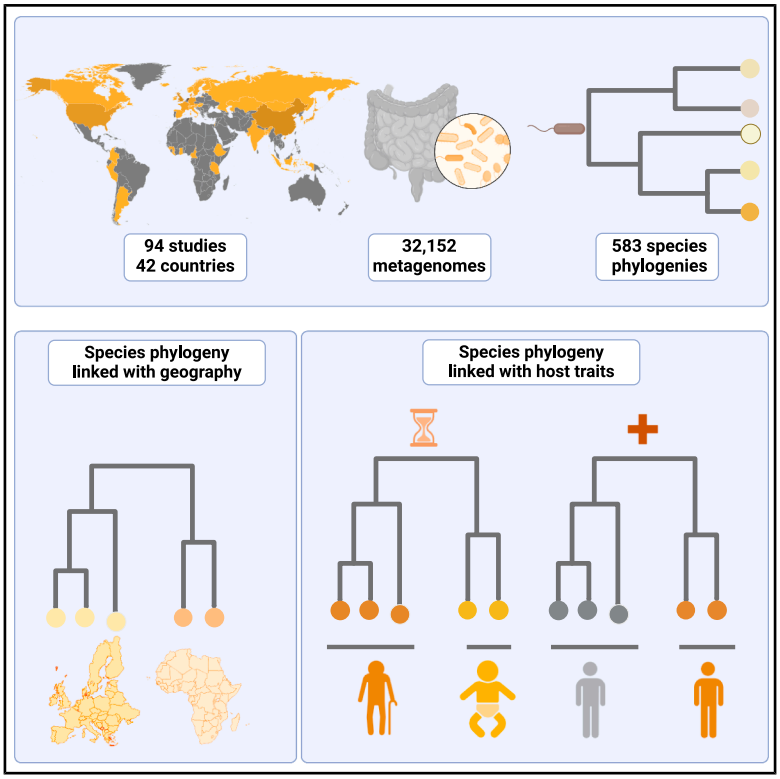
Article

Cell

Global genetic diversity of human gut microbiome species is related to geographic location and host health

Graphical abstract



Highlights

- Gut microbiome strain-level analysis for 583 species from 32,152 global samples
- Strain geographic stratification is related to species horizontal transmission
- Some human phenotypes are linked to microbial phylogenies
- Common cross-country *Collinsella* clades are more prevalent in cancer patients

Authors

Sergio Andreu-Sánchez, Aitor Blanco-Míguez, Daoming Wang, ..., Alexandra Zhernakova, Jingyuan Fu, Nicola Segata

Correspondence

j.fu@umcg.nl (J.F.), nicola.segata@unitn.it (N.S.)

In brief

32,152 gut metagenomes were analyzed, revealing strain-level geographic patterns shaped by horizontal transmission. Microbial phylogenies associate with human phenotypes, including *Collinsella* clades with melanoma and prostate cancer and *Ruminococcus gnavus* with aging.

Andreu-Sánchez et al., 2025, Cell *188*, 1–18
July 24, 2025
https://doi.org/10.1016/j.cell.2025.04.014

CellPress

Cell

CellPress
OPEN ACCESS

Article

Global genetic diversity of human gut microbiome species is related to geographic location and host health

Sergio Andreu-Sánchez,[1,2] Aitor Blanco-Míguez,[3,8] Daoming Wang,[1,2] Davide Golzato,[3] Paolo Manghi,[3] Vitor Heidrich,[3] Gloria Fackelmann,[3] Daria V. Zhernakova,[1] Alexander Kurilshikov,[1] Mireia Valles-Colomer,[3,4] Rinse K. Weersma,[5] Alexandra Zhernakova,[1] Jingyuan Fu,[1,2,*] and Nicola Segata[3,6,7,9,*]

[1]Department of Genetics, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands
[2]Department of Pediatrics, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands
[3]Department of CIBIO, University of Trento, Trento, Italy
[4]MELIS Department, Universitat Pompeu Fabra, Barcelona, Spain
[5]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands
[6]IEO, Istituto Europeo di Oncologia IRCSS, Milan, Italy
[7]Department of Twins Research and Genetic Epidemiology, King's College London, London, UK
[8]Present address: PreBiomics S.r.l., Trento, Italy
[9]Lead contact
*Correspondence: j.fu@umcg.nl (J.F.), nicola.segata@unitn.it (N.S.)
https://doi.org/10.1016/j.cell.2025.04.014

SUMMARY

The human gut harbors thousands of microbial species, each exhibiting significant inter-individual genetic variability. Although many studies have associated microbial relative abundances with human-health-related phenotypes, the substantial intraspecies genetic variability of gut microbes has not yet been comprehensively considered, limiting the potential of linking such genetic traits with host conditions. Here, we analyzed 32,152 metagenomes from 94 microbiome studies across the globe to investigate the human microbiome intraspecies genetic diversity. We reconstructed 583 species-specific phylogenies and linked them to geographic information and species' horizontal transmissibility. We identified 484 microbial-strain-level associations with 241 host phenotypes, encompassing human anthropometric factors, biochemical measurements, diseases, and lifestyle. We observed a higher prevalence of a *Ruminococcus gnavus* clade in nonagenarians correlated with distinct plasma bile acid profiles and a melanoma and prostate-cancer-associated *Collinsella* clade. Our large-scale intraspecies genetic analysis highlights the relevance of strain diversity as it relates to human health.

INTRODUCTION

The human gut microbiome is an intricate ecosystem of billions of microorganisms residing within the human gastrointestinal tract. The advent of DNA sequencing, followed by metagenomics, initiated an era of in-depth exploration into the variability of microbial taxa within and across host populations,[1–4] the complex assembly of microbial communities,[5,6] and the interplay between the gut microbiome and health and disease in humans.[7,8] The advancement of computational tools for the analysis of shotgun metagenomic sequencing has enabled researchers to not only profile microbes taxonomically and quantitatively but also to access their genetic information at subspecies and strain resolutions.[9–12] However, there is still much information to be gleaned from large-scale metagenomic association studies, which often still do not address the extensive genetic and functional diversity within microbial species.

The few studies exploring intraspecies metagenomic variation to date[13] have helped us to better understand host-microbe interactions. These studies have explored the person-to-person strain transmission[14] and biogeographic variability[9,15,16] of the gut microbiota. Previous research has also identified subspecies functional diversity linked to niche adaptation and microbial metabolic capabilities,[17–19] including adaptation to human genetic variants.[20] Strain variation has been found to be linked to different life-stages,[21] human lifestyle,[22] and phenotypic variability.[23] In the context of disease, it has been observed that a *Ruminococcus gnavus* clade[24] and a group of *E. coli* strains with an adherent invasive phenotype[25] commonly appear in inflammatory bowel disease (IBD) and that subspecies genetic variability has been related to the response to immune checkpoint cancer therapy[26,27] and microbial immunogenicity.[28] Most recently, within-species phylogenetic relationships between per-sample dominant strains have identified substantial

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

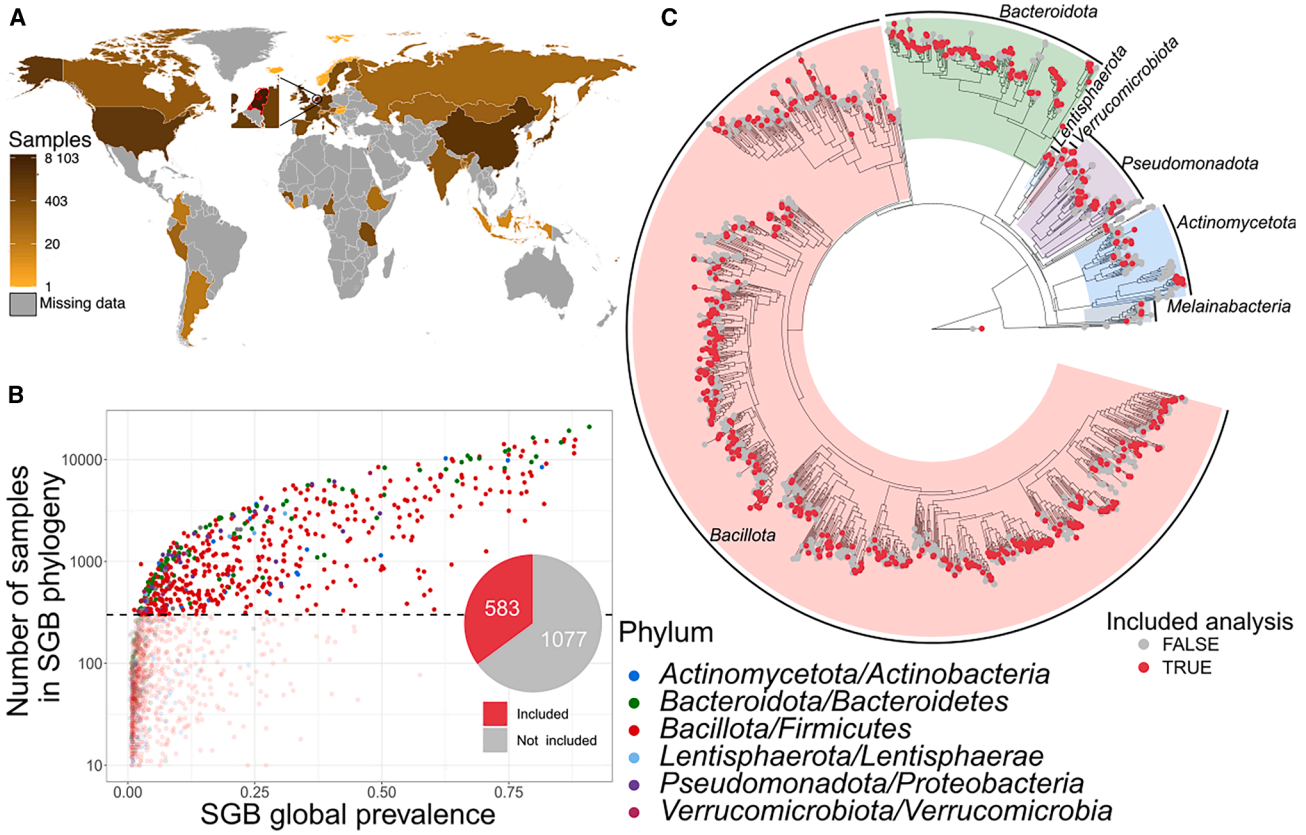


Figure 1. Global phylogenetic trees for 1,660 gut microbiome species using human gut metagenomic shotgun sequencing

(A) A global overview of the geographic origins and sample sizes of the gut metagenomes used in this study, which encompass a total of 32,152 samples from 42 countries.

(B) Prevalence of each species-level genome bin (SGB) in the metagenomic collection and the number of samples included in its phylogenetic tree. The 300-sample threshold (dashed line) is the cut-off used for inclusion in subsequent analyses.

(C) Overall phylogenetic diversity of the SGBs considered in our study, built using representative genomes for each SGB (STAR Methods). The 583 SGBs in red were included in the downstream analyses.

See also Figure S1.

intraspecies variation associated with type 2 diabetes,[29] IBD,[30] and colorectal cancer (CRC).[31]

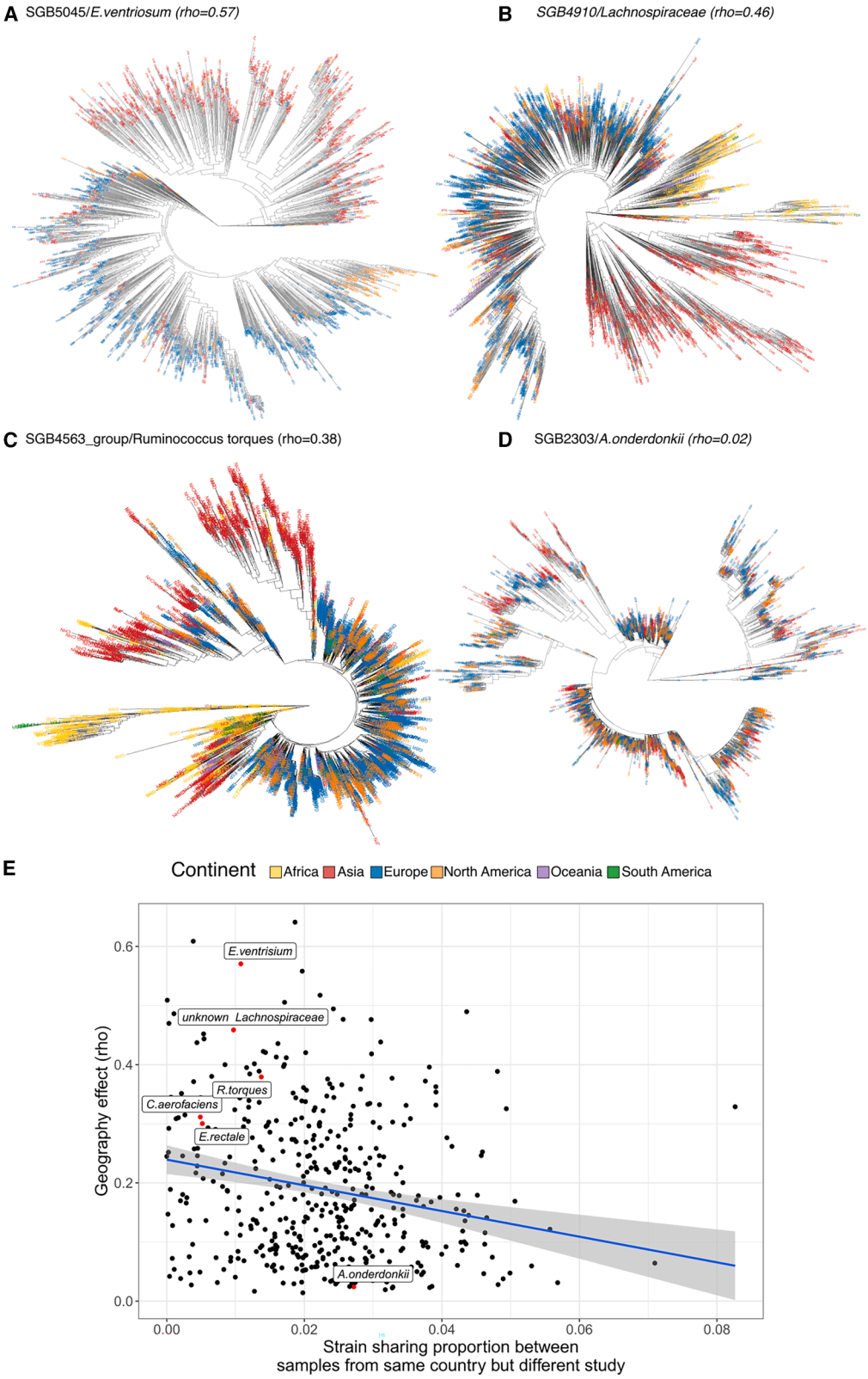
Despite these efforts, gut microbial intraspecies variability in the context of human health and disease remains relatively under-examined. This is partially due to the high resolution required for strain profiling, which can only be achieved for species with high coverage, and the substantial variability in microbiomes across individuals, which necessitates very large sample sizes for well-powered association studies. To address this, we carried out a global epidemiological investigation of specific microbiome members via strain-resolved metagenomics and phylogenetics. We incorporated 32,152 publicly accessible metagenomes, spanning samples from 94 distinct studies conducted across 42 countries and 6 continents, encompassing individuals from birth to 107 years of age. This integrated dataset covers over 241 phenotypes, including sample geographic origin, anthropometrics (n = 5), disease status (n = 102), medications and food supplements (n = 67), lifestyle factors and exposome (n = 24), and blood measurements (n = 48), among others. Of these phenotypes, 131 were only available for one cohort, primarily derived from the Lifelines cohort study, which comprises both the Dutch Microbiome Project (GacesaR_2022; 8,047 individuals) and Lifelines DEEP (ZhernakovaA_2016; 1,135 individuals) cohorts.[2,3] Another 110 phenotypes were available for at least two cohorts. We reconstructed phylogenetic trees for dominant strains within each sample for 1,660 microbial species. Using this expansive dataset, we analyzed intraspecies genetic variability across geographical locations and human phenotypic and environmental variation.

RESULTS

Phylogenetic reconstruction of 1,660 global microbial phylogenies

We assessed 32,152 gut metagenomics samples from 24,829 individuals. These were mainly European samples (65%), largely due to the contribution of datasets from individuals of Dutch origin (35% total), with 15.8% from North America, 14.3% from Asia, 3.83% from Africa, and 0.5% from Oceania and South America (Figure 1A; Table S1, 1). Using StrainPhlAn 4,[9] we

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014



(legend on next page)

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

reconstructed the phylogeny of the 1,660 species-level genome bins (SGBs, hereafter used as an operational definition synonymous to species)[32] for which enough samples with sufficient coverage were available (STAR Methods; Figure 1B; Table S1, 3).

StrainPhlAn 4 only considers the dominant strain in each sample, i.e., it uses the haplotype from the most abundant strain per individual sample. Although working with dominant strains might overlook the presence of multiple coexisting strains per species within each sample, we argue this is currently a necessary choice.[33] This is because profiling non-dominant strains is affected by strain-specific base-calling inaccuracies, and, at the current metagenomic sequencing depth, the log-normal abundance distribution and the observed within-species strain dominance cause non-dominat strains to be at very low abundance.[34] As a result, non-dominant allele frequencies are often indistinguishable from sequencing noise or can only rarely be reliably detected. In our data, we indeed identified a relatively small percentage of highly supported polymorphic positions per species (median percentage of polymorphic positions ranged from 1.79% in *Haemophilus parainfluenzae* to 0.026% in SGB53517 *Clostridia bacterium*). The median tended to increase in better-covered species, confirming that such within-sample variability likely exists ($r_{Pearson}$ = 0.3, $p = 2.5 \times 10^{-14}$) but would require greater sequencing efforts to uncover.

To ensure sufficient statistical power for analysis, we focused on the 583 SGB phylogenies with strains reconstructed from at least 300 samples (maximum of 20,956 samples for *Bacteroides uniformis*, median value of 929 samples per SGB) (Table S1, 3). Of these 583 SGBs, 567 belonged to 4 major human gut microbiota phyla: Bacillota (formerly known as Firmicutes; n = 412, 35% included), Bacteroidota (formerly known as *Bacteroidetes*; n = 98, 48% included), Pseudomonadota (formerly known as *Proteobacteria*; n = 29, 31.5% included), and Actinomycetota (formerly known as Actinobacteria; n = 21, 21.6% included). There were also 23 SGBs belonging to nine minor phyla, including two archaeal phyla (Crenarchaeota and Euryarchaeota) (Figure 1C).

We took several steps to assess the impact of different sample batches and studies in the phylogenies. We confirmed that the use of different DNA isolation methods in the same samples had minimal, if any, impact on SGB phylogeny. The phylogenetic placements of strains from the same samples were close regardless of whether DNA was extracted using the QIAmp Fast DNA Mini Kit (FSK) or the QIAGEN AllPrep DNA/RNA kit (APK) (Figure S1A; Table S2, 1), and their distance was consistent with sequencing noise (Figures S1B–1D; Table S2, 2). However, as expected and as a comparison, FSK and APK yielded comparably less similar species-level abundance profiles.[35] In addition, we focused on the yogurt-derived strain *Bifidobacterium animalis* as a positive control because all strains from this species that are found in humans should be identical due to their origin from food, which was determined in previous StrainPhlAn-based work.[14,33] We observed that the profiled strain was indeed the same in all samples, independent of study and country, which is in contrast to natural *B. animalis* strains in mice. This supports previous findings that technical biases in sample processing and sequencing do not substantially affect the genetic composition of this strain when it is found in humans (Figure 1D).[14] These results support the accuracy and robustness of our method for assessing the genetic diversity of microbial species.

Intraspecies genetic distance correlates with the hosts' geographic distance

Given the wide geographic distribution of the samples, we tested for geographical variation to determine whether greater geographic distances between samples do indeed correlate with higher phylogenetic distances of their respective microbial strains. Our analysis revealed significant associations for 456 of the 583 SGBs (Mantel test, Pearson's correlation geographical distance vs. phylogenetic distance, false discovery rate [FDR] < 0.05) (Table S3, 1).

Continent-specific clades were evident in many SGBs, although European and North American samples tended to cluster together. Examples of SGBs with strong continent-specificities include *Eubacterium ventriosum* ($rho_{MantelPearson}$ = 0.57, FDR < 7.9×10^{-4}, 2,582 samples), a Lachnospiraceae species ($rho_{MantelPearson}$ = 0.46, FDR < 7.9×10^{-4}, 6,553 samples), and *Ruminococcus torques* ($rho_{MantelPearson}$ = 0.38, FDR < 7.9×10^{-4}, 7,063 samples) (Figures 2A–2C). We further examined the phylogenies of *Collinsella aerofaciens* ($rho_{MantelPearson}$ = 0.31, FDR < 7.9×10^{-4}, 8,431 samples), which was previously reported to be co-adapted to humans,[15] and of *Eubacterium rectale* ($rho_{MantelPearson}$ = 0.3, FDR < 7.9×10^{-4}, 15,810 samples), which was previously shown to exhibit clear geographical stratification.[16] In both cases, we observed strong geographic correlations. Conversely, continent stratification was not obvious for phylogenies with a weak geographic effect, such as *Alistipes onderdonkii* ($rho_{MantelPearson}$ = 0.02, FDR < 7.9×10^{-4}, 12,042 samples) (Figure 2D).

Next, we identified specific taxonomic groups that exhibited stronger geographic effects than others. A rank-based enrichment analysis (STAR Methods) revealed that SGBs from the family *Lachnospiraceae* generally demonstrated a higher geographic effect compared with other families (set enrichment analysis [SEA], normalized enrichment score [NES] = 1.44, $p = 3.56 \times 10^{-5}$, FDR = 0.01) (Figures 2A; Table S3, 2). This enrichment was also evident for the order to which *Lachnospiraceae*

Figure 2. Geographic stratification of phylogenetic trees

(A–D) Representative SGB phylogenies, chosen due to their geographic diversity and sample size, with annotations for the continent where the samples were collected. For visualization purposes, we excluded samples from the Netherlands from the phylogeny due to the large amount of leaves.

(E) Geographic genetic effects and species transmissibility estimates are inversely correlated, showing that microbes that are more easily transmitted between people are less likely to be genetically stratified by geography. Only SGBs with significant geographic associations are displayed. Species reported in the text are highlighted in red. The linear smooth displays the fitted relationship between the features on the x and y axes. The gray shading represents the 95% confidence interval around the fitted line.

See also Figure S2.

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

belongs, Clostridiales (NES = 1.33, $p = 1.8 \times 10^{-4}$, FDR = 0.037). Interestingly, Clostridiales is part of the phylum Bacillota (formerly known as Firmicutes), which was previously observed to be the top phylum showing co-diversification with human populations.[15]

Thus, we do find evidence to support geographical variation linked to within-species genetic variation in the human gut microbiota, with the greatest links between microbial phylogeny and geography to likely be found in clades suspected to have the strongest shared evolutionary history with humans.

Geographic stratification is related to microbial horizontal transmission rates

Subsequently, we investigated whether SGBs with geographic associations shared common SGB-specific features, such as transmissibility rates (indicating transmission of microbes between hosts), average genome size, metabolic capabilities, and phenotypic traits (predicted from core genes from each SGB using Traitar[36]), and presence in different ecological niches. We identified 11 microbial phenotypes associated with geographic stratification (Figure 2B; Table S3, 3), and 3 positive associations with niche presence, including non-westernized microbiomes, ancient stool samples, and non-human primates (FDR < 0.05) (Table S3, 4).

We hypothesized that highly horizontally transmissible clades would be less likely to be stratified by geography because they may be more likely to overcome geographic barriers and spread across populations, thus decoupling phylogeny from geography. To test this hypothesis, we calculated strain-sharing events for the 583 SGBs among individuals from the same country but different studies. The SGB transmissibility rate (total number of sample-to-sample comparisons with the same strain divided by total number of sample-to-sample comparisons) ranged between 0% and 8.2%, with a median of 2.3%. As hypothesized, the geographic effect was smaller for SGBs with higher transmissibility ($r_{Pearson} = -0.21$, $p = 3 \times 10^{-7}$), and this trend persisted even when only considering the 456 SGBs significantly associated with geography ($r_{Pearson} = -0.20$, $p = 8.85 \times 10^{-6}$) (Figure 2E).

Phylogenetic association with human phenotypes

Next, we set out to test whether host phenotypic characteristics were associated with intraspecies SGB genetic variability. For each combination of the 583 SGB phylogenies and 256 phenotypes, we fitted a phylogenetic generalized linear mixed model (PGLMM) using the anpan R package,[37] which assesses the difference in predictive performance between models with and without phylogenetic information, quantified as the difference in generic (expected) log predictive density (elpd_diff; STAR Methods).

We performed an initial analysis including only age and continents as covariates. Then, for the associations identified, we conducted a continent-stratified model (i.e., a separate model per continent) while accounting for the geographic effect of different countries. This revealed 484 global associations in six continents, which accounted for 389 unique associations between 54 different phenotypes and 272 unique SGBs, with 80 associations reproduced in more than one continent (Table S4, 1). To ensure that possible population stratification of species did not confound these results, we associated the number of associations identified per species with their geographic stratification coefficient from the previous section. This did not find support for strong geographic confounding (linear model, number of supported associations vs. geographic effect, effect = 0.46, p = 0.23). In total, 272 unique species-phenotype associations (counted once if repeated by continent) were related to anthropometric factors (253 age, 12 BMI, 6 sex, and 1 waist-hip ratio), 52 to disease (13 IBD, 9 schizophrenia, 7 melanoma, and 23 associations to 14 other diseases), 28 to lifestyle and exposome (8 NO_2, 7 urbanicity, 5 current smoking, and 5 other phenotypes), 19 to biochemical associations (4 bilirubin associations, 3 high-density lipoprotein [HDL], 3 insulin, and 9 other phenotypes), 12 to medication and supplements intake (3 proton pump inhibitors [PPIs], 2 malazines, 2 vitamin B12, and 5 other phenotypes), and 6 to other uncategorized phenotypes (Figure S3A). Phenotypes available per cohort are listed in Table S1, 2.

The number of associations correlated positively with the number of samples included per continent ($rho_{Spearman} = 0.81$, p = 0.04) (Figures S3C and S3D), with an equal sample size yielding a similar proportion of SGB-phenotype associations per geographic region (Figure S3D). Thus, most associations were detected in European samples (n = 335), followed by Asian (n = 95), African (n = 48), North American (n = 8), South American (n = 1), and Oceanian (n = 1) samples (Figure S3B). A subsequent downsampling analysis of the European samples for a few species-age associations revealed a linear dependence between elpd_diff scores and sample size for associated taxa (a trend not present in non-associated taxa). This indicates that considerable power is required to identify phenotype-species associations, explaining the disparity in the number of associations between the continents (Figure S3C).

Despite the differences in the number of SGB-phenotype associations between continents, 80 such associations were reproducible between continents, even if not necessarily driven by the same clades (Figure S3B). The majority were reproduced between Asian and European studies (47 for age and 2 for IBD), followed by Europe and Africa (37 age), and finally between Europe and North America (6 age, 1 BMI, and 1 melanoma). Several associations were repeatedly observed in more than two continents, including ten associations with age seen in Africa, Asia, and Europe; one association with age seen in North America, Asia, and Europe; and two age associations reproduced in North America, Africa, Asia, and Europe (*Collinsella aerofaciens* and *Faecalibacterium prausnitzii*) (Table S4, 1). We thus observed evidence of wide associations between the genetic structure of subspecies and relevant host conditions and phenotypes that span across continents and can be used as hypotheses for follow-up experiments.

Phylogenetic association confirms three *Bifidobacterium longum* subspecies present in early life

Among anthropometric associations, age was the phenotype most commonly associated with SGB intraspecies diversity (253/583 SGBs), followed by BMI (12/583) and sex (6/583). For 14 SGBs, the age associations were driven by the presence of infant-specific clades. We identified 13 cases of species with phylogenetic signals from infants/toddlers (defined as

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

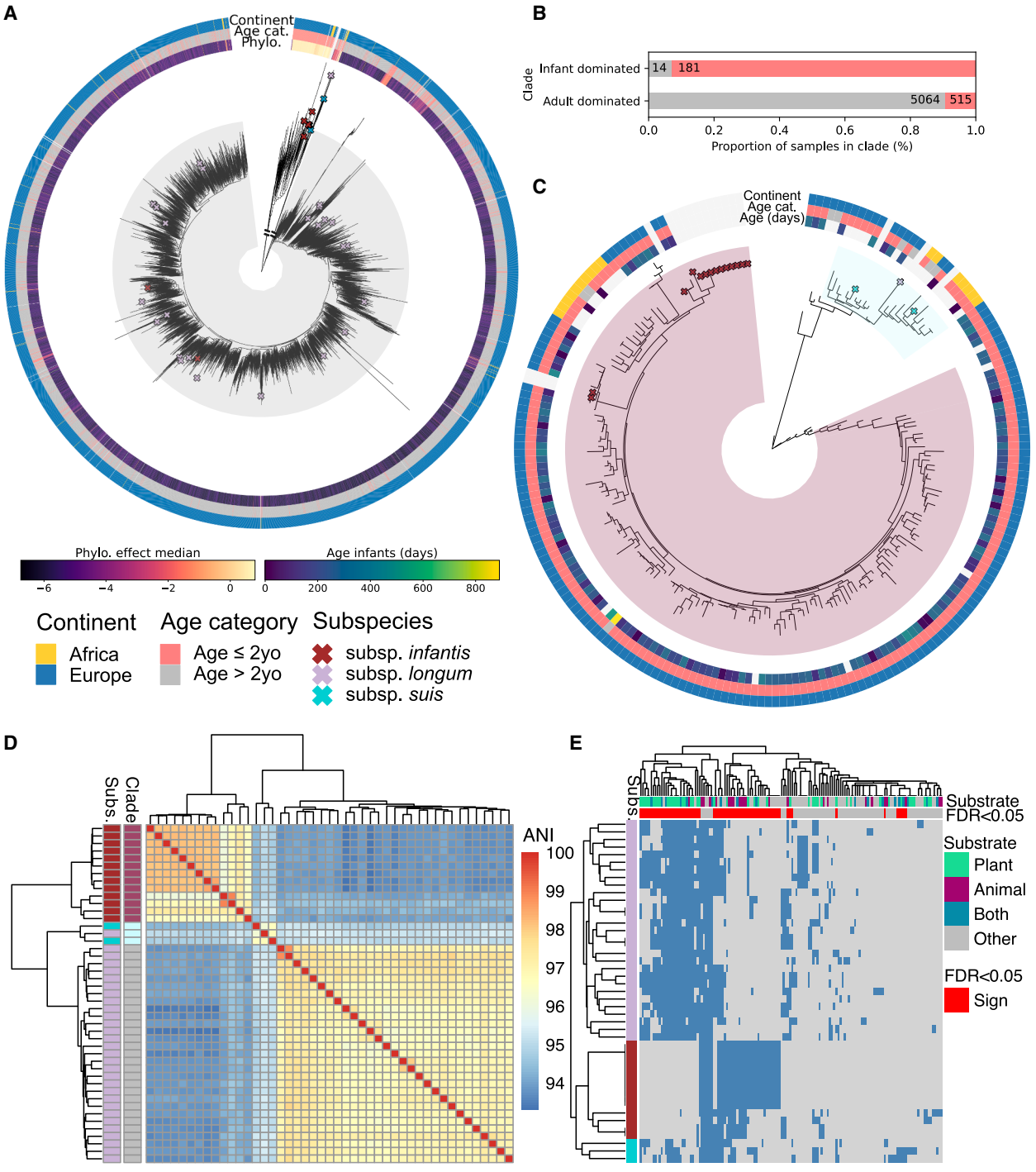


Figure 3. *Bifidobacterium longum* subspecies are associated with age and continent

(A) Phylogenetic tree of SGB17248 *B. longum* in a subset of European and African samples shows a clade dominated by individuals younger than 2 years of age (tips in red) and a clade of individuals older than 2 years of age (tips in gray). The phylogenetic signal of the association (phylo effect median), as given by the anpan model (see STAR Methods), the age category, and the continent are annotated in the outer rings.

(B) The infant-dominated clade is almost exclusively dominated by samples from individuals younger than 2 years, whereas the adult-dominated clade contains samples from individuals both older and younger than 2 years.

(C) Subset of the phylogenetic tree depicted in (A). The infant-dominated clade highlights the presence of two distinct infant clades: one dominated by references from *B. longum* subsp *suis* (blue) and one dominated by references of *B. longum* subsp. *infantis* (red). Annotation highlights the age of the infant (in days), the age category, the continent of the sample, and the NCBI subspecies for each reference genome.

(legend continued on next page)

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

individuals aged ≤2 years) in European samples, and 2 cases in African samples. The strongest association involved *Bifidobacterium longum*, the only species to show an infant-specific clade in both European (elpd_diff = −288) and African (elpd_diff = −9.1) populations. Indeed, analysis of the *B. longum* phylogeny revealed this was linked to a clear-cut clade separation between infants and older individuals (Figures 3A and 3B).

Previous research has pointed to the existence of three major *B. longum* subspecies in the gut of humans: *B. longum* subsp. *infantis*, which thrives in infants during breast milk feeding; a transitional subspecies closely related to *B. longum* subsp. *suis*, which was highly enriched in Bangladeshi infants after solid food introduction; and *B. longum* subsp. *longum*, which dominates after the age of 1.[21,40] Exploiting public *B. longum* genomes from isolates (STAR Methods), we observed that the adult-dominated clade clustered with references of *B. longum* subsp. *longum*, with infant samples falling in the adult clade being predominantly European (odds ratio [OR] European infant samples in adult-enriched clade = 14.2, $p_{Fisher} = 1.37 \times 10^{-7}$).

We also observed two clades largely exclusive to infants (Figure 3C). The major infant-specific clade was characterized by the presence of multiple *B. longum* subsp. *infantis* references and included both European and African infants (Figure 3C, in red). The smaller infant-specific clade (Figure 3C, in blue) where *B. longum* subsp. *suis* references clustered was significantly enriched in African samples (generalized linear model [GLM] Europe log-odds on clade = −3.32, $p = 8.13 \times 10^{-8}$). This minor infant-specific clade may represent the *B. longum suis* subspecies and/or its close relative transitional *B. longum*, and it might be more prevalent in non-westernized populations, as previously observed in Bangladesh.[21] Although we observed a reference of *B. longum* subsp. *longum* (GCA_000092325) within this clade, this genome was consistently closer to *B. longum* subsp. *suis* references, both when using whole-genome average nucleotide identity (ANI) values (Figure 3D) and when using the pangenome gene presence-absence matrix, suggesting that the GCA_000092325 genome might be mislabeled as *B. longum* subsp. *longum* (average ANI to genomes annotated as *B. longum* subsp. *suis* 97.7, average ANI to genomes annotated as *B. longum* subsp. *infantis* 95.1, average ANI to genomes annotated as *B. longum* subsp. *longum* 95.3).

B. longum subspecies are known to contain distinct carbohydrate-active enzymes (CAZymes) that allow them to adapt to the dietary transitions in infant diet. In agreement with this, we observed an enrichment of CAZymes with substrates for milk ($p_{Fisher} = 0.005$) and mucin ($p_{Fisher} = 0.02$) in genomes from the *B. longum* subsp. *infantis* clade (Figure 3D). Taken together, these results confirm that the statistical framework we used can also be used to identify subspecies variability associated with human phenotype.

Enrichment of a *Ruminococcus gnavus* clade in older Asian and European individuals

Following the discovery of specific *B. longum* subspecies enriched in infants, we investigated whether microbial clades might also be enriched in other age groups. We identified five Asian phylogenies with clades enriched in individuals over 90 years of age, including *Ruminococcus gnavus* (elpd_diff = −65.5) (Figure 4A), *Segatella copri* (elpd_diff = −14.2), a *Klebsiella* species (epld_diff = −5.2), *Alistipes onderdonkii* (elpd_diff = −5.4), and *F. prausnitzii* (elpd_diff = −4.2). These nonagenarian individuals were predominantly from XuQ_2021 (47/48), a study focused on Chinese nonagenarians and centenarians.[41]

We successfully confirmed the association between the *R. gnavus* phylogeny and age when only samples from XuQ_2021 were used, a setting in which possible between-cohort batch effects cannot play a role (elpd_diff = −20.1, nonagenarians = 47/96 samples). Within this phylogenetic tree, comprising 96 samples from individuals aged 56–105 years old, we observed an enrichment of nonagenarians in a major clade. We observed an enrichment of older individuals within such a major clade (GLM in XuQ_2021 $p_{xuQ} = 3.14 \times 10^{-5}$, generalized linear mixed model [GLMER] in all Asian studies [accounting for study] $p_{Asian\_tree} = 1.75 \times 10^{-11}$), independent of anthropometric, family, or technical confounders (GLM $p_{xuQ} = 6.8 \times 10^{-4}$). Intriguingly, although there were no nonagenarians in the European samples (21 octogenarians out of 1,171 in the phylogeny), we could replicate a significant phylogenetic signal between *R. gnavus* and age (elpd_diff = −34.99). By including all European samples in the phylogeny (Figure 4B) and comparing the average age of the European samples within the clade enriched in older individuals (defined using the Asian samples) with those outside of it, we detected a notable increase in the average age of the European samples within the clade (GLMER accounting for study, log-odds age = 0.03, $p = 1.88 \times 10^{-9}$). This age increase persisted even when restricting the analysis to samples from individuals aged >18 years (GLMER, log-odds age = 0.04, $p = 7.5 \times 10^{-11}$) (Figure 4C). In addition, in a small number of individuals ($n = 48$, mean age = 61.1 years old, 5 in the clade enriched in older individuals) from GacesaR_2022,[3] we saw no confounding effect of urbanicity in the age relationship with the presence of this clade ($p_{age} = 0.03$, p_{age} controlling for urbanicity = 0.02). Overall, this indicates that strains of *R. gnavus* within this clade tend to be more often seen in older individuals, independent of geographic origin.

R. gnavus has previously been related to inflammation,[42] but it was also associated with longevity in a Chinese study.[43] Previous research also highlighted the involvement of *R. gnavus* in bile acid metabolism,[44] particularly the production of iso-bile acids, which were elevated in Japanese centenarians.[45] Although bile acid data were not available for the nonagenarians in our merged

(D) Average nucleotide identity (ANI), between references obtained with fastANI,[38] shows three distinct clades and highlights a missannotation of one genome within the *suis* clade. NCBI subspecies annotation is displayed in the row annotation. The clade where the genome clusters in the phylogeny is displayed in the row annotation (gray: adult-like subsp. *longum* dominated clade, blue: infant-like subsp. *suis* dominated clade, reddish-brown: infant-like subsp. *infantis* dominated clade).

(E) Carbohydrate-active enzyme (CAZyme) subfamilies presence-absence matrix for *B. longum* references built with dbCAN3[39] highlights differences in carbohydrate degradation potential between clades. Row annotation displays subspecies annotation for each genome (with re-annotation of the mislabeled genome). Column annotation indicates the substrate origin of each CAZyme and whether it is significantly different between subspecies (Fisher's exact test).

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

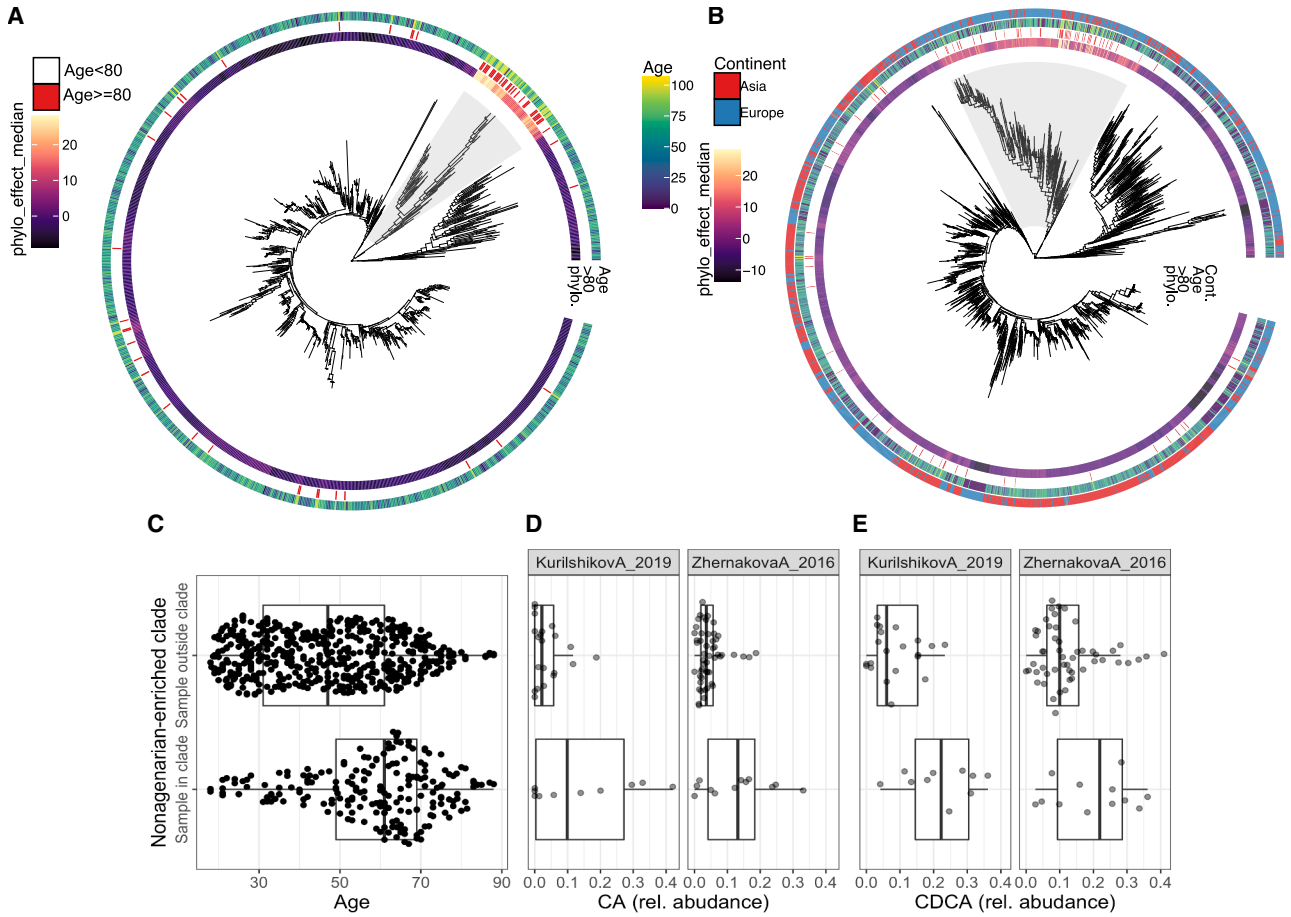


Figure 4. *Ruminococcus gnavus* clade characterized by higher age and distinct bile acid abundances

(A) Phylogenetic tree of SGB4584 *R. gnavus* in Asian samples identified a clade enriched in older individuals with high phylogenetic signal (highlighted in gray). The phylogenetic effect median (phylo_effect_median) indicating the strength of the phylogenetic signal on age prediction per sample (STAR Methods), the age category, and the age are annotated in the outer rings.

(B) Phylogenetic tree after including European samples shows an increase in the number of leaves in the clade enriched in older individuals. The phylogenetic effect median (phylo_effect_median), the age category, the age, and the continent are annotated in the outer rings.

(C) Comparison of age between samples in the highlighted clade in (B) (clade enriched in older individuals) and all others, using samples from adults residing in Europe. European adults whose strain belongs to the clade enriched in older individuals show an increased average age.

(D and E) Increased relative abundances of (D) plasma colic acid (CA) and (E) chenodeoxycholic acid (CDCA) in participants from KurilshikovA_2019 and ZhernakovaA_2016 for samples within the clade enriched in older individuals highlighted in (B) compared with those outside of the clade.

cohort, plasma measurements of 15 bile acids were available for samples from the Dutch cohorts ZhernakovaA_2016[17] and KurilshikovA_2019[46] (note that iso-bile acids were not measured and thus not included in our analysis). Despite a limited sample size, with 69 samples from ZhernakovaA_2016 (13 in the clade enriched in older individuals) and 31 samples from KurilshikovA_2019 (10 in the clade), we identified two age- and sex-independent associations in the *R. gnavus* phylogeny that were statistically significant after FDR-correction. Specifically, we observed associations of this clade with the abundances of two unconjugated primary bile acids: cholic acid (CA; linear mixed model, estimate = 0.09, FDR = 1.39×10^{-4}) (Figure 4D) and chenodeoxycholic acid (CDCA; linear mixed model, estimate = 0.1, FDR = 2.33×10^{-4}) (Figure 4E; Table S5, 1). These findings suggest that this *R. gnavus* clade enriched in older individuals is involved in bile acid metabolism, which supports previous findings.[44,45]

To further explore any links between bile acid metabolism and the *R. gnavus* clade enriched in older individuals, we reconstructed high-quality *R. gnavus* metagenome-assembled genomes (MAGs; >90% completeness, <5% contamination) and focused on the predicted presence of 13 known bile-acid-related genes (including *bai* genes, bile salt hydrolases [BSH], and hydroxysteroid dehydrogenases [HSDH]). Among 36 (out of 308) MAGs from samples within the clade enriched in older individuals, we identified a depletion of 7βHSDH (FDR = 3×10^{-8}, $\text{prevalence}_{\text{clade}} = 0.28$, $\text{prevalence}_{\text{rest}} = 0.8$) and an overrepresentation of 3αHSDH (FDR = 0.02, $\text{prevalence}_{\text{clade}} = 0.556$, $\text{prevalence}_{\text{rest}} = 0.324$), BaiA (FDR = 0.02, $\text{prevalence}_{\text{clade}} = 0.55$ $\text{prevalence}_{\text{rest}} = 0.32$), and BSH (FDR = 3×10^{-8},

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

prevalence$_{clade}$ = 0.44, prevalence$_{rest}$ = 0.06). Because some of these genes are usually found in operons, we also identified biosynthetic gene clusters related to metabolism. Among these, we observed a depletion of biosynthetic gene clusters for flavoenzyme sugar catabolism (FDR = 1.99×10^{-6}, prevalence$_{clade}$ = 0.04, prevalence$_{rest}$ = 0.75). Overall, this seems to indicate differential encoding of bile acid genes and a different sugar catabolism within the *R. gnavus* clade enriched in older individuals when compared with younger individuals.

A gut microbial intraspecies phylogenetic signal in melanoma and prostate cancer patients

Based on prior knowledge of the links between certain diseases and microbial strain variation in the gut microbiome,[13,29,30] we tested for associations between SGB strain-level phylogeny and a wide range of diseases contained in our multi-cohort dataset. This revealed that the strain-level genetic phylogeny of 37 SGBs was associated with human diseases. Notably, some SGBs were linked to more than one condition. For instance, in European samples, *C. aerofaciens* was associated with four different conditions (melanoma, IBD, hypertension, and asthma) and *E. rectale* was associated with three conditions (celiac disease, IBD, and any self-reported serious or chronic mental health disorder).

Overall, 17 different diseases were associated with at least one SGB strain-level phylogeny. Among these, CRC, hypertension, IBD, and melanoma exhibited associations across multiple continents. In samples from Asian individuals, CRC was associated with *Clostridium fessum* (elpd_diff = −8.87), whereas it was associated with *Lachnospira eligens* in samples from European individuals (elpd_diff = −4.2), as recently confirmed in a study comprising several additional CRC case/control cohorts.[31] It is worth noting that CRC has often been associated with *Fusobacterium* species and clades, as have several other bacterial species,[31,47,48] but the low abundance of these CRC-associated species did not allow for large phylogenetic reconstructions in this study (Table S1, 3).

Several strain-level gut microbial phylogenies were associated with other conditions. For example, 2 SGBs in Asian samples and 13 SGBs in European samples were linked to IBD, which might be related to the large sample sizes of IBD cohorts in European studies. Both associations in samples from Asian individuals were replicated in European samples and included *Blautia wexlerae* (elpd_diff$_{Europe}$ = −53.9, elpd_diff$_{Asia}$ = −4.74) and *F. prausnitzii* (elpd_diff$_{Europe}$ = −6.32, elpd_diff$_{Asia}$ = −4.44).

Notably, two of the strongest associations were observed between samples from melanoma patients and *C. aerofaciens* phylogeny in both European (elpd_diff = −141) and North American (elpd_diff = −49.9) gut microbiomes. Despite originating from different continents and studies, melanoma patients were enriched for the same clade in both European and North American samples. We integrated the European and North American samples into the *C. aerofaciens* phylogeny and delineated the melanoma-enriched clade (Figure 5). This clade predominantly comprised samples with a documented melanoma diagnosis (123 out of 181 samples in the clade had melanoma from a total of 170 samples from melanoma patients present in the phylogenetic tree, OR$_{melanoma\ in\ the\ clade}$ = 131.843, $p < 2 \times 10^{-16}$). These melanoma samples originated from three different countries across five melanoma-exclusive studies: LeeKA_2022[49] (85 samples: 33 Dutch and 52 British), McCullochJA_2022[50] (19 US samples), WindTT_2020[51] (8 Dutch samples), FrankeIAE_2017[52] (10 US samples), and PetersBA_2019[53] (1 US sample). We further investigated whether samples from LeeKA_2022 and McCullochJA_2022, which included information about participants who had undergone previous treatment, exhibited any significant association between previous treatment and clade membership. However, with the current sample size, no significant association was observed for either study (logistic regression log-odds$_{cancer_clade}$ with previous treatment = 0.3, p = 0.36). The clade was also not associated with the response evaluation criteria in solid tumors (RECIST) classification of the response to cancer therapy, indicating no evidence for this specific *Collinsella* clade being linked to cancer therapy efficacy (likelihood ratio test clade logistic regression, with and without RECIST term, p = 0.46).

In addition to melanoma, we found prostate cancer to be associated with the *C. aerofaciens* phylogeny at the strain level. The vast majority (12 out of 14) of participants from PernigoniN_2021[54] present in the phylogeny clustered within this same clade (Fisher's exact test for overrepresentation, $p = 4.97 \times 10^{-14}$), including 7 samples from the UK and 5 from Switzerland. Notably, all individuals from this study had prostate cancer,[13] either hormone-sensitive prostate cancer (four in the phylogeny, all also found within the cancer-enriched clade) or castration-resistant prostate cancer (ten in the phylogeny, eight also within the cancer-enriched clade). However, there was no association between the clade and treatment with antiandrogens (enzalutamide/abiraterone) or antineoplastics (docetaxel/cabazitaxel). The original study indicated that some bacteria could produce androgens, such as testosterone, which are associated with adverse cancer outcomes.[54] To investigate whether bacteria in this clade might be related to higher circulating testosterone levels, we leveraged plasma testosterone measurements available for healthy participants from SchirmerM_2016,[55] for which we identified 15/298 participants within the melanoma-enriched clade. However, we found no significant association between this clade and plasma concentrations of testosterone (p = 0.42).

To delve into the functionality of this clade, we reconstructed 44 high-quality MAGs from *C. aerofaciens* using the samples present in the phylogenetic tree. Among these, seven belonged to samples within the melanoma-enriched clade: four from the LeeKA_2022 study, two from McCullochJA_2022, and one from WindTT_2020. From these reconstructed MAGs we identified 332 UniRef90 annotations as differentially enriched between the melanoma clade and the rest of the phylogeny, with 218 genes more frequently present in the melanoma-enriched clade (Table S5, 4). For 56% of the tested genes, we could infer the MetaCyc pathway(s) to which they belong. Using these annotated genes, we performed an additional enrichment analysis (SEA) to uncover over- or under-represented pathways within the cancer-enriched clade. We identified 10 over-represented pathways in the melanoma-enriched clade (FDR < 0.05), including five related to the cobalamin (vitamin B12) biosynthetic process. High vitamin B12 levels have previously been related to

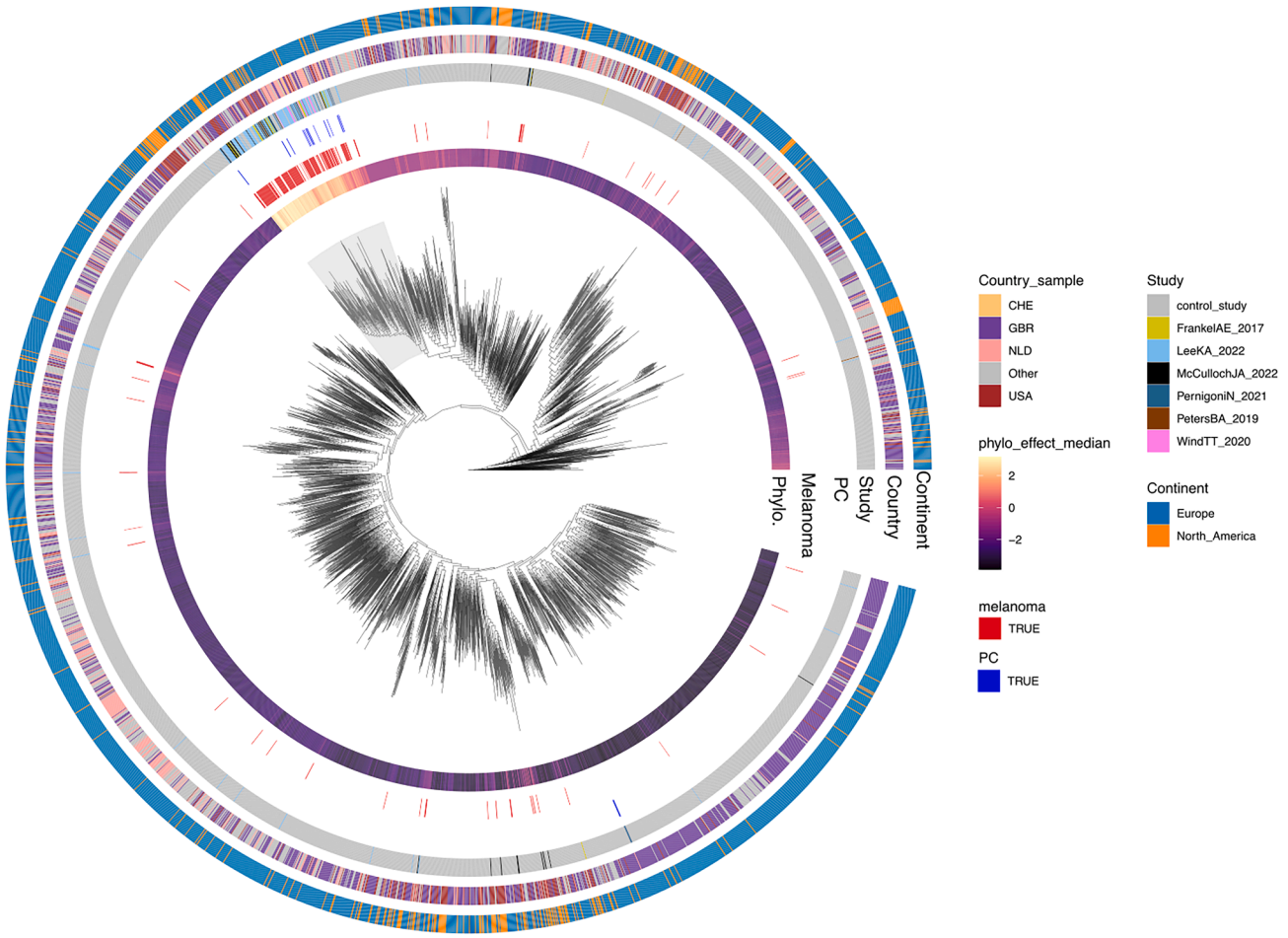


Figure 5. Phylogenetic tree of *Colinsella aerofaciens* shows a clade enriched in melanoma and prostate cancer

Gut metagenomic samples from individuals with melanoma (red annotation) and prostate cancer (PC, blue annotation) cluster together in a cancer-enriched clade (gray shading) with a high phylogenetic signal. Samples in the clade come from different studies, countries, and continents (colored around the phylogeny). The phylogenetic effect median (phylo_effect_median) indicating the strength of the phylogenetic signal on the melanoma prediction per sample (see STAR Methods), the cancer type, the study, the country, and the continent are annotated in the outer rings.

cancer risk, although these associations are rather inconsistent among cohorts.[56,57]

In summary, these results indicate that this *C. aerofaciens* clade is common in two different types of cancer, melanoma and prostate cancer, across several different studies and countries. Based on the data investigated in this study, the presence of this clade does not appear to be related to clinical outcomes in melanoma or to testosterone production in prostate cancer; however, further functional analyses are necessary to better determine the role of this *C. aerofaciens* clade in disease development and progression.

DISCUSSION

In this study, we investigated the genetic variability of the human gut microbiome at the strain level on a global scale. Our analysis involved reconstructing 1,660 phylogenetic trees, each representing the inter-individual diversity of the dominant microbial strains within a species derived from a large dataset of human gut metagenomes sourced from 42 different countries. This extensive dataset enabled us to explore intraspecies genetic diversity of microbes in relation to a wide range of human phenotypes and diseases. Although identifying single mutations associated with phenotypes can be daunting due to the inherent linkage of genetic variants in microbial genomes and their overall number, we associated the phylogenetic relationships between dominant strains in different metagenomes directly with human phenotypes, as it has also been attempted in previous studies.[29,30]

Our results support geographical variation of microbial strains, showing that the phylogenetic distances of most conspecific strains are correlated with host geographic distance. This trend could be caused by isolation-by-distance and subsequent genetic drift,[58,59] or by environmental selection for particular strains. Previous studies observed that, in some bacterial species, "non-western" populations tend to have more similar strains, which might support the lifestyle selection hypothesis.[33] Interestingly, we often observed gut microbial clades from European and North American samples to cluster together. Although

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

this might represent lifestyle-related selection, it could also reflect past human migration events, as has been discussed previously.[16] Moreover, it has also been observed that multiple bacterial species' phylogenies resemble their human host populations' phylogenies, highlighting that geographical stratification in those species is unlikely to be driven by lifestyle alone.[15] The different degrees of phylogeny-geography correlation that we observe across various microbial species, and their relationship with species transmissibility, might point to the presence of both environmental selection and isolation-by-distance but with different relative effects depending on the microbial species.

Subspecies microbial variation was linked not only to geography but also to a wide range of host characteristics. The associations we focused on seemed not to be confounded by major geographic factors, such as continent and country, although there could still be factors unaccounted for such as social networks.[60] Overall, most phylogeny-phenotype associations seem to be driven by small clades rather than major ones. This agrees with observations from a previous study focused on the phylogenetic relationship of gut microbes in IBD and healthy controls.[30] Although Kumbhari et al. describe a larger number of associations with a similar IBD sample size (20% of their associations), these associations are generally in small clades, and their top StrainPhlAn association agrees with one of the top associations described in our study, *Fusicatenibacter saccharivorans*.

Although the majority of associations between subspecies' microbial variability and host anthropometrics were related to host age (Figures 3 and 4), BMI was also linked with subspecies diversity. Interestingly, in European samples, we identified associations between two *F. prausnitzii* SGBs and BMI. In a recent association study between bacterial single-nucleotide variants (SNVs) and human BMI, the authors observed that most of the SNVs that could be replicated between Israeli and Dutch populations (15/17 SNV) belonged to *Faecalibacterium* and *F. prausnitzii*.[23] In the same vein, the authors report high inflation in the quantile-quantile (Q-Q) plots obtained from *F. prausnitzii* associations, suggesting tight linkage of significant genetic variations associated with BMI. These results may suggest that such an effect is representative of a clade-effect, as captured by our associations. Another possibility is that the identified SNVs are widespread across the phylogeny in the different BMI-associated clades. To discriminate between these two possibilities, microbial strain information can be controlled for during SNV association. Overall, both studies' results suggest a relationship between *F. prausnitzii* genetic makeup and host BMI.

Several gut microbial clades linked with specific host phenotypes could be replicated across studies and/or geography. A noteworthy example is the association between melanoma and the SGB14546_group, which corresponds to a set of closely related SGBs from the *Collinsella* genus. This association was detected in gut metagenomic samples from both North American and European individuals, including patients with stage III or IV melanoma (metastatic). In a further step to explore links between disease progression and gut microbial phylogeny, we utilized host information on previous treatment for melanoma but found sample clustering in such a clade to be unrelated, at least considering the smaller sample sizes available with this type of information. Nonetheless, we identified that participants with metastatic prostate cancer also cluster in the same cancer-associated clade as those with melanoma.[54] Although we did not identify a clear difference in gene content that might explain the clustering (with the exception of vitamin B12 biosynthesis), we hypothesize that the strains within the clade might be adapted to some common exposure between the diseases, such as immune dysregulation in cancer, which might drive the selection of specific bacterial subspecies. To the best of our knowledge, abundance levels of this species have not been previously identified as disease biomarkers for these cancers nor have they been associated with treatment response.

Our work explores the links between human phenotypic variation and intraspecies genetic variability in gut microbes and, in doing so, underscores the phylogenetic complexity of the gut microbiome at the strain level. We show that phylogenetic associations are generally related to small microbial clades, rather than larger subspecies entities being associated with a human phenotype or disease. We encourage further studies to focus on even larger sample sizes to increase the statistical power needed to identify these small clades associated with phenotypes of interest. To further boost statistical power, we recommend including publicly available datasets that represent similar populations, which are unlikely to contain geographical stratification and wherein similar clades might drive the association. Exploring and understanding the common metabolic capabilities of disease-associated clades will be fundamental to fostering our understanding of the relationship between gut microbes, health, and disease.

Limitations of the study

Because this study is associative and observational by design, it does not try to infer causal relationships between gut microbial subspecies and human health. At best, it points out subspecies-level clades that could be the target of mechanistic, *in vitro*, or *in vivo* experimental validations. In exploring the trade-off between accuracy and profiling intra-sample and inter-sample comprehensiveness, our analysis focuses on the most abundant strain for each SGB in each sample and cannot profile strains below a certain coverage level, which depends both on species abundance and total per-sample sequencing depth. In addition, most phenotypes were absent from most cohorts, leading to an overrepresentation of associations found in Dutch cohorts. Large sample sizes are required for identification of significant associations, which hampers the identification of continent-stratified signals. Exploration of isolation-by-distance in this dataset is inherently very noisy due to the presence of multiple confounding factors, such as traveling, lifestyle, population mixture, or practical distances that are different from map distances (e.g., islands), among others.

Some of the associations we identified, such as the clade of *R. gnavus* found in older individuals, might be driven by other unknown processes for which we do not account or which we do not measure. Aging is a complex process that is interrelated with metabolic and immunological changes and the appearance of disease, and it might be confounded by behavioral and environmental differences. In the absence of a more comprehensive study, we cannot draw a conclusion about which process

underlies the acquisition of these strains. Further work should focus on more deeply phenotyped trans-ethnic cohorts to answer questions related to the role of lifestyle and geography in subspecies microbial variability of the gut microbiome.

RESOURCE AVAILABILITY

Lead contact

For correspondence and material requests, please contact Nicola Segata (nicola.segata@unitn.it).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Datasets included in this study are publicly available. All datasets, with the exception of GacesaR_2022,[3] ZhernakovaA_2016,[2,61] KurilshikovA_2019,[62] and StražarM_2021,[63] have been collected previously and are curated in CuratedMetagenomicData. Accessions collected in that study are included in Table S1, 1. Phenotypes are available as part of Curated MetagenomicData.

Datasets not included in CuratedMetagenomicData are as follows:

- VichVilaA_2018: raw metagenomics, host genomics, shotgun sequencing, and phenotypic data used in this study are available from the EGA data repository: 1,000 inflammatory bowel disease (IBD) cohort (EGA: EGAD00001004194).
- KurilshikovA_2019: metagenomics data can be requested from EGA (EGA: EGAS00001003508). Phenotype data can be requested via: http://www.humanfunctionalgenomics.org.
- StražarM_2021: metagenomics data can be accessed from the European Nucleotide Archive (ENA: PRJNA686265). Phenotype data are included in the original publication.
- Lifelines datasets: metagenomics data can be requested from the European Genome-Phenome Archive (EGA) for GacesaR_2022 (EGA: EGAS00001005027) and ZhernakovaA_2016 (EGA: EGAD00001001991). To protect participants' privacy and respect the research agreements in the informed consent, participant metadata are not publicly available and cannot be deposited in public repositories. The Lifelines data can be accessed by all bona fide researchers with a scientific proposal by contacting the Lifelines Biobank (instructions at https://www.lifelines.nl/researcher/how-to-apply). Researchers will need to fill in an application form, which will be reviewed within 2 working weeks. If the proposed research complies with Lifelines regulations (for example, noncommercial use and guarantee of participants' privacy), researchers will then receive a financial offer and a data and materials transfer agreement to sign. In general, data will be released within 2 weeks after signing the offer and data and materials transfer agreement. The data will be released in a remote system (the Lifelines workspace) running on a high-performance computer cluster to ensure data quality and security. As Lifelines is a non-profit organization dependent on (governmental) subsidies, a fee is required to cover the costs of controlled data access and supporting infrastructure. The fee for data access on the high-performance computer is €3,500 for 1 year, and the fee for the Lifelines workspace environment is €4,500 for 1 year—or less for shorter periods of time. There are no restrictions on the downstream re-use of aggregated, non-identifiable results (as approved by Lifelines) nor are there authorship requirements, but Lifelines does request that it is acknowledged in publications using these data. The data access policy, data access fees, and an example data and materials transfer agreement (which includes details on how to acknowledge the use of Lifelines data in publications) are described in detail at https://www.lifelines.nl/researcher/how-to-apply. Note that data access for replication can be arranged through Lifelines. Lifelines will not charge an access fee for controlled access to the full dataset used in the manuscript (including phenotype and sequencing data), for the specific purpose of replication of the results presented in this article or for further assessment by the reviewers, for a period of 3 months. Researchers interested in such a replication study or review assessment can contact Lifelines at research@lifelines.nl.

Phylogenetic trees generated in this work, and analysis code, are available in the Zenodo repository: https://doi.org/10.5281/zenodo.14651412.

ACKNOWLEDGMENTS

We would like to acknowledge Kate McIntyre for manuscript editing.

This research was supported by a European Molecular Biology Organization Scientific Exchange Grant (#10263) to S.A.-S. J.F. was supported by a Netherlands Heart Foundation grant (IN-CONTROL CVON 2012-03); a Dutch Research council (NWO) VICI grant (VI.C.202.022); a European Research Council (ERC) Consolidator Grant (grant agreement no. 101001678); an AMMODO Science Award 2023 for Biomedical Sciences from Stichting Ammodo; and the Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the Government of the Netherlands. D.W. was supported by NWO VENI grant 222.016. N.S. was supported by the ERC (ERC-STG project MetaPG-716575 and ERC-CoG microTOUCH-101045015), the European Union's Horizon 2020 programme (ONCOBIOME-825410 project and IHMCSA-964590), the European Union NextGenerationEU (INEST), the National Cancer Institute of the National Institutes of Health (1U01CA230551), and the Premio Internazionale Lombardia e Ricerca 2019. A.Z. and A.K. were supported by NWO Gravitation grant Exposome-NL 024.004.017. A.Z. was also supported by NWO VICI grant VI.C.232.074 and EU Horizon Europe Program grant INITIALISE (101094099). G. F. was funded by the European Union under the Marie Sklodowska-Curie grant agreement no. 101152592–plasticOME.

In addition, we wish to thank all participants and co-authors from all the different studies used here. The Lifelines Biobank initiative has been made possible by a subsidy from the Dutch Ministry of Health, Welfare and Sport; the Dutch Ministry of Economic Affairs; the University Medical Centre Groningen (UMCG, the Netherlands); the University of Groningen; and the Northern Provinces of the Netherlands. The authors wish to acknowledge the services of the Lifelines Cohort Study, the contributing research centers delivering data to Lifelines, and all the study participants.

AUTHOR CONTRIBUTIONS

S.A.-S.: conception of the study, analysis, and writing; N.S.: conception of the study and supervision; J.F.: conception of the study and supervision; R.K.W. and A.Z.: discussion and cohort construction; A.B.-M.: data generation and discussion; D.V.Z.: analysis; D.W.: phenotype curation, analysis, and discussion; A.K.: discussion; M.V.-C.: discussion; P.M.: phenotype curation, data generation, and discussion; D.G.: data generation and discussion; G.F.: revision and discussion; V.H.: revision and discussion.

DECLARATION OF INTERESTS

R.K.W. acted as consultant for Takeda Pharmaceuticals; received unrestricted research grants from Takeda, Johnson & Johnson, Tramedico, and Ferring; and received speaker's fees from MSD, Abbvie, and Janssen Pharmaceuticals.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Microbial phylogenetic tree reconstruction
  - Analysis of within-sample polymorphisms
  - Comparison of dominant strains of StrainPhlAn with GT-Pro
  - Food references in phylogenetic tree and *B. animalis* analysis

- Geographic association with microbial phylogeny
- Species transmissibility and geographic effect
- Microbial characteristics and geographic effect
- Phylogenetic association
- Downsampling analysis
- Testing the influence of sequencing protocol
- Enrichment of taxonomic levels
- Clade-specific analyses
- Functional analysis of MAGs

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.04.014.

Received: August 6, 2024
Revised: January 23, 2025
Accepted: April 7, 2025
Published: April 30, 2025

REFERENCES

1. Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M.J., Valles-Colomer, M., Vandeputte, D., et al. (2016). Population-level analysis of gut microbiome variation. Science *352*, 560–564. https://doi.org/10.1126/science.aad3503.
2. Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science *352*, 565–569. https://doi.org/10.1126/science.aad3369.
3. Gacesa, R., Kurilshikov, A., Vich Vila, A., Sinha, T., Klaassen, M.A.Y., Bolte, L.A., Andreu-Sánchez, S., Chen, L., Collij, V., Hu, S., et al. (2022). Environmental factors shaping the gut microbiome in a Dutch population. Nature *604*, 732–739. https://doi.org/10.1038/s41586-022-04567-7.
4. Integrative HMP (iHMP) Research Network Consortium (2019). The Integrative Human Microbiome Project. Nature *569*, 641–648. https://doi.org/10.1038/s41586-019-1238-8.
5. Sprockett, D., Fukami, T., and Relman, D.A. (2018). Role of priority effects in the early-life assembly of the gut microbiota. Nat. Rev. Gastroenterol. Hepatol. *15*, 197–205. https://doi.org/10.1038/nrgastro.2017.173.
6. Shao, Y., Forster, S.C., Tsaliki, E., Vervier, K., Strang, A., Simpson, N., Kumar, N., Stares, M.D., Rodger, A., Brocklehurst, P., et al. (2019). Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature *574*, 117–121. https://doi.org/10.1038/s41586-019-1560-1.
7. Masenga, S.K., Hamooya, B., Hangoma, J., Hayumbu, V., Ertuglu, L.A., Ishimwe, J., Rahman, S., Saleem, M., Laffer, C.L., Elijovich, F., et al. (2022). Recent advances in modulation of cardiovascular diseases by the gut microbiota. J. Hum. Hypertens. *36*, 952–959. https://doi.org/10.1038/s41371-022-00698-6.
8. Fan, Y., and Pedersen, O. (2021). Gut microbiota in human metabolic health and disease. Nat. Rev. Microbiol. *19*, 55–71. https://doi.org/10.1038/s41579-020-0433-9.
9. Blanco Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., Nickols, W.A., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn. Nat. Biotechnol. *41*, 1633–1644.
10. Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol. *39*, 727–736. https://doi.org/10.1038/s41587-020-00797-0.
11. Van Rossum, T., Costea, P.I., Paoli, L., Alves, R., Thielemann, R., Sunagawa, S., and Bork, P. (2022). metaSNV v2: detection of SNVs and subspecies in prokaryotic metagenomes. Bioinformatics *38*, 1162–1164. https://doi.org/10.1093/BIOINFORMATICS/BTAB789.
12. Shi, Z.J., Dimitrov, B., Zhao, C., Nayfach, S., and Pollard, K.S. (2022). Fast and accurate metagenotyping of the human gut microbiome with GT-Pro. Nat. Biotechnol. *40*, 507–516. https://doi.org/10.1038/s41587-021-01102-3.
13. Yan, Y., Nguyen, L.H., Franzosa, E.A., and Huttenhower, C. (2020). Strain-level epidemiology of microbial communities and the human microbiome. Genome Med. *12*, 71. https://doi.org/10.1186/s13073-020-00765-y.
14. Valles-Colomer, M., Blanco, M., Manghi, P., Asnicar, F., Dubois, L., Golzato, D., Armanini, F., Cumbo, F., Huang, K.D., Manara, S., et al. (2023). The person-to-person transmission landscape of the gut and oral microbiomes. Nature *614*, 125–135. https://doi.org/10.1038/s41586-022-05620-1.
15. Suzuki, T.A., Fitzstevens, J.L., Schmidt, V.T., Enav, H., Huus, K.E., Mbong Ngwese, M., Grießhammer, A., Pfleiderer, A., Adegbite Bayode, R., Zinsou, J.F., et al. (2022). Codiversification of gut microbiota with humans. Science *377*, 1328–1332. https://doi.org/10.1126/science.abm7759.
16. Karcher, N., Pasolli, E., Asnicar, F., Huang, K.D., Tett, A., Manara, S., Armanini, F., Bain, D., Duncan, S.H., Louis, P., et al. (2020). Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biol. *21*, 138. https://doi.org/10.1186/s13059-020-02042-y.
17. Wang, D., Doestzada, M., Chen, L., Andreu-Sánchez, S., van den Munckhof, I.C.L., Augustijn, H.E., Koehorst, M., Ruiz-Moreno, A.J., Bloks, V.W., Riksen, N.P., et al. (2021). Characterization of gut microbial structural variations as determinants of human bile acid metabolism. Cell Host Microbe *29*, 1802–1814.e5. https://doi.org/10.1016/j.chom.2021.11.003.
18. Vatanen, T., Plichta, D.R., Somani, J., Münch, P.C., Arthur, T.D., Hall, A.B., Rudolf, S., Oakeley, E.J., Ke, X., Young, R.A., et al. (2019). Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nat. Microbiol. *4*, 470–479. https://doi.org/10.1038/s41564-018-0321-5.
19. Costea, P.I., Coelho, L.P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., Hildebrand, F., Kushugulova, A., Zeller, G., and Bork, P. (2017). Subspecies in the global human gut microbiome. Mol. Syst. Biol. *13*, 960. https://doi.org/10.15252/msb.20177589.
20. Zhernakova, D.V., Wang, D., Liu, L., Andreu-Sánchez, S., Zhang, Y., Ruiz-Moreno, A.J., Peng, H., Plomp, N., Del Castillo-Izquierdo, Á., Gacesa, R., et al. (2024). Host genetic regulation of human gut microbial structural variation. Nature *625*, 813–821. https://doi.org/10.1038/s41586-023-06893-w.
21. Vatanen, T., Ang, Q.Y., Siegwald, L., Sarker, S.A., Le Roy, C.I., Duboux, S., Delannoy-Bruno, O., Ngom-Bru, C., Boulangé, C.L., Stražar, M., et al. (2022). A distinct clade of Bifidobacterium longum in the gut of Bangladeshi children thrives during weaning. Cell *185*, 4280–4297.e12. https://doi.org/10.1016/j.cell.2022.10.011.
22. Shridhar, S.V., Beghini, F., Alexander, M., Singh, A., Juárez, R.M., Brito, I.L., and Christakis, N.A. (2024). Environmental, socioeconomic, and health factors associated with gut microbiome species and strains in isolated Honduras villages. Cell Rep. *43*, 114442. https://doi.org/10.1016/J.CELREP.2024.114442.
23. Zahavi, L., Lavon, A., Reicher, L., Shoer, S., Godneva, A., Leviatan, S., Rein, M., Weissbrod, O., Weinberger, A., and Segal, E. (2023). Bacterial SNPs in the human gut microbiome associate with host BMI. Nat. Med. *29*, 2785–2792. https://doi.org/10.1038/s41591-023-02599-8.

24. Hall, A.B., Yassour, M., Sauk, J., Garner, A., Jiang, X., Arthur, T., Lagoudas, G.K., Vatanen, T., Fornelos, N., Wilson, R., et al. (2017). A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. Genome Med. *9*, 103. https://doi.org/10.1186/s13073-017-0490-5.

25. Martinez-Medina, M., and Garcia-Gil, L.J. (2014). Escherichia coli in chronic inflammatory bowel diseases: An update on adherent invasive Escherichia coli pathogenicity. World J. Gastrointest. Pathophysiol. *5*, 213–227. https://doi.org/10.4291/wjgp.v5.i3.213.

26. Liu, R., Zou, Y., Wang, W.-Q., Chen, J.-H., Zhang, L., Feng, J., Yin, J.-Y., Mao, X.-Y., Li, Q., Luo, Z.-Y., et al. (2023). Gut microbial structural variation associates with immune checkpoint inhibitor response. Nat. Commun. *14*, 7421. https://doi.org/10.1038/s41467-023-42997-7.

27. Gunjur, A., Shao, Y., Rozday, T., Klein, O., Mu, A., Haak, B.W., Markman, B., Kee, D., Carlino, M.S., Underhill, C., et al. (2024). A gut microbial signature for combination immune checkpoint blockade across cancer types. Nat. Med. *30*, 797–809. https://doi.org/10.1038/s41591-024-02823-z.

28. Geva-Zatorsky, N., Sefik, E., Kua, L., Pasman, L., Tan, T.G., Ortiz-Lopez, A., Yanortsang, T.B., Yang, L., Jupp, R., Mathis, D., et al. (2017). Mining the Human Gut Microbiota for Immunomodulatory Organisms. Cell *168*, 928–943.e11. https://doi.org/10.1016/j.cell.2017.01.022.

29. Mei, Z., Wang, F., Bhosle, A., Dong, D., Mehta, R., Ghazi, A., Zhang, Y., Liu, Y., Rinott, E., Ma, S., et al. (2024). Strain-specific gut microbial signatures in type 2 diabetes identified in a cross-cohort analysis of 8,117 metagenomes. Nat. Med. *30*, 2265–2276. https://doi.org/10.1038/s41591-024-03067-7.

30. Kumbhari, A., Cheng, T.N.H., Ananthakrishnan, A.N., Kochar, B., Burke, K.E., Shannon, K., Lau, H., Xavier, R.J., and Smillie, C.S. (2024). Discovery of disease-adapted bacterial lineages in inflammatory bowel diseases. Cell Host Microbe *32*, 1147–1162.e12. https://doi.org/10.1016/j.chom.2024.05.022.

31. Piccinno, G., Thompson, K.N., Manghi, P., Ghazi, A.R., Thomas, A.M., Blanco-Míguez, A., Asnicar, F., Mladenovic, K., Pinto, F., Armanini, F., et al. (2025). The gut microbiome in colorectal cancer: a cross-stage strain-level pooled analysis of 3,741 individuals from 18 cohorts. Nat. Med. https://doi.org/10.1038/s41591-025-03693-9.

32. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

33. Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. *27*, 626–638. https://doi.org/10.1101/gr.216242.116.

34. Chen-Liaw, A., Aggarwala, V., Mogno, I., Haifer, C., Li, Z., Eggers, J., Helmus, D., Hart, A., Wehkamp, J., Lamousé-Smith, E.S.N., et al. (2025). Gut microbiota strain richness is species specific and affects engraftment. Nature *637*, 422–429. https://doi.org/10.1038/s41586-024-08242-x.

35. Fernández-Pato, A., Sinha, T., Gacesa, R., Andreu-Sánchez, S., Gois, M. F.B., Gelderloos-Arends, J., Jansen, D.B.H., Kruk, M., Jaeger, M., Joosten, L.A.B., et al. (2024). Choice of DNA extraction method affects stool microbiome recovery and subsequent phenotypic association analyses. Sci. Rep. *14*, 3911. https://doi.org/10.1038/s41598-024-54353-w.

36. Weimann, A., Mooren, K., Frank, J., Pope, P.B., Bremges, A., and McHardy, A.C. (2016). From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. mSystems *1*, e00101-16. https://doi.org/10.1128/mSystems.00101-16.

37. Ghazi, A.R., Thompson, K.N., Bhosle, A., Mei, Z., Yan, Y., Wang, F., Wang, K., Franzosa, E.A., and Huttenhower, C. (2025). Quantifying metagenomic strain associations from microbiomes with anpan. Preprint at bioRxiv. https://doi.org/10.1101/2025.01.06.631550.

38. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. *9*, 5114. https://doi.org/10.1038/s41467-018-07641-9.

39. Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., and Yin, Y. (2023). dbCAN3: automated carbohydrate-active enzyme and substrate annotation. Nucleic Acids Res. *51*, W115–W121. https://doi.org/10.1093/nar/gkad328.

40. Underwood, M.A., German, J.B., Lebrilla, C.B., and Mills, D.A. (2015). Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. Pediatr. Res. *77*, 229–235. https://doi.org/10.1038/pr.2014.156.

41. Xu, Q., Wu, C., Zhu, Q., Gao, R., Lu, J., Valles-Colomer, M., Zhu, J., Yin, F., Huang, L., Ding, L., et al. (2022). Metagenomic and metabolomic remodeling in nonagenarians and centenarians and its association with genetic and socioeconomic factors. Nat. Aging *2*, 438–452. https://doi.org/10.1038/s43587-022-00193-0.

42. Crost, E.H., Coletto, E., Bell, A., and Juge, N. (2023). Ruminococcus gnavus: friend or foe for human health. FEMS Microbiol. Rev. *47*, fuad014. https://doi.org/10.1093/femsre/fuad014.

43. Wang, N., Li, R., Lin, H., Fu, C., Wang, X., Zhang, Y., Su, M., Huang, P., Qian, J., Jiang, F., et al. (2019). Enriched taxa were found among the gut microbiota of centenarians in East China. PLoS One *14*, e0222763. https://doi.org/10.1371/journal.pone.0222763.

44. Devlin, A.S., and Fischbach, M.A. (2015). A biosynthetic pathway for a prominent class of microbiota-derived bile acids. Nat. Chem. Biol. *11*, 685–690. https://doi.org/10.1038/nchembio.1864.

45. Sato, Y., Atarashi, K., Plichta, D.R., Arai, Y., Sasajima, S., Kearney, S.M., Suda, W., Takeshita, K., Sasaki, T., Okamoto, S., et al. (2021). Novel bile acid biosynthetic pathways are enriched in the microbiome of centenarians. Nature *599*, 458–464. https://doi.org/10.1038/s41586-021-03832-5.

46. Chen, L., van den Munckhof, I.C.L., Schraa, K., Ter Horst, R., Koehorst, M., van Faassen, M., van der Ley, C., Doestzada, M., Zhernakova, D.V., Kurilshikov, A., et al. (2020). Genetic and Microbial Associations to Plasma and Fecal Bile Acids in Obesity Relate to Plasma Lipids and Liver Fat Content. Cell Rep. *33*, 108212. https://doi.org/10.1016/j.celrep.2020.108212.

47. Zepeda-Rivera, M., Minot, S.S., Bouzek, H., Wu, H., Blanco-Míguez, A., Manghi, P., Jones, D.S., LaCourse, K.D., Wu, Y., McMahon, E.F., et al. (2024). A distinct Fusobacterium nucleatum clade dominates the colorectal cancer niche. Nature *628*, 424–432. https://doi.org/10.1038/s41586-024-07182-w.

48. Borozan, I., Zaidi, S.H., Harrison, T.A., Phipps, A.I., Zheng, J., Lee, S., Trinh, Q.M., Steinfelder, R.S., Adams, J., Banbury, B.L., et al. (2022). Molecular and Pathology Features of Colorectal Tumors and Patient Outcomes Are Associated with Fusobacterium nucleatum and Its Subspecies animalis. Cancer Epidemiol. Biomarkers Prev. *31*, 210–220. https://doi.org/10.1158/1055-9965.EPI-21-0463.

49. Lee, K.A., Thomas, A.M., Bolte, L.A., Björk, J.R., de Ruijter, L.K., Armanini, F., Asnicar, F., Blanco-Miguez, A., Board, R., Calbet-Llopart, N., et al. (2022). Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. Nat. Med. *28*, 535–544. https://doi.org/10.1038/s41591-022-01695-5.

50. McCulloch, J.A., Davar, D., Rodrigues, R.R., Badger, J.H., Fang, J.R., Cole, A.M., Balaji, A.K., Vetizou, M., Prescott, S.M., Fernandes, M.R., et al. (2022). Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1. Nat. Med. *28*, 545–556. https://doi.org/10.1038/s41591-022-01698-2.

51. Wind, T.T., Gacesa, R., Vich Vila, A., de Haan, J.J., Jalving, M., Weersma, R.K., and Hospers, G.A.P. (2020). Gut microbial species and metabolic pathways associated with response to treatment with immune checkpoint inhibitors in metastatic melanoma. Melanoma Res. *30*, 235–246. https://doi.org/10.1097/CMR.0000000000000656.

52. Frankel, A.E., Coughlin, L.A., Kim, J., Froehlich, T.W., Xie, Y., Frenkel, E. P., and Koh, A.Y. (2017). Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients. Neoplasia *19*, 848–855. https://doi.org/10.1016/j.neo.2017.08.004.

53. Peters, B.A., Wilson, M., Moran, U., Pavlick, A., Izsak, A., Wechter, T., Weber, J.S., Osman, I., and Ahn, J. (2019). Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. Genome Med. *11*, 61. https://doi.org/10.1186/s13073-019-0672-4.

54. Pernigoni, N., Zagato, E., Calcinotto, A., Troiani, M., Mestre, R.P., Calì, B., Attanasio, G., Troisi, J., Minini, M., Mosole, S., et al. (2021). Commensal bacteria promote endocrine resistance in prostate cancer through androgen biosynthesis. Science *374*, 216–224. https://doi.org/10.1126/science.abf8403.

55. Sánchez-Maldonado, J.M., Cáliz, R., Canet, L., Ter Horst, R., Bakker, O., den Broeder, A.A., Martínez-Bueno, M., Canhão, H., Rodríguez-Ramos, A., Lupiañez, C.B., Soto-Pino, M.J., et al. (2019). Steroid hormone-related polymorphisms associate with the development of bone erosions in rheumatoid arthritis and help to predict disease progression: Results from the REPAIR consortium. Sci. Rep. *9*, 14812.

56. Essén, A., Santaolalla, A., Garmo, H., Hammar, N., Walldius, G., Jungner, I., Malmström, H., Holmberg, L., and Van Hemelrijck, M. (2019). Baseline serum folate, vitamin B12 and the risk of prostate and breast cancer using data from the Swedish AMORIS cohort. Cancer Causes Control *30*, 603–615. https://doi.org/10.1007/s10552-019-01170-6.

57. Obeid, R. (2022). High Plasma Vitamin B12 and Cancer in Human Studies: A Scoping Review to Judge Causality and Alternative Explanations. Nutrients *14*, 4476. https://doi.org/10.3390/nu14214476.

58. Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., Horner-Devine, M.C., Kane, M., Krumins, J.A., Kuske, C.R., et al. (2006). Microbial biogeography: putting microorganisms on the map. Nat. Rev. Microbiol. *4*, 102–112. https://doi.org/10.1038/nrmicro1341.

59. Moeller, A.H., Suzuki, T.A., Lin, D., Lacey, E.A., Wasser, S.K., and Nachman, M.W. (2017). Dispersal limitation promotes the diversification of the mammalian gut microbiota. Proc. Natl. Acad. Sci. USA *114*, 13768–13773. https://doi.org/10.1073/pnas.1700122114.

60. Beghini, F., Pullman, J., Alexander, M., Shridhar, S.V., Prinster, D., Singh, A., Matute Juárez, R., Airoldi, E.M., Brito, I.L., and Christakis, N.A. (2025). Gut microbiome strain-sharing within isolated village social networks. Nature *637*, 167–175. https://doi.org/10.1038/s41586-024-08222-1.

61. Chen, L., Wang, D., Garmaeva, S., Kurilshikov, A., Vich, V.A., Ranko, G., Sinha, T., Lifelines Cohort Study, Eran, S., Weersma, R.K., et al. (2021). The long-term genetic stability and individual specificity of the human gut microbiome. Cell *184*, 2302–2315.e12.

62. Kurilshikov, A., Van Den Munckhof, I.C.L., Chen, L., Bonder, M.J., Schraa, K., Rutten, J.H.W., Riksen, N.P., De Graaf, J., Oosting, M., Sanna, S., et al. (2019). Gut Microbial Associations to Plasma Metabolites Linked to Cardiovascular Phenotypes and Risk. Circ. Res. *124*, 1808–1820. https://doi.org/10.1161/CIRCRESAHA.118.314642.

63. Stražar, M., Temba, G.S., Vlamakis, H., Kullaya, V.I., Lyamuya, F., Mmbaga, B.T., Joosten, L.A.B., van der Ven, A.J.A.M., Netea, M.G., de Mast, Q., et al. (2021). Author Correction: Gut microbiome-mediated metabolism effects on immunity in rural and urban African populations. Nat. Commun. *12*, 5818. https://doi.org/10.1038/s41467-021-26145-7.

64. Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat. Commun. *11*, 2500. https://doi.org/10.1038/s41467-020-16366-7.

65. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M. T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

66. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

67. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

68. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

69. Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J.B., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. Nat. Methods *14*, 1023–1024. https://doi.org/10.1038/nmeth.4468.

70. Ianiro, G., Punčochář, M., Karcher, N., Porcari, S., Armanini, F., Asnicar, F., Beghini, F., Blanco-Míguez, A., Cumbo, F., Manghi, P., et al. (2022). Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases. Nat. Med. *28*, 1913–1923. https://doi.org/10.1038/s41591-022-01964-3.

71. Asnicar, F., Manara, S., Zolfo, M., Truong, D.T., Scholz, M., Armanini, F., Ferretti, P., Gorfer, V., Pedrotti, A., Tett, A., et al. (2017). Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. mSystems *2*, e00164-16. https://doi.org/10.1128/mSystems.00164-16.

72. Asnicar, F., Berry, S.E., Valdes, A.M., Nguyen, L.H., Piccinno, G., Drew, D.A., Leeming, E., Gibson, R., Le Roy, C., Khatib, H.A., et al. (2021). Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat. Med. *27*, 321–332. https://doi.org/10.1038/s41591-020-01183-8.

73. Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe *17*, 690–703. https://doi.org/10.1016/j.chom.2015.04.004.

74. Bedarf, J.R., Hildebrand, F., Coelho, L.P., Sunagawa, S., Bahram, M., Goeser, F., Bork, P., and Wüllner, U. (2017). Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. Genome Med. *9*, 39. https://doi.org/10.1186/s13073-017-0428-y.

75. Bengtsson-Palme, J., Angelin, M., Huss, M., Kjellqvist, S., Kristiansson, E., Palmgren, H., Larsson, D.G.J., and Johansson, A. (2015). The Human Gut Microbiome as a Transporter of Antibiotic Resistance Genes between Continents. Antimicrob. Agents Chemother. *59*, 6551–6560. https://doi.org/10.1128/AAC.00933-15.

76. Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W., Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. Nature *535*, 435–439. https://doi.org/10.1038/nature18927.

77. Brooks, B., Olm, M.R., Firek, B.A., Baker, R., Thomas, B.C., Morowitz, M. J., and Banfield, J.F. (2017). Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nat. Commun. *8*, 1814. https://doi.org/10.1038/s41467-017-02018-w.

78. Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., He, Z., Zhong, W., Fan, Y., Zhang, L., et al. (2017). Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. Genome Biol. *18*, 142. https://doi.org/10.1186/s13059-017-1271-6.

79. Chu, D.M., Ma, J., Prince, A.L., Antony, K.M., Seferovic, M.D., and Aagaard, K.M. (2017). Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. Nat. Med. *23*, 314–326. https://doi.org/10.1038/nm.4272.

80. David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J. E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559–563. https://doi.org/10.1038/nature12820.

81. De Filippis, F., Pasolli, E., Tett, A., Tarallo, S., Naccarati, A., De Angelis, M., Neviani, E., Cocolin, L., Gobbetti, M., Segata, N., et al. (2019). Distinct Genetic and Functional Traits of Human Intestinal Prevotella copri Strains Are Associated with Different Habitual Diets. Cell Host Microbe *25*, 444–453.e3. https://doi.org/10.1016/j.chom.2019.01.004.

82. Dhakan, D.B., Maji, A., Sharma, A.K., Saxena, R., Pulikkan, J., Grace, T., Gomez, A., Scaria, J., Amato, K.R., and Sharma, V.K. (2019). The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigaScience *8*, giz004. https://doi.org/10.1093/gigascience/giz004.

83. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat. Commun. *6*, 6528. https://doi.org/10.1038/ncomms7528.

84. Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe *24*, 133–145.e5. https://doi.org/10.1016/j.chom.2018.06.005.

85. Gopalakrishnan, V., Spencer, C.N., Nezi, L., Reuben, A., Andrews, M.C., Karpinets, T.V., Prieto, P.A., Vicente, D., Hoffman, K., Wei, S.C., et al. (2018). Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. Science *359*, 97–103. https://doi.org/10.1126/science.aan4236.

86. Gupta, A., Dhakan, D.B., Maji, A., Saxena, R., P K, V.P., Mahajan, S., Pulikkan, J., Kurian, J., Gomez, A.M., Scaria, J., et al. (2019). Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. mSystems *4*, e00438-19. https://doi.org/10.1128/mSystems.00438-19.

87. Hannigan, G.D., Duhaime, M.B., Ruffin, M.T., 4th, Koumpouras, C.C., and Schloss, P.D. (2018). Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. mBio *9*, e02248-18. https://doi.org/10.1128/mBio.02248-18.

88. Hansen, L.B.S., Roager, H.M., Søndertoft, N.B., Gøbel, R.J., Kristensen, M., Vallès-Colomer, M., Vieira-Silva, S., Ibrügger, S., Lind, M.V., Mærkedahl, R.B., et al. (2018). A low-gluten diet induces changes in the intestinal microbiome of healthy Danish adults. Nat. Commun. *9*, 4630. https://doi.org/10.1038/s41467-018-07019-x.

89. He, Q., Gao, Y., Jie, Z., Yu, X., Laursen, J.M., Xiao, L., Li, Y., Li, L., Zhang, F., Feng, Q., et al. (2017). Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. GigaScience *6*, 1–11.

90. Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. https://doi.org/10.1038/nature11234.

91. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662. https://doi.org/10.1038/s41586-019-1237-9

92. Ijaz, U.Z., Quince, C., Hanske, L., Loman, N., Calus, S.T., Bertz, M., Edwards, C.A., Gaya, D.R., Hansen, R., McGrogan, P., et al. (2017). The distinct features of microbial "dysbiosis" of Crohn's disease do not occur to the same extent in their unaffected, genetically-linked kindred. PLoS One *12*, e0172605. https://doi.org/10.1371/journal.pone.0172605.

93. Kaur, L., Gordon, M., Baines, P.A., Iheozor-Ejiofor, Z., Sinopoulou, V., and Akobeng, A.K. (2020). Probiotics for induction of remission in ulcerative colitis. Cochrane Database Syst. Rev. *3*, CD005573. https://doi.org/10.1002/14651858.CD005573.pub3.

94. Keohane, D.M., Ghosh, T.S., Jeffery, I.B., Molloy, M.G., O'Toole, P.W., and Shanahan, F. (2020). Microbiome and health implications for ethnic minorities after enforced lifestyle changes. Nat. Med. *26*, 1089–1095. https://doi.org/10.1038/s41591-020-0963-8.

95. Kieser, S., Sarker, S.A., Sakwinska, O., Foata, F., Sultana, S., Khan, Z., Islam, S., Porta, N., Combremont, S., Betrisey, B., et al. (2018). Bangladeshi children with acute diarrhoea show faecal microbiomes with increased Streptococcus abundance, irrespective of diarrhoea aetiology. Environ. Microbiol. *20*, 2256–2269. https://doi.org/10.1111/1462-2920.14274.

96. Kostic, A.D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., Peet, A., Tillmann, V., Pöhö, P., Mattila, I., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe *17*, 260–273. https://doi.org/10.1016/j.chom.2015.01.001.

97. Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. Nature *500*, 541–546. https://doi.org/10.1038/nature12506.

98. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. *32*, 834–841. https://doi.org/10.1038/nbt.2942.

99. Li, J., Zhao, F., Wang, Y., Chen, J., Tao, J., Tian, G., Wu, S., Liu, W., Cui, Q., Geng, B., et al. (2017). Gut microbiota dysbiosis contributes to the development of hypertension. Microbiome *5*, 14. https://doi.org/10.1186/s40168-016-0222-x.

100. Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xi, X., Liang, Z., Hou, Q., Zhou, B., et al. (2016). Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. Sci. Rep. *6*, 34826. https://doi.org/10.1038/srep34826.

101. Lokmer, A., Cian, A., Froment, A., Gantois, N., Viscogliosi, E., Chabé, M., and Ségurel, L. (2019). Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. PLoS One *14*, e0211139. https://doi.org/10.1371/journal.pone.0211139.

102. Loman, N.J., Constantinidou, C., Christner, M., Rohde, H., Chan, J.Z.-M., Quick, J., Weir, J.C., Quince, C., Smith, G.P., Betley, J.R., et al. (2013). A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA *309*, 1502–1510. https://doi.org/10.1001/jama.2013.3231.

103. Loomba, R., Seguritan, V., Li, W., Long, T., Klitgord, N., Bhatt, A., Dulai, P.S., Caussy, C., Bettencourt, R., Highlander, S.K., et al. (2017). Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab. *25*, 1054–1062.e5. https://doi.org/10.1016/j.cmet.2017.04.001.

104. Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. *32*, 822–828. https://doi.org/10.1038/nbt.2939.

105. Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. *6*, 6505. https://doi.org/10.1038/ncomms7505.

106. Olm, M.R., Brown, C.T., Brooks, B., Firek, B., Baker, R., Burstein, D., Soenjoyo, K., Thomas, B.C., Morowitz, M., and Banfield, J.F. (2017). Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. Genome Res. *27*, 601–612. https://doi.org/10.1101/gr.213256.116.

107. Pehrsson, E.C., Tsukayama, P., Patel, S., Mejía-Bautista, M., Sosa-Soto, G., Navarrete, K.M., Calderon, M., Cabrera, L., Hoyos-Arango, W., Bertoli, M.T., et al. (2016). Interconnected microbiomes and resistomes in low-income human habitats. Nature *533*, 212–216. https://doi.org/10.1038/nature17672.

108. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60. https://doi.org/10.1038/nature11450.

109. Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. Nature *513*, 59–64. https://doi.org/10.1038/nature13568.

110. Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., and Candela, M. (2015). Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr. Biol. *25*, 1682–1693. https://doi.org/10.1016/j.cub.2015.04.055.

111. Raymond, F., Ouameur, A.A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.-L., Giroux, R., Bérubé, È., et al. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. ISME J. *10*, 707–720. https://doi.org/10.1038/ismej.2015.148.

112. Rosa, B.A., Supali, T., Gankpala, L., Djuardi, Y., Sartono, E., Zhou, Y., Fischer, K., Martin, J., Tyagi, R., Bolay, F.K., et al. (2018). Differential human gut microbiome assemblages during soil-transmitted helminth infections in Indonesia and Liberia. Microbiome *6*, 33. https://doi.org/10.1186/s40168-018-0416-5.

113. Rubel, M.A., Abbas, A., Taylor, L.J., Connell, A., Tanes, C., Bittinger, K., Ndze, V.N., Fonsah, J.Y., Ngwang, E., Essiane, A., et al. (2020). Lifestyle and the presence of helminths is associated with gut microbiome composition in Cameroonians. Genome Biol. *21*, 122. https://doi.org/10.1186/s13059-020-02020-4.

114. Sankaranarayanan, K., Ozga, A.T., Warinner, C., Tito, R.Y., Obregon-Tito, A.J., Xu, J., Gaffney, P.M., Jervis, L.L., Cox, D., Stephens, L., et al. (2015). Gut Microbiome Diversity among Cheyenne and Arapaho Individuals from Western Oklahoma. Curr. Biol. *25*, 3161–3169. https://doi.org/10.1016/j.cub.2015.10.060.

115. Schirmer, M., Smeekens, S.P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E.A., Ter Horst, R.T., Jansen, T., Jacobs, L., Bonder, M.J., et al. (2016). Linking the human gut microbiome to inflammatory cytokine production capacity. Cell *167*, 1897. https://doi.org/10.1016/j.cell.2016.11.046.

116. Smits, S.A., Leach, J., Sonnenburg, E.D., Gonzalez, C.G., Lichtman, J.S., Reid, G., Knight, R., Manjurano, A., Changalucha, J., Elias, J.E., et al. (2017). Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science *357*, 802–806. https://doi.org/10.1126/science.aan4834.

117. Spencer, C.N., McQuade, J.L., Gopalakrishnan, V., McCulloch, J.A., Vetizou, M., Cogdill, A.P., Khan, M.A.W., Zhang, X., White, M.G., Peterson, C.B., et al. (2021). Dietary fiber and probiotics influence the gut microbiome and melanoma immunotherapy response. Science *374*, 1632–1640. https://doi.org/10.1126/science.aaz7015.

118. Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., et al. (2019). The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. Cell Host Microbe *26*, 666–679.e7.

119. Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat. Med. *25*, 667–678. https://doi.org/10.1038/s41591-019-0405-7.

120. Vincent, C., Miller, M.A., Edens, T.J., Mehrotra, S., Dewar, K., and Manges, A.R. (2016). Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and Clostridium difficile colonization or infection. Microbiome *4*, 12. https://doi.org/10.1186/s40168-016-0156-3.

121. Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A.Y., Hercog, R., Goedert, J.J., Shi, J., Bork, P., and Sinha, R. (2016). Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. PLoS One *11*, e0155362. https://doi.org/10.1371/journal.pone.0155362.

122. Wampach, L., Heintz-Buschart, A., Fritz, J.V., Ramiro-Garcia, J., Habier, J., Herold, M., Narayanasamy, S., Kaysen, A., Hogan, A.H., Bindl, L., et al. (2018). Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. Nat. Commun. *9*, 5091.

123. Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat. Med. *25*, 968–976. https://doi.org/10.1038/s41591-019-0458-7.

124. Yassour, M., Vatanen, T., Siljander, H., Hämäläinen, A.-M., Härkönen, T., Ryhänen, S.J., Franzosa, E.A., Vlamakis, H., Huttenhower, C., Gevers, D., et al. (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci. Transl. Med. *8*, 343ra81. https://doi.org/10.1126/scitranslmed.aad0917.

125. Yassour, M., Jason, E., Hogstrom, L.J., Arthur, T.D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S.M., et al. (2018). Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe *24*, 146–154.e4. https://doi.org/10.1016/j.chom.2018.06.007.

126. Ye, Z., Zhang, N., Wu, C., Zhang, X., Wang, Q., Huang, X., Du, L., Cao, Q., Tang, J., Zhou, C., et al. (2018). A metagenomic study of the gut microbiome in Behcet's disease. Microbiome *6*, 135. https://doi.org/10.1186/s40168-018-0520-6.

127. Yu, J., Feng, Q., Wong, S.H., Zhang, D., Liang, Q.Y., Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut *66*, 70–78. https://doi.org/10.1136/gutjnl-2015-309800.

128. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized Nutrition by Prediction of Glycemic Responses. Cell *163*, 1079–1094. https://doi.org/10.1016/j.cell.2015.11.001.

129. Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. *10*, 766. https://doi.org/10.15252/msb.20145645.

130. Zhu, F., Ju, Y., Wang, W., Wang, Q., Guo, R., Ma, Q., Sun, Q., Fan, Y., Xie, Y., Yang, Z., et al. (2020). Metagenome-wide association of gut microbiome features for schizophrenia. Nat. Commun. *11*, 1612. https://doi.org/10.1038/s41467-020-15457-9.

131. Mai, U., and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics *19*, 272. https://doi.org/10.1186/s12864-018-4620-2.

132. Katoh, K., Misawa, K., Kuma, K.-I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066. https://doi.org/10.1093/nar/gkf436.

133. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

134. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

135. Fu, J., Wang, D., Andreu-Sánchez, S., Peng, H., Ruiz-Moreno, A., Zhernakova, D., Kurilshikov, A., Gacesa, R., Temba, G., Kullaya, V., et al. (2024). Microbiome-wide PheWAS links gut microbial SNVs to human health and exposures. Preprint at Research Square. https://doi.org/10.21203/rs.3.rs-5063726/v1.

136. Pasolli, E., De Filippis, F., Mauriello, I.E., Cumbo, F., Walsh, A.M., Leech, J., Cotter, P.D., Segata, N., and Ercolini, D. (2020). Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. Nat. Commun. *11*, 2610. https://doi.org/10.1038/s41467-020-16438-8.

137. Blacher, E., Bashiardes, S., Shapiro, H., Rothschild, D., Mor, U., Dori-Bachash, M., Kleimeyer, C., Moresi, C., Harnik, Y., Zur, M., et al. (2019). Potential roles of gut microbiome and metabolites in modulating ALS in mice. Nature *572*, 474–480. https://doi.org/10.1038/s41586-019-1443-5.

138. Kim, M.-S., and Bae, J.-W. (2018). Lysogeny is prevalent and widely distributed in the murine gut microbiota. ISME J. *12*, 1127–1141. https://doi.org/10.1038/s41396-018-0061-9.

139. Rosshart, S.P., Herz, J., Vassallo, B.G., Hunter, A., Wall, M.K., Badger, J. H., McCulloch, J.A., Anastasakis, D.G., Sarshad, A.A., Leonardi, I., et al. (2019). Laboratory mice born to wild mice have natural microbiota and model human immune responses. Science *365*, eaaw4361. https://doi.org/10.1126/science.aaw4361.

140. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290. https://doi.org/10.1093/bioinformatics/btg412.

141. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2019). Fast gene set enrichment analysis. Preprint at bioRxiv, 060012. https://doi.org/10.1101/060012.

142. Ducarmon, Q.R., Karcher, N., Tytgat, H.L.P., Delannoy-Bruno, O., Pekel, S., Springer, F., Schudoma, C., and Zeller, G. (2024). Large-scale computational analyses of gut microbial CAZyme repertoires enabled by Cayman. Preprint at biorXiv. https://doi.org/10.1101/2024.01.08.574624.

143. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

144. Yang, Y., Gao, W., Zhu, R., Tao, L., Chen, W., Zhu, X., Shen, M., Xu, T., Zhao, T., Zhang, X., et al. (2025). Systematic identification of secondary bile acid production genes in global microbiome. mSystems *10*, e0081724. https://doi.org/10.1101/2024.06.08.598071.

145. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

146. Pascal Andreu, V., Augustijn, H.E., Chen, L., Zhernakova, A., Fu, J., Fischbach, M.A., Dodd, D., and Medema, M.H. (2023). gutSMASH predicts specialized primary metabolic pathways from the human gut microbiota. Nat. Biotechnol. *41*, 1416–1423. https://doi.org/10.1038/s41587-023-01675-1.

Please cite this article in press as: Andreu-Sánchez et al., Global genetic diversity of human gut microbiome species is related to geographic location and host health, Cell (2025), https://doi.org/10.1016/j.cell.2025.04.014

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Gut microbiome phylogenies	This paper	Zenodo: https://doi.org/10.5281/zenodo.14651412
Software and algorithms		
Python (v3.9)	Python Software Foundation	https://www.python.org/
R (v4.3)	https://www.R-project.org/	https://www.R-project.org/
MetaPhlAn 4/StrainPhlAn 4	Blanco-Míguez et al.[9]	https://github.com/biobakery/MetaPhlAn
PhyloPhlAn 3.0	Asnicar et al.[64]	https://github.com/biobakery/phylophlan
GT-Pro (v1.0.1)	Shi et al.[12]	https://github.com/zjshi/gt-pro
Anpan	Ghazi et al.[37]	https://github.com/biobakery/anpan
fastANI	Jain. et al.[38]	https://github.com/ParBLiSS/FastANI
Roary	Page et al.[65]	https://sanger-pathogens.github.io/Roary/
MEGAHIT (v1.1.1)	Li et al.[66]	https://github.com/voutcn/megahit
MetaBAT	Kang et al.[67]	https://linsalrob.github.io/ComputationalGenomicsManual/CrossAssembly/Metabat.html
checkM (v1.1.2)	Parks et al.[68]	https://github.com/Ecogenomics/CheckM
Traitar	Weimann et al.[36]	https://github.com/hzi-bifo/traitar
Other		
curatedMetagenomicData (v3)	Pasolli et al.[69]	https://github.com/waldronlab/curatedMetagenomicData

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We sampled 90 datasets[4,6,14,19,21,24,32,49–54,70–130] collected and curated by curatedMetagenomicData (as available in the Github commit 11 February 2023).[69] We removed studies that involved fecal microbiota transplantation and kept only gut microbiome samples. Three additional Dutch cohorts were included, KurilshikovA_2019,[63] GacesaR_2022,[3] and ZhernakovaA_2016,[2] as well as 338 4-year follow-up samples from the latter.[61] We also included an additional publicly available Tanzanian cohort, StražarM_2021.[63] In total, this resulted in 32,232 metagenomes from 94 studies. We then removed 80 samples from ThomasAM_2019_c because they overlapped with YachidaS_2019, resulting in 32,152 samples. Phenotypes belonging to curatedMetagenomicData were merged, and units were standardized with the additional cohorts.

METHOD DETAILS

Microbial phylogenetic tree reconstruction

We processed sequencing reads using MetaPhlAn 4[9] and the database mpa_vJan21_CHOCOPhlAnSGB_202103, with default parameters. To build phylogenetic trees, we used the StrainPhlAn 4 approach with the same database. Minimal code changes were implemented in StrainPhlAn 4 to allow processing of the large number of samples as input. For each SGB, samples were divided between primary, used for selection of marker genes used to build the phylogenies, and secondary, based on the SGB's estimated depth of coverage per sample. Estimated depth of coverage was calculated with [Equation 1].

$$Coverage_{SGB,Sample} = \frac{\left(AvgReadLength_{Sample} \times ReadNumber_{Sample} \times Rel.Abundance_{SGB,Sample}\right)}{AvgGenomeLength_{SGB}} \quad \text{(Equation 1)}$$

Equation 1. **Estimation of SGB read depth of coverage per sample.** Total coverage is computed as the average read length per sample used for mapping multiplied by the number of reads used for mapping times the estimated relative abundance for the SGB according to MetaPhlAn 4. MetaPhlAn's average genome length of the SGB is then used for division and computing the average depth of coverage per genomic position.

We used a threshold of $\geq$ 2X to define primary samples and a relative abundance >0 (depth of coverage <2X) to define secondary samples. We set the StrainPhlAn parameters to '–sample_with_n_markers 50 –secondary_sample_with_n_markers 50 –sample_with_n_markers_after_filt 33 –marker_in_n_samples 50', which is more restrictive than the current defaults, aiming to yield a more robust phylogeny reconstruction.

In addition, we included the tag argument '–treeshrink' to remove outliers from the resulting phylogenies. TreeShrink[131] was developed to identify abnormally large branches from a phylogeny, based on an identification of branches that increase the diameter of the phylogeny. This algorithm targets samples with high levels of polymorphisms (e.g., due to sequencing errors) or where reads map to the wrong marker genes. StrainPhlAn uses the marker genes defined in the database to identify polymorphisms per sample. Highly heterogeneous positions within one individual (due to a mixture of microbial strains) are masked, and the rest of the nucleotides are used to build a multiple sequence alignment between all the individuals using MAFFT.[132] Alignments were subsequently used to build maximum likelihood trees by RAxML[133] using the PhyloPhlAn default GTRCAT model, in most cases. Given the high computational needs of large phylogenies, for eight phylogenetic trees we used another maximum likelihood software, IQTree v2.3.0 (GTR model),[134] as summarized in Table S1, 3. To check whether differences in the maximum likelihood tree software used (IQTree or RAxML) had a major influence in the sample-to-sample distance matrix used by phylogenetic generalized linear mixed models (PGLMM), we reconstructed phylogenies with both RAxML and IQTree for nine random species. A Mantel test (Pearson correlation) between each showed high correspondence between distance metrics (0.72–0.99, mean of 0.93). A median of two samples were removed per phylogeny due to TreeShrink, with a maximum of 24 samples being removed.

Subsequently, we estimated the final number of samples available per phylogeny and removed phylogenetic trees with less than 300 leaves/samples from the analysis. As an additional analysis, we reconstructed *B. longum*'s phylogeny, including 48 publicly available references of *B. longum*, and retrieved subspecies annotation from NCBI [Table S1, 4]. For that, we used the '–reference' option from StrainPhlAn.

Analysis of within-sample polymorphisms

For each SGB, we re-treated information regarding the percentage of polymorphic sites (i.e. where more than one allele was observed) found per individual in the marker genes used for multiple sequence alignment (as generated per StrainPhlAn). For each SGB, we extracted the median and maximum percentage of polymorphic sites per sample. We performed a Pearson correlation between the number of samples per phylogeny (as a proxy of how well a species is covered in a particular sample) and the median percentage of polymorphic sites per SGB. Finally, we conducted a gene set enrichment analysis (GSEA) to identify which taxonomic levels were enriched in SGBs with a higher median of polymorphic sites, using all taxonomic levels at once.

Comparison of dominant strains of StrainPhlAn with GT-Pro

To assess the robustness of the methodology for determining within-species genetic variability, we compared the phylogenetic distances of the StrainPhlAn trees with genetic dissimilarities estimated from core genome SNVs. Core genome SNVs were metagenotyped with GT-Pro in 10,781 samples from GacesaR_2022, ZhernakovaA_2016, SchimerM_2016, KurilshikovA_2019, VichVilaA_2018, and StražarM_2021, as described elsewhere.[135] Manhattan distance (scaled by the number of shared SNVs) was used to compute sample-to-sample dissimilarities of SNV frequencies per species. We then assigned SGB categories for the reference genomes from GT-Pro using PhyloPhlAn's 'phylophlan_metagenomic' (v3.0.39) (SGB.Jan21 database). Common species with both GT-Pro distances and StrainPhlAn phylogenies were used for a Mantel test (vegan, v2.6-4) (Pearson correlation, 1000 permutations). We then discarded species with fewer than 50 paired samples between distances and estimated a Benjamini-Hochberg (BH) FDR. For all the remaining 243 common taxa (median of 1,316 samples) we identified correlations significantly different than 0 for all taxa (FDR<0.05), with a median correlation value of 0.7, suggesting a robust estimation of dominant strains [Table S2, 3].

Food references in phylogenetic tree and *B. animalis* analysis

As a positive control to assess phylogenetic trees built from samples stemming from different biological and technical backgrounds, we further focused on the *B. animalis* SGB17278. As this species is acquired from yogurt consumption and not a natural gut microbiome inhabitant, we would expect to observe a common strain in all samples. Within the MetaPhlAn 4 genomic database from January 21, SGB17278 contained at least one genome tagged as food.[136] We then rebuilt the phylogenetic tree from this SGB and included the reference genomes with the StrainPhlAn –references option. We further included metagenomes from three mouse studies[137–139] to assess species' variability in a natural population. We selected samples with a relative abundance of SGB17278/*B. animalis* >0.1% as primary samples in StrainPhlAn. Using the reconstructed phylogenetic tree, we then extracted phylogenetic distances between human metagenomes and the food MAG reference from the MetaPhlAn 4 database,[9] ten mouse strains, and human geographic locations and studies. To test whether there were significant differences in the distances, we performed permutations of the labels of interest. We first tested whether human–human distances were significantly different from the human–food distances. We calculated the t-statistic on the non-permuted data and the t-statistics of the data after 1,000 permutations of the mouse/human label. The P-value was estimated by computing the proportion of permutation t-statistics $\geq$ the non-permuted t-statistic. We performed the same approach to test whether distances between mouse strains were larger than distances between human strains.

Geographic association with microbial phylogeny

Phylogenetic distances were extracted from phylogenies using ape's v5.7-1 cophenetic.phylo.[140] The geographic location of a country's capital city was obtained from the dataset world.cities from the R package maps v3.4.1. We then applied distVincentySphere from the geosphere v1.5-18 package to obtain the geographic distances between the cities. Next, we computed the correlation between phylogenetic and geographic distances using a Mantel test (vegan, v2.6-4) (Pearson and Spearman correlations, 2,000

permutations) and estimated BH-FDRs. A comparison of the correlation estimates using the two different coefficients (Pearson's and Spearman's) showed high correlation ($r_{Pearson}$=0.93). Since most estimates seemed robust enough, we subsequently worked with the Pearson's results.

To test for overrepresentation of specific taxonomies within geographically associated SGBs, we performed a Fisher's exact test (with an alternative hypothesis for greater) for each taxon at taxonomic levels ranging from phylum to genera. FDRs were estimated using the BH procedure.

Species transmissibility and geographic effect

Using the phylogenetic trees, we inferred strain-sharing between all available samples using a previously validated approach.[14] The methodology involves identifying a phylogenetic distance threshold within a species-specific distribution of phylogenetic distances (normalized by the median distance within the species) that effectively distinguishes between distances among a distribution of distances from longitudinal samples (one pair per subject) collected within 6 months and unrelated samples. Since related samples collected during 6 months will likely exhibit a distribution with a peak around 0 due to strain retention, and unrelated samples will exhibit a peak far from 0 due to the fact that most unrelated people contain different strains, a cut-off that splits those distributions can act as a proxy for a strain-sharing cut-off.[14]

For samples with at least 30 pairs of longitudinal distances, a distance that maximizes the Youden index (defined as sensitivity + specificity – 1) is chosen. If this distance is larger than the 5th percentile of the distribution, the 5th percentile is used instead as an upper bound. If not enough longitudinal samples are available (<30 pairs), the 3rd percentile of the overall distribution is used as the distance cut-off, as previously defined.

We used previously computed percentile values.[70] For the SGBs that did not have a previous threshold value, we observed that no longitudinal samples were available in the phylogenies and used the 3rd percentile as the threshold. We then used those thresholds to determine strain-sharing between samples. Species transmissibility was computed as the fraction of samples with the same strain among all pairs of samples from the same country but different studies. This rate was used as a proxy for bacterial migration rates (dispersion). To avoid possible confounders due to same-study biases (unaccounted-for social networks, contamination events, other technical biases), we computed these rates in samples from different studies. Opposed to using sharing rates between unrelated individuals, within family strain-sharing rates could not distinguish between horizontal and vertical strain transmission. Vertically transmitted taxa, followed by strain retention, might be more commonly seen in geographically stratified taxa, where they may represent evolutionary co-adaptation.

Species transmissibility was associated with geography–phylogeny correlation values using linear models.

Microbial characteristics and geographic effect

For each SGB, we retrieved the average genome length from the genomes used for the construction of the SGB in the MetaPhlAn 4 Jan 21 database. We then associated the average genome length with the rho value obtained from a Mantel test between geographic and phylogenetic distances (geographic effect) using a linear model. We further assessed the proportion of available genomes present in each environment and associated it to the geographic effect using linear models. The BH FDR was computed.

We used Traitar (v1.1.12)[36] to predict microbial phenotypes. For each SGB, we extracted a set of core genes (genes present in 50% of genomes available in the MetaPhlAn 4 genomic database). We kept only annotations where the prediction of phypat and the phypat hypat + PGL classifiers (including additional evolutionary information on phenotype gains and losses) matched. Annotations were then associated with geographic effects using a linear model, and BH FDRs were estimated. For FDR < 0.05 results, we further added whether the family of the SGB was *Lachnospiraceae* as a covariate in a subsequent model, as this family was highly enriched in geographically stratified species and might act as a confounder.

Phylogenetic association

Phylogenetic generalized mixed models in anpan

We matched metadata information available for the different cohorts with the reconstructed phylogenies. If longitudinal sampling was available, a random sample from an individual was chosen for analysis. We focused on phenotype–phylogeny pairs of at least 20 samples. For categorical data, we required at least 15 observations per level.

We set up to run PGLMMs using anpan[37] to identify pairs of phylogenetic trees and phenotypes with a strong phylogenetic signal. Anpan models the phenotype of interest as a function of different covariates and a phylogenetic random effect where a covariance matrix of the phylogenetic relations between samples (assuming a Brownian motion) is multiplied by the phylogenetic effect term (σ_{phylo}). The fit is performed in a Bayesian framework, either in a binomial or Gaussian model, depending on the dependent phenotype. Evaluation of fit is done through a leave-one-out (loo) cross-validation approximation assessed using Pareto smoothing to obtain expected log predictive density (ELPD). Models that achieve better prediction of the dependent phenotypes would result in a higher ELPD. The ELPD obtained in this model is then compared to the ELPD obtained in a null model that does not introduce the phylogenetic random-effect term but does include all other covariate terms in the prediction. The ELPD difference (elp_diff) between both models (null-phylogenetic) is obtained, in addition to its standard error.

To select models with strong evidence of an existing phylogenetic effect, we focused on models where the confidence interval of the elp_diff (elp_diff ± 2SE) did not include the value 0 and where the elp_diff was below -4, considered indicative of an alternative

model predicting significantly better than the null. In addition, we removed associations where the diagnostic Pareto K value was $\geq$ 0.7 in at least 1% of the samples. High Pareto K values are indicative of a few samples having a big influence on the model performance, which might be related to either the presence of outliers or highly unbalanced binary dependent phenotypes. In model comparisons with high Pareto values, the interpretation of elpd_diff should be considered carefully, because it might be a bad approximation of the real loo estimation. In those cases, σ_{phylo} can be used instead as a metric indicating whether the phylogenetic effect on the dependent variable is different than 0, using the estimated posterior distribution of the term.

Simulation analysis to assess model performance

In order to assess whether anpan properly controls for false discoveries with its default priors, we performed 200 simulations, in which we generated random trees with 120 samples each (ape, rtree) and a random covariate following a normal distribution. We then created a residual term, with a standard error of 1 and a mean of 0. Finally, we created a phylogenetic term, which was specific as a σ_{phylo}=0 multiplying a multivariate normal distribution of 0 mean and a variance-covariance matrix given by the random phylogeny. The observed variable was then obtained as the sum of the random covariate, the residual term, and the phylogenetic term. This model was fitted using anpan. We then assessed whether the model uncovered any association (|elpd_diff| > 4, with a 95% CI not overlapping 0), and observed that out of the 200 simulations, in none of them the phylogenetic model was chosen above the model with no phylogenetic term, indicating that the regularization of the model (where the phylogenetic term σ_{phylo} is shrunk towards 0) makes the false positive rate asymptotically approaching 0 as the signal-to-noise ratio increases.

Phylogenetic generalized mixed models fit to data

We fitted a model in which the phenotype of interest was used as a dependent variable, with the participant's age and continent used as covariates. This model aims to uncover phylogenetic associations with human phenotypes that are independent of geographic differences between samples. For the statistically supported hits, we then conducted a subsequent analysis stratifying per continent. For each continent, we fitted a second model with the phenotype of interest as dependent variable and country, age, and a random intercept taking phylogeny into account as independent variables.

For individual phylogeny–phenotype pairs, we performed additional analysis controlling for specific covariates that might confound the phylogenetic signal and/or stratifying per study. A final model fitted for disease associations included a study-specific offset proportional to the prevalence of the disease per study, implemented in anpan.

From the fitted model, we extracted the leaf-wise (sample-wise) posterior distribution of the phylogenetic effect. This is estimated from the PGLMM and represents the sample-specific offset term introduced by the random effects, i.e., the effect the phylogeny has in the prediction of the dependent variable. We summarized the posterior distribution per sample as the distribution's median.

To visualize the results, we extracted the median of the posterior distribution of the phylogenetic effect per leaf and plotted it together with the other covariates used.

Downsampling analysis

To get an indication of the effect of sample size in elpd_diff estimates, we repeated the PGLMM models on several data subsamples from European cohorts in the age associations. We selected eight different taxa and ran anpan models between age and the phylogeny of these species for four subsamples of the European samples (75%, 50%, 25% and 10% of samples). The taxa selected included two species associated with North America, Asia and European analyses; three species associated in Europe and Asia; two taxa associated only to Asia; and one species not associated in any analysis as a negative control.

Testing the influence of sequencing protocol

433 samples from SchirmerM_2016[115] had been extracted with both the APK and FSK isolation protocols.[35] Per each available phylogenetic tree, we kept only the SchirmerM_2016 samples that were isolated with both protocols. For phylogenetic trees with at least 20 remaining samples, we ran an anpan model using the protocol as dependent variable, with no additional covariates.

Next, we assessed whether distances between replicates in different protocols resemble distances between pseudo-replicates generated by splitting sequencing reads from one sample into two. For that, first we subsampled to equal sequencing depth each replicate in each of the protocols (APK/FSK), so that if subject A_{APK} had 2 million reads, and subject A_{FSK} had 5 million reads, subject A_{FSK} was subsampled to 2 million reads. Then, we randomly split each sample in half. All samples were then processed through MetaPhlAn 4 and StrainPhlAn 4 as described above. Then, we extracted the phylogenetic distances from each of the phylogenies, and subsample only the subjects where all four replicates were available (2 pseudo-replicate per protocol). We performed a permutational multivariate analysis of variance (PERMANOVA) (adonis2, vegan), testing whether protocol could significantly explain the distances (restricting permutations within subject ID).

As an additional test we sought to identify whether differences in sequencing efforts on different species might also influence the identification of a common strain. For this, we used SchirmerM_2016 samples that were isolated with two different protocols and computed the depth of coverage per SGB per sample using Equation 1. Then, per subject and SGB, we determined the absolute difference (delta) between the depth of coverage under the two isolation methods. As isolation methods can result in widely different relative abundances,[35] which also result in different depth of coverage, we correlated those absolute delta values with the phylogenetic distance extracted from the species phylogeny for samples from the same subject isolated with the two different protocols. For

that, we assumed that if depth of coverage influences the strain identification, larger differences in delta would result in greater phylogenetic distances. We performed a Spearman correlation test (exact=False to deal with ties), using as an alternative hypothesis that the correlation should be greater than 0.

Enrichment of taxonomic levels

To test if geographic effects were enriched in specific taxonomic levels, we ran a gene set enrichment analysis (GSEA, here referred to as SEA) as implemented in R, fgsea v1.26.0.[141] We ran SEA instead of an overrepresentation analysis because most SGBs showed significant geographic effects. Each SGB was grouped into the taxonomic level to which it belonged (the same SGB overlapped in several categories as taxonomic levels are nested). We ran fgsea with standard parameters, using the rho values for ranking the associations.

For taxonomical enrichment in phenotype associations, we ran an overrepresentation analysis. For each taxonomic level, we retrieved how many statistically supported SGB–phenotype pairs were found out of how many SGB–phenotype pairs were tested. We then compared the proportions of pairs from SGBs from a particular taxonomy to all others. We ran a Fisher's exact test for over-enrichment (alternative hypothesis, greater). After running this procedure at all taxonomic levels, BH FDR was applied once on the resulting overall P-values.

Clade-specific analyses

Infant B. longum

B. longum reference genomes were included in the *B. longum* phylogeny to identify the most likely subspecies representing each clade. We used the NCBI-annotated taxonomy to identify their subspecies. To define subsp. *infantis*–like and subsp. *suis*–like clades, we obtained the common ancestor for the references annotated as *longum* and *suis* that fell within the infant-enriched clades. We observed a reference labeled as subsp. *longum* within the *suis*-like clade. We performed two independent analyses to confirm that this is the case beyond just the marker-gene-based phylogeny. First, we calculated ANI between all reference genomes using fastANI[38] with default parameters, visualized the clustering of reference genomes using a hierarchical clustering approach, and computed the average ANI between the potential mislabeled genome and the genomes annotated in each of the different clades. In addition, we obtained a gene presence-absence matrix for the pangenome generated from the reference genomes using roary, with default parameters.[65] CAZyme annotations for each reference genome were obtained using dbCAN3.[39] CAZyme presence/absence association between subspecies was done by means of a Fisher's exact test. Substrate origin enrichment analysis was carried out using a one-sided Fisher test between the number of significant associations that had a prevalence of 1 in subsp. *infantis*, for each possible substrate, as annotated in a previous work.[142]

Melanoma

Metadata from melanoma studies were extracted from curatedMetagenomicData. We then used the anpan phylogenetic effect results to define which samples belonged to the cancer-enriched clade we identified. To explore whether any further phenotypes associated with clade membership, we ran logistic regression models. First, we assessed the enrichment of melanoma samples in the melanoma-enriched clade by running a mixed-effect model with glmer of the form: *Melanoma* ~ *Melanoma_clade + Age + (1|Country)*. Then, we obtained information regarding previous treatment for the LeeKA_2022 and McCullochJA_2022 samples. Since this definition might vary per study, we ran two independent logistic models of the form: *Melanoma_clade* ~ *previous_therapy*. A third model was then fitted to test whether individuals with strains in this clade had a different response to checkpoint inhibitor treatment. For that, we used samples from LeeKA_2022, McCullochJA_2022, and FrankelAE_2017 as they had the most samples. We used the variable RECIST, which indicates classification for response to cancer therapy and has four levels (complete response, partial response, stable disease, and progressive disease), in a model of the form: *Melanoma_clade* ~ *Age + RECIST + (1|study) + (1|study:country)*. The last term was included since LeeKA_2022 was a study with two nested countries.

We further analyzed samples from PernigoniN_2021. First, we performed a Fisher's exact test to explore whether samples of this cohort were over-enriched in the clade of interest. Then, we associated the clade with treatment while accounting for participant age. To study the hormonal levels associated with this clade, we retrieved data from the SchimerM_2016 participants in the clade. We obtained their previously published circulating hormonal levels[55] and inverse rank transformed them, effectively converting data ranks into normally distributed values, using [Equation 2].

$$x_{i'} = \Phi^{-1}\left(\frac{r_i - 0.5}{n}\right) \quad \text{(Equation 2)}$$

Equation 2. Inverse rank normal transformation. $x_{i'}$ is the transformed value for the i-th observation. Φ^{-1} is the inverse of the cumulative distribution function (CDF) of the standard normal distribution (i.e., the quantile function). r_i is the rank of the i-th observation in the data. n is the total number of non-missing observations in the dataset.

Finally, we ran a linear model of the form: *Invrank(testosterone)* ~ *Sex + Age + Clade*, where *Invrank* refers to the rank-based inverse normal transformation of the data [Equation 2].

Elder individuals

We pruned the SGB4584/*R. gnavus* phylogenetic tree to include only European and Asian samples. We then identified the clade associated with older individuals by identifying all the descendant leaves from the most common recent ancestor from XuQ_2021

samples with a high phylogenetic signal using a nonagenarian–SGB4584 anpan analysis within XuQ_2021 samples. We defined this clade as the 'elder-enriched' clade. Next, using only European samples, we compared the average age of the elder-enriched clade to that of the rest of the phylogeny. For that, we fitted a logistic regression model of the form: elder_*clade* ∼ *Age + (1|Country)*.

We extracted all baseline samples from ZhernakovaA_2016 and KurilshikovA_2018 that were available in the phylogeny. We retrieved previously measured concentrations of 15 bile acids,[17,46] including primary CA and deoxycolic acid, and their secondary bile acid forms, lithocholic acid, and ursodeoxycholic acid and CDCA, including their taurine and glycine conjugates. Using their relative concentrations (absolute concentration/sum concentrations within a sample), we associated the relative concentration to whether the sample belonged to the elder-enriched clade with the model: Bile_acid ∼ elder_*clade + Age + Sex + (1|study)*. BH FDR were estimated from the resulting P-values.

Functional analysis of MAGs

MAGs were built using the following pipeline. First, assemblies were reconstructed from sequencing reads using MEGAHIT (v1.1.1),[66] using the parameters –min-counts 2 –k-list 21,31,41,51,61,71,81,91,99. We then filtered out contigs shorter than 1,500bp. Second, we used MetaBAT2[67], to bin contigs per sample. Third, we estimated MAG completeness and contamination per MAG using the 'lineage_wf' workflow with default parameters in CheckM (v1.1.2).[68] We then selected the high-quality MAGs, i.e., MAGs with at least 90% completeness and maximum 5% contamination. Fourth, we ran PhyloPhlAn 3.0[64] with the script 'phylophlan_metagenomic' (v 3.0.39) to assign each high-quality MAG to a putative SGB, using the January 21 MetaRefSGB database. Fifth, we used Prokka[143] v1.14 with default parameters for open reading frame prediction, and the fasta files were annotated for UniRef90 and UniRef50 (version 2019/06) using uniref_annotator (https://github.com/biobakery/uniref_annotator).

We merged the UniRef90 annotations for all MAGs to create a pan-genomic gene content matrix of presence-absence and then tested for enrichment of different UniRef90 annotations in specific clades. For this, we ran a Fisher's exact test for enrichment of a specific UniRef90 term within the genes present in a clade versus the genes present in the whole pangenome. We added a pseudocount of 1 to test for genes that were totally absent in a clade.

To run an enrichment analysis on the pathways from the associated UniRef90 genes, we annotated each gene to an Enzyme Commission (EC) number and to the MetaCyc pathways each EC belongs to. We then performed a GSEA analysis on MetaCyc pathways.

We investigated the presence of 13 previously curated bile-acid-related genes.[144] We built HMMER profiles using publicly available seeds for each of the genes. For that, we built multiple sequence alignments per gene seed using MUSCLE[145] and ran three cycles of hmmalign and hmmbuild. Then, we ran each of the gene profiles against each of the amino acid sequences predicted by Prokka from each of the *R. gnavus* MAGs (evalue=$1x10^{-65}$).

To determine the presence of specific gene clusters, we used gutSmash v1.0.0[146] on MAGs of interest.

Supplemental figures

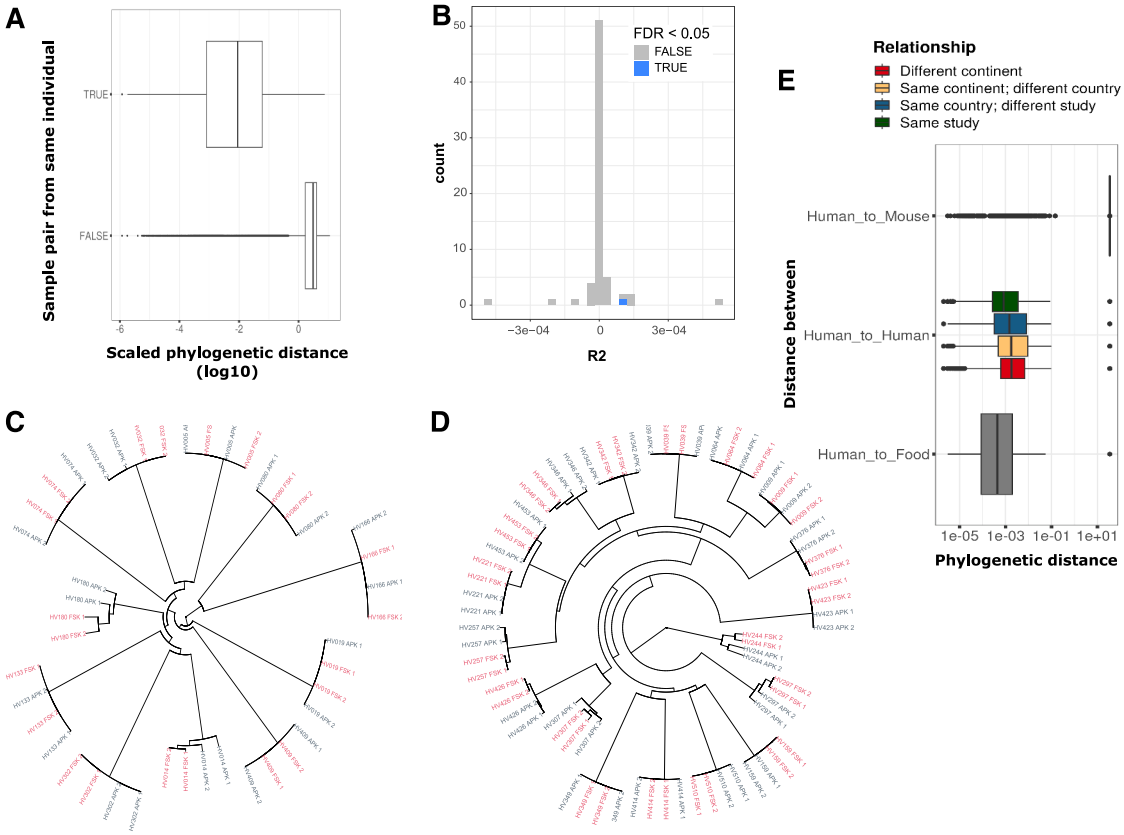


Figure S1. Batch effects in phylogenetic analyses, related to Figure 1

(A) Samples from the same individuals that were isolated with different methods (QIAGEN AllPrep DNA/RNA kit [APK], originally published in SchimerM_2016, or QIAmp Fast DNA Mini Kit [FSK], later resequenced) tend to have lower phylogenetic distance between their dominant strains than strains from different individuals (where some samples may share the same strains), showing that sequencing protocol largely does not confound strain phylogenies. Distances include all 152 tested species and are standardized by each phylogenetic tree's standard deviation of the phylogenetic distance.

(B) Distribution of R^2 estimated for isolation protocol from a permutational multivariate analysis of variance (PERMANOVA) on genetic distances for different taxa. Samples from the same subjects isolated with different protocols were subsampled to equal sequencing depth and split in two to generate pseudo-replicates.

(C and D) Phylogenetic trees of species SGB5809_group and *Faecalibacterium prausnitzii* (SGB15332_group) for samples from SchimerM_2016 under two different protocols, with pseudo-replicates. These phylogenies represent the two species where the PERMANOVA test identified significant differences between protocols. No major differences between protocols can be observed.

(E) Phylogenetic distances between samples in the tree of SGB17278/*Bifidobacterium animalis*, showing distances between human samples from different continents, countries, and/or studies; distances between human samples and a single yogurt-isolated strain; and distances between human and strains found in mice gut microbiota.

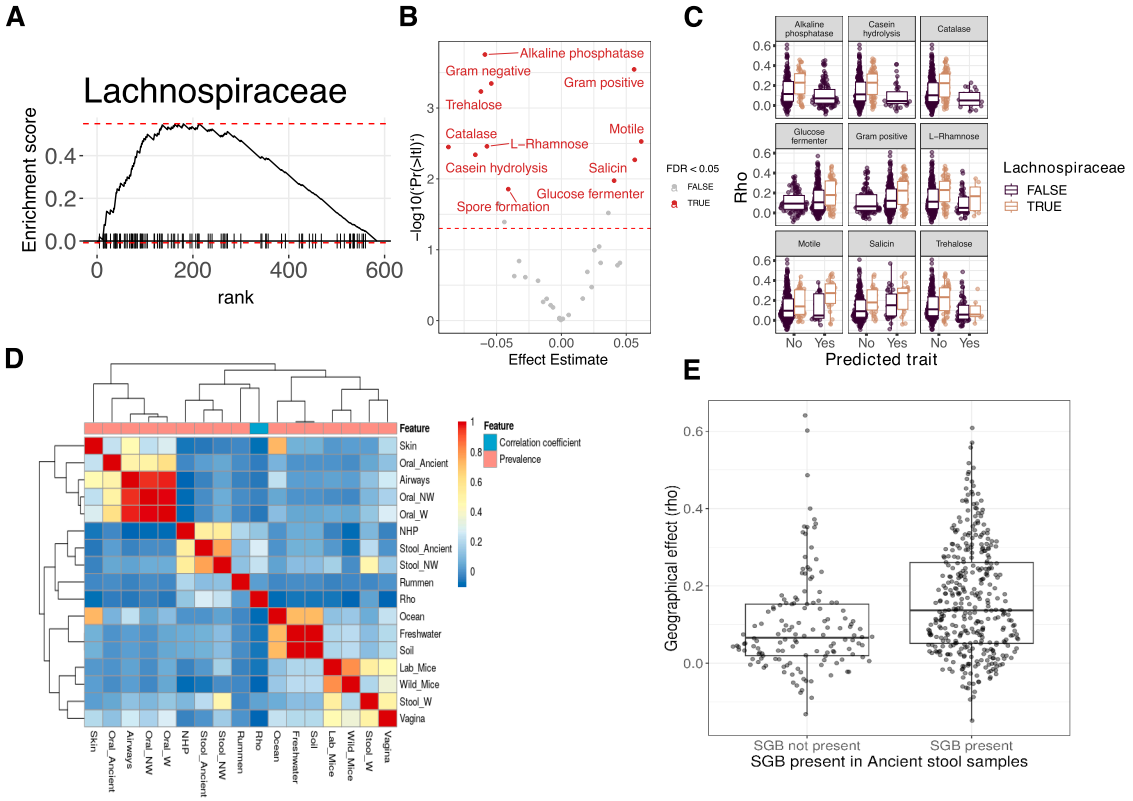


Figure S2. Factors related to geographic effects, related to Figure 2

(A) Gene set enrichment analysis of the *Lachnospiraceae* family, including 55 members. x axis indicates the rank of the geographic effect, i.e., the rho value obtained from a Mantel test between the geographic and phylogenetic distances. Highlighted in black is the rank of *Lachnospiraceae*. y axis presents the cumulative enrichment score.

(B) Volcano plot showing the associations between predicted microbial phenotypes from Traitar and geographic effects. x axis indicates the estimated effect from the linear model, y axis its ($\log_{10}$ transformed) corresponding *p* value. Dashed line indicates the nominal *p* value threshold of 0.05. Color indicates whether the FDR for the association was estimated to be <0.05.

(C) Boxplot comparing the distribution of rho values in species-level genome bins (SGBs) predicted to have several features. Coloring indicates whether the SGB belongs to the *Lachnospiraceae* family.

(D) Heatmap showing the correlation between SGB prevalence in different environments and geographic effect.

(E) Comparison of geographic effects of SGBs present in at least one ancient human stool sample vs. those not present.

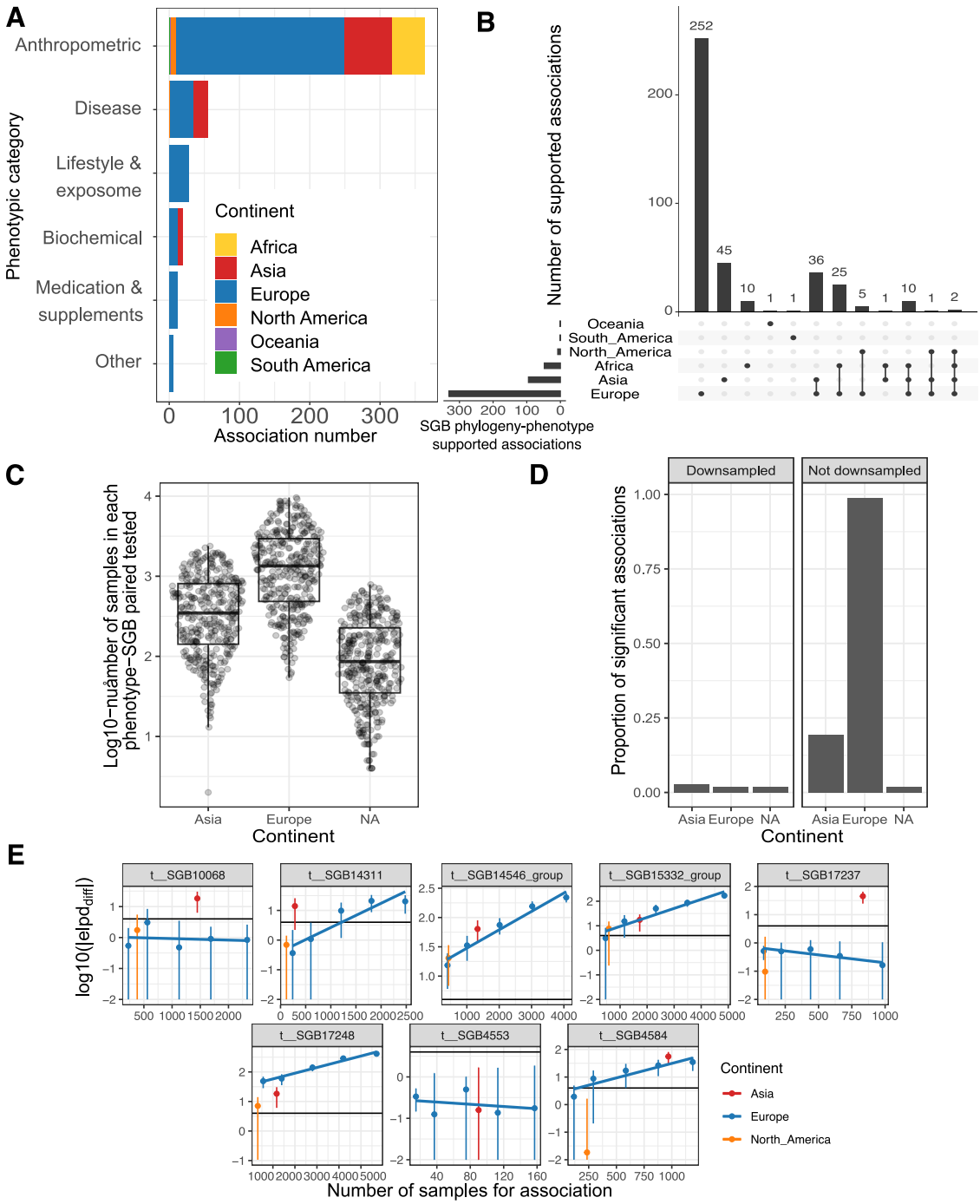


Figure S3. PGLMMs identify association of phylogenetic trees and human phenotypic variation, related to STAR Methods, Phylogenetic association

(A) Bar plot displaying the number of associations per phenotypic category. Color indicates the continent where the association was found.

(B) Upset plot displaying the number of common SGB-phenotype associations between different geographies.

(C) Number of samples in each phenotype-SGB pair tested on the phylogenetic framework is different between continents.

(D) The number of associations decreases when all continents are downsampled to equal numbers of samples.

(E) ELPD relationship with sample number. The number of samples in a phylogeny is linearly related with the elpd_diff score obtained from anpan. This linear relationship is not seen in associations with no signal. Horizontal line indicates the elpd_diff cut-off used to reject the null hypothesis.