicking the “Data” section leads to a broken link, which is a bit worrying for data availability for the current paper. Also, importantly, the paper does not address whether and how processed data tables will also be made accessible.

4. Population genetics analysis:

I am not an expert in population genetics modeling, so this is not a major issue, but I found several of the analyses and metrics used in the paper challenging to fully interpret. I believe that some clarification would benefit readers who are not experts in this area. First, it would be helpful to include a clearer explanation of the rationale and methodology behind key analyses such as the use of the L99 metric and the demographic inference using SNV frequencies, including what each metric represents, how it is calculated, why it is appropriate for inferring gene flow or split times, etc. It would also be helpful to state explicitly the main assumptions underlying population genetic models (e.g. regarding recombination, mutation rate, or population structure), explain how realistic these assumptions are for the data at hand, and any impact that possible violation of these assumptions might have on interpretation of the results. Also, it would be good to include additional detail on how model fit to the data was assessed, the robustness of parameter choices (related to my first point above), and how uncertainty or confidence intervals for key estimates was incorporated. Lastly, as only a subset of species with sufficient data were included in some analyses, some discussion of the implications for the generalizability of findings to less well-sampled taxa would be beneficial.

(Remarks on code availability)

(Remarks to the Author)

The authors employed population genetic methods to infer evolutionary history from genetic diversity patterns in contemporary microbial genomes, utilizing advances in strain-resolved metagenomics. They specifically focused on two geographically distant non-industrialized populations: the Tsimane horticulturalists of Bolivia and the Hadza hunter-gatherers of Tanzania, whose ancestors separated thousands of years ago.

They did deep metagenomic sequencing on 133 fecal samples from 85 Tsimane adults, recovering 12,746 metagenome-assembled genomes (MAGs) representing 1,408 bacterial and archaeal species. These were then compared to 32,034 MAGs from 137 Hadza hunter-gatherers. This allowed for the quantification of how these populations' gut microbial species have diverged over the last 100,000 years.

Key findings:

1. Despite vast geographic separation and no direct historical contact, 87.4% (1,231) of species in the Tsimane cohort were also found in Hadza individuals. Crucially, the majority (60.2%, 848) of these shared species were undetected in the fecal microbiomes of industrialized populations.
2. The authors calculated the average nucleotide identity (ANI) between MAGs. The typical ANI between Tsimane and Hadza MAGs (around 98%) suggests they must have shared a common ancestor within the last 100-200,000 years, based on estimated mutation rates.
3. While average ANI showed little geographic differentiation, a stronger trend was observed in the high ANI tail (>99.75%) of the distribution, showing a significant enrichment of within-population pairs. This suggests restricted strain transmission between populations in recent evolutionary history, consistent with divergence due to historical separation.
4. The study defined clonal strains as MAG pairs with >10% identical genes. 85.6% of shared species (545 out of 636) showed no recent Tsimane-Hadza strain sharing, indicating very limited inter-population strain exchange in the past ~5,000 years.
5. To focus on gene flow via HGT in MAG pairs with less than 10% identical genes, an orthogonal approach was used. By quantifying the 99th percentile of tract lengths, the authors found that most species had shorter L99 lengths between Tsimane and Hadza compared to within populations, indicating genetic isolation in the recent past.
6. The authors tried to estimate the timing of genetic isolation between microbiomes. Genetic exchange was notably more recent between industrialized populations compared to non-industrialized populations (Tsimane-Hadza: median ~43,563 years). This estimated onset of genetic isolation between the Tsimane and Hadza microbiomes roughly aligns with the timeframe of initial human migration out of Africa (60-70,000 years ago) and the settlement of the Americas (16-23,000 years ago).

Overall, the article is well written and clearly structured; the following are a few specific comments that I believe could further improve it.

Questions and suggestions:

1. The authors should clarify that ANI is robust for species level identification, but it cannot be used naively for dating. In gut microbes with high recombination, horizontal gene transfer creates mosaic genomes, so it does not always reflect vertical evolutionary history. As a result, divergence time estimates such as "Tsimane and Hadza strains last shared a common ancestor ~100–200k years ago" may be biased. The clonal-segment analysis presented later is therefore the more appropriate framework for inferring divergence times in this context.
2. The authors state that ~60% of species in Tsimane and Hadza are absent from industrialized populations, which may be biased by the smaller sample sizes in these cohorts. Did the authors perform any sample size correction to account for potential undersampling of these isolated populations?
3. The authors use recombination tract lengths to infer stronger genetic isolation between Hadza and Tsimane compared to industrialized populations. However, recombination rates vary widely by species, genome size, and mobile elements. If industrialized datasets are enriched for high-recombining taxa (producing long tracts) while Hadza/Tsimane datasets contain more low-recombining taxa (producing short tracts), the observed differences in gene flow may partly reflect species composition rather than true host-associated isolation. This possibility should be considered in the interpretation.
4. Line 117: For the estimate of molecular clock and mutation rate for gut bacteria – isn't this expected to differ between species? How much does it vary?
5. Line 119: ANI values – this differs between species, as shown in the Supplementary Table. Please clarify whether you report averages across all species.
6. Line 121: "Most Tsimane species" – what does "most" refer to? Providing a percentage or distribution would be clearer.
7. Figure 1b: The Hadza had lower sample numbers than the Tsimane. Would it be possible to assess whether Hadza is comparable to Tsimane with respect to unique species after sample size correction?

8. Humans lost substantial diversity during the “Out of Africa” bottleneck. Could a similar reduction also be observed in microbiomes?

9. Line 156: Why was 5,000 years chosen as the cutoff for recent strain sharing? Please justify. Given that the Out of Africa migration occurred ~70 kya, why not use an older date?

10. Line 175: *Agathobacter rectalis* is not shown in the figures, while other species discussed in this section are. Is there a reason for this?

11. Line 213: Please refer here to the Supplementary Text where the simulations are described in more detail.

12. Line 233: Please specify the criteria for selecting the 15 species used for estimating isolation times. Without clear justification, this could be perceived as cherry picking.

13. Line 273: “The loss of VANISH taxa with changes in lifestyle, which are linked to an increased risk for a number of chronic metabolic and inflammatory disorders” – it would be helpful to indicate the fraction or extent to which each of these three classes is involved in such disorders.

14. Sample collection: The Hadza collection conditions should also be described for consistency with the Tsimane. Even if this is from a previous publication, please clarify whether the same procedures were used, as differences could influence results. It would also be good to restate the ethical clearance, since this article is authored by the same group.

15. It would strengthen the study to include environmental sampling, as the authors did not sequence soil, water, or animal microbiomes from Tanzania or Bolivia to test for potential environmental reservoirs and account for likely ecologically acquired taxa.

(Remarks on code availability)

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

I thank the authors for their comprehensive responses to my comments. I support publication of the manuscript.

(Remarks on code availability)

The code appears to be well documented and usable. The major limitation precluding full reproducibility will be the restricted access to the raw data, but this may be necessary for ethical reasons.

Referee #2

(Remarks to the Author)

I thank the authors for their comprehensive and thoughtful revision, which has substantially strengthened the manuscript. The authors have addressed nearly all of my concerns, including the addition of sensitivity analyses for parameter choices, the independent mapping-based validation pipeline, improved statistical reporting with multiple testing correction, and the commitment to releasing processed data via Zenodo. However, two issues remain that I believe should be addressed:

First, after seeing a figure of the sequencing coverage of the data (Figure R2.5), I remain concerned about the potential impact of differential sequencing coverage between the Hadza and Tsimane cohorts on the key findings. As shown in Figure R2.5, while average depth slightly differs between cohorts (slightly higher for the Tsimane), the Hadza cohort also exhibits a long tail of extremely high coverage samples, with some samples above 200 Gbp. This coverage heterogeneity could potentially introduce systematic biases in several places. Most critically, ultra-deep sequencing of a subset of Hadza samples enables assembly of rare, low-abundance organisms that would fail to assemble at lower depth. These rare species, detected only in deeply sequenced Hadza samples, would then appear absent from both the Tsimane and industrialized cohorts, not because they represent ancestral VANISH taxa, but simply because insufficient sequencing depth precludes their detection elsewhere. Importantly, the rarefaction analysis provided by the authors controls for number of subjects but not for sequencing depth per subject, leaving this artifact unaddressed. The estimated asymptote difference between the Hadza and Tsimane may therefore be partially depth-driven rather than reflecting true biological differences in gamma diversity.

Beyond the VANISH taxa finding, the uneven coverage distribution could also bias the L99-based genetic isolation analysis. Within-Hadza comparisons that involve two high-coverage samples would yield more complete MAGs with more accurate SNV calls and fewer assembly-induced tract breaks, producing artificially elevated within-population L99 values. Mixed-coverage Tsimane-Hadza between-population comparisons, by contrast, would be more prone to tract fragmentation, producing relatively shorter between-population L99 values. This could artifactually inflate the apparent contrast between

within- and between-population isolation that forms the core evidence for co-migration. One approach to address this would be to remove the very highly sequenced outliers and repeat the analysis. In addition, a more rigorous approach would be to perform a depth-rarefied analysis in which all samples are downsampled to a common sequencing depth prior to assembly, and to repeat the key analyses (including the species catalog, VANISH taxa prevalence, and L99 calculations) on this depth-equalized dataset. This would directly test whether the central findings are robust to the coverage heterogeneity currently present in the data.

Second, I appreciate the authors' commitment to indigenous data sovereignty and their detailed explanation of the CARE principles framework. At the same time, sharing the data is important: the data generated in this study represent an exceptionally rare and scientifically valuable resource that cannot easily be regenerated, making access especially important for the field. Without the ability to independently validate the results, or integrate these data into broader comparative studies, the community cannot fully evaluate or build upon what the paper presents as landmark findings about human-microbiome co-evolution.

This is why I remain concerned that the current data sharing mechanism does not meet the standards expected for a Nature publication. The THLHP data sharing webpage (this is the website readers are referred to for data access) explicitly commits to the FAIR principles (Findable, Accessible, Interoperable, Reusable), yet the access mechanism described on that same page fails to meet several of these in practice; what is described is a multi-stage discretionary review process with no defined timeline, no accession number, and no stated decision criteria, which does not constitute meaningful accessibility or reusability. Moreover, the access process as described on the THLHP website is entirely controlled by the co-Directors of the project, which creates a structural conflict of interest in which the data generators are also the gatekeepers, effectively allowing the authors to maintain a de facto monopoly on follow-up analyses using this dataset. This is qualitatively different from standard controlled-access repositories such as NIH dbGaP, where access decisions are made by independent committees with explicit and publicly stated criteria. I would also note that the level of restriction applied here appears disproportionate relative to the sensitivity of the data. The dataset in question consists of metagenomic sequences from stool samples, and the associated re-identification risks are substantially lower than for human genomic data. I think that it would be reasonable to deposit the raw data (after removing host reads) in a standard controlled-access repository such as dbGaP, where access decisions are made independently of the authors.

(Remarks on code availability)

Referee #3

(Remarks to the Author)

The authors have addressed all my comments satisfactorily. Thank you for the detailed responses and additional analyses.

I have just one small suggestion. The rarefaction analyses are quite nice and, as far as I can see, they are only included in the rebuttal letter. It would be useful to add them to the supplementary material if possible.

(Remarks on code availability)

Version 2:

Reviewer comments:

Referee #2

(Remarks to the Author)

I thank the authors for their thoughtful revision. The EGA deposition with a formal DAC and publicly posted access policy, together with the Zenodo release of processed files, adequately addresses the data sharing concerns.

Regarding differences in depth between populations, the cross-species correlation and the coverage-restricted reanalysis adequately address some of my concerns. However, regarding my first point, the new analysis does not fully address per-sample depth heterogeneity within cohorts, which could be the actual driver of differential rare-taxon recovery. The rigorous way to address this would be a depth-rarefied reassembly, or a statistical approach that controls for depth as a covariate. However, short of that, I would suggest adding an explicit acknowledgement in the discussion of this limitation in the manuscript text so readers can evaluate the finding accordingly.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

We appreciate the reviewer's thoughtful critiques and appreciation of the novelty, rigor, and clarity of the study. We have made extensive revisions to strengthen the framing and interpretation, as well as added data and analyses. We include point-by-point responses below.

Referee #1 (Remarks to the Author):

This fascinating, well-conducted study investigates the evolutionary divergence between gut microbiota from two human populations (Tsimane and Hadza) that have been isolated from one another for tens of thousands of years. Through a series of sophisticated modeling analyses based on robust population genetic theory, the authors conclude that these two human populations have inherited gut bacteria from a common source (likely their common ancestors in Africa), that these gut bacteria have co-migrated with the populations as they spread from Africa to the Americas, that gene flow between geographically separated co-migrating lineages has been minimal or nonexistent, and that many of these co-migrating lineages have been lost from industrialized populations. The approach—assessing gut bacterial genome divergence between two indigenous human populations through deep metagenomic sequencing—is clever and novel, and it expands our understanding of the evolution of the human gut microbiota. A major novel finding of the study pertains to the absence of gene flow between related bacteria from these human populations. Whereas past work on co-migration / parallel diversification of gut bacteria and hosts (eg in humans and in humans and non-human primate species) has focused on core genes (only a subset of the most conserved loci in bacterial genomes), in the present study the authors present strong evidence of genome-wide genetic isolation between bacteria from these divergent human populations. Thus, the present study shows that the tight relationships between gut symbionts and their host lineages has led to genome-wide divergence of bacterial strains mediated by reduced gene flow following host migration. The authors further employ rigorous molecular clock analyses that corroborate the timing of co-divergence between gut bacteria and hosts.

An additional finding of the paper is that many of the co-migrating taxa have been lost from industrialized human populations. This finding adds evolutionary context to the authors prior work on these VANISH taxa, as well as the work of other groups showing that industrialized human populations harbor reduced gut bacterial diversity, including reduced diversity of gut bacterial lineages inferred to be ancestral to humans. The loss of ancestral taxa may have implications for human health, since the symbionts that have co-evolved with humans may be those whose loss is most costly to hosts.

The paper is well written and fun to read. The figures are beautiful, clear, and compelling. Overall, the methods are sound and easy to follow.

I had questions / comments about the framing, methods, and results that could be addressed.

Specific Comments:

Framing: One critique of the manuscript as written is that it could better highlight the novel aspects of the present study, given what was known previously from other comparative studies. In the introduction and in several places in the text, the authors state or imply that it was not known whether gut bacteria in extant humans were recently acquired from environmental sources or inherited from ancestral host populations. There is a large body of comparative studies that have provided strong evidence that gut bacteria have existed within their host lineages for millions of years. These prior studies include recent genome-resolved metagenomics studies but also date back to > 15 years of comparative microbiome studies. Major novel contributions of the current study appear to be 1) the focus on indigenous, geographically isolated populations of known divergence time (a major strength) and 2) the sophisticated analyses of gene flow and molecular dating showing genome-wide evidence of genetic isolation whose timing lines up with human migration events. Because this study is not the first to establish that gut bacterial associations with humans are ancestral, I would suggest that the authors tone down some of the claims about the novelty of the results re: establishing that extant gut bacteria represent ancestral associations. The introduction could be easily rewritten to better incorporate knowledge from past work yet still highlight the novel and exciting aspects of the present study.

We thank the reviewer for raising this important point. We have revised the Introduction to provide better context of prior comparative work (e.g., (Ley et al., 2008; Moeller et al., 2016; Ochman et al., 2010; Sprockett et al., 2025)). We now place the novelty of our approach and findings within the context of previous studies that have shown co-divergence of microbes with their hosts. We believe the additions to the text provide important framing based on prior literature and underscores the unique contributions of our study.

Contiguousness of MAGs: The approach based on L99 employed to assess gene flow is conceptually sound. However, it was not clear to me how contiguousness of MAGs (or lack thereof) may affect these analyses. Are there systematic differences in N50 between MAGs from different populations that could bias these results? How was variation in contiguousness incorporated?

We thank the reviewer for this important methodological consideration. It is worth emphasizing that our L99-based gene flow analysis is based on the SNV catalog described in the Methods, which maps the assembled MAGs to a common reference genome for each species. As a result, differences in MAG contiguousness within a given species do not directly affect this analysis, as long as the underlying contigs are still binned correctly.

However, missing regions of the genome can still be an issue, whether due to errors in assembly or true differences in gene content compared to the reference genome. For this reason, our L99 analysis only considers sites in the reference genome that are present in both MAGs that are being compared (i.e. missing sites are neither counted towards the length of a tract nor used as evidence to split up an existing tract). This same approach is used in our calibration simulations, which convert the L99 score to a corresponding split time. Together, these steps ensure that missing sites – whether from assembly errors or gene content variation

– have a minimal impact on our split time estimates. We realize that this latter point was not sufficiently emphasized in the previous version of our manuscript, so we have revised the corresponding section of the Supplemental Methods (Section 4) to make this important methodological control more clear.

We also sympathize with the reviewer’s broader point that large differences in MAG assembly quality could influence our results in other ways. To address this concern, we have also performed a comprehensive comparison of N50 values across all five populations in our study (Tsimane, Hadza, Asia, Europe, and North America).

Our analysis reveals that while statistically significant differences in N50 exist across populations (Kruskal-Wallis test: $p < 0.001$), the effect size is negligible (eta-squared effect size, $\eta^2 = 0.0016$), indicating that population membership explains only 0.16% of the total variance in N50. The median N50 values are remarkably consistent across all populations (Tsimane median: 21,513 bp; Hadza median: 24,275 bp; Asia: 25,921 bp; Europe: 25,149 bp; North America: 30,509 bp) (**Table R1.1, Figure R1.1**). We have added **Fig. R1.1** as a new supplementary figure (Fig. S3) in the Supplementary Information.

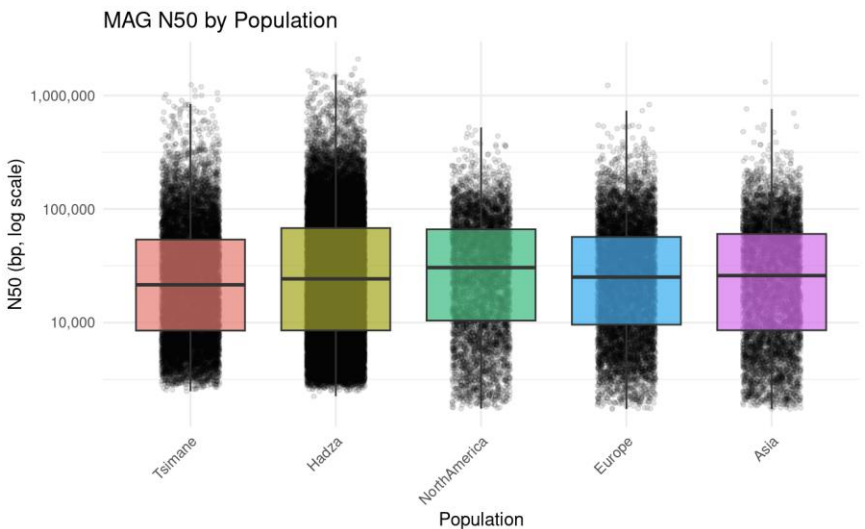


Figure R1.1. Comparison of MAG N50 values across populations. Boxplots showing the distribution of N50 values for metagenome-assembled genomes (MAGs) from five human populations: Tsimane ($n = 12,746$), Hadza ($n = 32,034$), Asia ($n = 6,426$), Europe ($n = 8,647$), and North America ($n = 5,463$). Individual data points are shown as jittered dots behind boxplots. Box centers represent medians, box limits represent the 25th and 75th percentiles (interquartile range), and whiskers extend to $1.5\times$ the interquartile range. N50 values are displayed on a log scale.

Table R1.1. Summary statistics of MAG N50 values by population.

Study	N	Median N50 (bp)	Q25 (bp)	Q75 (bp)	Mean N50 (bp)	St. Dev. (bp)
Tsimane	12,746	21,513	10,000	40,000	21,513	~10,000
Hadza	32,034	24,275	10,000	40,000	24,275	~10,000
North America	5,463	30,509	10,000	40,000	30,509	~10,000
Europe	8,647	25,149	10,000	40,000	25,149	~10,000
Asia	6,426	25,921	10,000	40,000	25,921	~10,000

Asia	6,426	25,920	8,551	60,370	41,937	50,515
Europe	8,647	25,149	9,577	56,850	42,229	52,236
North America	5,463	30,509	10,406	66,286	45,753	48,441
Tsimane	12,746	21,512	8,495	53,879	44,852	72,351
Hadza	32,034	24,275	8,521	67,984	56,073	96,174

Critically, when directly comparing non-industrialized (Tsimane and Hadza) versus industrialized populations (Asia, Europe, and North America), we find no statistically significant difference in N50 (Wilcoxon test: $p = 0.253$), with a negligible effect size (Cohen's $d = 0.123$) and minimal absolute difference in medians (3,408 bp, representing approximately 13% of the median N50). This confirms that the assembly contiguity of MAGs from the Tsimane and Hadza populations is essentially equivalent to that of MAGs from the industrialized comparator populations.

Given the negligible effect size of N50 differences across populations (eta-squared effect size, $\eta^2 = 0.0016$) and the absence of significant differences between non-industrialized and industrialized populations specifically ($p = 0.253$), we conclude that variation in MAG contiguousness does not meaningfully confound our inferences of genetic isolation between the Tsimane and Hadza microbiomes.

L99 calculations: Similarly, I wondered how the authors controlled for relatedness of hosts when conducting the within-population L99 measures. For example, were comparisons of hosts from the same family (i.e., close kin) excluded? Given that close kin are known to share strains, this could confound these analyses, which are aimed at quantifying population-level divergence.

We thank the reviewer for raising this important point. We excluded all identifiable mother-child pairs from our analyses, which represent the only level of kin relationship available in our de-identified dataset.

In addition, our L99 analysis includes a key filtering step that removes all strain comparisons with more than 10% identical genes. The remaining strain pairs are estimated to have diverged for at least several thousand years (**Extended Data Fig. 6**), ensuring that within-population L99 values primarily reflect historical gene flow rather than recent host relatedness.

To illustrate why this filter is essential for controlling potential confounding by host relatedness, we have added a new analysis that repeats the L99 analysis without the 10% identical-gene filter (**Figs. S5-S6** in the Supplementary Methods). In this case, within-population L99 values for a given species become markedly more variable across populations, presumably because the inclusion of nearly identical strains makes the analysis sensitive to the relatedness structure of specific populations. In the revised manuscript, we have added a discussion about this new result in the Supplementary Methods.

Line 381-389: Were any steps taken to account for strain heterogeneity of MAGs? If not, I wondered whether high levels of strain heterogeneity within individual hosts could affect measures of recombination and divergence times.

We thank the reviewer for raising this important point. Strain-level diversity within individual hosts is a rich area of study that we have not examined in depth here. Within the scope of our analyses, however, we believe that strain heterogeneity has limited impact on our main findings for the following reasons.

At the conceptual level, all of our population genetic analyses, including the inference of recombination and population divergence parameters, are modeling long-term processes that span multiple host-colonization cycles (i.e. at least hundreds of years) and thus make no explicit assumption of the colonization structure of any individual hosts. While the presence of diverse strains within individual hosts might impact the parameters of these long term processes, for example by facilitating gene flow within a species, such biological effects will be captured by the inferred population genetic parameters. Thus, under the assumption that each MAG faithfully represents the genotype of one strain in a given host, our population genetic modeling should be self-consistent regardless of the level of strain heterogeneity within individual hosts.

Nevertheless, we acknowledge that high within-host strain diversity could lead to technical artifacts in the MAG reconstruction process, which in turn could influence the downstream population genetic analyses. For example, if two strains of a species are present at comparable frequencies in a sample, the assembled MAG for one strain may inadvertently include SNVs originating from the other. Such chimeric SNVs could break otherwise long tracts of identical sequence, potentially inflating the apparent genetic isolation between strains from different hosts. However, any interruption of shared segments caused by these artifacts would affect both within-population and between-population strain comparisons in the same direction. Critically, this effect cannot create the markedly shorter between-population L99 values in non-industrialized populations.

To empirically test whether strain heterogeneity affects our results, we leveraged a new mapping-based metagenomic pipeline we introduced in the revised manuscript (see also Referee #2 comment 1 below) to derive a simple measure of within-host heterogeneity. For each sample-species pair, we computed the fraction of high-coverage sites with a major-allele frequency <80% (this fraction should be low for samples dominated by a single strain and higher in samples containing mixtures of divergent strains). This analysis revealed that samples from non-industrialized populations tend to exhibit higher within-host heterogeneity than those from industrialized populations, and that species with high-quality MAGs within a sample are biased toward lower heterogeneity (**Fig R1.2 Left**).

We then asked whether such systematic differences could influence our key findings. After restricting to samples with low heterogeneity (score <10⁻³), we repeated the L99 analysis using the resulting MAG subset. The resulting L99 values showed good correlation (Pearson's $r=0.90$)

with our original estimates (**Fig R1.2 Right**) and continued to show the same qualitative pattern of significantly shorter between-population L99 values in **Fig. 3c** (**Fig. S9** in the Supplementary Methods). These results indicate that, although levels of strain heterogeneity differ across populations, our main population-genetic inferences are robust to these differences.

We have added this heterogeneity analysis and the discussion above as a new section in the **Supplementary Methods** (Section 4.6), and we hope these additional results address the reviewer's concerns.

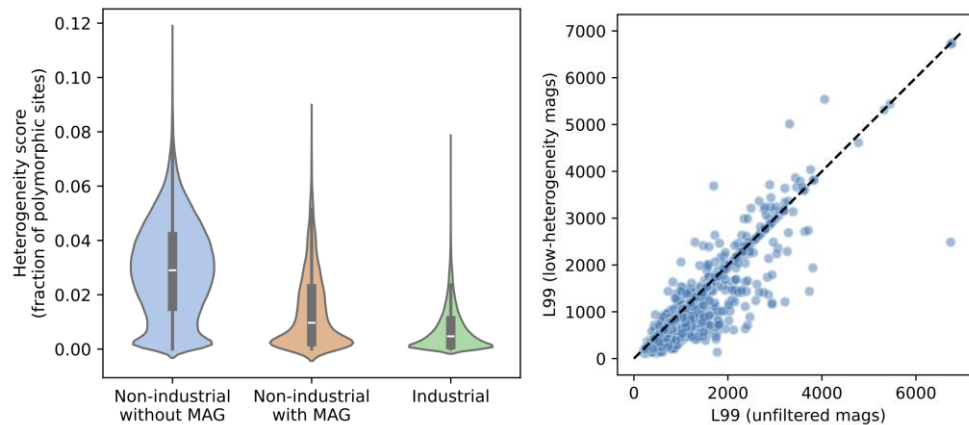


Figure R1.2. Effect of strain heterogeneity on L99 analysis. (Now Fig. S9 in the Supplementary Methods.)

Left: Violin plots showing the distribution of the strain-level heterogeneity score (defined as the fraction of high-coverage sites with a major-allele frequency <80%) across all species-sample pairs. Orange denotes cases in which a high-quality MAG was reconstructed for that species in that sample; blue denotes cases without a corresponding MAG. Green denotes species-sample pairs from individuals living in industrialized societies.

Right: Scatter plot comparing L99 values for each species computed using the full set of MAGs versus MAGs from only low-heterogeneity species-sample pairs.

Molecular Clock analyses: The authors use a classic population genetic equation that models the genetic divergence between lineages as a function of the divergence time, mutation rate, and rate of gene flow. This equation is well suited to estimate divergence times. However, I wondered whether the bacterial lineage divergence time is really the important quantity to establish / test for co-migration, given that related strains from Tsimane and Hadza may coalesce before the divergence time of the host populations. Consider, for example, a case in which the ancestral host population harbored multiple strains of a species, and when the host population diverged (eg leading to the Tsimane and the Hadza), the two descendant populations each inherited a different strain (or different sets of strains) (in a manner analogous to incomplete lineage sorting). In this case, the MRCA for strains in Tsimane and Hadza would predate the timing of the ancestral host population, and estimating the timing of the ancestral host population using bacterial genetic divergence would require some estimate of the ancestral bacterial effective population size (N_e). Given that ancestral N_e when the bacterial lineages split is likely unknown, I do not think the authors need to redo or extend any analyses (what they

have done is very reasonable). However, this consideration does affect the interpretation of results and could be discussed somewhere. For example, if the ancestral host population contained strain diversity, then the divergence times of bacterial strains between Tsimane and Hadza may predate the divergence of the host populations, even if strains co-migrated with hosts.

We appreciate this thoughtful comment and fully agree that lineage divergence between Tsimane and Hadza strains may predate the host population split due to ancestral strain diversity (analogous to incomplete lineage sorting). Indeed, as shown in **Fig. 1**, within- and between-population strain pairs show overlapping ANI values, and their inferred TMRCAs often predate the host divergence time. These observations indicate that although Eq. S1 is useful for i) establishing upper bounds on divergence between pairs of strains or ii) for estimating divergence times between bacterial *species*, it cannot be directly applied to date the recent within-species isolation (such as the Tsimane and Hadza microbiomes). For exactly this reason, we developed additional approaches (**Figs. 2-4**) that examine complementary signatures of genetic isolation due to co-migration, including recent clonal sharing, horizontal gene transfer, and SNV-frequency patterns.

In response to this and related comments, we have added a short discussion in the **Supplementary Methods** (end of Section 2) to clarify the interpretation of divergence times inferred from ANI differences.

Figure S9: It would be ideal to show the underlying data points in this figure.

Thank you for the suggestion. We have updated this figure (now **Fig. S11**) to display the underlying data points. In addition, we have revised the supplementary table of Moments model parameters to include all 71 species (after sample size filtering but prior to quality filtering) along with their residual sizes (**Extended Data Table 6**).

Absence of species from industrialized populations: It was not clear to me how absence was assessed. Is this based on lack of MAGs of these species from industrialized populations, or is it based on mapping analyses? Mapping analyses may be more sensitive to detect low abundance bacteria that do not assemble well into MAGs and therefore represent a more conservative approach for testing for absence.

The reviewer is correct that absence was assessed by MAG recovery and that a mapping-based approach would be more sensitive. We have now supplemented our MAG-based analysis with a mapping-based species prevalence quantification. To do this, we generated a species representative genome database by choosing a representative genome from each of the 3,335 species in our data set, and then mapped the reads from each metagenomic sample onto these species-representative genomes. Any genome that achieved a breadth of coverage greater than 50% in a given metagenome was deemed “present”. In **Figure R1.3** below we compare species prevalence between a MAG-recovery and mapping-based approach (now included as a new **Extended Data Fig. 3**). Indeed the mapping-based approach is much more

conservative. That said, we also show that most of the species we deem “absent” according to our MAG recovery approach are indeed quite rare in the industrialized cohorts. Of the 739 species with MAGs in the Hadza and Tsimane for which we did not recover any MAGs from industrialized samples, 703 (95%) have a prevalence less than 1% in industrialized samples using our mapping-based approach. We have rephrased the wording in the text from “absent” to “rare or absent”.

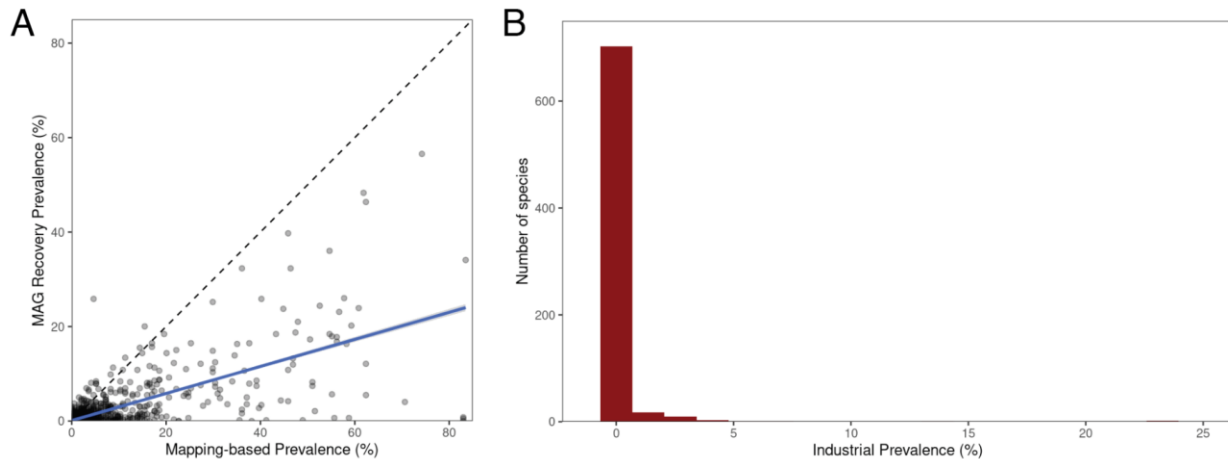


Figure R1.3. Comparison of MAG recovery-based and read mapping-based prevalence metrics.

- (A) Comparison of species prevalence metrics using a MAG recovery based approach and a read mapping based approach. Pearson correlation: 0.74, $p < 0.001$.
- (B) Read-mapping based prevalence of species in industrial samples that were deemed to be absent from industrialized samples using MAG recovery based approach (703 of 759 species have a prevalence of less than 1%).

Abstract: “Understanding which microbes have been long-term residents of the human gut, and may have co-evolved with humans, has implications for the importance of microbial biodiversity loss for health. However, the genetic complexities of microbial evolution and the plasticity of gut microbiome composition have made it challenging to define these long-term associations.” I agree with this statement, but several prior studies demonstrated parallel phylogenetic tress between host lineages and gut bacterial lineages/communities, including studies focused on human populations (such as the Suzuki/Ley paper cited) as well as studies focused on human and non-human primate species (the earliest of these go back 15 years, eg, Ochman et al., 2010 PLOS Biology: <https://doi.org/10.1371/journal.pbio.1000546>.) From these studies, it has been established that lineages of gut bacteria (including lineages identified in the present study) have been ‘long-term residents of the human gut’. Many of these microbes not only predate split of human populations from one another but also the split of Homo sapiens from other primate species.

We agree that previous comparative studies, including those between human populations (Suzuki et al., 2021) and primate species (Moeller et al., 2016; Ochman et al., 2010), have presented strong evidence for long-term host microbe associations. To better contextualize our

findings and emphasize how our study expands upon these fundamental observations, we have updated the Abstract and Introduction with this background and several new references (see our response above). We highlight that our contribution is recombination-aware genome-wide analyses in deeply sequenced microbiomes from two geographically isolated human populations of known divergence time, providing a higher resolution analysis of these long-term associations between humans and gut microbes for a more recent time window.

Introduction: “However, the comparatively recent ages of existing paleofeces (~1000-2000 years old), combined with the challenges of analyzing ancient DNA due to its degradation over time, makes it difficult to determine whether these microbes were also present in more ancestral human populations (~10-100kya) or whether they reflect a more recent acquisition from a common environmental source.”

Even without very old ancient DNA samples, comparative studies have established that many lineages of gut bacteria were not recently acquired from common environmental sources. It would not be correct to imply that DNA material from ancient specimens is required to understand ancestral states (most macroevolutionary and molecular evolutionary studies do not have very old DNA material). For example, it is known from cophylogenetic studies that *Buchnera* first formed endosymbiosis with aphid hosts hundreds of millions of years ago, even though we have no ancient DNA samples dating back 100+ million years. Similarly, the results of Suzuki and Ley’s paper are not consistent with recent symbiont acquisition, nor are the results of comparative studies of nonhuman primates.

The reviewer makes a good point. Indeed there is compelling evidence from comparative and cophylogenetic studies that many gut bacterial lineages are long-term symbionts rather than recent environmental acquisitions. We have updated the Introduction to make it clear that comparative methods that are not reliant upon ancient specimens have established long-term host-microbe relationships, and that ancient DNA analyses are merely one method of determining ancestral states. We have also revised the sentence above to focus more on faithful transmission of lineages within the host species, rather than species presence alone.

Modification to the text:

However, the comparatively recent ages of existing paleofeces (~1000-2000 years old), combined with the challenges of analyzing ancient DNA due to its degradation over time, makes it difficult to determine whether these same microbial lineages have been faithfully transmitted over longer evolutionary timescales (~10-100kya) or whether they were acquired or transmitted more recently.

Introduction: “potentially consistent with a past history of co-migration”

I don’t think the adverb ‘potentially’ is needed, unless there are specific points that the authors can make to call into question these past results.

Thanks for bringing this up. Previous work by our group (Good, 2022) and others (Perez-Lamarque & Morlon, 2024) has argued that the methodology employed in the cited study by Suzuki et al (2022) has important limitations. Specifically, these studies have demonstrated that

significant phylogenetic correlations (with magnitudes similar to those observed by Suzuki et al) can arise even in the absence of co-migration. Given these results, we believe that “potentially consistent” is the most accurate summary of the Suzuki et al paper, which gives credit to this landmark study while not oversimplifying their main results. In the revised manuscript, we have now cited the two papers above [along with a related recent review article (Moeller et al., 2023)] to provide context for this claim. With our broader reorganization of this paragraph, it should also now be more apparent that these caveats refer to within-species co-migration in particular, rather than to the broader literature on co-diversification between different host species.

Introduction: However, interpreting these correlations remains challenging due to the complex ways that microbial genomes evolve. Traditional phylogenetic methods break down for many species of gut bacteria, where horizontal gene transfer and homologous recombination produce mosaic genomes that cannot be represented by a single phylogenetic tree. Moreover, existing strain-level analyses have been biased toward species in industrialized populations, which are rarely observed in ancient paleofeces. These sampling biases have hampered our ability to infer the evolutionary history of microbial taxa that are the strongest potential candidates for ancestral co-migration.

I do not fully follow this line of reasoning. While it is true that core-gene phylogenies do not fully capture HGT/recombination events, this does not mean that core-gene phylogenies cannot be used to understand how core genomes of bacteria diversify or co-migrate with hosts. The past phylogenetic analyses from which evidence of co-migration was gleaned were based on single-copy core gene alignments, which contain genes that are generally not horizontally transmitted between distantly related bacteria. Past work (eg Suzuki and Ley et al) has shown that the lineage trees of the bacteria are associated with the lineage trees of the host – that result supports long-term associations between the bacteria and the host, even if it does not fully capture the history of recombination and HGT that those bacteria have experienced since. If HGT of these core genes was rampant, we would not expect to see signals of parallel diversification between hosts and symbionts.

Thanks for this comment, we agree that the previous version of this paragraph failed to spell out our logic in sufficient detail. We have now revised the entire paragraph to better explain our main points.

A quick point of clarification: in this context, we are using HGT and homologous recombination to refer to recombination within (rather than between) bacterial species. The studies cited in this paragraph (particularly Liu & Good, PLoS Bio 2024) show that this within-species recombination occurs at high rates in many gut bacteria even for the single-copy core genes referenced by the Reviewer. This implies that recombination potentially complicates the interpretation of the within-species phylogenies inferred in works like Suzuki and Ley et al.

Regarding the specific effects recombination has on the ability to observe phylogenetic correlations within species, we have investigated this effect in detail in the preprint referenced above (Good, 2022), which is currently under review at *mBio*. To briefly summarize the main results here, we simulated a standard population genetic model with high rates of

recombination, but in a geographically structured population with no history of co-migration. We found that genome-wide phylogenies inferred from simulations of this spatial null model exhibit statistically significant correlations with the observed human phylogeny – with effect sizes comparable to those observed by Suzuki et al – despite no underlying co-migration. Moreover, these correlations hold even for high spatial migration rates ($m = 10-30$ per Tmrca), in which most microbial strains trace back to an ancestor on a different continent within the last 5-10k years. These findings are reproduced in **Figure R1.4** below.

This counterexample demonstrates our point that on shorter within-species timescales, significant correlations between genome-wide phylogenies do not necessarily imply long-term associations between bacteria and their host, suggesting that additional care is required when interpreting these signals in the context of co-migration. They also address the Reviewer's question above, showing that these apparent phylogenetic correlations can be preserved even if within-species HGT is rampant – and regardless of whether co-migration has occurred or not. While the specific details of these simulations are beyond the scope of our present study, we hope that our description of them here provides some concrete justification for our phrasing in the Introduction. In the revised manuscript, we have added a reference to the preprint above to provide more context for this subtle point, and we have also revised the overall phrasing to better emphasize the specific gaps addressed by our present study.

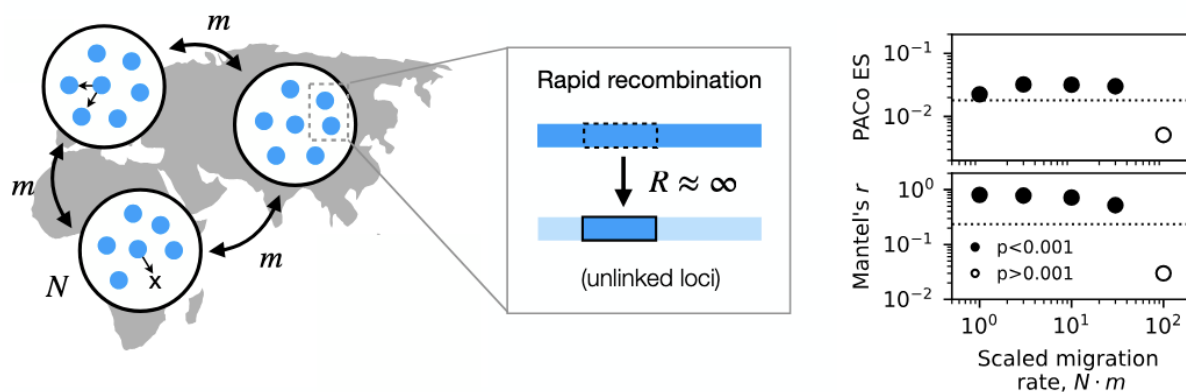


Figure R1.4. Geographic structure can produce statistically significant phylogenetic correlations even in the absence of co-migration (reproduced from the updated version of Fig. 1B from Good, 2022) A spatial null model with finite rates of migration (left) and widespread within-species recombination. Recombination occurs infinitely quickly within each continent (center), leading to rapid unlinking of mutations. Genome-wide phylogenies inferred from simulations of this model exhibit significant correlations with the observed human phylogeny (right), even when most strains recently migrated from another continent ($Nm \gg 1$). Points denote simulation results, while the dashed lines show the corresponding correlations observed by Suzuki et al (2022) for *Prevotella copri*.

Modifications to the text:

However, interpreting these correlations remains challenging on the shorter timescales corresponding to human dispersal out of Africa. Traditional phylogenetic methods²⁰ break down

on these timescales, where horizontal gene transfer and homologous recombination between strains within the same species produce mosaic patterns of ancestry that cannot be represented by a single phylogenetic tree^{21–24}. These high rates of recombination make it difficult to distinguish ancient co-migration from other sources of population structure including geography or host variation in lifestyle^{18,19}. Moreover, existing strain-level analyses have been biased toward species in industrialized populations²⁵. These sampling biases have hampered our ability to infer the evolutionary history of microbial taxa that are the strongest potential candidates for ancestral co-migration.

Line 120: “therefore requires that they must have shared a common ancestor within the last ~100-200k year”

Given the uncertainty around bacterial molecular clocks, this statement may be too strong. As the authors note in the supplemental material, the estimate could be off by an order of magnitude.

We thank the reviewer for this comment. We have revised the text to soften this statement.

Modification to the text:

The typical ANI between Tsimane and Hadza MAGs (~98%; Fig. 1c) therefore suggests that they likely shared a common genetic ancestor within the last ~100-200k years (Extended Data Fig. 5).

Line 131: “The broad overlap between these distributions indicates that most of the diversity within the Tsimane population is much older than the onset of geographic isolation”

This could be edited to make clear that you mean diversity “within bacterial species”. Most gut bacterial diversity within the Tsimane must be older than the onset of their geographic isolation from the Hadza, given that the microbiota contains multiple phyla. The result reported in this section is consistent with the possibility that bacterial species contain multiple ancestral strains, each of which co-migrated with hosts.

We thank the reviewer for this comment. We have revised the text to address this potential confusion.

Modification to the text:

The broad overlap between these distributions indicates that most of the strain-level diversity within the Tsimane microbial populations is much older than the onset of geographic isolation.

Lines 172-177:

This conclusion seems self-evident, ie, if two host populations share a species it does not mean they share the same strain. If there is a broader point here, it could be made more clearly.

We agree with the reviewer that this section requires clarification. The examples in the previous paragraph suggested that the industrialized populations might be serving as a vector for globally prevalent species to the Hadza and Tsimane, so we wanted to test whether this was always

true. To our surprise, this analysis revealed that some globally prevalent species that exhibit high strain sharing between industrialized populations do not show recent strain sharing with the Hadza and Tsimane. Our intention was to draw a contrast between these two groups of species: those that have been shared recently across all populations (e.g., *Akkermansia muciniphila*) and others that, while shared recently in industrialized populations show no recent sharing between the Hadza and Tsimane (e.g., *Agathobacter rectalis*, *Ruminococcus bromii*). Therefore, global prevalence of a species does not necessarily imply that the species is being transmitted (e.g., via globalization, etc.) across all populations or lifestyles. We have honed the wording to clarify the two groups of species that we are referring to more explicitly.

Modifications to the text:

Intriguingly, the comparison with industrialized MAGs also revealed another class of species that are prevalent across lifestyles, but exhibit little recent strain transmission between the Tsimane and the Hadza (Fig. 2e, Extended Data Fig. 8). These taxa, which include common species like Agathobacter rectalis and Ruminococcus bromii, suggest that Tsimane and Hadza strains can remain isolated despite the species' global prevalence and recent sharing within industrialized populations.

Referee #2 (Remarks to the Author):

This is an interesting paper from Carter and colleagues. The paper is well written and attempts to answer an exciting question in microbiome research: how connected are specific gut microbial species with human evolutionary history, and to what extent have industrialization and lifestyle changes affected this association? The paper uses deep metagenomic sequencing and population genetic analyses to show that gut microbial species (some of which absent from industrialized populations) have likely co-migrated with humans. I believe this is an important paper with a proof of concept that deep, strain-resolved metagenomic sequencing combined with population genetic analyses is a powerful approach for reconstructing the evolutionary history of the human gut microbiome across diverse populations. In addition, the biological results, although maybe not surprising, are exciting. At the same time, there are a few issues:

1. Justification of parameter choice:

One big question is regarding the parameter choices for MAG generation and subsequent analyses. The analysis relies on established thresholds for MAG quality (e.g. 50% completeness and 10% contamination), which are in line with current standards. However, they remain arbitrary to some degree. This is important because variations in these thresholds can affect the inclusion or exclusion of certain MAGs, potentially affecting the observed diversity and the number of shared microbial species or strains. Similarly, the use of a 10% identical gene cutoff to define recent clonal relationships is somewhat subjective. The age boundary linked to this threshold varies across species and is calibrated only through estimates of bacterial mutation rates, which themselves have broad confidence intervals. Small changes to this cutoff might shift the inferred rates of recent strain sharing, especially in comparisons across populations. Other steps (filtering criteria for core gene identification, clustering thresholds, and minimum gene alignment lengths) can also introduce parameter-dependent variation. Together, these choices could change the resolution and sensitivity of downstream analyses. It is unclear if conclusions in the paper change under alternative parameter settings. This can potentially be addressed by testing different parameters in the MAG assembly and strain-sharing analyses, performing sensitivity analyses to determine how varying these cutoffs affect the qualitative results, and reporting the degree of stability or variability in key findings as parameters are adjusted.

We thank the reviewer for raising this important point. Like many large-scale metagenomic analyses, our pipeline involves numerous parameter choices at different steps. We agree that, in principle, varying some of these thresholds could lead to differences in the inclusion of particular MAGs or strains, which might in turn influence downstream results. To directly test the stability of our findings with respect to parameter choices, we have introduced two major analyses that address the two specific stages of the pipeline the reviewer highlighted: i) MAG assembly and ii) strain-sharing analysis.

First, rather than attempting to independently vary every parameter in the MAG assembly pipeline, which is computationally challenging given the large number of interacting steps, we sought to bypass this complexity by introducing an entirely independent analysis pipeline based

on read mapping (new **Supplementary Method** Section 4.5). This approach circumvents the parameter dependencies inherent to the MAG reconstruction process. We adapted our gene flow analysis to this mapping-based framework and extended this analysis to the full genome to further reduce dependence on core-genome related parameters. These new mapping-based estimates were strongly correlated with the previous MAG-based values across the vast majority of species and reproduced the key qualitative patterns in **Fig. 3** (**Fig. R2.1** and **Fig. S10-11**). This mapping-based analysis also enabled us to more robustly detect rare taxa, which is discussed in more detail in our other response below (**Fig. R2.4**). Overall, the close agreement between the MAG-based and mapping-based estimates demonstrates that our main conclusions are robust to the details of MAG reconstruction.

Second, we examined the effect of varying the threshold used to define recent clonal relationships. As the reviewer noted, this threshold is inherently somewhat arbitrary and directly affects the precise values of clonal sharing rates in **Fig 2**. However, because nearly all Tsimane-Hadza strain pairs exhibit close to zero identical genes (**Fig. 2b**), the key population-level pattern of minimal strain sharing is expected to be robust to the specific cutoff. To demonstrate this, we scanned a broad range of thresholds (5-50%) and repeated the strain-sharing analysis. Across this full range, we observed little to no clonal sharing between Tsimane and Hadza in almost all species (**Fig R2.2**). However, as the reviewer points out, the specific choice of clonal relationship does determine the associated age boundary, which will also vary across bacterial species. Therefore, this result should be interpreted as a qualitative pattern of reduced strain sharing over an approximate timescale of ~5,000 years, rather than a precise estimate for each species. We have revised a sentence in the main text to highlight this parameter dependence and added a corresponding discussion in the **Supplementary Methods** (Section 3).

Taken together, we believe that these results, along with the expanded discussion in the **Supplementary Methods**, directly address the reviewer's concerns about parameter sensitivity and reinforce the robustness of our key findings.

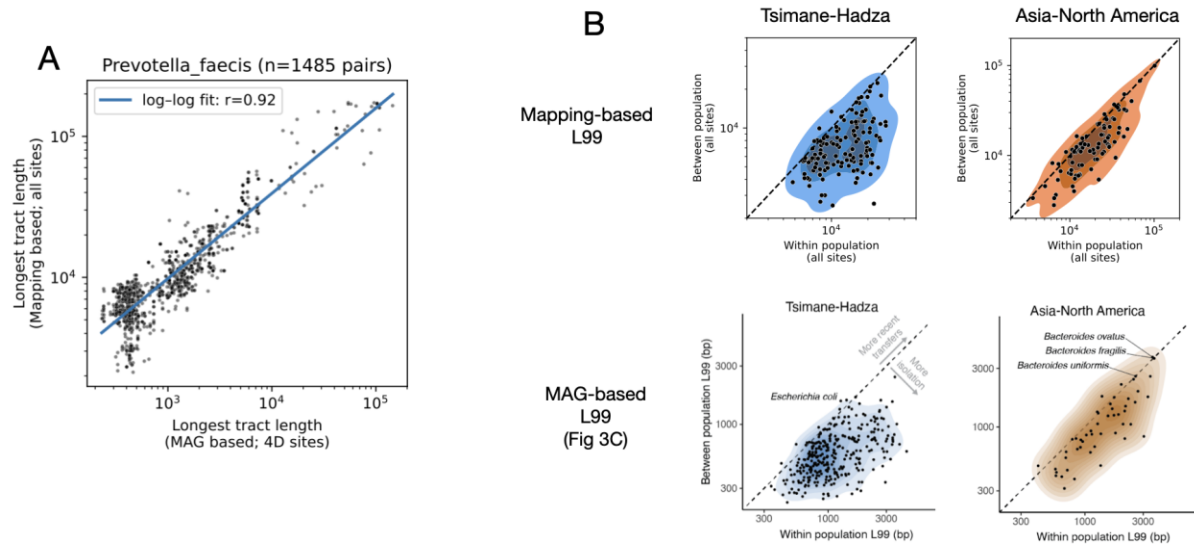


Figure R2.1. Mapping-based validation of IBS tract estimates.

- (A) Comparison of the longest IBS tract lengths inferred from MAG-based and mapping-based approaches for one representative species. Each point corresponds to a pair of samples for which both methods produced an estimate. Pearson's r denotes the correlation between the log-transformed tract lengths. Correlations for additional species are provided in **Fig. S7** of the Supplementary Methods.
- (B) Scatter plots comparing between-population and within-population L99 values across species. The mapping-based analysis reproduces the qualitative pattern of stronger recent isolation between Asia and North America than between the Tsimane and Hadza.

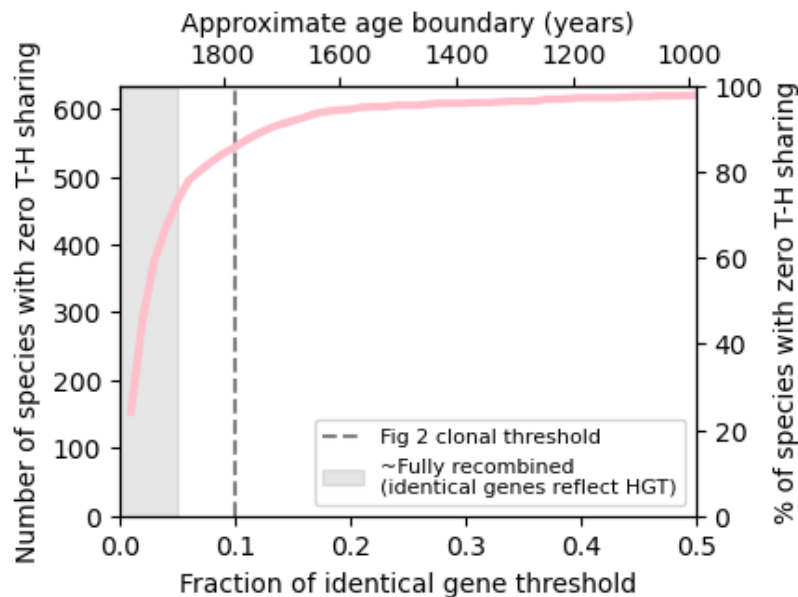


Figure R2.2. Lack of clonal strain sharing between Tsimane and Hadza is robust to the definition of clonality. The clonal strain-sharing rates shown in **Fig. 2** were recalculated using

a range of thresholds for the fraction of identical genes between genomes. The solid line indicates the number of species with zero strain sharing at each threshold. The secondary x-axis shows the estimated age boundary corresponding to each threshold for defining clonal relationships, as inferred from recombination rate analysis. The shaded region marks thresholds below 5% identical genes, which no longer reflect clonality and are thus not informative for this analysis. This sensitivity analysis demonstrates that the vast majority of species show no recent strain sharing between Tsimane and Hadza, regardless of the specific threshold used to define clonal strains.

Along the same lines, one specific area is the section on microbial biodiversity loss, which relies heavily on the parameters chosen to define the presence or absence of a taxon in industrialized populations. The sample size required to confidently call a taxon as “absent” is not clearly specified, and increasing the number of samples could alter detection rates. This is especially true for rare taxa. In addition, details on the sequencing depth and the computational strategy for taxon identification (whether through MAG assembly, read mapping, reference-based profiling) can influence which species are detected and classified as “vanishing” or persistently absent. All of these methodological choices can have a big effect on the observed patterns and conclusions. Similar to the above, an evaluation across a range of parameters and approaches would help establish the robustness of the findings. Sensitivity analyses and modeling the probability of taxon detection as a function of sample size, using data from other rare taxa, would be especially helpful.

We agree with the reviewer that it is important to clarify. Below we present several analyses that address the question of species richness and taxon detection using multiple methods.

First, we performed a rarefaction curve analysis to understand how cohort size affects overall species recovery. Each of the 5 studies used in this meta-analysis had a different number of subjects: Hadza (137 subjects), Tsimane (85 subjects), Asia (361 subjects), North America (253 subjects), and Europe (258 subjects). To determine the extent to which this different subject number contributes to the observed differences in unique species detected in each population, we generated a rarefaction curve (**Figure R2.3**). To generate this curve we randomly selected subjects at different levels of sampling (increments of 5 subjects up to 85). At each sampling we determined the number of unique species found in that collection of subjects. We then fit a species accumulation curve [according to $S = (S_{max} \times N) / (k + N)$, where S is the number of species, N is the number of subjects and S_{max} is the maximum number of species] for each population's rarefaction curve. Using these curves we can estimate the asymptote, or the maximal number of species present in the population. For the Hadza, the estimated number of species is 2,443 (95% C.I.: 2,385-2,503 species); for the Tsimane the estimated number of species is 1,590 (95% C.I.: 1,561-1,620); for Europe the estimated number of species is 999 (95% C.I.: 952-1052 species); for North America the estimated number of species is 582 (95% C.I.: 545-620 species); for China the estimated number of species is 490 (95% C.I.: 464-519 species). These results indicate that differences in unique species detected in these two cohorts reflect actual differences in underlying gamma diversity, rather than an artifact of sampling effort.

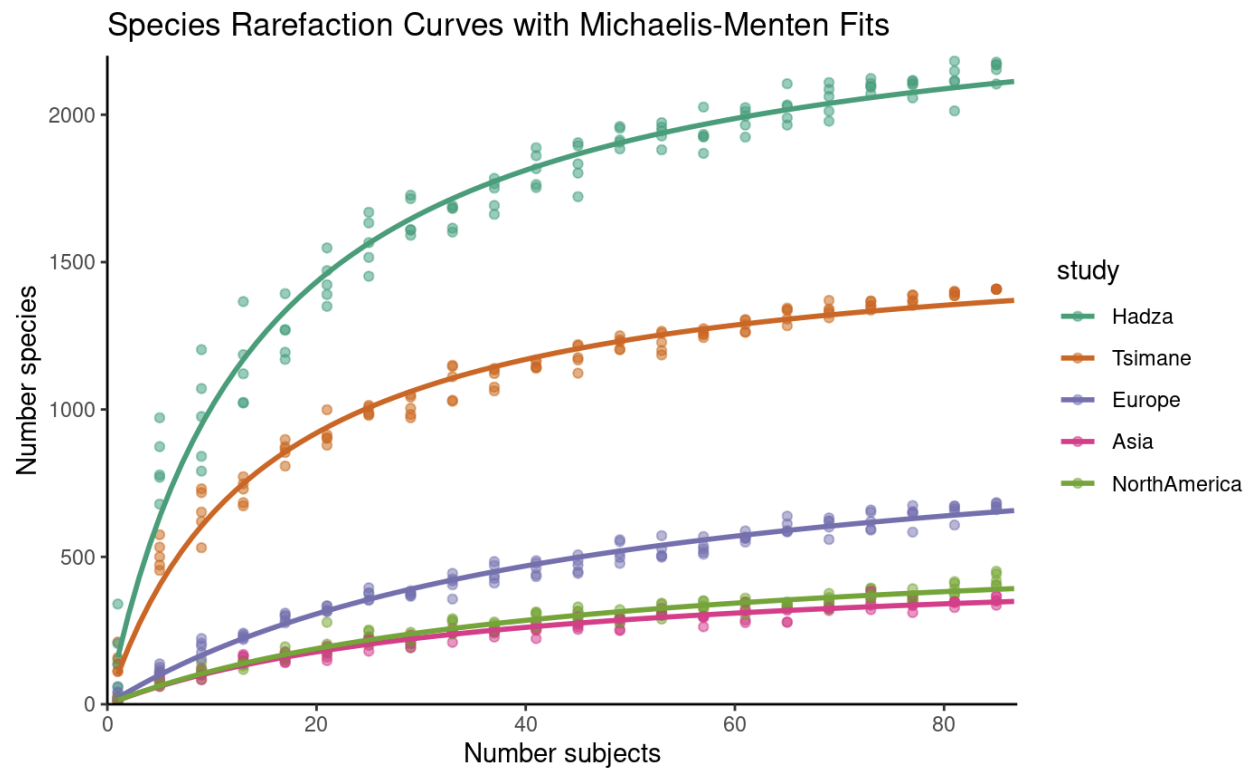


Figure R2.3. Rarefaction curve showing the number of unique species detected in each population given some number of randomly selected subjects.

As the reviewer points out, the computational approach used can influence taxon detection. In the analysis above we are just using MAG recovery from each sample to determine species presence or absence (of note: for the industrialized populations, we did not do *de novo* MAG generation, rather we simply used the MAG collections as published). We therefore performed an orthogonal analysis where we mapped reads from the industrialized cohorts onto our MAG database. To do this, we generated a species representative genome database by choosing a representative genome from each of the 3,335 species in our data set according to the following score: $\text{Completeness} - 5 \times \text{Contamination} + 3 \times \log_{10}(\text{N50}) + \text{Length}/1\text{e}6$. We then mapped reads from each metagenome onto these species-representative genomes. Any genome that achieved a breadth of coverage greater than 50% in a given metagenome was deemed “present”. In **Figure R1.3** we compare species prevalence between a MAG-recovery and mapping-based approach (now included as a new **Extended Data Fig. 3**). Indeed the mapping-based approach is much more conservative. That said, we also show that most of the species with no recovered MAGs are indeed quite rare in the industrialized cohorts. Of the 739 species with MAGs in the Hadza and Tsimane for which we did not recover any MAGs from industrialized samples, 703 (95%) have a prevalence less than 1% in industrialized samples using our mapping-based approach.

Taken together, we believe that these two orthogonal approaches show that the large differences in gamma diversity between the Hadza/Tsimane and industrialized populations is

robust to sample size. Furthermore many of the species that are shared between the Hadza and Tsimane are indeed rare or absent from the industrialized cohorts.

2. Statistical and computational issues:

There are a few questions regarding the statistical and computational analysis that I outline below.

Multiple testing: The manuscript makes hundreds of comparisons across different microbial species, but in most places I could not find a mention of correction for multiple hypothesis testing (e.g., FDR. This can potentially cause inflation of false positive rates in the presented results. This is relevant throughout the paper, but especially in the large-scale pairwise ANI calculation, strain sharing rates and percentage of clonally related pairs (Fig 2), gene flow metrics (Fig 3).

We apologize for the confusion. To clarify, for most of our analyses, we controlled for multiple testing by comparing aggregate distributions of individual datapoints or summary statistics across species (e.g. **Fig. 1a**), rather than looking for a small number of outlier species. However, we agree that these statistical tests were either omitted or not well described in the previous version of the manuscript. We have therefore revised the text to add in explicit P-values where appropriate, including in **Figs 2 and 3**. Here is a brief summary of the tests that we now describe in the manuscript:

- We now perform a permutation-based test to determine how likely it is that the Tsimane-Hadza strain sharing rates observed for each species could be observed by chance (e.g. the right-most plots of **Figs. 2a,b**). We calculate a p-value separately for each species by randomly permuting population labels 10^4 times and calculating the T-H strain sharing rate in each permutation. The empirically determined p-value, then, is the fraction of permuted T-H strain sharing rates that is greater than the observed value. We then perform multiple-testing corrections using the Benjamini-Hochberg method.
- We also perform a test in which we randomly permute the population labels for all MAGs 10^4 times (within each species) and calculate the number of species for which there is a strain sharing rate of 0%. We use this to empirically determine the P-value of the number of species with 0% species that we observed in our data (545 species out of 636). This yielded a P-value $<10^{-4}$.
- We perform an additional species-level permutation test of L99 differences, in which P-values are computed from within-species label permutations and adjusted for multiple testing using the Benjamini-Hochberg procedure. This analysis identifies species exhibiting significant between-population isolation as measured by the L99 metric; notably, the vast majority of species (295 out of 319) show significant isolation between the Tsimane and Hadza populations. The resulting P-values have been added to **Fig. 3b** and **Extended Data Table 5**.

Together, these analyses confirm that our overall conclusions are robust after appropriate multiple-testing correction. We thank the reviewer for raising this issue.

Potential confounders and batch effect: It was not clear how potential confounders (such as sample collection approach, sample preservation and handling, sequencing platform, sampling year, sequencing coverage, QC values, sequencing batch of course, etc) were controlled for in the statistical analysis. This could lead to systematic errors between cohorts, so it would be helpful to incorporate this across analyses throughout the paper. This issue appears throughout the Results (Fig 1, 2, 3) when comparing different populations (Tsimane vs. Hadza vs. Industrialized).

We agree that differences in sample collection, preservation, and sequencing pipelines across cohorts can introduce potential confounders. The Hadza and Tsimane metagenomes were generated in separate field campaigns and sequenced at different facilities several years apart, so complete harmonization of technical variables is not possible. However, we did attempt to process these stool samples in similar ways using Qiagen Power Soil kits, Illumina Nextera Flex library prep kits and Illumina NovaSeq 600 sequencers. Additionally, we attempted to sequence these samples to similar depth (we achieved ~21 Gbp/sample for the Hadza and ~28 Gbp/sample for the Tsimane). All metagenomes from non-industrial populations were pre-processed, assembled and binned using the same computational pipelines and databases. In **Figure R2.5** below we show the quantification of various quality metrics related to metagenomic sequencing, assembly and binning.

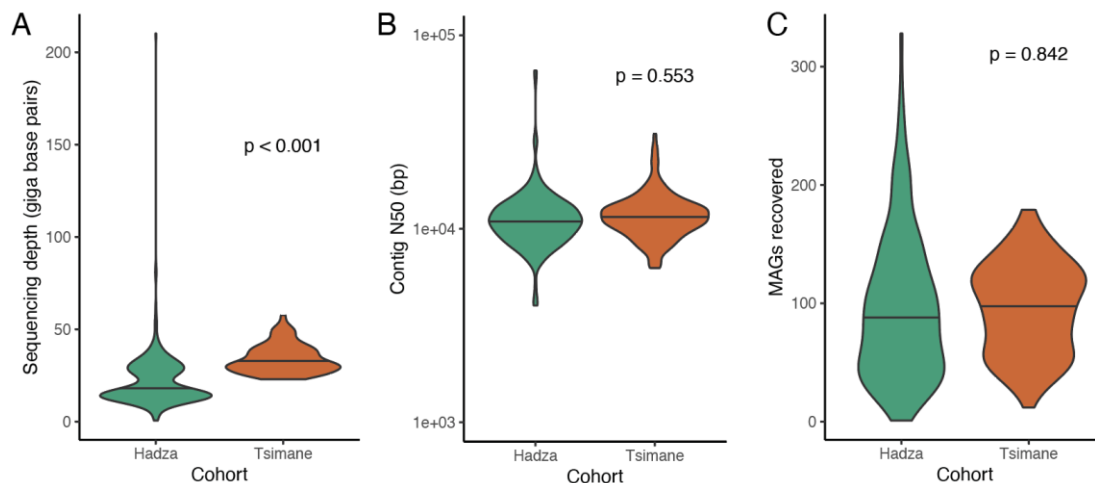


Figure R2.5. Comparison of sequencing metrics between the Hadza and Tsimane cohorts.

- Sequencing depth (measured in giga base pairs) across the Hadza and Tsimane samples used in the study.
- Weighted median (N50) of contig lengths in assemblies from Hadza and Tsimane metagenomes..
- The number of medium-quality and higher MAGs recovered from each Hadza and Tsimane sample.

All P-values were calculated using a t-test comparing the two cohorts.

These analyses show that, while there is some difference in the average sequencing depth per sample, this does not result in significantly different assembly sizes nor differences in the number of MAGs we could recover per sample.

In our response to Reviewer 1 (Comment #2), we also show that while there are some differences in MAG N50 across populations, the effect size is small. We also note that the MAGs from the American, European and Chinese cohorts were downloaded directly from online databases, namely the UHGG (Almeida et al., 2021).

Together these analyses show that while there are some minor differences between populations for these various QC metrics, the effect sizes are small and unlikely to bias our results. Moreover, since the bulk of our analyses focus on SNP differences between pairs of strains (typically supported by many independent reads in both samples), they are less likely to be biased by batch effects compared to other MAG statistics.

Uncertainty estimation: many population-level estimates (e.g. prevalence, strain sharing rates, split time estimates) are presented as single numbers, and lack estimation of uncertainty (confidence or credible intervals) making it challenging to evaluate the reliability of these estimates. This is also relevant in the context of point 1 above.

Thanks for this suggestion. As with the reviewer's above comments regarding multiple testing, we have sought to expand and make more explicit the statistical tests we use with regards to various estimates that we make. While the large number of species makes it difficult to illustrate these confidence intervals on many plots, we have now created a new **Extended Data Table 1** that reports the prevalence, inter- and intra-population strain sharing rates and split time estimates (**Fig. 3**) along with their uncertainties, where relevant. We also now perform two-tailed Mann-Whitney U-tests on the estimated population split time distributions in **Fig. 3d** comparing the Hadza-Tsimane split times and Asia-North America split times, as well as the Hadza-Tsimane split times and Europe-North America split times and show that the results are statistically significant (both $P < 10^{-4}$).

Together with the p-values discussed above, these two different ways of quantifying uncertainty should provide some basis for interpreting the numerical values we report.

3. Data sharing:

One of the major strengths of the paper is the deep metagenomic data generated. As such, a significant limitation of this study is the lack of immediate public availability of the data. This includes both the raw sequencing data and the processed data tables underlying the analysis. Open access to raw data is critical for independent validation, reproducibility, and reanalysis as new computational tools and reference databases emerge. Equally important are the processed data tables, which include quality-controlled MAGs, strain-sharing matrices, species presence/absence profiles, and other summary statistics that support key results and figures in the paper. Without access to these, the scientific community is unable to evaluate the analytical

pipelines, test alternative parameter choices, or benchmark findings against other datasets. Most importantly, making both raw and processed data available enables meta-analyses that put findings in a broader context.

Although it is stated that the raw sequencing data will be made available upon publication through the data access procedures established by the Tsimane Health and Life History Project, this approach still presents challenges. Restricted or conditional access can impede timely and widespread data reuse, and may slow down validation or meta-analyses. In addition, sharing protected data can be done through standard databases (e.g. NIH dbGap) - as is, it is not clear where and how the data will be available, what is the process for getting accesses, what are the criteria for determining whether access will be granted, etc. Following the link in the paper (<https://tsimane.anth.ucsb.edu/data.html>) and clicking the “Data” section leads to a broken link, which is a bit worrying for data availability for the current paper. Also, importantly, the paper does not address whether and how processed data tables will also be made accessible.

We thank the reviewer for raising these important points. We have updated our Data Availability section which is included below for convenience. Additionally, we have verified that on our end the link referred to by the reviewer works fine, so we apologize for any inconvenience, and suspect that this was a temporary glitch. As described below, for ethical reasons data access is monitored by the THLHP group, who actively works with external groups on matters of data accessibility for other projects.

We agree that the processed data files and tables have fewer ethical and privacy concerns than the raw data, and are important for enhancing rigor and reproducibility. Upon publication of the manuscript, we will release the processed data files from our analyses in a Zenodo archive (DOI: 10.5281/zenodo.17967214). These files will include the SNV catalogs and intermediate processing files of the 636 species which were used in our population genetic analysis. This will enable others to fully reproduce our results. We have also updated the manuscript to refer to Extended Data Tables which contain more metadata for all of the recovered MAGs, species prevalence profiles as well as other summary statistics from our population genetics analyses that support key results and figures in the paper.

Data Availability

Individual-level data are stored in the Tsimane Health and Life History Project (THLHP) Data Repository, and are available through restricted access for ethical reasons. THLHP's highest priority is the safeguarding of human subjects and minimization of risk to study participants. The THLHP adheres to the “CARE Principles for Indigenous Data Governance” (Collective Benefit, Authority to Control, Responsibility, and Ethics), which assure that the Tsimane 1) have sovereignty over how data are shared, 2) are the primary gatekeepers determining ethical use, 3) are actively engaged in the data generation, and 4) derive benefit from data generated and shared for use whenever possible. The THLHP is also committed to the “FAIR Guiding Principles for scientific data management and stewardship” (Findable, Accessible, Interoperable, Reusable). Requests for individual-level data should take the form of an application that details the exact uses of the data and the research questions to be addressed,

procedures that will be employed for data security and individual privacy, potential benefits to the study communities, and procedures for assessing and minimizing stigmatizing interpretations of the research results (see the following webpage for links to the data sharing policy and data request forms: <https://tsimane.anth.ucsb.edu/data.html>). Requests for individual-level data will require institutional IRB approval (even if exempt) and will be reviewed by an Advisory Council composed of Tsimane community leaders, community members, Bolivian scientists, and the THLHP leadership. The study authors and the THLHP leadership are committed to open science and are available to assist interested investigators in preparing data access requests.

Processed data files for population genetics analysis can be found on Zenodo (DOI: [10.5281/zenodo.17967214](https://doi.org/10.5281/zenodo.17967214)). All Hadza data is publicly available as specified in the manuscript (DOI: [10.5281/zenodo.8072245](https://doi.org/10.5281/zenodo.8072245)).

4. Population genetics analysis:

I am not an expert in population genetics modeling, so this is not a major issue, but I found several of the analyses and metrics used in the paper challenging to fully interpret. I believe that some clarification would benefit readers who are not experts in this area. First, it would be helpful to include a clearer explanation of the rationale and methodology behind key analyses such as the use of the L99 metric and the demographic inference using SNV frequencies, including what each metric represents, how it is calculated, why it is appropriate for inferring gene flow or split times, etc.

We thank the reviewer for bringing this up. Due to space limitations, we placed these pedagogical descriptions in the **Supplementary Methods** (36 page pdf including 16 figures), which contain detailed explanations of the rationale, calculation, assumptions, and interpretation of each metric. However, we recognize that this may not have been sufficiently clear from the Methods section in the main text. In the revised manuscript, we have also added explicit pointers in the Main Text and Methods to guide readers to the relevant Supplementary sections. In response to the reviewer's suggestions, we have also expanded these Supplementary sections to address the specific points below.

It would also be helpful to state explicitly the main assumptions underlying population genetic models (e.g. regarding recombination, mutation rate, or population structure), explain how realistic these assumptions are for the data at hand, and any impact that possible violation of these assumptions might have on interpretation of the results.

Thank you for this suggestion. We have revised the relevant sections of the **Supplementary Methods** (e.g. Section 4.3 and 4.5) to make the underlying model assumptions more explicit and to clarify how they relate to the data and interpretation of our results.

Also, it would be good to include additional detail on how model fit to the data was assessed, the robustness of parameter choices (related to my first point above), and how uncertainty or confidence intervals for key estimates was incorporated.

Thank you for this suggestion. In the revised **Supplementary Methods**, we have included the new robustness analyses above (Section 3 and 4.5). The procedures we use to assess model fit are described in Sections 5.2 and 5.3, and the approach for estimating uncertainty in divergence time parameters is described in Section 4.4.

Lastly, as only a subset of species with sufficient data were included in some analyses, some discussion of the implications for the generalizability of findings to less well-sampled taxa would be beneficial.

Thank you for this suggestion. We agree that the analysis in **Fig. 4** is limited to a much smaller subset of species compared to **Figs. 2 and 3**. In the revised **Supplementary Methods**, we have added a new section to discuss the extent to which the patterns observed in **Fig. 4** may generalize to less well-sampled taxa, as well as taxa with more complex population structure (Section 5.4).

Referee #3 (Remarks to the Author):

The authors employed population genetic methods to infer evolutionary history from genetic diversity patterns in contemporary microbial genomes, utilizing advances in strain-resolved metagenomics. They specifically focused on two geographically distant non-industrialized populations: the Tsimane horticulturalists of Bolivia and the Hadza hunter-gatherers of Tanzania, whose ancestors separated thousands of years ago.

They did deep metagenomic sequencing on 133 fecal samples from 85 Tsimane adults, recovering 12,746 metagenome-assembled genomes (MAGs) representing 1,408 bacterial and archaeal species. These were then compared to 32,034 MAGs from 137 Hadza hunter-gatherers. This allowed for the quantification of how these populations' gut microbial species have diverged over the last 100,000 years.

Key findings:

1. Despite vast geographic separation and no direct historical contact, 87.4% (1,231) of species in the Tsimane cohort were also found in Hadza individuals. Crucially, the majority (60.2%, 848) of these shared species were undetected in the fecal microbiomes of industrialized populations.
2. The authors calculated the average nucleotide identity (ANI) between MAGs. The typical ANI between Tsimane and Hadza MAGs (around 98%) suggests they must have shared a common ancestor within the last 100-200,000 years, based on estimated mutation rates.
3. While average ANI showed little geographic differentiation, a stronger trend was observed in the high ANI tail (>99.75%) of the distribution, showing a significant enrichment of within-

population pairs. This suggests restricted strain transmission between populations in recent evolutionary history, consistent with divergence due to historical separation.

4. The study defined clonal strains as MAG pairs with >10% identical genes. 85.6% of shared species (545 out of 636) showed no recent Tsimane-Hadza strain sharing, indicating very limited inter-population strain exchange in the past ~5,000 years.

5. To focus on gene flow via HGT in MAG pairs with less than 10% identical genes, an orthogonal approach was used. By quantifying the 99th percentile of tract lengths, the authors found that most species had shorter L99 lengths between Tsimane and Hadza compared to within populations, indicating genetic isolation in the recent past.

6. The authors tried to estimate the timing of genetic isolation between microbiomes. Genetic exchange was notably more recent between industrialized populations compared to non-industrialized populations (Tsimane-Hadza: median ~43,563 years). This estimated onset of genetic isolation between the Tsimane and Hadza microbiomes roughly aligns with the timeframe of initial human migration out of Africa (60-70,000 years ago) and the settlement of the Americas (16-23,000 years ago).

Overall, the article is well written and clearly structured; the following are a few specific comments that I believe could further improve it.

Questions and suggestions:

1. The authors should clarify that ANI is robust for species level identification, but it cannot be used naively for dating. In gut microbes with high recombination, horizontal gene transfer creates mosaic genomes, so it does not always reflect vertical evolutionary history. As a result, divergence time estimates such as “Tsimane and Hadza strains last shared a common ancestor ~100-200k years ago” may be biased. The clonal-segment analysis presented later is therefore the more appropriate framework for inferring divergence times in this context.

We thank the reviewer for this helpful comment and fully agree that ANI is robust for species-level identification but cannot be naively used for dating divergence times due to recombination. To clarify this, we have updated the Main Text and **Extended Data Fig. 5** to emphasize that ANI-based divergence times represent an *upper bound* on the time to the most recent common ancestor. We have also revised the Supplementary Methods to make this limitation explicit.

2. The authors state that ~60% of species in Tsimane and Hadza are absent from industrialized populations, which may be biased by the smaller sample sizes in these cohorts. Did the authors perform any sample size correction to account for potential undersampling of these isolated populations?

Unfortunately all of these sampling campaigns were done independently from one another so there was no explicit attempt to control for sample size. However, it is worth noting that the

industrial cohorts were the biggest cohorts of the 5 that we studied in this meta-analysis (summarized in **Table R3.1** below), decreasing concern that the detection of species in the non-industrial cohorts and lack of detection in the industrial cohorts could be caused by differences in sample size.

Table R3.1. Summary of sample sizes per cohort

Cohort	Sample size (number of subjects)
Hadza	137
Tsimane	85
Europe	258
China	361
North America	253

We also would like to refer the reviewer to **Figure R1.2** regarding additional analyses regarding species detection in industrialized cohorts using a more sensitive read mapping-based approach. We have adjusted wording from “absent” to “rare or absent” to reflect the difficulty in establishing true absence. Additionally, we perform a rarefaction analysis below in **Figure R3.3** to address whether species richness observed in the Tsimane is affected by sampling effort.

3. The authors use recombination tract lengths to infer stronger genetic isolation between Hadza and Tsimane compared to industrialized populations. However, recombination rates vary widely by species, genome size, and mobile elements. If industrialized datasets are enriched for high-recombining taxa (producing long tracts) while Hadza/Tsimane datasets contain more low-recombining taxa (producing short tracts), the observed differences in gene flow may partly reflect species composition rather than true host-associated isolation. This possibility should be considered in the interpretation.

We thank the reviewer for raising this important point. To control for potential confounding effects of variation in species composition between populations, we also examined our gene flow metric (L99) within species that are prevalent across all populations. As shown in **Extended Data Fig. 9**, even when restricted to the same species, the Tsimane and Hadza strains exhibit stronger genetic isolation compared to those from industrialized populations. We have previously noted this result in the main text (“This same trend was also observed within species that were prevalent across lifestyles...”). To clarify the significance of this analysis as an important control, we have revised the caption of **Extended Data Fig. 9** accordingly.

As an additional analysis, we directly tested whether inferred recombination rates systematically bias the inferred population split times (**Figure R3.1**). Specifically, we examined the relationship between the ratio T_{mrca}/T_{mosaic} , which quantifies the relative strength of recombination versus

mutation (Liu & Good 2024), and the inferred population split time. As shown in the updated analysis, we observed no significant correlation between this ratio and split time, indicating that variation in recombination rates does not systematically influence our split-time estimates.

Together, we hope these analyses address the reviewer's concern.

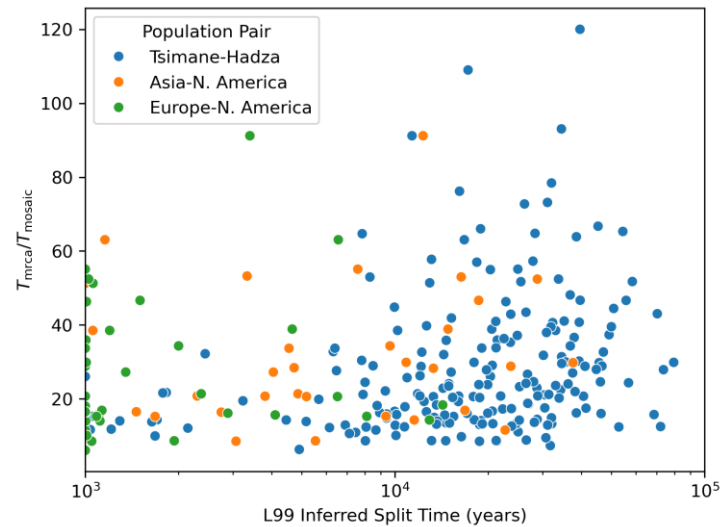


Figure R3.1. Relationship between inferred recombination parameters and population split time. Scatter plot comparing the relative strength of recombination versus mutation (T_{mrca}/T_{mosaic}) and the inferred population split time for the species included in **Fig. 3c**. No significant correlation was observed.

4. Line 117: For the estimate of molecular clock and mutation rate for gut bacteria – isn't this expected to differ between species? How much does it vary?

We thank the reviewer for this comment; a similar point was also raised by Reviewer 1. In Section 1.2 of the **Supplementary Methods**, we include a dedicated discussion of mutation-rate variation across gut bacterial species. Although yearly mutation rates can vary by several fold among species, the purpose of the relevant sentences in the text is to provide an approximate timescale for the age of a typical gut bacterial species, which serves as an upper bound on how long the Tsimane and Hadza MAGs have shared a common ancestor. In the revised text, we have softened the language to clarify that the 100-200ky value is not intended as a precise estimate.

5. Line 119: ANI values – this differs between species, as shown in the Supplementary Table. Please clarify whether you report averages across all species.

We thank the reviewer for this helpful request for clarification. The reported ANI value of ~98% reflects the median of the pooled distribution across all species shown in **Fig. 1c**. Importantly, this value is also representative at the species level: most species show an average Hadza-Tsimane ANI close to 98%. To illustrate this pattern, we have added a new **Extended Data Fig.**

6 (also included below), which summarizes several diversity statistics at the species level. In particular, the right panel of **Fig R3.2a** shows that the distribution of species-specific average Hadza-Tsimane ANI values clusters around a median of ~98%. We hope this addition helps address the reviewer's comment.

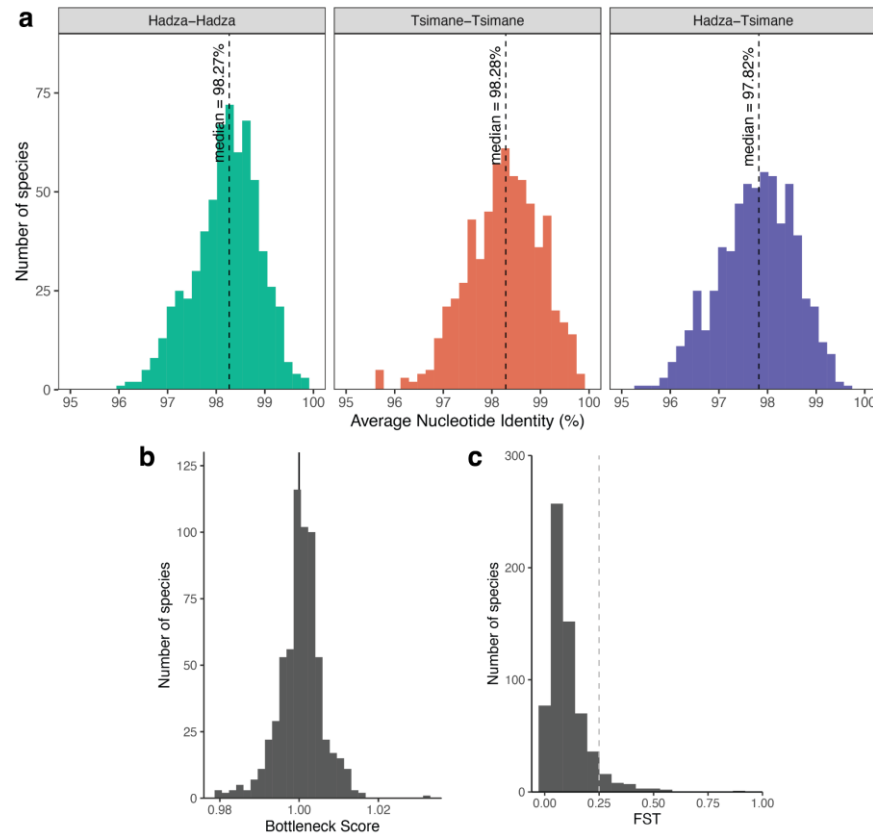


Figure R3.2.

(A) Distribution of average within-species average nucleotide identity for each of the three population contrasts.

(B) Distribution of bottleneck scores for each of the 636 species. The bottleneck score was calculated for each species as: [Avg. T-T ANI / Avg. H-H ANI]. The median of the distribution is 1.00, represented by the vertical black line.

(C) Distribution of fixation indices (Fst) for each species. The vertical dashed line represents the approximate fixation index between host genomes for the Hadza and Tsimane (Fst = Verdu et al., 2014).

6. Line 121: “Most Tsimane species” – what does “most” refer to? Providing a percentage or distribution would be clearer.

We thank the reviewer for the suggestion. We have now provided the corresponding distribution in **Extended Data Fig. 6**, as described in our response to the previous comment and changed the wording to “most species present in the Tsimane samples”.

7. Figure 1b: The Hadza had lower sample numbers than the Tsimane. Would it be possible to assess whether Hadza is comparable to Tsimane with respect to unique species after sample size correction?

The reviewer raises an excellent point that the number of unique subjects is different between the Hadza (137 subjects) and Tsimane (85 subjects) cohorts. To determine the extent to which this different subject number contributes to the observed differences in unique species detected in each population, we generated a rarefaction curve (**Figure R3.3**). To generate this curve we randomly selected subjects at different levels of sampling (increments of 5 subjects up to 85). At each sampling we determined the number of unique species found in that collection of subjects. We then fit a species accumulation curve [according to $S = (S_{\text{max}} \times N) / (k + N)$, where S is the number of species, N is the number of subjects and S_{max} is the maximum number of species] for each population's rarefaction curve. Using these curves we can estimate the asymptote, or the maximal number of species present in the population. For the Hadza, the estimated number of species is 2,443 (95% C.I.: 2,385-2,503 species); for the Tsimane the estimated number of species is 1,590 (95% CI: 1,561-1,620). These results indicate that differences in unique species detected in these two cohorts reflect actual differences in underlying gamma diversity, rather than an artifact of sampling effort.

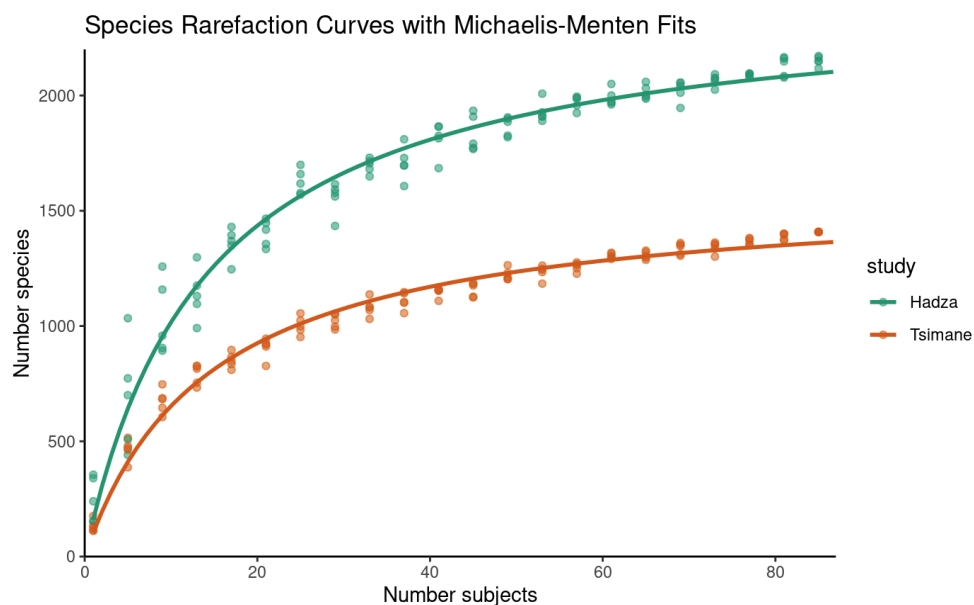


Figure R3.3. Rarefaction curve showing the number of unique species detected in each population given some number of randomly selected subjects.

8. Humans lost substantial diversity during the “Out of Africa” bottleneck. Could a similar reduction also be observed in microbiomes?

We thank the reviewer for raising this fascinating question. Consistent with an Out-of-Africa (OOA) bottleneck, we did observe a smaller number of VANISH species in the Tsimane

compared to the Hadza (**Fig 1b**). However, it is difficult to rule out alternative explanations such as lifestyle differences.

To further probe this question at the level of within-species diversity, we compared the average within-population ANI for each species between the Tsimane and the Hadza. Specifically, we defined a bottleneck score for each species as the ratio between the mean Tsimane ANI and the mean Hadza ANI. Interestingly, the distribution of this score across species centers around one, indicating no systematic bias toward bottlenecks (or expansions). We have included this new analysis in **Extended Data Fig. 6**, alongside other species-level diversity statistics. Consistent with this result, we also found no bias in the distribution of post-split population sizes inferred in the Moments analysis (**Extended Data Table 6**).

Overall, while these simple metrics do not reveal a strong signal, we agree that this is an intriguing direction that warrants further investigation and methodological development.

9. Line 156: Why was 5,000 years chosen as the cutoff for recent strain sharing? Please justify. Given that the Out of Africa migration occurred ~70 kya, why not use an older date?

We thank the reviewer for raising this point. In short, the 5,000-year cutoff reflects the approximate upper limit of timescales over which *clonal* strain sharing can be detected, given the rate at which bacterial recombination erases shared ancestry. Beyond this timescale, recombination is expected to fully overwrite the clonal backbone, making it impossible to distinguish recently diverged strains from distantly related strains.

More specifically, the timescale of clonal relationships depends on both (i) the threshold used to define clonal identity (here, 10% identical genes) and (ii) the rate of recombination, which determines how quickly identical genomic segments are overwritten over time. We chose a relatively low threshold to maximize the temporal window accessible by this approach. We note that this threshold cannot be lowered any further in order to avoid confounding effects from horizontal gene transfer (HGT), which can produce isolated identical genes between otherwise unrelated strains. Based on our empirical estimates of recombination rates across these species, this 10% threshold corresponds to a divergence time of roughly 1,000-5,000 years, depending on the species (new **Extended Data Fig. 7**). We agree that microbial isolation could extend much further back in time than the ~5,000-year window accessible to this approach. For precisely this reason, we employed complementary analyses in **Figs. 3 and 4** to probe deeper timescales of population isolation. In the revised manuscript, we have added some additional text at the beginning of **Supplementary Methods 4** to better emphasize this fact.

As Reviewer 2 also noted, the choice of threshold might appear arbitrary, so we performed a sensitivity analysis across a wide range of cutoffs. This analysis (now shown in **Fig. R2.2**) demonstrates that the main result—that the vast majority of species show no recent T-H strain sharing—holds robustly across thresholds.

In the revised manuscript, we have added this sensitivity analysis and expanded the **Supplementary Methods** to clarify the rationale for the chosen threshold (Section 3). We hope these additions address the reviewer's comment.

10. Line 175: *Agathobacter rectalis* is not shown in the figures, while other species discussed in this section are. Is there a reason for this?

We thank the reviewer for catching this oversight. Due to space limitations, the *A. rectalis* plot is shown in **Extended Data Fig. 8**. We realized that the figure reference was missing in the main text and have now added it when discussing *A. rectalis*.

11. Line 213: Please refer here to the Supplementary Text where the simulations are described in more detail.

Thank you for this suggestion. We have updated the text accordingly.

12. Line 233: Please specify the criteria for selecting the 15 species used for estimating isolation times. Without clear justification, this could be perceived as cherry picking.

We thank the reviewer for raising this point. The criteria for selecting the 15 species used in the Moments analysis are described in detail in Section 5.2 of the **Supplementary Methods**. These criteria involve several steps that are difficult to fully specify in the main text, so we have added a note in that line directing readers to the Supplementary Methods for the full description.

In addition, following Reviewer 1's comment, we have expanded the supplementary table of Moments model parameters to include all 71 species (after sample size filtering but prior to quality filtering) to ensure full transparency (**Extended Data Table 6**).

13. Line 273: "The loss of VANISH taxa with changes in lifestyle, which are linked to an increased risk for a number of chronic metabolic and inflammatory disorders" – it would be helpful to indicate the fraction or extent to which each of these three classes is involved in such disorders.

We apologize for the linguistic ambiguity here. We intended this sentence to mean that changes in lifestyle are associated with increased risks for a number of chronic diseases and that changes in lifestyle are concomitant with loss of VANISH taxa. We have now refined this wording in the text.

14. Sample collection: The Hadza collection conditions should also be described for consistency with the Tsimane. Even if this is from a previous publication, please clarify whether the same procedures were used, as differences could influence results. It would also be good to restate the ethical clearance, since this article is authored by the same group.

We thank the reviewer for raising these important points. Given that the Hadza and Tsimane sampling campaigns were performed independently of one another, there was not an explicit attempt to perform similar sampling and preservation procedures. That said, in both campaigns stool samples were directly flash frozen in liquid nitrogen as quickly as possible after collection, no special preservation kits were used. We have added the following text to the Methods in the Sample Collection section to describe the collection of stool samples from the Hadza cohort:

For the Hadza cohort, samples were collected as described previously (Smits et al., 2017). Briefly, at the time of sample collection when available we recorded the age (either provided by the individual or estimated based on information provided in their recall), gender, weight and related information. After samples were retrieved from the individuals they were immediately placed into liquid nitrogen containers in the field, transported to Arusha frozen and mailed on dry ice by air to the United States and maintained at -80°C until further processing.

Second, we have included an expanded ethics and inclusion statement below, which we also now include in the manuscript:

Ethics & Inclusion Statement

Informed consent among Tsimane was collected at three levels: the Gran Consejo Tsimane (Tsimane governing body), community leaders, and study participants. All study protocols were approved by the Institutional Review Boards of the University of California Santa Barbara (#3-21-0652) and the University of New Mexico (#07-157). Within Bolivia, study protocols were approved by the Universidad Mayor de San Simón in Cochabamba Bolivia, and the Gran Consejo Tsimane. Specimen export was approved by local, regional, and national authorities, with complete export documentation. The binational (Bolivia-US) THLHP team has worked in Bolivia for more than two decades and trained dozens of Tsimane research assistants employed by the team. All of our research is conducted in collaboration with local Tsimane communities and the Tsimane governing body. The original impetus for collecting fecal samples was to analyze microbial composition. However, as a benefit to participants, samples were also analyzed by field microscopy for applicable parasitic species, which could be treated with antihelminthics or anti-protozoan medications. The aim of the study was explained verbally to participants. The THLHP has also engaged in capacity building in terms of education (along with the One Pencil Project, a partnered 501(c)(3) NGO that provides educational supplies for primary schools, and scholarships to Tsimane who go on to university or medical school), as well as investing in health infrastructure at the municipal hospital. The THLHP has employed more than a dozen Bolivian doctors over the years providing direct medical care, has trained local health officials, built an official medical and dental clinic, and donated expertise, personnel, and equipment to two local hospitals.

*In this study, we also analyze previously published metagenomic data from stool samples that were collected from Hadza hunter-gatherers in 2013/2014 (*REF). The samples were legally and ethically collected in consultation with Tanzanian field-guides and after receiving informed consent from Hadza participants who were provided a translated verbal description of the research project. 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 and the Tanzania Commission for Science and Technology (MR/53i 100/83, NIMR/HQ/R.8a/Vol.IX/1542). The results of a recent study(Carter et al., 2023) were converted to an infographic using the local language and conveyed back to the Hadza via an anthropologist and Tanzanian field-guide.

15. It would strengthen the study to include environmental sampling, as the authors did not sequence soil, water, or animal microbiomes from Tanzania or Bolivia to test for potential environmental reservoirs and account for likely ecologically acquired taxa.

Unfortunately this was not in the scope of this study for the Tsimane and will be an interesting topic for future research.

References

- Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 39(1), 105–114.
- Carter, M. M., Olm, M. R., Merrill, B. D., Dahan, D., Tripathi, S., Spencer, S. P., Yu, F. B., Jain, S., Neff, N., Jha, A. R., Sonnenburg, E. D., & Sonnenburg, J. L. (2023). Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell*, 186(14), 3111–3124.e13.
- Good, B. H. (2022). Limited codiversification of the gut microbiota with humans. In *bioRxiv* (p. 2022.10.27.514143). <https://doi.org/10.1101/2022.10.27.514143>
- Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D., Knight, R., & Gordon, J. I. (2008). Evolution of mammals and their gut microbes. *Science (New York, N.Y.)*, 320(5883), 1647–1651.
- Moeller, A. H., Caro-Quintero, A., Mjungu, D., Georgiev, A. V., Lonsdorf, E. V., Muller, M. N., Pusey, A. E., Peeters, M., Hahn, B. H., & Ochman, H. (2016). Cospeciation of gut microbiota with hominids. *Science*, 353(6297), 380–382.
- Moeller, A. H., Sanders, J. G., Sprockett, D. D., & Landers, A. (2023). Assessing co-diversification in host-associated microbiomes. *Journal of Evolutionary Biology*, 36(12), 1659–1668.
- Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J.-B. N., Peeters, M., Hahn, B. H., & Hugenholtz, P. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biology*, 8(11), e1000546.
- Perez-Lamarque, B., & Morlon, H. (2024). Distinguishing cophylogenetic signal from phylogenetic congruence clarifies the interplay between evolutionary history and species interactions. *Systematic Biology*, 73(3), 613–622.

- Smits, S. A., Leach, J., Sonnenburg, E. D., Gonzalez, C. G., Lichtman, J. S., Reid, G., Knight, R., Manjurano, A., Chagalucha, J., Elias, J. E., Dominguez-Bello, M. G., & Sonnenburg, J. L. (2017). Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science*, 357(6353), 802–806.
- Sprockett, D. D., Dillard, B. A., Landers, A. A., Sanders, J. G., & Moeller, A. H. (2025). Recent genetic drift in the co-diversified gut bacterial symbionts of laboratory mice. *Nature Communications*, 16(1), 2218.
- Suzuki, T. A., Fitzstevens, L., Schmidt, V. T., Enav, H., Huus, K., Mbong, M., Adegbite, B. R., Zinsou, J. F., Esen, M., Velavan, T. P., Adegnika, A. A., Song, L. H., Spector, T. D., Muehlbauer, A. L., Marchi, N., Blekhman, R., Ségurel, L., Youngblut, N. D., Kremsner, P., & Ley, R. E. (2021). Codiversification of gut microbiota with humans. In *bioRxiv* (p. 2021.10.12.462973). <https://doi.org/10.1101/2021.10.12.462973>

Referees' comments:

Referee #1 (Remarks to the Author):

I thank the authors for their comprehensive responses to my comments. I support publication of the manuscript.

Referee #1 (Remarks on code availability):

The code appears to be well documented and usable. The major limitation precluding full reproducibility will be the restricted access to the raw data, but this may be necessary for ethical reasons.

Referee #2 (Remarks to the Author):

I thank the authors for their comprehensive and thoughtful revision, which has substantially strengthened the manuscript. The authors have addressed nearly all of my concerns, including the addition of sensitivity analyses for parameter choices, the independent mapping-based validation pipeline, improved statistical reporting with multiple testing correction, and the commitment to releasing processed data via Zenodo. However, two issues remain that I believe should be addressed:

2.2.1 First, after seeing a figure of the sequencing coverage of the data (Figure R2.5), I remain concerned about the potential impact of differential sequencing coverage between the Hadza and Tsimane cohorts on the key findings. As shown in Figure R2.5, while average depth slightly differs between cohorts (slightly higher for the Tsimane), the Hadza cohort also exhibits a long tail of extremely high coverage samples, with some samples above 200 Gbp. This coverage heterogeneity could potentially introduce systematic biases in several places. Most critically, ultra-deep sequencing of a subset of Hadza samples enables assembly of rare, low-abundance organisms that would fail to assemble at lower depth. These rare species, detected only in deeply sequenced Hadza samples, would then appear absent from both the Tsimane and industrialized cohorts, not because they represent ancestral VANISH taxa, but simply because insufficient sequencing depth precludes their detection elsewhere. Importantly, the rarefaction analysis provided by the authors controls for number of subjects but not for sequencing depth per subject, leaving this artifact unaddressed. The estimated asymptote difference between the Hadza and Tsimane may therefore be partially depth-driven rather than reflecting true biological differences in gamma diversity.

We thank the reviewer for this important comment. We agree that the very deeply sequenced Hadza samples warrant additional attention. For reference, below is **Figure R2.5** from the previous round of responses.

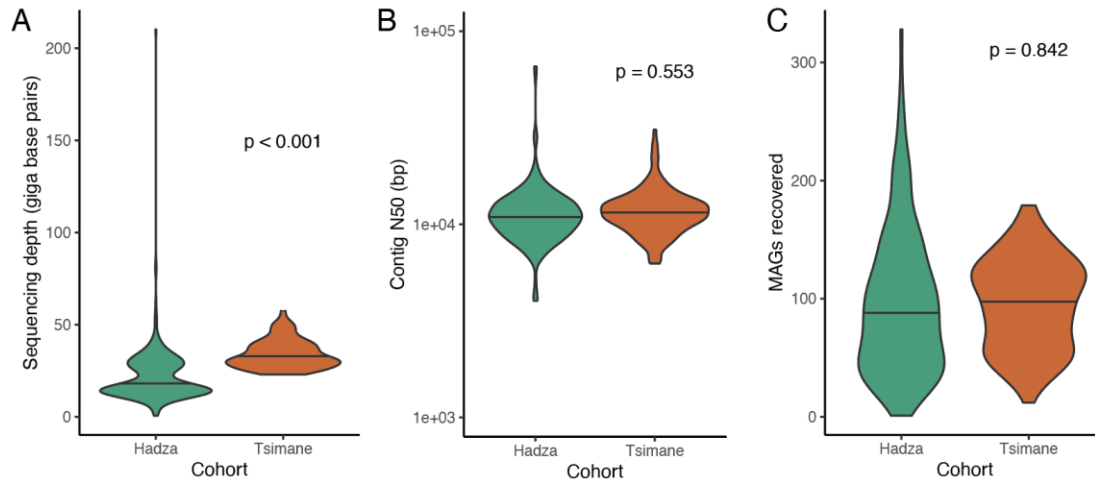


Figure R2.5. Comparison of sequencing metrics between the Hadza and Tsimane cohorts.

- (a) Sequencing depth (measured in giga base pairs) across the Hadza and Tsimane samples used in the study.
- (b) Weighted median (N50) of contig lengths in assemblies from Hadza and Tsimane metagenomes..
- (c) The number of medium-quality and higher MAGs recovered from each Hadza and Tsimane sample.

All P-values were calculated using a t-test comparing the two cohorts.

To address the concern we now have tested whether very deeply sequenced Hadza samples could preferentially contribute rare MAGs and thereby influence species-level comparisons. We subsetting the Hadza samples by removing any Hadza samples that were sequenced more deeply than the deepest-sequenced Tsimane sample (this removed 6 Hadza samples). We determined the number of unique species found in the Hadza and Tsimane, and how many species were shared between both populations, comparing the original (no filter) to this depth-filtered dataset (**Figure R2.1**). Note that these data are essentially the same data we show in **Figure 1b**, except in stacked bar format to make comparisons between subsetting approaches easier.

The result from this depth-filtering analysis is highly similar to that in the full dataset. In the depth-based filtering, we remove 6 Hadza samples (which is 1.8% of samples) and we see a 1.9% reduction in the number of species unique to the Hadza. We conclude from this analysis that when controlling for sequencing depth we do not see a disproportionate drop in Hadza-unique species and that the observed differences in gamma diversity between the Hadza and Tsimane are likely to be genuine.

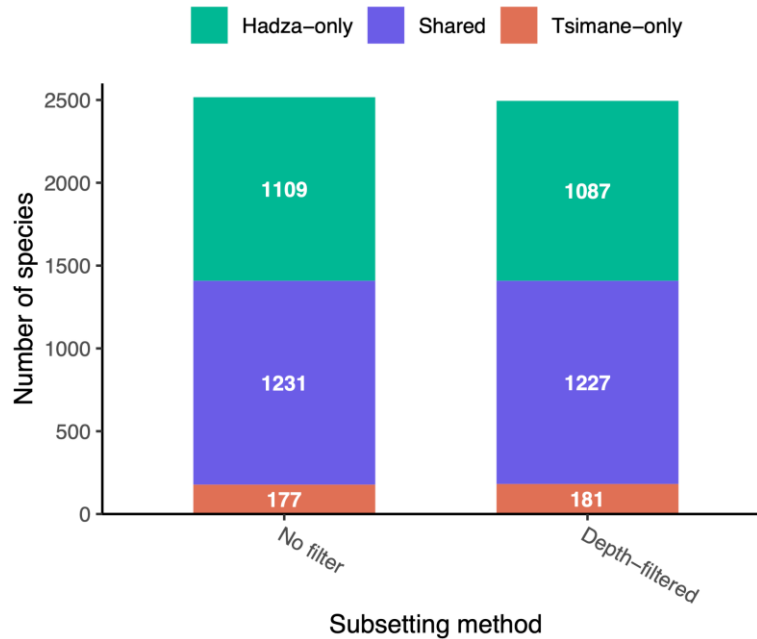


Figure R2.1. Comparisons of MAG recovery when removing high-depth Hadza samples.

Stacked bar charts show the number of unique species recovered from Hadza-only (green), Tsimane-only (orange), or both populations (purple) under three approaches: 1) no subsampling filter applied (“No filter” – identical data found in **Figure 1b**) and 2) removal of Hadza samples sequenced to greater depth than the most deeply sequenced Tsimane sample included in the study, as measured by post-deduplication bases sequenced (“Depth-filtered”).

2.2.2. Beyond the VANISH taxa finding, the uneven coverage distribution could also bias the L99-based genetic isolation analysis. Within-Hadza comparisons that involve two high-coverage samples would yield more complete MAGs with more accurate SNV calls and fewer assembly-induced tract breaks, producing artificially elevated within-population L99 values. Mixed-coverage Tsimane-Hadza between-population comparisons, by contrast, would be more prone to tract fragmentation, producing relatively shorter between-population L99 values. This could artifactually inflate the apparent contrast between within- and between-population isolation that forms the core evidence for co-migration. One approach to address this would be to remove the very highly sequenced outliers and repeat the analysis. In addition, a more rigorous approach would be to perform a depth-rarefied analysis in which all samples are downsampled to a common sequencing depth prior to assembly, and to repeat the key analyses (including the species catalog, VANISH taxa prevalence, and L99 calculations) on this depth-equalized dataset. This would directly test whether the central findings are robust to the coverage heterogeneity currently present in the data.

We thank the reviewer for raising this important concern. We agree that coverage heterogeneity between the Hadza and Tsimane cohorts is a potential confounder worth examining carefully, and we address it with two complementary lines of evidence.

First, we note that an analysis directly relevant to this concern is already presented in the supplementary materials (**Figure S5**). This figure compares within-population L99 values across population pairs on a per-species basis. If sequencing depth were a primary driver of L99 values, we would expect the Hadza to exhibit systematically elevated within-population L99 values relative to the Tsimane. Instead, we observe a strong correlation in within-population L99 lengths across species between the two populations, consistent with the interpretation that this statistic reflects biological properties of each species (e.g., recombination parameters) rather than technical variation in sequencing depth.

Second, to directly address the reviewer's concern, we performed a coverage-restricted analysis in which we retained only MAGs with median coverage below 50x in both the Hadza and Tsimane cohorts, thereby reducing the influence of high-coverage outliers (**Figure R2.1**). Repeating the L99 analysis on this subset yielded results that were quantitatively similar to those obtained from the full dataset. This finding confirms that the highly sequenced samples are not driving the observed patterns of genetic isolation. Together, these analyses indicate that coverage heterogeneity does not substantially bias the L99-based co-migration inference.

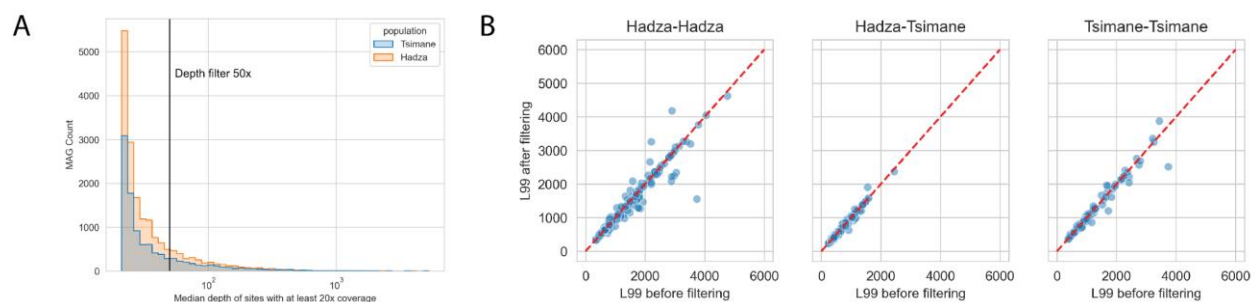


Figure R2.1. Coverage filtering does not substantively affect L99-based genetic isolation estimates.

- (A) Distribution of median sequencing depth across MAGs in the Hadza and Tsimane cohorts, computed over sites with at least 20x coverage. The dashed vertical line indicates the 50x threshold used for coverage filtering.
- (B) Comparison of within- and between-population L99 values before and after excluding MAGs with median coverage exceeding 50x, shown separately for Hadza-Hadza (HH), Hadza-Tsimane (HT), and Tsimane-Tsimane (TT) population pairs. Each point represents a single species. The dashed line indicates the line of equality ($y = x$). The strong concordance across all three comparison types indicates that high-coverage MAGs do not drive the observed patterns of genetic isolation.

2.2.3. Second, I appreciate the authors' commitment to indigenous data sovereignty and their detailed explanation of the CARE principles framework. At the same time, sharing the data is important: the data generated in this study represent an exceptionally rare and scientifically valuable resource that cannot easily be regenerated, making access especially important for the field. Without the ability to independently validate the results, or integrate these data into broader comparative studies, the community cannot fully evaluate or build upon what the paper presents as landmark findings about human-microbiome co-evolution.

This is why I remain concerned that the current data sharing mechanism does not meet the standards expected for a Nature publication. The THLHP data sharing webpage (this is the website readers are referred to for data access) explicitly commits to the FAIR principles (Findable, Accessible, Interoperable, Reusable), yet the access mechanism described on that same page fails to meet several of these in practice; what is described is a multi-stage discretionary review process with no defined timeline, no accession number, and no stated decision criteria, which does not constitute meaningful accessibility or reusability. Moreover, the access process as described on the THLHP website is entirely controlled by the co-Directors of the project, which creates a structural conflict of interest in which the data generators are also the gatekeepers, effectively allowing the authors to maintain a de facto monopoly on follow-up analyses using this dataset. This is qualitatively different from standard controlled-access repositories such as NIH dbGaP, where access decisions are made by independent committees with explicit and publicly stated criteria. I would also note that the level of restriction applied here appears disproportionate relative to the sensitivity of the data. The dataset in question consists of metagenomic sequences from stool samples, and the associated re-identification risks are substantially lower than for human genomic data. I think that it would be reasonable to deposit the raw data (after removing host reads) in a standard controlled-access repository such as dbGaP, where access decisions are made independently of the authors.

We thank the reviewer for this important comment. We agree that these data are a rare and scientifically valuable resource, and we share the goal of enabling independent validation and reuse to the greatest extent possible. We also agree that, for a resource of this importance, the access pathway must be clear, transparent, and workable for the broader community.

To address these concerns, the data are being deposited in the European Genome-Phenome Archive (EGA) under Data Access Committee EGAC50000000982 ("The Tsimane Health and Life History Project"), with a formal Data Access Policy (EGAP50000000936) that is publicly visible on the EGA website. This directly addresses the reviewer's concerns regarding accessioning, explicit access criteria, and transparency of the review process. EGA is a purpose-built controlled-access repository for human genomic data jointly managed by EMBL-EBI and the Centre for Genomic Regulation. We have updated the Data Availability Statement accordingly. This revised process enables scientists to obtain access for independent validation and broader comparative analyses through a defined, transparent request process.

Though we value open science, we respectfully note that deposit in NIH dbGaP is not feasible for these data. dbGaP's model delegates access decisions to independent review committees with no mechanism for community consultation, which would remove the Tsimane community from any meaningful role in governing how their data are used. This requirement follows directly from the ethical and governance framework under which these samples were collected. For Indigenous-partnered research, responsible stewardship extends beyond re-identification risk to include protecting against stigmatizing or non-consensual secondary uses, and the THLHP governance structure, which includes Tsimane and Bolivian representation, must retain oversight of these decisions. Although the THLHP DAC includes study co-authors, the publicly

posted Data Access Policy makes the decision criteria explicit and transparent, and access decisions are made within this broader community governance framework rather than unilaterally by manuscript authors.

To further maximize reproducibility and reuse, we will also release the processed data files underlying our population-genetic analyses via Zenodo, including the SNV catalogs and intermediate processing files for the 636 species analyzed in the manuscript. Together, these revisions provide a substantially more explicit, transparent, and functional access framework while preserving the protections owed to the Tsimane communities.

Referee #3 (Remarks to the Author):

The authors have addressed all my comments satisfactorily. Thank you for the detailed responses and additional analyses.

2.3.1. I have just one small suggestion. The rarefaction analyses are quite nice and, as far as I can see, they are only included in the rebuttal letter. It would be useful to add them to the supplementary material if possible.

Thanks for this suggestion, we have added **Figure R2.3** from our previous rebuttal letter to **Extended Data Figure 2**.

We appreciate the reviewer's continued thoughtfulness in improving the rigor of the manuscript.

Reviewer #2

Regarding differences in depth between populations, the cross-species correlation and the coverage-restricted reanalysis adequately address some of my concerns. However, regarding my first point, the new analysis does not fully address per-sample depth heterogeneity within cohorts, which could be the actual driver of differential rare-taxon recovery. The rigorous way to address this would be a depth-rarefied reassembly, or a statistical approach that controls for depth as a covariate. However, short of that, I would suggest adding an explicit acknowledgement in the discussion of this limitation in the manuscript text so readers can evaluate the finding accordingly.

From our previous revision, after removing the deeply sequenced Hadza outliers the average depth per sample is slightly higher in the Tsimane samples (**Fig. R2.5a**), yet the Hadza gamma diversity (i.e., species recovered) remains higher. Therefore, while the reviewer's concern is reasonable in principle, the observed result tilts in the opposite direction of how we would expect such sequencing depth differences to manifest in the data. In addition, the depth-rarefied or statistical approaches would be time consuming and unlikely to be informative, since equalizing depth does not resolve potential population differences in species abundance or strain heterogeneity. We do acknowledge that, as is often the case with cross-cohort metagenomic studies, there are limitations posed by data structure interacting with biological heterogeneity to impact recovery of rare MAGs. Because this point is most relevant to interpretation of **Fig. 1b**, we believe the clearest approach is to note it directly after describing that figure panel, so the caveat co-occurs with the relevant result.

We have included the following limitation/caveat sentence to the Results: "Similar results were observed when controlling for sequencing depth outliers between populations (**Extended Data Fig. 2d**) or using a mapping-based analysis to rule out common assembly biases (**Extended Data Fig. 3**); however, it remains possible that differences in strain-level heterogeneity and/or sequencing depth distributions between cohorts contributed to the differential recovery of some rare MAGs."