

Prehistoric global migration of vanishing gut microbes with humans

<https://doi.org/10.1038/s41586-026-11106-1>

Received: 22 May 2025

Accepted: 1 September 2026

Published online: 07 October 2026

Open access

 Check for updates

Matthew M. Carter^{1,18}, Zhiru Liu^{2,18}, Matthew R. Olm^{1,3,18}, Melanie Martin^{4,5}, Daniel D. Sprockett⁶, Parsa Ghadermazi³, Benjamin C. Trumble^{7,8,9}, Hillard Kaplan¹⁰, Jonathan Stieglitz¹¹, Daniel Eid Rodriguez¹², David A. Relman^{13,14}, Erica D. Sonnenburg¹, Michael Gurven⁴, Benjamin H. Good^{2,15,16}✉ & Justin L. Sonnenburg^{1,16,17}✉

The gut microbiome is crucial for health and is affected strongly by lifestyle¹. Many microorganisms commonly found in non-industrialized populations are disappearing or have become extinct in industrialized populations^{2–6}. Studying which microorganisms have been long-term residents of the human gut and may have co-evolved with humans^{2,7,8} could provide insights into how microbial biodiversity loss affects human health. However, the genetic complexities of microbial evolution and the plasticity of gut microbiome composition have made it challenging to resolve the evolutionary history of these long-term associations. Here we performed deep metagenomic sequencing of the Tsimane horticulturalists of Bolivia and compared their gut microbiomes with those of the Hadza hunter-gatherers of Tanzania³. These two populations, whose ancestors have been separated for tens of thousands of years, share 1,231 microbial species, most of which are rare in or absent from industrialized populations. Population genetic analyses of 636 of the shared species revealed patterns of microbial divergence and gene flow consistent with prehistoric human co-migration, with estimated split times that approximately align with human migration out of Africa and into the Americas. Our findings indicate that a diverse gut microbiome co-migrated with humans worldwide and has persisted over millennia. However, many of these species are now vanishing from industrialized populations and the consequences for human health remain uncertain.

Commensal gut bacteria play a complex part in human biology¹, but the evolutionary history of these microbial relationships remains poorly characterized⁹. Identifying long-term associations between humans and microorganisms is important for understanding their co-evolution^{7,10,11}, and this could reveal microbial species and functions integral to human health and physiology.

Direct observations of historical microbial relationships have been obtained by sequencing palaeofaeces¹². This study suggested that ancient gut microbiomes were more similar to those of contemporary non-industrialized populations than to those of industrialized ones (the term industrialized here refers to post-industrial populations as well as those in advanced and ongoing stages of industrialization). However, the comparatively recent ages of existing palaeofaeces (around 1,000–2,000 years old), combined with the challenges of analysing degraded ancient DNA, makes it difficult to determine whether these microbial lineages have been faithfully transmitted over long

evolutionary timescales (of around 10,000–100,000 years) or were acquired or transmitted more recently.

Comparative genomics methods provide an alternative approach to inferring evolutionary history from the patterns of genetic diversity among contemporary microbial genomes^{8,13–19}. Building on previous work in non-human primates^{2,8,19–21}, advances in strain-resolved metagenomics have made it possible to use this approach on gut microbial species sampled from human populations worldwide. Previous work has identified phylogenetic correlations between gut microbial genomes and their human populations of origin⁷, potentially consistent with long-term associations^{18,22,23}. However, interpreting these correlations remains challenging because they require genome comparisons at the subspecies, or strain, level. Conventional phylogenetic methods²⁴ often fail in such settings because intra-species gene transfer and homologous recombination between strains produce mosaic patterns of ancestry that cannot be represented by a single phylogenetic

¹Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. ²Department of Applied Physics, Stanford University, Stanford, CA, USA.

³Department of Integrative Physiology, University of Colorado Boulder, Boulder, CO, USA. ⁴Department of Anthropology, University of California Santa Barbara, Santa Barbara, CA, USA.

⁵Department of Anthropology, University of Washington, Seattle, WA, USA. ⁶Department of Microbiology and Immunology, Wake Forest University School of Medicine, Winston-Salem, NC, USA.

⁷School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA. ⁸Center for Evolution and Medicine, Arizona State University, Tempe, AZ, USA. ⁹Institute of Human Origins, Arizona State University, Tempe, AZ, USA.

¹⁰Economic Science Institute, Chapman University, Orange, CA, USA. ¹¹Toulouse School of Economics, Toulouse, France. ¹²Universidad Mayor de San Simón, Cochabamba, Bolivia. ¹³Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA.

¹⁴Infectious Diseases Section, Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA. ¹⁵Department of Biology, Stanford University, Stanford, CA, USA. ¹⁶Chan Zuckerberg Biohub, San Francisco, CA, USA. ¹⁷Center for Human Microbiome Studies, Stanford University School of Medicine, Stanford, CA, USA.

¹⁸These authors contributed equally: Matthew M. Carter, Zhiru Liu, Matthew R. Olm. ✉e-mail: bhgood@stanford.edu;

jsonnenburg@stanford.edu

tree^{25–28}. These high rates of recombination make it difficult to distinguish ancient co-migration from other sources of population structure, such as geography or host lifestyle variation^{22,23}. Moreover, existing strain-level analyses have been biased towards gut microbial species in industrialized populations²⁹. These sampling biases have hindered our ability to infer the evolutionary history of the microbial taxa that are the most likely candidates for ancestral co-migration.

To address these challenges, we conducted deep metagenomic sequencing on faecal samples from the Tsimane horticulturalists of Bolivia (previously characterized using 16S ribosomal RNA sequencing⁶) and compared their microbiomes with those of the Hadza hunter-gatherers of Tanzania³. These two cohorts represent unique and geographically distant contemporary populations whose ancestors separated thousands of years ago, providing a rare opportunity for examining human–microorganism co-migration patterns across a wide range of candidate ancestral microbial taxa. We identified more than 1,200 microbial species shared between the two populations, most of which (60%) were rare in or absent from industrialized populations. Using population genetic analyses that explicitly account for microbial recombination, we provide evidence of strain-level reproductive isolation that is stronger than that observed in microorganisms from comparable industrialized populations. These findings suggest the transgenerational maintenance of gut microbial communities throughout ancient human migrations, providing insights into the roles of horizontal gene transfer and host–microorganism co-evolution in the long-term structuring of the human microbiome.

Sequencing and genome recovery

The Tsimane people are Indigenous forager-horticulturalists living in the Bolivian Amazon. The individuals in this cohort live in six villages along the Maniqui River and three villages in a nearby forest region⁶. We performed deep metagenomic sequencing on 133 faecal samples collected from 85 Tsimane adults in 2009 and 2012–2013 (median depth = 31.9 gigabases; Extended Data Fig. 1a). We recovered 12,746 metagenome-assembled genomes (MAGs) representing 1,408 bacterial and archaeal species (Fig. 1a, Methods and Supplementary Table 1). We then compared these genomes with an analogous collection of 32,034 MAGs previously obtained from deeply sequenced microbiome samples of 137 Hadza hunter-gatherers in Tanzania³ (Extended Data Fig. 1b) to quantify how the gut microbial species of these populations have diverged over the past 100,000 years.

The Tsimane and Hadza share many species

Despite their large geographical separation and lack of historical direct contact, 87.4% (1,231) of the species in the Tsimane cohort were also found in Hadza individuals (Fig. 1b). Some of these shared species (for example, *Ruminococcus bromii*) are also prevalent worldwide, but most (60.2%; 848 species) are rare in or absent from the faecal microbiomes of industrialized populations (Extended Data Figs. 2 and 3 and Supplementary Table 2). Similar results were obtained when controlling for sequencing depth outliers between populations (Extended Data Fig. 2d) or when using a mapping-based analysis to rule out common assembly biases (Extended Data Fig. 3). However, differences in strain-level heterogeneity and/or sequencing depth distributions between cohorts may have contributed to the differential recovery of some rare MAGs. Species sharing between distant populations could result from ancient co-migration or more-recent acquisition from the environment or human contact. We sought to distinguish between these scenarios by examining the genetic variation in each of these shared species and comparing these patterns with expectations from historical human migrations.

As a first step, we calculated the average nucleotide identity (ANI) between all pairs of MAGs in each of the 636 shared species with at least

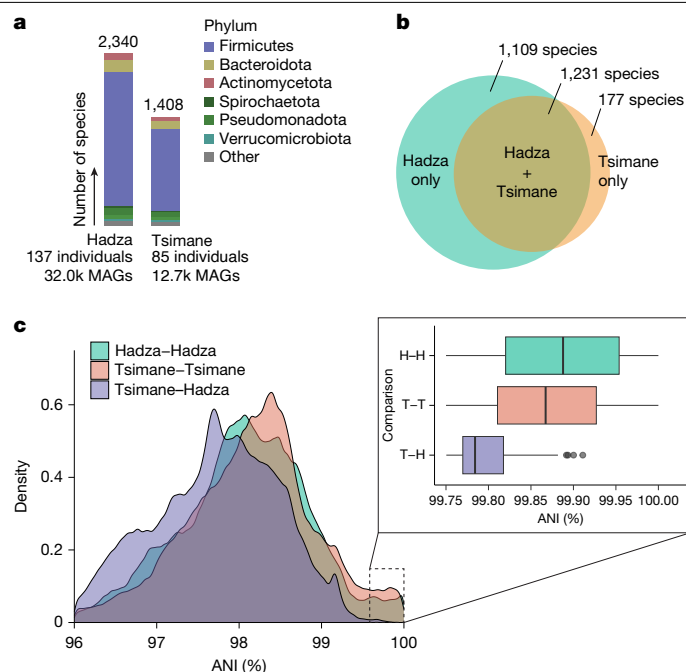


Fig. 1 | A large fraction of gut microbial species found in the Tsimane are also found in the Hadza. a, The taxonomic diversity of microbial species recovered from Tsimane and Hadza microbiomes, coloured according to bacterial phylum. **b**, Microbial species shared between populations. **c**, The ANI for pairs of conspecific genomes from the same versus different populations, aggregated across the 636 shared species with at least four MAGs in each population ($n = 1,012,124$ intra-species pairwise genome ANI comparisons). Inset, the corresponding distributions for ANI values greater than 99.75% ($n = 10,927$ intra-species pairwise genome comparisons). H–H, Hadza–Hadza; T–H, Tsimane–Hadza; T–T, Tsimane–Tsimane. The centre line represents the median; the box bounds represent the 25th and 75th percentiles (interquartile range, IQR); whiskers extend to the most extreme values within $1.5 \times$ IQR; points are individually plotted outliers. The results for each species are shown in Extended Data Fig. 6.

four MAGs each in both the Hadza and Tsimane cohorts (Extended Data Fig. 4). These ANI values provide an upper bound on the time to the most-recent common ancestor for each MAG pair. Estimates of the molecular clock in gut bacteria suggest that mutations accumulate at a rate of at least 10^{-7} per site per year (Extended Data Fig. 5 and Supplementary Methods 1). The typical ANI between Tsimane and Hadza MAGs (around 98%; Fig. 1c) therefore suggests that they shared a common genetic ancestor in the past 100,000–200,000 years (Extended Data Fig. 5). The fact that most species present in Tsimane samples share a common ancestor with those in the Hadza in this timeframe (Extended Data Fig. 6) places strong constraints on their previous evolutionary history and suggests that the ancestors of the Tsimane did not acquire a completely new suite of gut microorganisms during or after the population's migration to the Americas.

Further insights can be obtained by comparing the ANI values between Tsimane–Hadza genome pairs and their corresponding within-population counterparts (Tsimane–Tsimane, Hadza–Hadza). As expected, these data revealed that pairs of conspecific genomes from the same human population were, on average, more similar to each other than to genomes in different populations ($P < 0.001$, analysis of variance, Fig. 1c). However, the effect sizes were smaller than the variation within each population (mean fixation index (F_{ST}) = 0.10; Extended Data Fig. 6). The broad overlap between these distributions indicates that most of the strain-level diversity in the Tsimane microbial populations pre-dates the onset of geographical isolation.

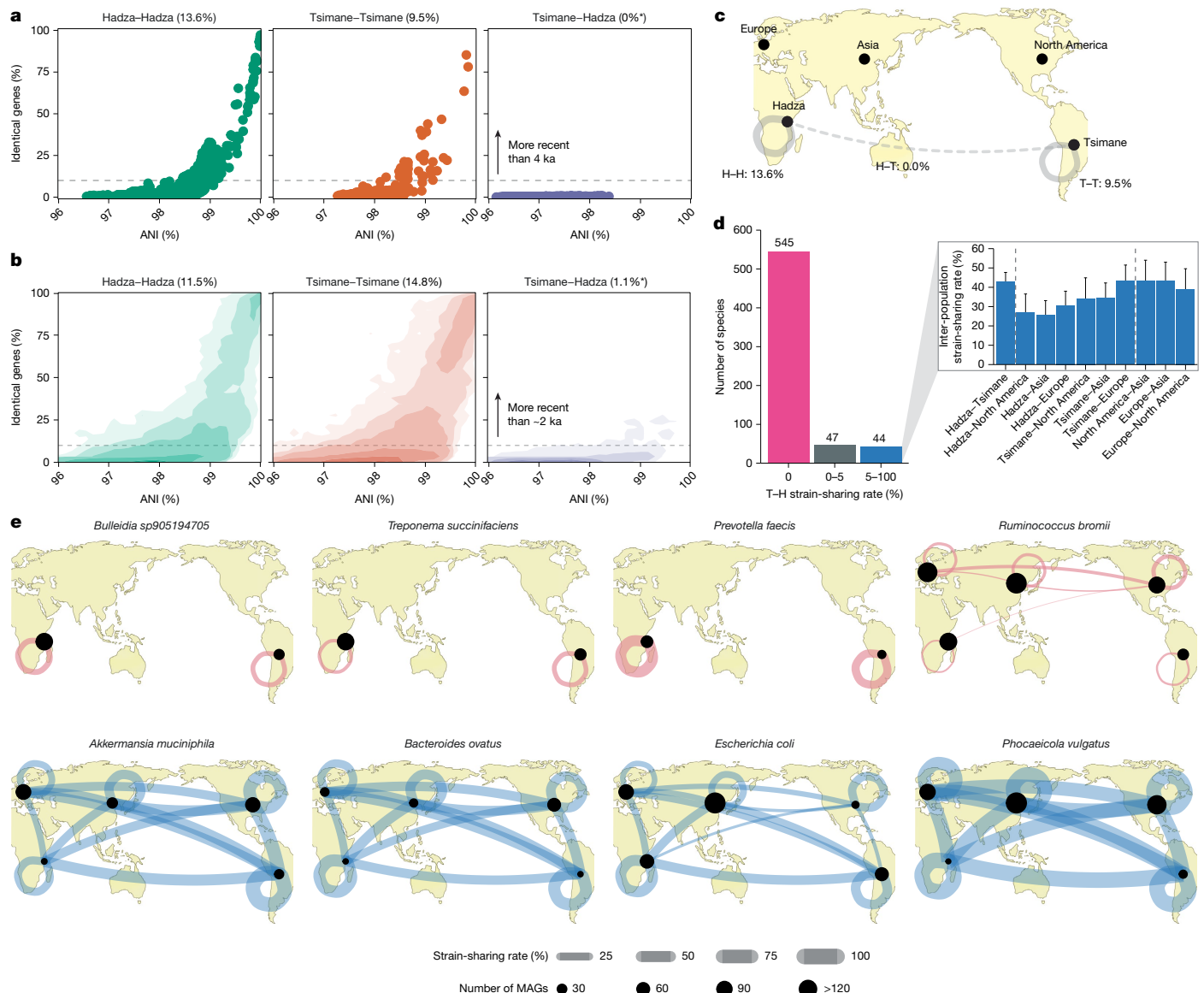


Fig. 2 | Most species shared between the Tsimane and Hadza are genetically isolated, except for some globally dispersed species. a, b, Relationship between ANI and identical gene sharing between MAG pairs for the species *Bulleidia sp905194705* (**a**) and for the 636 abundant bacterial species shared between the Tsimane and Hadza (**b**). The grey dashed lines represent the 10% identical gene threshold used to define recently separated lineages; the threshold varies across species, from approximately 4,000 years for *Bulleidia sp905194705* to a median of around 2,000 years (95% interpercentile range: 500–5,000 years) for the species group in **b** (Extended Data Fig. 8). The percentage of MAG pairs that fall above the grey dashed lines is shown. The asterisks indicate that the percentage of MAGs that share more than 10% identical genes is significantly higher than expected by chance for both *Bulleidia sp905194705* (adjusted $P < 0.001$, one-sided permutation test with Benjamini–Hochberg correction) and all species ($P < 10^{-4}$, one-sided permutation test). **c**, Location of the five cohorts used for a global analysis of

inter-population strain sharing. Data from Asia³², Europe³³, North America³⁴ and the Hadza people³ are from previously published studies. Grey circles and grey dashed lines represent the recent lineage-sharing rates within populations (H–H and T–T) and between populations (T–H), respectively. **d**, Number of species (total $n = 636$) with T–H strain-sharing rates of 0%, 0–5% and 5–100%. The number of species with T–H strain-sharing rates of 0% is significantly higher than expected by chance ($P < 10^{-4}$, one-sided permutation test). Inset, the strain-sharing rates for all inter-population comparisons of species in which 5–100% of lineages were shared recently. Data are mean $\pm$ s.e.m. across species. **e**, The within- and between-population strain-sharing rates for eight species with a range of different behaviours. Top, four species with T–H strain sharing rates of 0%. Bottom, four species with T–H strain sharing rates of >5%. Dot size indicates the number of MAGs in each population, with dots positioned as in **c** (no dot means no MAGs); line thickness indicates the recent lineage-sharing rate.

Reduced recent strain sharing

Although the average ANI values showed little geographical differentiation, we observed a much stronger trend in the high-ANI tail of the distribution ($>99.75\%$), indicating a marked enrichment of pairs within populations compared with between populations ($P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test; Fig. 1c, inset). This within-population enrichment suggests that there was restricted strain transmission between these

populations in recent evolutionary history, consistent with divergence driven by historical separation rather than ongoing exchange. However, estimating the timing of this separation can be challenging, because these high ANI values are often strongly influenced by recombination with other strains^{25,28,30}. Previous work has shown that these high-ANI pairs represent nearly clonal strains, which have genomes that have not yet been overwritten by recombination^{25,28,31} (Extended Data Fig. 5 and Supplementary Methods 3). Their clonal ancestry can be inferred from

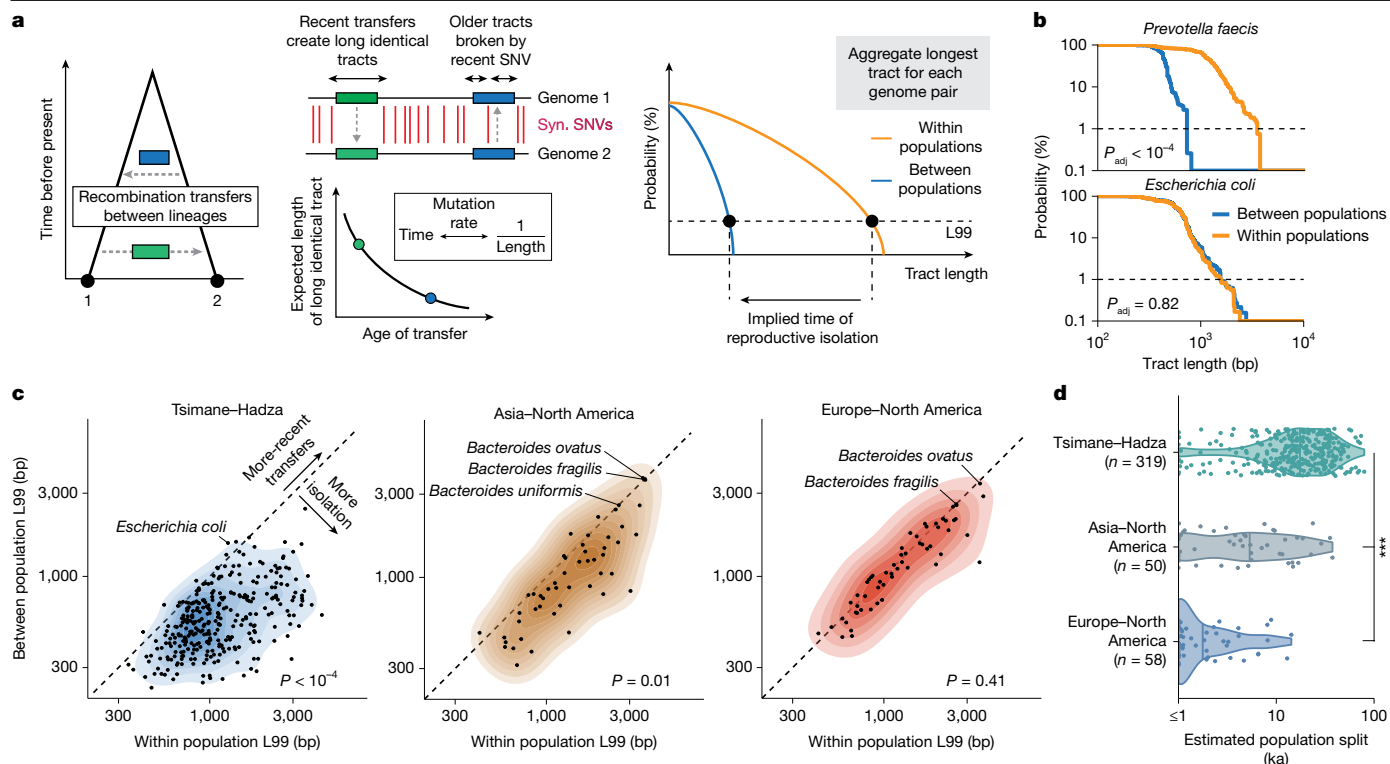


Fig. 3 | The absence of recent horizontal transfer of core genes indicates genetic isolation between Tsimane and Hadza gut bacterial populations for more than 10,000 years. a, Recombination creates long identical genome tracts between strains that are shortened by subsequent mutations. This produces a characteristic relationship between tract length and transfer age (Supplementary Methods 4). Reproductive isolation causes long tracts between populations to derive from older, pre-isolation transfers, leading to shorter aggregated tract lengths compared with those of within-population comparisons. This difference can be used to infer the timing of genetic isolation. Syn., synonymous; SNV, single-nucleotide variant. **b**, Empirical cumulative distribution function plots of tract length for two example species, comparing between and within the Tsimane and Hadza populations. bp, base

pair. Dashed lines indicate the L99 lengths. Adjusted $P(P_{adj})$ values correspond to one-sided permutation test results (with Benjamini-Hochberg correction, Supplementary Methods 4). **c**, Comparison of the L99 metrics for several population pairs (Tsimane-Hadza, Asia-North America and North America-Europe). P values indicate the results of Wilcoxon rank-sum tests comparing the results of between- and within-population L99 values for each pair of populations. **d**, The estimated population split time for each species. The 95% confidence intervals for each species (Supplementary Methods 4) are listed in Supplementary Table 3. *** $P < 10^{-4}$ for the comparisons of the Tsimane-Hadza split time estimates with those of both Asia-North America and Europe-North America (two-sided Mann-Whitney U -test).

the large stretches of nearly identical DNA sequences, which are interspersed with recombined segments from other strains. We estimated the total amount of clonal ancestry for each pair of MAGs using the fraction of core genes with identical DNA sequences (Fig. 2a). By calibrating these estimates with the rate of mutation accumulation in the vertically inherited genome regions, we determined that pairs of MAGs with more than 10% identical genes correspond to clonal strains that shared a recent common ancestor within the past approximately 5,000 years (Fig. 2a,b, Extended Data Fig. 7, Supplementary Table 4 and Supplementary Methods 3). Using this threshold, we identified all such pairs of clonal strains among the 636 shared species in Fig. 1c to investigate the landscape of recent strain transmission between the Tsimane and the Hadza.

Aggregated comparisons across all 636 species revealed that, in total, 78,746 MAG pairs (7.8%) were clonally related by this criterion, implying that they shared a most-recent common ancestor at most around 5,000 years ago (5 ka; Fig. 2b,c). Most of these clonally related pairs (93.8%) were found in the intra-population comparisons, indicating that most recent relationships occurred in the same human population ($P < 10^{-4}$, permutation test). Moreover, the few clonal strain pairs shared between the Tsimane and Hadza were unevenly distributed across species, with 545 of the 636 species (85.6%) showing no recent strain sharing between the two populations (Fig. 2d; $P < 10^{-4}$, permutation test). These results indicate that, in most shared species, very limited strain sharing has occurred between the Tsimane and Hadza over the past approximately 5,000 years.

However, there were some notable exceptions to this trend of limited inter-population strain sharing. Some of the species had levels of Tsimane-Hadza strain sharing comparable to their within-population baselines, consistent with a more recent spread. These species, which include *Akkermansia muciniphila* and *Bacteroides ovatus* (Fig. 2e, bottom), tend to be enriched in industrialized populations. To explore this pattern further, we expanded our analysis to include an additional 8,855 MAGs from Europe, Asia and North America³²⁻³⁴. The results revealed that these recently shared Tsimane-Hadza species also had closely related MAGs distributed globally (Fig. 2c-e and Extended Data Fig. 8). This suggests that species with high Tsimane-Hadza strain-sharing rates have been introduced into the Tsimane and Hadza through more-recent interactions with other human populations.

Notably, the comparison with industrialized MAGs also revealed another class of species that is prevalent across lifestyles but showed little recent strain transmission between the Tsimane and Hadza (Fig. 2e and Extended Data Fig. 8). These taxa, which include common species such as *Agathobacter rectalis* and *Ruminococcus bromii*, indicate that Tsimane and Hadza strains can remain isolated despite the global prevalence and recent sharing of these species in industrialized populations.

Reduced gene flow via horizontal gene transfer

One limitation of strain-sharing metrics is that the total number of clonal pairs is shaped by several factors, such as host contact

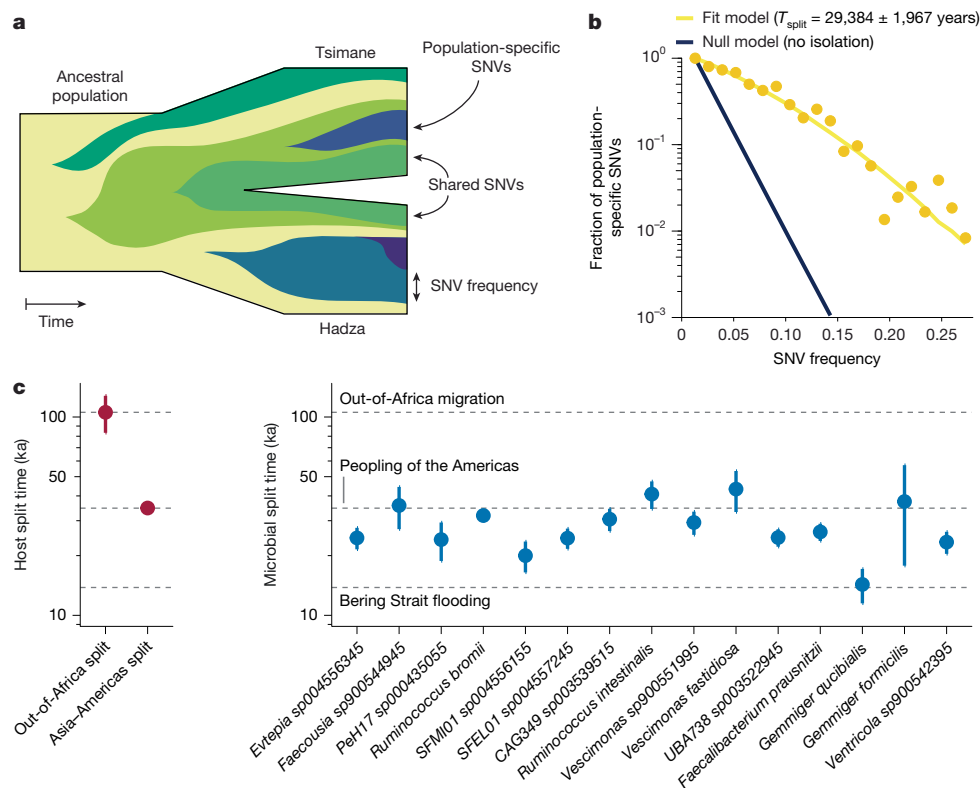


Fig. 4 | Gut bacteria demographic patterns mirror ancient human migration timelines. **a**, Illustration of a population split; each coloured segment represents the allele frequency trajectory of an SNV. More-recent mutations are less likely to spread between genetically isolated populations. **b**, Comparison of demographic scenarios for one example species (*Vesicomonas sp900551995*). The plot shows part of the two-population SNV frequency distribution (Supplementary Methods S), illustrating the relationship between SNV population frequency and the fraction of SNVs unique to one population. Observed data from the Tsimane and Hadza cohorts (yellow dots) are compared with predictions from two scenarios: the inferred demographic model (yellow curve; the 95% confidence interval represents uncertainty in T_{split}) derived using the complete SNV frequency data and a no-separation scenario assuming free migration (black curve; Supplementary Methods S). **c**, Right, inferred

population split times of 15 gut bacterial species shared between the Tsimane and Hadza. Left, the inferred timings of two human population splits—the out-of-Africa migration and the Asia–Americas split—are shown for comparison. These splits (shown with ± 2 s.e.) are based on a previously published study⁴⁸ that used the same demographic modelling method⁴⁹. The date of the Bering Strait flooding is based on a previously published study⁵⁰. Each point represents the maximum-likelihood split-time estimate for one species (centre value); error bars show ± 2 s.e. The s.e. for each species was derived from the site-frequency spectrum for that species as the asymptotic s.e. obtained from the Fisher information matrix of the Moments composite-likelihood fit (Methods); the exact number of biologically independent samples per species and the number of synonymous SNVs are listed in Supplementary Table 6.

networks^{35,36} and genome-wide selective sweeps³⁷, that can be difficult to compare across populations. To further corroborate our observation of microbial isolation between the Tsimane and Hadza (Fig. 2), we used an orthogonal approach that included information from the remaining pairs of MAGs (that is, those with <10% identical genes). Although their original clonal ancestry has been overwritten by recombination with other strains, recent recombination events between the ancestors of these more-distantly related strains generate shorter segments of nearly identical DNA against a backdrop of lower genome-wide ANI (95–99%) (Fig. 3a). Subsequent mutation and recombination events gradually degrade these segments over time, creating a relationship between the length of an identical stretch of DNA and the age of the transfer event. The timing of these events can be used to measure the rates of historical microbial gene flow within versus between human populations (Fig. 3a)—the conventional genetic signature indicating that they were part of the same microbial population³⁸.

We quantified this signal by identifying the longest tract of identical synonymous sites between each pair of MAGs for each of the 636 shared species (representative species are shown in Fig. 3b). We excluded clonal pairs (>10% identical genes) to focus on gene flow rather than on DNA acquired through common ancestry (as analysed in Fig. 2). Differences in tract lengths for within-population versus

between-population comparisons were most pronounced in the longer tracts, coinciding with recent recombination events (Fig. 3b and Supplementary Methods 4). Thus, to quantify this trend systematically, we calculated the 99th percentile of tract lengths (L99 length) for each of the 636 species.

Consistent with our other results, most species had shorter L99 lengths for comparisons between the Tsimane and Hadza than for comparisons within populations, suggesting genetic isolation in the recent past (Fig. 3c; $P < 10^{-4}$, Wilcoxon rank-sum test). However, some exceptions—including *Escherichia coli*—had similar L99 lengths across populations, reflecting their high levels of more recent strain-sharing (Fig. 2). We repeated this analysis for MAGs from Asia, Europe and North America (Fig. 3c). The differences in L99 lengths for within-population versus between-population comparisons were smaller for these industrialized populations than for the Tsimane and Hadza populations (Wilcoxon rank-sum test; Asia–North America, $P = 0.01$; Europe–North America, $P = 0.41$). This suggests that there is a stronger barrier to gene flow between the Tsimane and Hadza populations than between the industrialized ones. The same trend was also observed in species that were prevalent across lifestyles (Extended Data Fig. 9, $P = 0.018$, Wilcoxon rank-sum test), suggesting that it is not driven by systematic differences in the underlying microbial taxa being compared.

Split times mirror human migration

Next we used the observed L99 lengths (Fig. 3c) to estimate the timing of genetic isolation between the Tsimane and Hadza microbiomes. For each species, we added simulated mutations to pairs of within-population MAGs until their L99 lengths matched their observed between-population levels. By calibrating these *in silico* population splits with the estimated mutation rates of gut commensals²⁸, we can estimate the number of years since the gut strains in a given pair of populations last exchanged DNA (Fig. 3d, Supplementary Table 5 and Supplementary Methods 4). These estimates revealed that genetic exchange was considerably more recent between industrialized populations (Europe–North America: median = 1,109 years, s.e.m. = 377 years; Asia–North America: median = 4,637 years, s.e.m. = 1,267 years) than between non-industrialized populations (Tsimane–Hadza: median = 17,090 years, s.e.m. = 863 years, $P < 10^{-4}$, two-tailed Mann–Whitney *U*-test). Notably, the estimated onset of genetic isolation between the Tsimane and Hadza microbiomes roughly aligns with the timeframe between the initial migration of humans out of Africa (around 60–70 ka)^{39,40} and the settlement of the Americas (around 16–23 ka)^{41,42}, considering the uncertainties in our molecular clock estimates (Supplementary Methods 1).

To obtain an additional independent estimate of microbial split times, we used an existing demographic inference method⁴³. Rather than using stretches of identical DNA, this method uses an entirely orthogonal approach, analysing the distribution of single-nucleotide variant frequencies across populations (Fig. 4a,b, Supplementary Methods 5). Although applying this model to microorganisms has challenges, such as small sample sizes and complex population structures (for example, the presence of clades or subspecies), we identified 15 species with sufficient data and an appropriate population structure to estimate isolation times (Supplementary Table 6; see Supplementary Methods 5 for detailed selection criteria). For each species, the inferred models broadly align with those based on L99 lengths (Fig. 3d and Extended Data Fig. 10) and clearly support an ancient population split in microbial demography rather than two populations with no isolation (Fig. 4b,c and Supplementary Methods 5). Together, these results indicate that many gut microbial species have cohabited with humans for at least as long as humans have been migrating worldwide.

Discussion

Using several complementary analyses, we show that hundreds of commensal species inhabited the human gut for tens of thousands of years. These commensals have been transmitted across generations in Africa and during human migration worldwide. The findings shed light on some of the historical dynamics that have shaped the present-day geographical distribution of gut microbiome species. Many of these species were previously identified as VANISH taxa (volatile and/or associated negatively with industrialized societies of humans) based on their enrichment in Hadza hunter-gatherers^{3,44,45}. Our genetic analyses demonstrate that many VANISH species were also present in the common ancestors of the Tsimane and Hadza, suggesting that widespread losses might explain the absence of these species from industrialized populations. We also observed examples of BloSSUM species (bloom or selected in societies of urbanization or modernization)⁴ with extensive strain-sharing across industrialized and non-industrialized populations; this probably resulted from species transmission from industrialized populations to the Tsimane and Hadza. Finally, we identified a third class of species (stable independent of lifestyle (StIL)) that is prevalent across lifestyles but has signatures of Tsimane–Hadza isolation consistent with prehistoric co-migration. Notably, in this subset of StIL species, the degree of Tsimane–Hadza isolation was generally stronger than that observed between similarly separated industrialized populations (Fig. 2e and Extended Data Fig. 8). These observations

suggest that ongoing exchange in industrialized populations has probably degraded signals of co-migration; this may partly explain why previous observations of these signals have been limited.

Our results highlight that, by focusing on the microbiomes of contemporary populations with limited exposure to industrialization, we can better understand long-term human–microbiome associations that extend to early human history. Although these are contemporary populations, the genetic relationships between their gut bacteria can shed light on their past evolutionary history, similar to the benefits of including diverse populations in human genome studies⁴⁶. As we expand our knowledge of microbiomes in rural and Indigenous populations, it is crucial to adhere to ethical research practices and foster collaboration with local communities, respecting their autonomy throughout the research process while ensuring mutual benefit, fostering open dialogue and building trust. Notably, the loss of VANISH taxa with industrialization, which is linked to an increased risk of several chronic metabolic and inflammatory disorders, raises the possibility that reintroducing these species may help to restore microbial functions that co-evolved with human biology and improve people's health⁴⁷. This endeavour will require thoughtful and inclusive dialogue between researchers, ethicists and Indigenous communities.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-026-11106-1>.

- Sommer, F. & Bäckhed, F. The gut microbiota — masters of host development and physiology. *Nat. Rev. Microbiol.* **11**, 227–238 (2013).
- Sanders, J. G. et al. Widespread extinctions of co-diversified primate gut bacterial symbionts from humans. *Nat. Microbiol.* **8**, 1039–1050 (2023).
- Carter, M. M. et al. Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell* **186**, 3111–3124 (2023).
- Sonnenburg, J. L. & Sonnenburg, E. D. Vulnerability of the industrialized microbiota. *Science* **366**, eaaw9255 (2019).
- Blaser, M. J. The theory of disappearing microbiota and the epidemics of chronic diseases. *Nat. Rev. Immunol.* **17**, 461–463 (2017).
- Sprockett, D. D. et al. Microbiota assembly, structure, and dynamics among Tsimane horticulturalists of the Bolivian Amazon. *Nat. Commun.* **11**, 3772 (2020).
- Suzuki, T. A. et al. Codiversification of gut microbiota with humans. *Science* **377**, 1328–1332 (2022).
- Moeller, A. H. et al. Cospeciation of gut microbiota with hominids. *Science* **353**, 380–382 (2016).
- Amato, K. R. & Carmody, R. N. Gut microbial intersections with human ecology and evolution. *Annu. Rev. Anthropol.* **52**, 295–311 (2023).
- Nyholm, S. V. & McFall-Ngai, M. J. The winnowing: establishing the squid–*Vibrio* symbiosis. *Nat. Rev. Microbiol.* **2**, 632–642 (2004).
- Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* **42**, 165–190 (2008).
- Wibowo, M. C. et al. Reconstruction of ancient microbial genomes from the human gut. *Nature* **594**, 234–239 (2021).
- Moodley, Y. et al. The peopling of the Pacific from a bacterial perspective. *Science* **323**, 527–530 (2009).
- Linz, B. et al. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**, 915–918 (2007).
- Falush, D. et al. Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585 (2003).
- Thorpe, H. A. et al. Repeated out-of-Africa expansions of *Helicobacter pylori* driven by replacement of deleterious mutations. *Nat. Commun.* **13**, 6842 (2022).
- Mah, J. C., Lohmueller, K. E. & Garud, N. R. Inference of the demographic histories and selective effects of human gut commensal microbiota over the course of human history. *Mol. Biol. Evol.* **42**, msaf010 (2025).
- Moeller, A. H., Sanders, J. G., Sprockett, D. D. & Landers, A. Assessing co-diversification in host-associated microbiomes. *J. Evol. Biol.* **36**, 1659–1668 (2023).
- Nishida, A. H. & Ochman, H. Captivity and the co-diversification of great ape microbiomes. *Nat. Commun.* **12**, 5632 (2021).
- Amato, K. R. et al. Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. *ISME J.* **13**, 576–587 (2019).
- Ochman, H. et al. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol.* **8**, e1000546 (2010).
- Perez-Lamarque, B. & Morlon, H. Distinguishing cophylogenetic signal from phylogenetic congruence clarifies the interplay between evolutionary history and species interactions. *Syst. Biol.* **73**, 613–622 (2024).

23. Good, B. H. Limited codiversification of the gut microbiota with humans. *mBio* **17**, e03727-25 (2026).
24. Washburne, A. D. et al. Methods for phylogenetic analysis of microbiome data. *Nat. Microbiol.* **3**, 652–661 (2018).
25. Sakoparnig, T., Field, C. & van Nimwegen, E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *eLife* **10**, e65366 (2021).
26. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
27. Yahara, K. et al. Chromosome painting in silico in a bacterial species reveals fine population structure. *Mol. Biol. Evol.* **30**, 1454–1464 (2013).
28. Liu, Z. & Good, B. H. Dynamics of bacterial recombination in the human gut microbiome. *PLoS Biol.* **22**, e3002472 (2024).
29. Abdill, R. J., Adamowicz, E. M. & Blekman, R. Public human microbiome data are dominated by highly developed countries. *PLoS Biol.* **20**, e3001536 (2022).
30. Dixit, P. D., Pang, T. Y., Studier, F. W. & Maslov, S. Recombinant transfer in the basic genome of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **112**, 9070–9075 (2015).
31. Milkman, R. & Bridges, M. M. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* **126**, 505–517 (1990).
32. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
33. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
34. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
35. Beghini, F. et al. Gut microbiome strain-sharing within isolated village social networks. *Nature* **637**, 167–175 (2025).
36. Valles-Colomer, M. et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).
37. Yu, X. A. et al. Genome-wide sweeps create fundamental ecological units in the human gut microbiome. *Nature* **655**, 202–209 (2026).
38. Arevalo, P., Vaninsberghe, D., Elsherbini, J., Gore, J. & Polz, M. F. A reverse ecology approach based on a biological definition of microbial populations. *Cell* **178**, 820–834 (2019).
39. Haber, M. et al. A rare deep-rooting D0 African Y-chromosomal haplogroup and its implications for the expansion of modern humans out of Africa. *Genetics* **212**, 1421–1428 (2019).
40. Soares, P. et al. The expansion of mtDNA haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* **29**, 915–927 (2012).
41. Potter, B. A. et al. Current evidence allows multiple models for the peopling of the Americas. *Sci. Adv.* **4**, eaat5473 (2018).
42. Bennett, M. R. et al. Evidence of humans in North America during the Last Glacial Maximum. *Science* **373**, 1528–1531 (2021).
43. Jouganous, J., Long, W., Ragsdale, A. P. & Gravel, S. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics* **206**, 1549–1567 (2017).
44. Fragiadakis, G. K. et al. Links between environment, diet, and the hunter-gatherer microbiome. *Gut Microbes* **10**, 216–227 (2019).
45. Smits, S. A. et al. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802–806 (2017).
46. Ragsdale, A. P. et al. A weakly structured stem for human origins in Africa. *Nature* **617**, 755–763 (2023).
47. Li, F. et al. Cardiomitabolic benefits of a non-industrialized-type diet are linked to gut microbiome modulation. *Cell* **188**, 1226–1247 (2025).
48. Medina-Muñoz, S. G. et al. Demographic modeling of admixed Latin American populations from whole genomes. *Am. J. Hum. Genet.* **110**, 1804–1816 (2023).
49. Ragsdale, A. P. & Gravel, S. Models of archaic admixture and recent history from two-locus statistics. *PLoS Genet.* **15**, e1008204 (2019).
50. Pico, T., Mitrovica, J. X. & Mix, A. C. Sea level fingerprinting of the Bering Strait flooding history detects the source of the Younger Dryas climate event. *Sci. Adv.* **6**, eaay2935 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

Methods

Ethics and inclusion statement

Informed consent was obtained from the Tsimane 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). In 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 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 has trained dozens of Tsimane research assistants employed by the team. All of our research was conducted in collaboration with local Tsimane communities and the Gran Consejo Tsimane. The original impetus for collecting faecal samples was to analyse their microbial composition. However, as a benefit to participants, samples were also analysed by field microscopy for applicable parasitic species that could be treated with antihelminthic or anti-protozoan medications. The aim of the study was explained verbally to the participants. The THLHP has also engaged in capacity building through education (partnering with the One Pencil Project, a US non-governmental organization that provides educational supplies to primary schools and scholarships for Tsimane students pursuing university or medical school). The THLHP has invested in health infrastructure at the municipal hospital, has employed more than a dozen Bolivian physicians over the years, providing direct medical care during more than 65,000 patient visits, and has trained local health officials, built an official medical and dental clinic and provided expertise, personnel and equipment to two local hospitals.

We also analysed previously published metagenomic data³ from stool samples that were collected from Hadza hunter-gatherers in 2013–2014. The samples were legally and ethically collected in consultation with Tanzanian field guides after obtaining informed consent from Hadza participants who were given a translated verbal description of the research project. A material transfer agreement with the National Institute for Medical Research in Tanzania specifies that the collected stool samples will be 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 and NIMR/HQ/R.8a/Vol.IX/1542). The results of a recent study³ were converted into an infographic using the local language and conveyed to the Hadza by an anthropologist and a Tanzanian field guide.

Sample collection

The Tsimane stool samples were part of a previously published study⁶ and, as part of the current work, were subjected to additional sequencing. In brief, the two sets of samples (2009 and 2012–2013) were collected from participants across nine villages under the auspices of the THLHP⁵¹, which collects health and demographic data from Tsimane participants and provides primary health care to their communities. The set of stool samples used in this study was from Tsimane individuals who were at least 3 years old (mean age = 28.4 years, s.d. = 13.2 years). Field specimen collection protocols were devised considering the logistical challenges and were consistent across the 2009 and 2012–2013 cohorts. During data collection, none of the families in the participating villages had plumbing, used pit toilets or nappies, or had access to refrigeration⁶. Faecal samples were collected in sterile urine specimen cups given to the participants the day before sample collection. Researchers visited the participants' homes between 7 a.m. and 9 a.m. to collect the specimens. Samples were transported to a field laboratory in coolers with reusable ice packs within 1–2 h of

collection. Samples were homogenized in the collection cup and then divided into 2-ml sterile cryotubes using non-sterile wooden tongue depressors. Cryotubes were immediately stored in liquid nitrogen and then transferred to –20 °C freezers before being transported to the USA on dry ice. Samples were stored in –80 °C laboratory freezers in the USA until analysis. Sample collection information for the Hadza cohort is provided in the previous study⁴⁵. In brief, at the time of sample collection, when available, we recorded the age (either provided by the individual or estimated based on their recall), gender, weight and related information. After retrieval from the participants, the samples were immediately placed into liquid-nitrogen containers in the field, transported frozen to Arusha and then mailed by air on dry ice to the USA, where they were maintained at –80 °C until further processing.

Library preparation and sequencing

DNA was extracted from stool samples using MoBio PowerSoil kits (Qiagen) following the manufacturer's instructions. Libraries were prepared using Nextera Flex kits with a target of 10 ng of DNA with 12 bp unique dual-indexed barcodes for 12 cycles to minimize amplification bias. Paired-end sequencing (2 × 140 bp) was performed on a NovaSeq 6000 using S4 flow cells at the UCSD Microbiome & Metabolomics Core. Samples were randomized across runs and sequenced repeatedly until the target depth was reached. The minimum target depth for each sample was 100 million paired-end reads (around 28 gigabases). A total of 5.25 tera-base pairs of metagenomic data were generated.

Metagenome quality control and assembly

Raw sequencing reads were demultiplexed and processed using a custom processing pipeline (<https://github.com/MrOlm/nf-genomeresolvedmetagenomics>). In brief, adapters were trimmed using FastP⁵². Reads that aligned with the human and PhiX genomes were removed using Bowtie2 (ref. 53). FastQC was used to ensure read quality⁵⁴. Metagenomes were assembled using SPAdes⁵⁵, and reads were mapped against the assemblies using Bowtie2. Contig coverage was determined using CoverM⁵⁶. Assembly size and contig metrics were evaluated using QUAST⁵⁷, and assemblies were filtered to contigs of ≥1,500 bp for all downstream analyses.

Metagenome-assembled genome recovery

Genome binning was performed using METABAT2⁵⁸, and genome bin quality was assessed using CheckM⁵⁹. Genomes with ≥50% completeness and <10% contamination according to CheckM were retained, in accordance with MIMAG standards⁶⁰. Genomes with ≥50% completeness and <10% contamination were classified as medium-quality genome bins, and those with ≥90% completeness and <5% contamination as high-quality genome bins. We next used GTDB-Tk⁶¹ (r220) to determine the taxonomic classification of each medium- and high-quality genome bin. Genome bins derived from publicly available studies^{32–34} were also re-run using GTDB-Tk r220 to ensure up-to-date taxonomic classification.

Mapping-based prevalence calculations

To determine species presence across samples, quality-filtered metagenomic reads were mapped against the species-level representative genome database using Bowtie2. Alignment files were processed using inStrain quick_profile (v.1.2.14) and CoverM (v.0.4.0) to calculate per-genome coverage statistics. A species was considered present in a given sample if it achieved a breadth of coverage of ≥0.5 (at least 50% of the representative genome was covered by mapped reads). Prevalence for each species was then calculated as the percentage of metagenomes in which it was detected.

Sequence divergence and clonal strain-sharing analyses

To determine the all-versus-all ANI in each GTDB-assigned species, we used dRep^{62,63} (v.3.5.1) with the following parameters: (dRep compare

--S_algorithm goANI --SkipMash). This command ensures that all genomes in a GTDB-assigned species are compared using the goANI algorithm^{64,65}. In brief, goANI identifies open reading frames using Prodigal (v.2.6.3)⁶⁶, aligns their nucleotide sequences with NSimScan⁶⁷ and calculates ANI as the average sequence identity of all aligned genes. We used these ANI values to estimate an upper bound on the corresponding divergence times using molecular clock analysis described in Supplementary Methods 1.

To evaluate the statistical significance of the ANI differences within versus between populations, we performed a one-way analysis of variance for each species comparing the ANI distributions of intra- and inter-population strain pairs. We then applied Benjamini–Hochberg corrections to the *P* values for each species to obtain a false discovery rate-corrected *q* value for each species. We next used Fisher's sumlog method to combine these *q* values across species into a single final *P* value. We also calculated the fixation index F_{ST} (a standard measure of population differentiation) using the procedure described in Supplementary Methods 2.

To perform the strain-sharing analysis (Fig. 2), we first calculated the number of identical genes shared by each genome pair. The set of genes aligned between each genome pair was filtered to include only those with at least 500 bp aligned. The percentage of identical genes was then calculated as the number of gene pairs with 100% nucleotide identity divided by the total number of aligned gene pairs with at least 500 bp aligned.

We defined strains as clonally related when they shared more than 10% identical genes. To estimate the divergence time between such clonal pairs, we used a hidden Markov model-based method²⁸ to identify recent recombination events and obtain recombination-corrected divergence time estimates. Further details on this analysis and the corresponding inference procedure are provided in Supplementary Methods 3.

We used a permutation test to determine whether within-population genome pairs are more clonally related than are between-population genome pairs. For each species, we computed the test statistic $(p_{HH} + p_{TT})/2 - p_{HT}$, where p_{XX} denotes the proportion of genome pairs with >10% identical genes in each population combination. We generated a null distribution by permuting population labels across genomes within each species ($n = 10,000$ permutations) and calculated empirical *P* values as $(n^* + 1)/(n + 1)$, where n^* is the number of permutations with a test statistic of at least the observed value. For species-specific permutation tests, *P* values were corrected for multiple comparisons using the Benjamini–Hochberg procedure. For the global test, we permuted labels within species and pooled pairs across all species to calculate a single test statistic and *P* value.

SNV catalogues and gene flow analysis

For each species sufficiently represented in the Hadza and Tsimane, we generated a catalogue of SNVs present in the core genomes of that species (core genomes are defined as the core genes present in >90% of MAGs in a species). To create an SNV catalogue for a given species, we first chose one genome in the species as the representative genome based on the highest score determined using the following formula: Completeness $- 5 \times$ Contamination $+ 3 \times \log_{10}(N_{50}) + \text{Length}/10^6$ (N_{50} is the contig length at which 50% of the total assembled sequence length is contained in contigs of that length or longer). We then used a previously published pipeline to annotate the sites in the representative genome that had SNVs compared with all of the other genomes in the species⁶⁸ (https://github.com/zjshi/snv_analysis_almeida2019). We next filtered this whole-genome SNV catalogue to include only sites located in core genes. To identify core genes in the pangenome of each species, we called open reading frames for all genomes in the species using Prodigal (v.2.6.3)⁶⁶, clustered all open reading frames in the pangenome at 90% identity using MMseqs2⁶⁹ and determined which gene clusters were present in at least 90% of the genomes. On the

basis of these core genes, we created an additional mask that selected fourfold degenerate sites, thus focusing on sites where all mutations would be synonymous. The resulting SNVs were used to perform the gene flow analysis shown in Fig. 3. Further details of these analyses, including the underlying assumptions and motivations, are provided in Supplementary Methods 4.

Demographic inference using Moments

We used the SNV catalogues to perform the demographic inference shown in Fig. 4. For each species, we computed the frequency of each SNV in the Tsimane and Hadza populations to obtain a joint frequency spectrum across all fourfold degenerate sites. We then used the Moments software package⁴³ to fit these data to a demographic model of a historical population split. Further details of these analyses, including the quality-control criteria used to select the species in Fig. 4, are provided in Supplementary Methods 5.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Individual-level data are stored in the European Genome-Phenome Archive (EGA), governed by the Tsimane Health and Life History Project (THLHP), and available through restricted access for ethical reasons. The highest priority of the THLHP is to protect the study participants and minimize risk. The THLHP adheres to the CARE Principles for Indigenous Data Governance (collective benefit, authority to control, responsibility and ethics), which ensure that the Tsimane people (1) have sovereignty over how the data are shared; (2) are the primary gatekeepers determining ethical use; (3) are actively included in data collection; and (4) benefit from the data generated and shared whenever possible. The THLHP is also committed to the FAIR Guiding Principles for scientific data management and stewardship (findable, accessible, interoperable and reusable). Requests for individual-level data must be submitted as an application detailing the exact uses of the data and the research questions to be addressed, the procedures that will be used for data security and privacy, potential benefits to the study communities and procedures used to assess and minimize stigmatized interpretations of the research results (the data-sharing policy and data-request forms can be found at <https://tsimane.anth.ucsb.edu/data.html>). Requests for individual-level data may require approval from the EGA under Data Access Committee EGAC50000000982; the Data Access Policy is accessible publicly from the EGA (EGAP50000000936). The authors and 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 and other data are available at Zenodo (<https://doi.org/10.5281/zenodo.21765848>)⁷⁰. All Hadza data are publicly available at Zenodo as described previously (<https://doi.org/10.5281/zenodo.8072245>)³. Source data are provided with this paper.

Code availability

All custom code used for data analysis and figure generation is available at GitHub and indexed at Zenodo (<https://github.com/SonnenburgLab/comigration-metagenomics> and <https://doi.org/10.5281/zenodo.21206858>)⁷¹ or at GitHub (<https://github.com/MrOlm/nf-genomeresolvedmetagenomics>).

51. Gurven, M. et al. The Tsimane Health and Life History Project: integrating anthropology and biomedicine. *Evol. Anthropol.* **26**, 54–73 (2017).
52. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107 (2023).

53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
54. Babraham Bioinformatics. FastQC: a quality control tool for high throughput sequence data. v.0.12.0 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (Babraham Institute, 2023).
55. Pribelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes DE Novo Assembler. *Curr. Protoc. Bioinform.* **70**, e102 (2020).
56. Aroney, S. T. N. et al. CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**, btaf147 <https://doi.org/10.1093/bioinformatics/btaf147> (2025).
57. More, S. & More, A. Assessment the quality of genome assemblies by using QUAST tool for metagenomics. *Int. J. Recent Technol. Eng.* **8**, 4253–4259 (2020).
58. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
59. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
60. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
61. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
62. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
63. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
64. Varghese, N. J. et al. Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43**, 6761–6771 (2015).
65. Olm, M. R. et al. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems* **5**, e00731-19 (2020).
66. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
67. Novichkov, V., Kaznadzey, A., Alexandrova, N. & Kaznadzey, D. NSimScan: DNA comparison tool with increased speed, sensitivity and accuracy. *Bioinformatics* **32**, 2380–2381 (2016).
68. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
69. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
70. Carter, M. M. Prehistoric global migration of vanishing gut microbes with humans [Data set]. *Zenodo* <https://doi.org/10.5281/zenodo.21765848> (2026).
71. Carter, M. M. Prehistoric global migration of vanishing gut microbes with humans [Computer software]. *Zenodo* <https://doi.org/10.5281/zenodo.21206858> (2026).

Acknowledgements We thank the participants from the Hadza and Tsimane host villages and their families; the numerous people and organizations providing logistical support and collecting samples in Bolivia and Tanzania, including the THLHP staff members and researchers who provided assistance during field data collection; the Gran Consejo Tsimane; and D. Safaris, J. Chungalucha, A. Manjurano, M. G. Domiguez-Bello, M. St. Onge, A. Weakley, B. Merrill, S. Smits, Y. Gautam, D. Bhandari, S. Tandukar, G. P. Gautam, and J. B. Sherchand.

Author contributions Conceptualization: M.M.C., Z.L., M.R.O., B.H.G. and J.L.S. Methodology: M.M.C., Z.L., M.R.O., P.G., B.H.G. and J.L.S. Software and analysis: M.M.C., Z.L., M.R.O., P.G. and B.H.G. Writing, review and editing: M.M.C., Z.L., M.R.O., M.M., D.D.S., B.C.T., H.K., J.S., D.E.R., D.A.R., E.D.S., M.G., B.H.G. and J.L.S. Visualization: M.M.C., Z.L., M.R.O., B.H.G. and J.L.S. Administration of the THLHP: B.C.T., H.K. and M.G. Project administration: B.H.G. and J.L.S. Supervision: B.H.G. and J.L.S. Funding acquisition: D.A.R., B.H.G. and J.L.S. All authors read and approved the manuscript.

Funding This work was supported by Coefficient Giving, a Stanford Bio-X Bowes Fellowship to Z.L., the National Institutes of Health/National Institute on Aging (R01-AG054442 to H.K.), Alex and Susie Algard, the National Science Foundation (BCS 0422690 to M.G. and DDIG 1232370 to M.M.), the Wenner–Gren Foundation to M.M., the Thomas C. and Joan M. Merigan Endowment at Stanford University to D.A.R., the National Institutes of Health/National Institute of General Medical Sciences (R35-GM146949 to B.H.G.), and the French National Research Agency under the Investments for the Future (Investissements d’Avenir) programme (ANR-17-EURE-0010 to J.S.). J.L.S. and B.H.G. are Chan Zuckerberg Biohub – San Francisco Investigators.

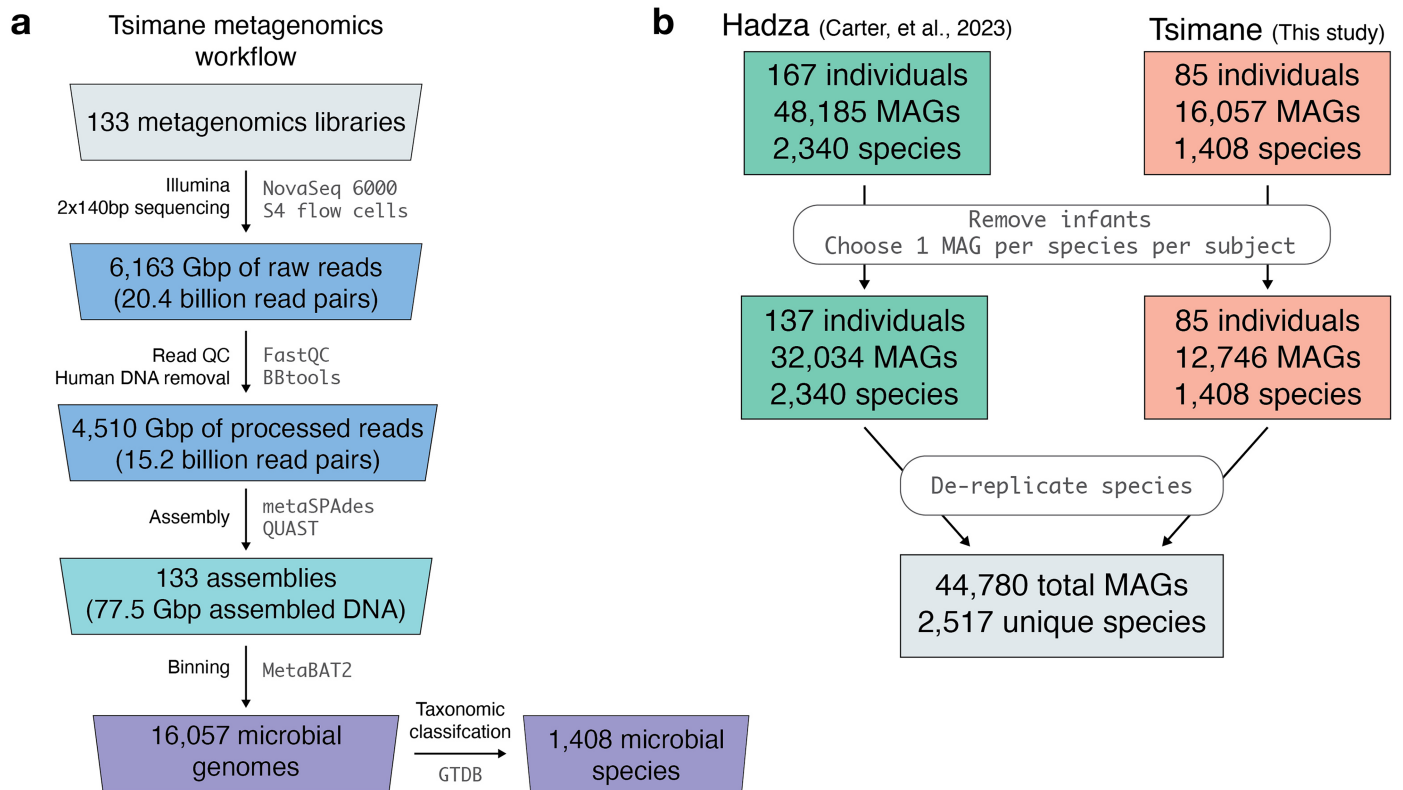
Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-026-11106-1>.

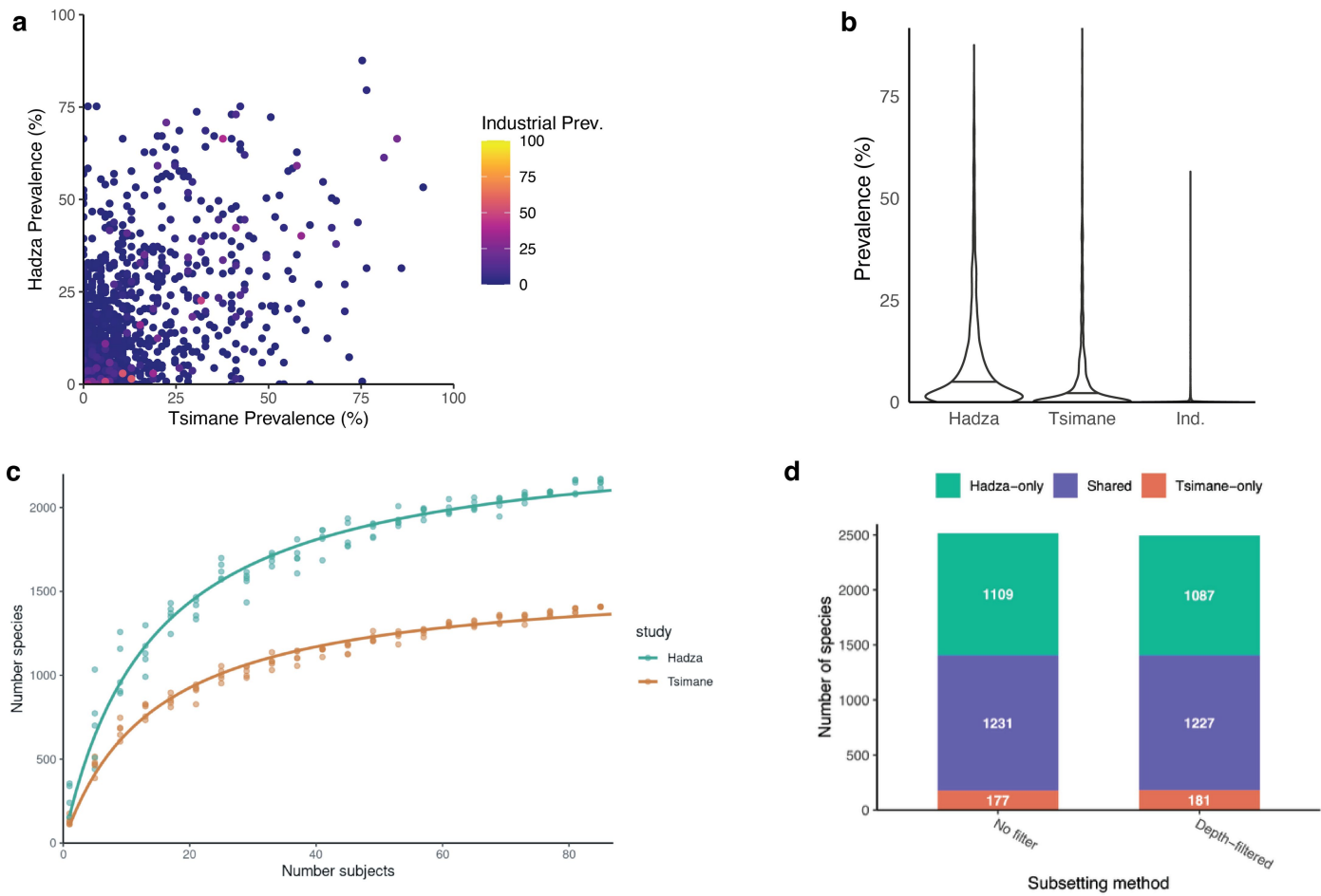
Correspondence and requests for materials should be addressed to Benjamin H. Good or Justin L. Sonnenburg.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

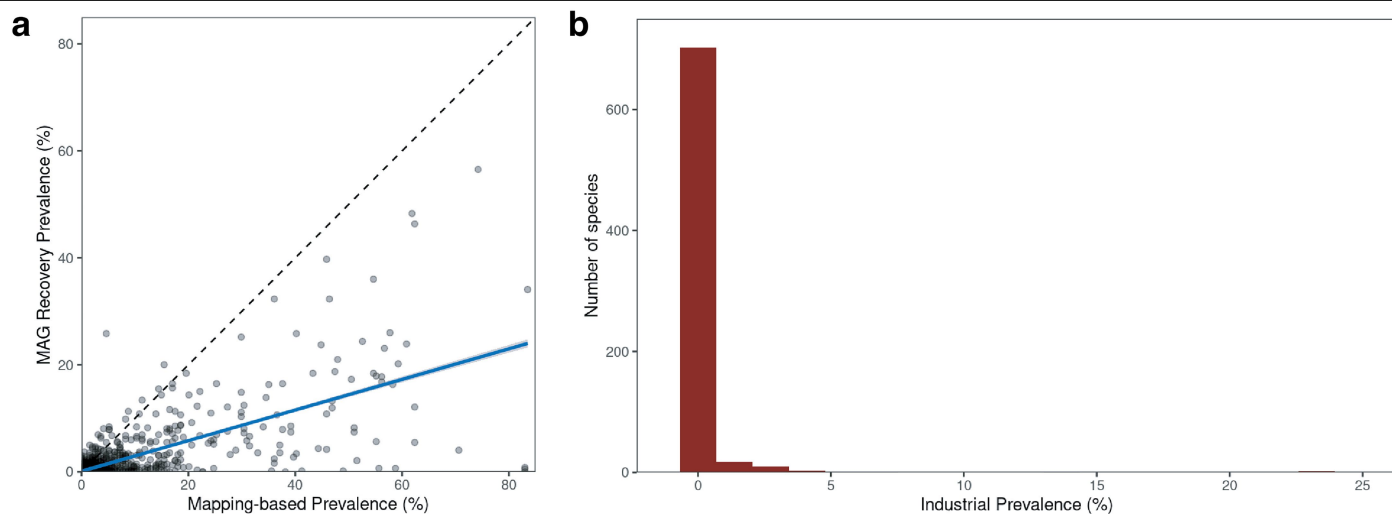


Extended Data Fig. 1 | Metagenomics workflow and curation of Tsimane and Hadza data sets. a, Overview of computational workflow, tools used and primary data generated from Tsimane stool samples. **b**, Overview of genome database curation for Tsimane and Hadza cohorts.



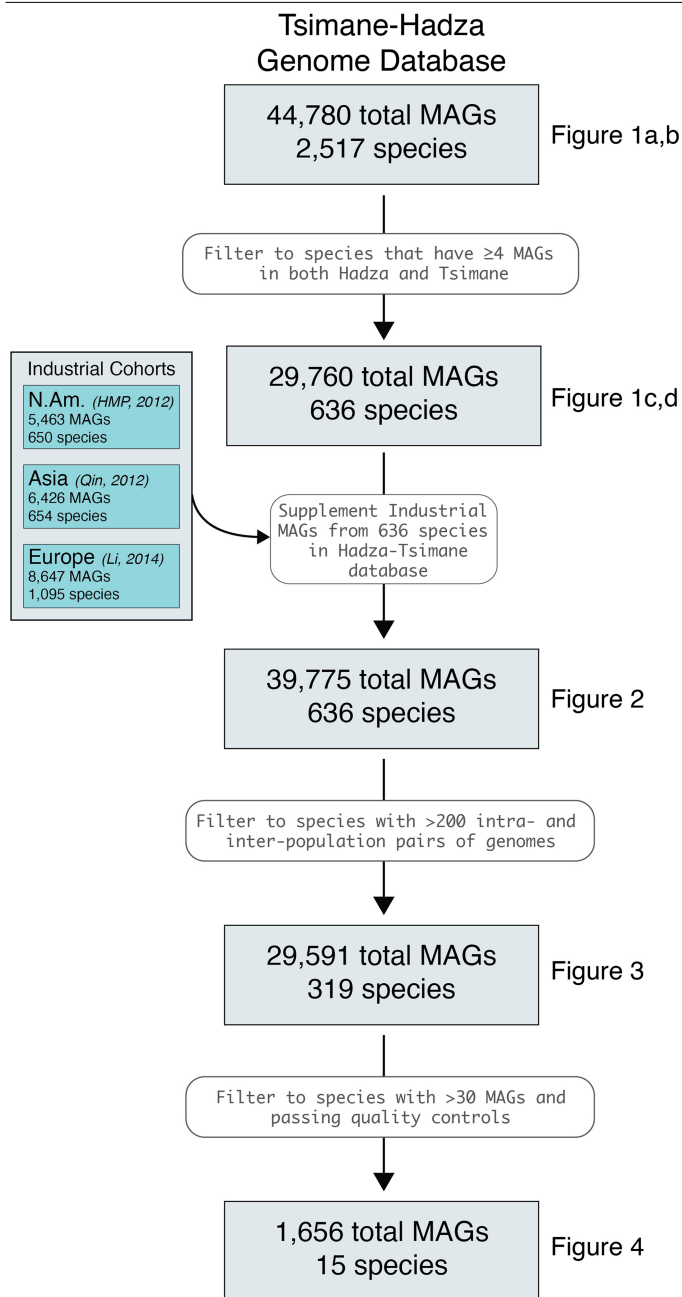
Extended Data Fig. 2 | Prevalence of gut bacterial species in the Hadza, Tsimane and Industrial populations. **a**, Scatter plot showing the prevalence of the 1,440 species recovered in the Tsimane (x axis) and in the Hadza (y-axis) as well as industrial populations (color scale). **b**, Violin plots summarize the prevalence of these gut microbial species in the Hadza, Tsimane and Industrial populations. **c**, Rarefaction curves showing the number of unique species detected in each population given some number of randomly selected 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). **d**, Depth-controlled comparison of species recovery between the two cohorts. Bars show the number of gut bacterial species detected only in the Hadza, shared between both cohorts, or detected only in the Tsimane, under two subsetting schemes: “No filter” (all samples) and “Depth-filtered”, in which the 4 Hadza samples (1.2%) sequenced more deeply than the deepest Tsimane sample were removed to match sequencing depth between cohorts. Matching depth changes the number of Hadza-only species by only 2.0% (1,109 to 1,087) and the number of shared species by 0.3% (1,231 to 1,227), indicating that the larger set of species recovered from the Hadza is not an artefact of their greater sequencing depth. $n = 331$ Hadza and 133 Tsimane biologically independent fecal samples.

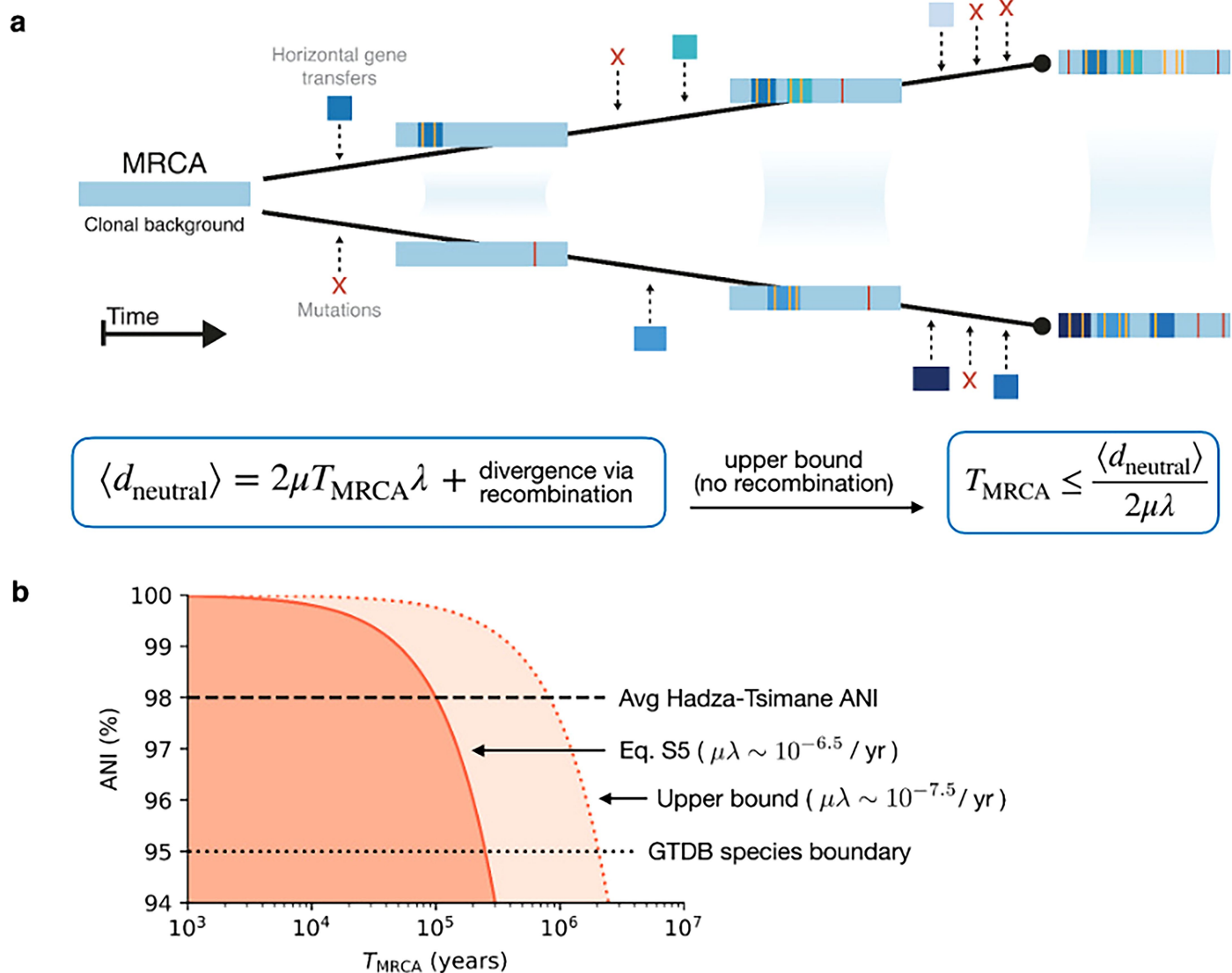


Extended Data Fig. 3 | Comparison of MAG recovery-based and read mapping-based prevalence metrics. a, Comparison of species prevalence metrics using a MAG recovery based approach and a read mapping based approach (Methods). Pearson correlation: 0.74, $p = 0.001$, two-sided.

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%).

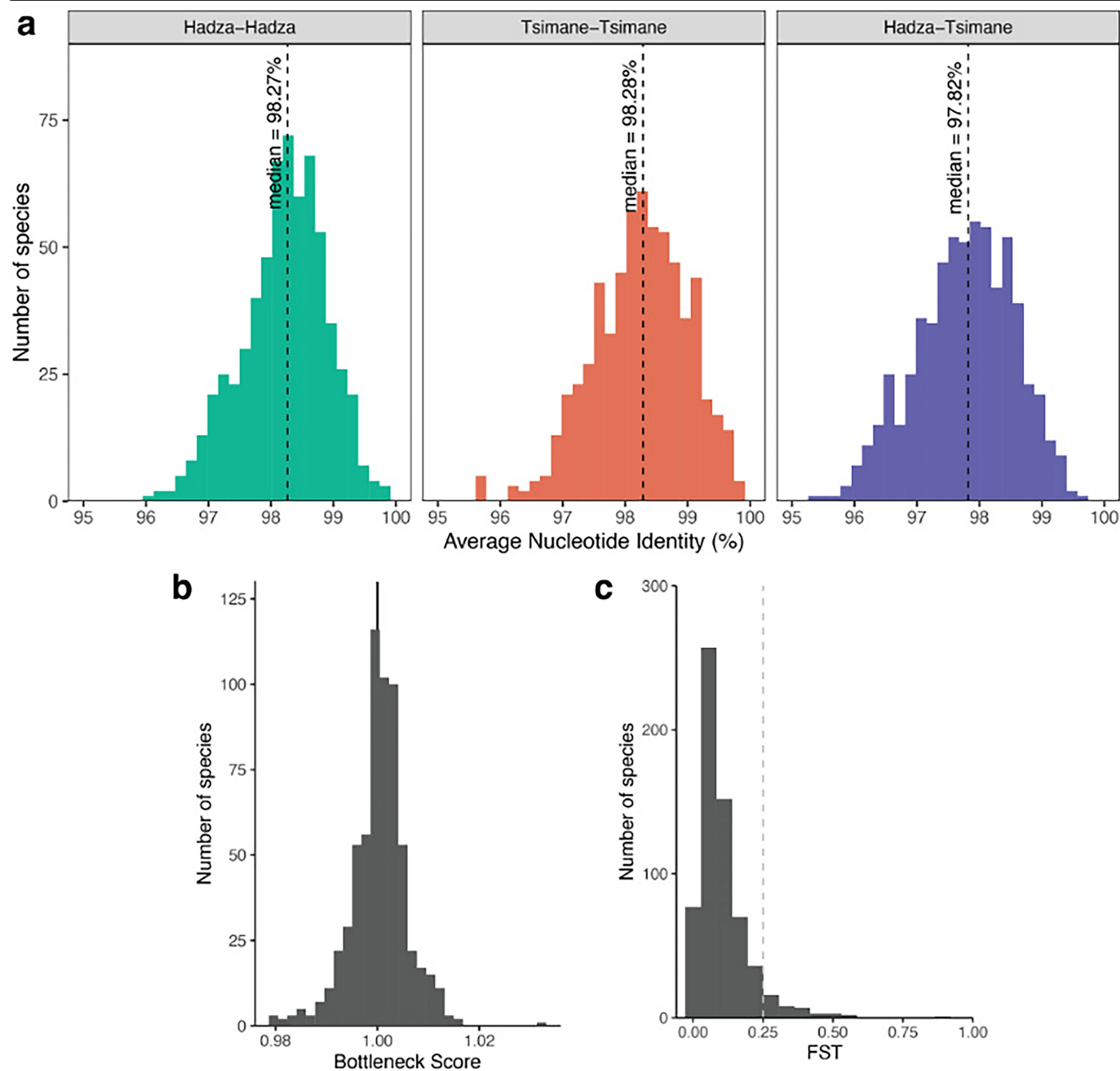


Extended Data Fig. 4 | A schematic demonstrating which cohorts, MAGs and species were used for each analysis and the filtering criteria that were deployed. See Supplementary Methods 5 for a detailed description of the quality control metrics used for the demographic inference approach in Fig. 4.



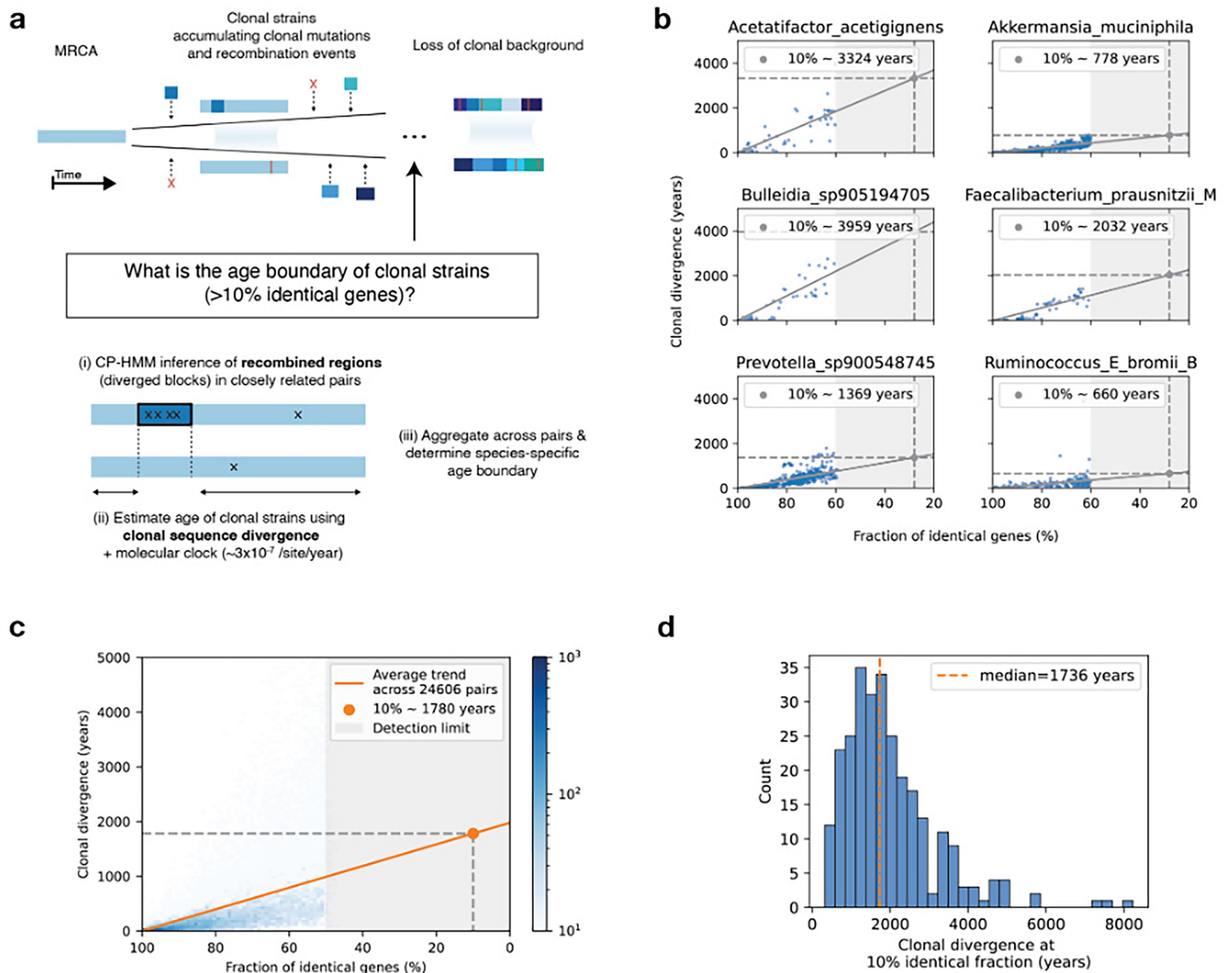
Extended Data Fig. 5 | **a**, Schematic illustration of the bacterial molecular clock. A pair of strains descended from a clonal ancestor at time $t = 0$ accumulate genetic differences via (i) de novo mutations (red) and (ii) homologous recombination events from other strains (blue blocks), which can introduce multiple genetic differences at the same time (yellow lines). These differences accumulate as the divergence time increases, reducing the overall ANI as well as the fraction of identical genes (light blue). Neglecting recombination leads

to an upper bound on the expected time to the most recent common ancestor of a pair of MAGs (Supplementary Methods 1). **b**, Implied upper bound of divergence times based on ANI. Shaded regions denote the T_{MRCA} values that are consistent with a given ANI level. The solid red curve uses the midpoint estimate of the yearly mutation rate, while the dashed red curve denotes an approximate upper bound.



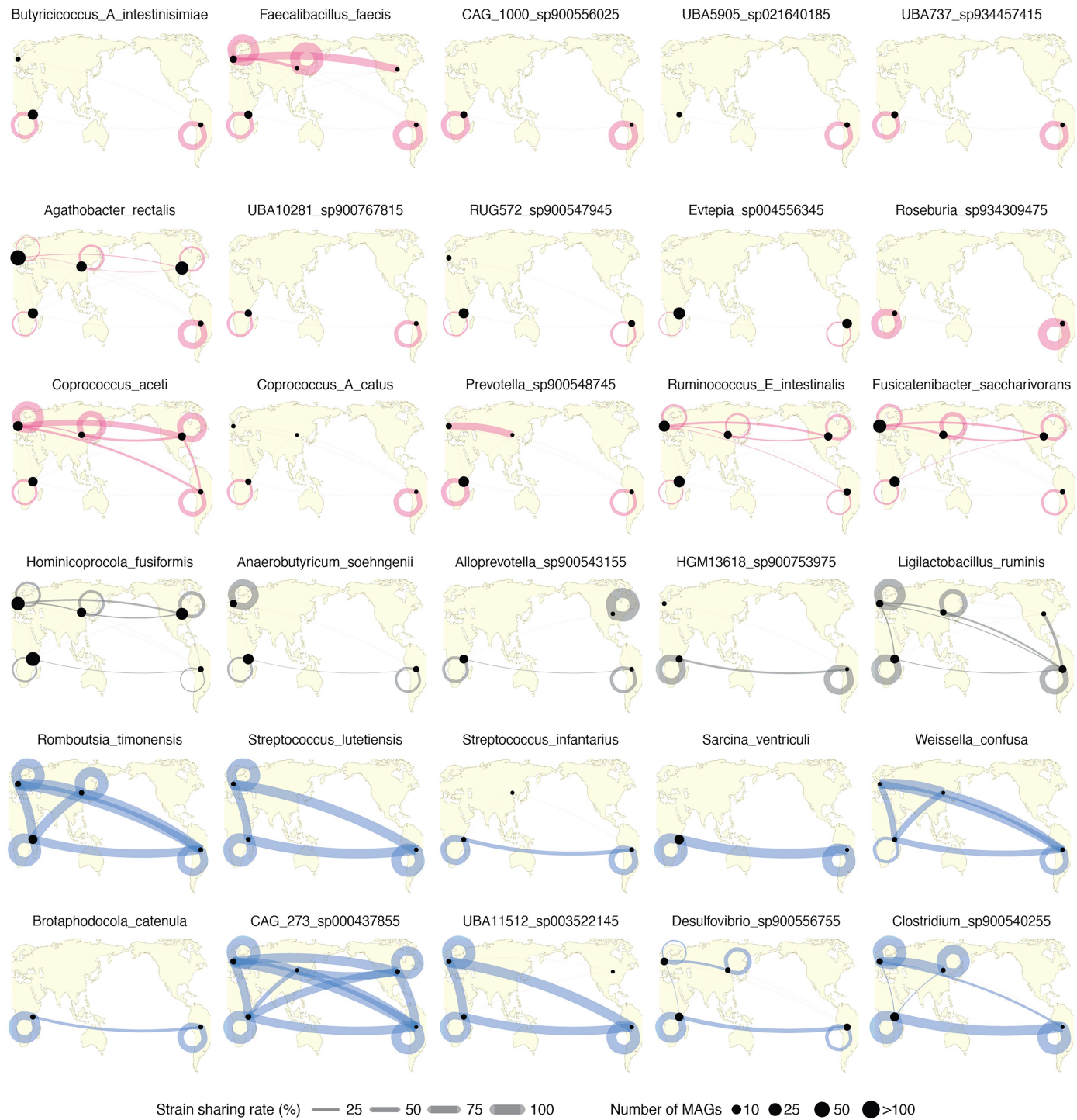
Extended Data Fig. 6 **a**, Distribution of average within-species ANI for each of the three population contrasts. **b**, Distribution of bottleneck scores for each of the 636 species. The bottleneck score was calculated for each species as: [Avg. T-T ANI / Avg. H-H ANI]. The median of the distribution is 1.00,

represented by the vertical black line. **c** Distribution of population fixation indices (F_{ST}) for each microbial species (Supplementary Methods 2). The vertical dashed line represents the approximate fixation index between host genomes for the Hadza and Tsimane (F_{ST} = 0.25, Verdu et al., 2014).



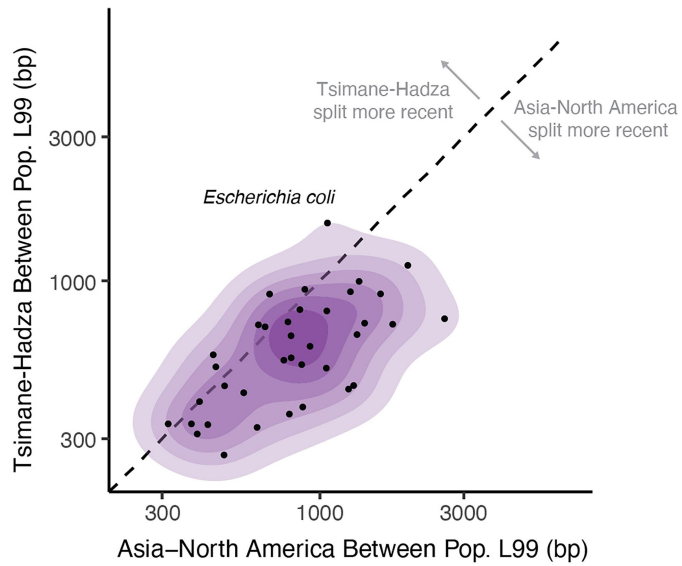
Extended Data Fig. 7 | a, Schematic illustration of the age boundary of clonal strains. A pair of strains descended from a clonal ancestor gradually lose their shared clonal background over time through (i) de novo mutations (red lines) and (ii) homologous recombination events from other strains (blue blocks). To determine a species-specific age boundary of clonal strains, recombination events and the clonal divergence times are inferred from closely related strain pairs (> 50% identical genes). The average trend across the aggregated data per species is then used to extrapolate the divergence time corresponding to the clonal strain boundary, defined as 10% identical genes (Supplementary Methods 3). **b**, Scatterplots showing the linear trend between the inferred

clonal divergence time and the fraction of identical genes for six example species. Each blue dot denotes one pair of MAGs. Grey symbols denote the linearly extrapolated divergence time corresponding to 10% identical genes. **c**, 2-D histogram (density heatmap) analogous to panel a, showing the linear trend for data aggregated across species. Solid line shows the linear fit using the data of all 44,780 MAG pairs with more than 50% identical genes. Color bar indicates the number of MAG pairs within each bin. **d**, Histogram of the extrapolated clonal divergence times corresponding to 10% identical genes, for each microbial species. Results of recombination inference for each pair of closely related strains are included in Supplementary Table 4.

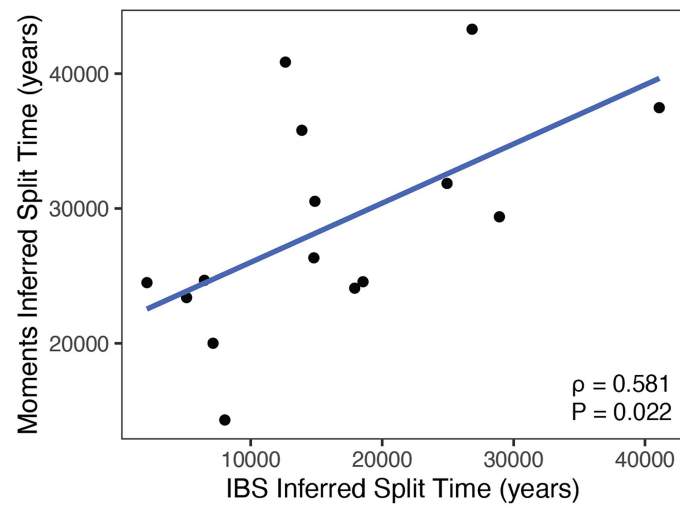


Extended Data Fig. 8 | A larger collection of map plots, similar to the ones shown in Fig. 2e. The top three rows (magenta lines) have Tsimane-Hadza (T-H) strain sharing rates of 0%, the fourth row (gray lines) shows species with T-H strain sharing rates between 0% and 5%, the bottom two rows (blue lines) show

species that have T-H strain sharing rates >5%. Size of dots represent the number of MAGs of that species derived from each population (no dot indicates no MAGs from that population). Thickness of connecting lines indicates the recent lineage sharing rate.



Extended Data Fig. 9 | Species-specific comparisons of the gut bacterial gene flow metric (L99 score, Fig. 3a) for species that are present in both the Tsimane and Hadza cohorts, as well as industrialized populations from Asia and North America. These comparisons show that for many individual species, the genetic isolation between the Tsimane and Hadza strains is stronger than that observed between comparable industrialized populations, even after controlling for differences in species composition (Wilcoxon test, $P = 0.018$, two-sided).



Extended Data Fig. 10 | Scatter plot showing the correlation between the identity-by-state (IBS) inferred split time in years with the split time inferred by Moments demographic modeling for the 15 species for which we have data for both analyses (Supplementary Methods 5). Pearson correlation coefficient = 0.581, $P = 0.022$, two-sided.

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.