

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	DNA was extracted using MoBio PowerSoil kits (Qiagen). Library preparation used Nextera Flex kits. Paired-end sequencing (2x140bp) was performed on a NovaSeq 6000 using 54 flow cells at the UCSD Microbiome & Metabolomics Core.
Data analysis	Custom metagenomics pipeline (https://github.com/MrOlm/nf-genomeresolvedmetagenomics). numpy v1.19.2, scipy v1.10.0, fastp v1.0 (adapter trimming), Bowtie2 v2.14.1 (read mapping), FastQC v0.11.5 (quality control), SPAdes v3.14.1 (metagenome assembly), CoverM v0.4.0 (contig coverage), QUAST v5.0.2 (assembly QC), MetaBAT2 v2.15 (genome binning), CheckM 1.1.3 (genome quality), GTDB-Tk r220 (taxonomic classification), dRep v3.5.1 (ANI calculations, goANI algorithm), Prodigal v2. 6.3 (gene calling), NSimScan v1.1.84 (nucleotide alignment), MMseqs2 v15.6 (gene clustering), inStrain vl.2.14 (coverage profiling), Moments software package (no version number, https://github.com/MomentsLD/moments , demographic inference), rnaturalearthdata v1.0.0 (map plotting), custom SNVanalysis pipeline (https://github.com/zjshi/snv_analysis_almeida2019). All custom analysis code and figure generation scripts available at https://github.com/Sonnenburglab/comigration-metagenomics (DOI: 10.5281/zenodo.21206858)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Individual-level data are stored in the European Genome-Phenome Archive (EGA) and governed by the Tsimane Health and Life History Project (THLHP), 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 approval through 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. 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.21765848). All Hadza data is publicly available as specified in the manuscript (DOI: 10.5281/zenodo.8072245)3.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Sex/gender of participants was recorded when available during sample collection. The study analyzes microbiome composition and microbial population genetics; sex was not a variable in the primary analysis of microbial co-migration patterns. Both male and female participants were included.

Reporting on race, ethnicity, or other socially relevant groupings

Reporting on race, ethnicity, or other socially relevant groupings: The study compares gut microbiomes from two specific non-industrialized populations (Tsimane of Bolivia and Hadza of Tanzania) with publicly available data from industrialized populations (Europe, Asia, North America). These groupings are based on geographic ancestry, subsistence strategy, and degree of industrialization, which are directly relevant to the research question of microbial co-migration with human populations. The terms "industrialized" and "non-industrialized" are defined in the manuscript.

Population characteristics

Tsimane: 85 adults sampled across 9 villages along the Maniqui River and nearby forest region in the Bolivian Amazon. Mean age = 28.4 years, s.d. = 13.2 years. All participants at least 3 years of age. The Tsimane are indigenous forager-horticulturalists. Hadza: 137 hunter-gatherers from Tanzania (previously published cohort). Age (at least 3 years old), gender, and weight recorded when available. We do not use these features as co-variables in our analysis. Industrialized comparators: publicly available MAGs from European, Asian, and North American cohorts.

Recruitment

Tsimane participants were recruited through the Tsimane Health and Life History Project (THLHP), which has operated in Bolivia for over two decades. Informed consent was collected at three levels: the Gran Consejo Tsimane (governing body), community leaders, and individual study participants. The aim of the study was explained verbally. Samples were collected as part of a previously published study. Hadza participants provided informed consent through a translated verbal description of the research project, facilitated by Tanzanian field-guides. No self-selection bias is anticipated as recruitment was community-based; however, the study is limited to the villages and communities reached by the THLHP and Hadza field teams.

Ethics oversight

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; #15-133; #17-230). Within Bolivia, study protocols were approved by the Universidad Mayor de San Simón in Cochabamba Bolivia, and the Gran Consejo Tsimane. For specimen export, approval was provided by the main Tsimane governing body (Gran Consejo Tsimane), the Universidad Mayor de San Simón Medical School (Cochabamba, Bolivia), and the National Center for Tropical Diseases (CENETROP) in Santa Cruz, Bolivia. The binational (Bolivia-US) Tsimane Health and Life History Project (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 during more than 65,000 patient visits, 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 2013/20143. 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 study³ were converted to an infographic using the local language and conveyed back to the Hadza via an anthropologist and Tanzanian field-guide.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	133 fecal samples from 85 Tsimane adults (median sequencing depth = 31.9 Gbp, minimum target 100 million paired-end reads per sample). 137 Hadza individuals (previously published). 8,855 additional MAGs from public European, Asian, and North American cohorts. Sample sizes were chosen to achieve sufficient sequencing depth for metagenome-assembled genome recovery and strain-level population genetic analyses across hundreds of shared microbial species. No formal power analysis was performed; sample sizes were determined by the available cohorts and sequencing resources.
Data exclusions	Metagenome assemblies were filtered to contigs ≥ 1500 bp. Genome bins with $<50\%$ completeness or $>10\%$ contamination (per CheckM) were excluded, following MIMAG standards. For population genetic analyses, only species with at least 4 MAGs each in the Hadza and Tsimane cohorts were included (636 of 1,231 shared species). Gene alignments shorter than 500 bp were excluded from the identical gene analysis. These criteria were pre-established based on standard quality thresholds in the field.
Replication	Tsimane samples were collected across two independent time points (2009 and 2012–2013). Population genetic analyses were performed using two independent and orthogonal approaches: (1) identical DNA tract length analysis (L99 lengths, Fig. 3) and (2) demographic inference from SNV frequency spectra using Moments (Fig. 4). Both yielded consistent split time estimates. Strain sharing and gene flow analyses were performed across all 636 qualifying species, providing hundreds of independent replicates. Within-population comparisons served as internal controls.
Randomization	This is an observational/comparative genomics study, not an experimental intervention. Samples were randomized across sequencing runs to minimize batch effects. Population labels were permuted in permutation tests to generate null distributions ($n = 10,000$ permutations per test).
Blinding	Blinding was not relevant to this study as it is a computational analysis of metagenomic data. All analyses were performed using pre-established, automated computational pipelines applied uniformly across all samples and populations.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.