Article

Expanding the human gut microbiome atlas of Africa

https://doi.org/10.1038/s41586-024-08485-8

Received: 6 February 2024

Accepted: 3 December 2024

Published online: 29 January 2025

Open access

Check for updates

Dylan G. Maghini[1,2,15], Ovokeraye H. Oduaran[1,15], Luicer A. Ingasia Olubayo[1], Jane A. Cook[3], Natalie Smyth[1], Theophilous Mathema[1], Carl W. Belger[1,4], Godfred Agongo[5,6], Palwendé R. Boua[1,7], Solomon S. R. Choma[8], F. Xavier Gómez-Olivé[9], Isaac Kisiangani[10], Given R. Mashaba[8], Lisa Micklesfield[11], Shukri F. Mohamed[10], Engelbert A. Nonterah[6], Shane Norris[11,12], Hermann Sorgho[7], Stephen Tollman[9], Floidy Wafawanaka[9], Furahini Tluway[1], Michèle Ramsay[1], Jakob Wirbel[2], the AWI-Gen 2 Collaborative Centre*, Ami S. Bhatt[3,13,16] ✉ & Scott Hazelhurst[1,14,16] ✉

Population studies provide insights into the interplay between the gut microbiome and geographical, lifestyle, genetic and environmental factors. However, low- and middle-income countries, in which approximately 84% of the world's population lives[1], are not equitably represented in large-scale gut microbiome research[2–4]. Here we present the AWI-Gen 2 Microbiome Project, a cross-sectional gut microbiome study sampling 1,801 women from Burkina Faso, Ghana, Kenya and South Africa. By engaging with communities that range from rural and horticultural to post-industrial and urban informal settlements, we capture a far greater breadth of the world's population diversity. Using shotgun metagenomic sequencing, we identify taxa with geographic and lifestyle associations, including *Treponema* and *Cryptobacteroides* species loss and *Bifidobacterium* species gain in urban populations. We uncover 1,005 bacterial metagenome-assembled genomes, and we identify antibiotic susceptibility as a factor that might drive *Treponema succinifaciens* absence in urban populations. Finally, we find an HIV infection signature defined by several taxa not previously associated with HIV, including *Dysosmobacter welbionis* and *Enterocloster* sp. This study represents the largest population-representative survey of gut metagenomes of African individuals so far, and paired with extensive clinical biomarkers and demographic data, provides extensive opportunity for microbiome-related discovery.

Large population studies can identify lifestyle, genetic and environmental factors that drive gut microbiome composition. Indeed, early studies[5,6] established baseline human gut microbiome measurements, and more recent studies have related the gut microbiome to disease and lifestyle factors[7–9]. However, because these studies typically focus on high-income populations with relatively homogeneous resource access and disease profiles, their results often do not translate to populations with different lifestyle practices, health challenges and access to healthcare, and varied environmental exposures. In addition, large microbiome studies have typically relied on facility-based convenience-sampling models, which generalize less well to the population level than more resource-intensive cross-sectional sampling.

Low- and middle-income countries (LMICs) account for approximately 84% of the world's population[1] but are extremely under-represented in gut microbiome research[2–4]. China, an upper-middle income country, is an outlier with several recent large population studies[10–12]. Targeted studies in LMICs have found microbiome associations with infectious disease[13,14] and compositional differences between microbiomes of specific LMIC and high-income country (HIC) populations[15–18], including in the context of non-communicable diseases[19]. Several studies have also evaluated early childhood microbiomes in LMICs, especially in relation to malnutrition[20–24]. However, comprehensive measurement of global lifestyle diversity that affects gut microbiome composition requires large, population-representative studies. It is essential to work within frameworks that support representative measurements from

[1] Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa. [2] Department of Medicine (Hematology), Stanford University, Stanford, CA, USA. [3] Department of Genetics, Stanford University, Stanford, CA, USA. [4] School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa. [5] Department of Biochemistry and Forensic Sciences, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana. [6] Navrongo Health Research Centre, Ghana Health Science, Navrongo, Ghana. [7] Clinical Research Unit of Nanoro, Institut de Recherche en Sciences de la Santé, Nanoro, Burkina Faso. [8] DIMAMO Population Health Research Centre, University of Limpopo, Polokwane, South Africa. [9] MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of the Witwatersrand, Johannesburg, South Africa. [10] African Population Health and Research Center, Nairobi, Kenya. [11] SAMRC/Wits Developmental Pathways for Health Research Unit, University of the Witwatersrand, Johannesburg, South Africa. [12] School of Human Development and Health, University of Southampton, Southampton, UK. [13] Department of Medicine (Hematology, Blood and Marrow Transplantation), Stanford University, Stanford, CA, USA. [14] School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa. [15] These authors contributed equally: Dylan G. Maghini, Ovokeraye H. Oduaran. [16] These authors jointly supervised this work: Ami S. Bhatt, Scott Hazelhurst. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: asbhatt@stanford.edu; Scott.Hazelhurst@wits.ac.za

populations, facilitate input and leadership from local stakeholders, and clearly identify community needs[25–27].

The Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen)[28] provides a powerful framework for population-representative and community-engaged research (Extended Data Fig. 1). Nested within the Human Heredity and Health in Africa Consortium (H3Africa), AWI-Gen studies genomic and environmental factors affecting the changing disease burden among adults in six communities in four African countries. The study is a partnership between the University of the Witwatersrand and the International Network for the Demographic Evaluation of Populations and Their Health (INDEPTH), a network of health and demographic surveillance systems (HDSSs) in LMICs. Together, the five AWI-Gen HDSSs and the Developmental Pathways for Health Research Unit in Soweto, South Africa support random cross-sectional population sampling. By contrast, most extant microbiome and genomics data are based on non-random recruitment of self-selecting participants, which is not optimal for capturing population-level trends. Each HDSS has engaged with host communities for over a decade, conducting engagement before study approval and through study conclusion[29,30] that ensures high participant retention while focusing on community needs. AWI-Gen collected blood and urine biomarkers, captured extensive participant data—demographic, health history, environment and lifestyle—and genotyped all participants on the H3Africa Custom SNP Array[31]. Emphasizing genomics capacity-building and equitable collaborations, AWI-Gen presents a unique opportunity for microbiome research in understudied populations and holds immense potential for associating the microbiome with rich genotype and phenotype data. The first phase of AWI-Gen ran from 2012 to 2017, during which we conducted pilot microbiome projects at two South African sites[32,33].

Here we present the second phase of the AWI-Gen Microbiome Project. From 2018 to 2023, we randomly sampled 1,820 adults (1,801 women and 19 men) from well-characterized populations in six research centres in Burkina Faso, Ghana, Kenya and South Africa. These centres have widely different population densities, subsistence strategies, income levels and disease profiles. Leveraging extensive clinical and demographic data, we find that geography has the strongest effect on microbiome variation. We assemble thousands of prokaryotic and phage genomes, including hundreds for *Treponema succinifaciens*, a hallmark bacterial species previously described as absent in industrial populations. Finally, we find HIV-associated differences in microbiome composition that differ from those described in HIC populations. Altogether, this study demonstrates the importance of investigating the gut microbiome in undersampled populations, provides a framework for equitable microbiome research and represents the largest population-representative profile of African gut microbiomes so far.

The AWI-Gen 2 Microbiome Project enrolled participants from rural villages in Nanoro, Burkina Faso[34] ($n = 384$), Navrongo, Ghana[35] ($n = 235$), the Agincourt-Bushbuckridge subdistrict in South Africa[36] ($n = 533$) and Dikgale, South Africa[37], in which the HDSS is now called DIMAMO ($n = 203$), from the township of Soweto, South Africa ($n = 226$) and from the Korogocho and Viwandani urban informal settlements in Nairobi, Kenya[38] ($n = 239$) (Fig. 1a). Participants were a cross-sectional representation of the adults in the HDSS catchment areas (Supplementary Methods). The study communities span rural, peri-urban and urban areas, and therefore have drastic differences in population density, water sanitation, access to healthcare and disease profiles (Table 1 and Supplementary Data 1). Briefly, the Nanoro and Navrongo study centres are in primarily horticultural rural regions of western Africa where subsistence farming and cattle-keeping are dominant subsistence strategies. The Agincourt and DIMAMO centres largely consist of semi-rural villages that are undergoing rapid epidemiological transition and industrialization. Soweto is a district within the city of Johannesburg, which under apartheid was designated as an area for black people to live; as a post-industrial area, employment in Soweto is often related to business, retail and industry, but unemployment among women remains above 60% (ref. 39). The Nairobi centre captures two urban informal settlements where population density is very high and residents have limited access to piped water and sanitation.

AWI-Gen 2 is a population cross-sectional study of adults aged 32–98 (99% were between 41 and 84). Pregnant women and people who had been resident for fewer than 10 years were excluded. At the point of recruitment into AWI-Gen, only one person per household was included. Most participants were women ($n = 1{,}801$), although a small number of men ($n = 19$) were sampled as well. The focus on women was motivated by downstream interest in combining these data with a companion menopause study. Samples from men had significantly lower alpha-diversity ($P = 0.027$) (Extended Data Fig. 2), and were excluded for all site comparisons presented below; however, given the poor representation of these populations in existing microbiome studies, samples from men were included in the genome catalogues presented below. Participants completed a questionnaire with guidance from a field worker, and donated blood, urine and single stool samples. Stool samples were collected in temperature-stable buffer and processed at a single time in the same location to minimize handling bias (Methods). DNA was extracted from each stool sample, followed by 2 × 150 base pair paired-end sequencing. We generated a median of 44.16 million (M) (range 27.48M to 104.79M) reads per sample, with a median of 31.20M (range 17.80M to 72.95M) reads remaining after quality control and removal of human reads (Methods, Extended Data Fig. 3 and Supplementary Data 2). For an extended description of each study centre and recruitment methodology, see the Supplementary Methods.

Taxonomic composition across sites

We first characterized the overall taxonomic composition in the study populations. We performed taxonomic classification with mOTUs3 using an updated database that incorporates new genomes found in this study, and summarized features with the GTDB taxonomy (Methods and Extended Data Fig. 3). After clustering samples by overall microbiome composition, the primary axis of variation captures a trade-off in relative abundance between Bacteroidota and Bacillota A (Fig. 1b and Extended Data Fig. 4), and correlates with the abundance of the archaeal phylum Methanobacteriota. The second principal coordinate (PCo2) captures site differences (Fig. 1b), generally ordering samples along a gradient corresponding to site population densities, subsistence strategies, environments and sociodemographic factors. The exception to this gradient is Nairobi, Kenya, which is a dense urban site yet it falls in the middle of the gradient. This second axis is correlated with the abundance of Spirochaetota and Elusimicrobiota, phyla that are described to decrease in relative abundance with industrialization[40]. Relative to external cohorts from western Europe and Japan, individuals in the AWI-Gen 2 cohort have higher relative abundance of Verrucomicrobiota and Spirochaetota and lower relative abundance of Actinomycetota (Extended Data Fig. 4).

Seeking to identify the geography, disease and lifestyle factors that have the greatest effect on compositional variation, we performed distance-based redundancy analysis with the available covariates, excluding highly correlated variables (Extended Data Fig. 5 and Methods). Site explains the greatest amount of compositional variation (7.92%), followed by other variables inherently related to the microbiome, such as recency of antibiotic use (0.79%), recency of diarrhoea (0.59%), use of deworming medication (0.51%) or probiotics use (0.46%). Interestingly, HIV status is the only disease-related variable explaining a sizable amount of variation (0.52%). Other disease variables included arthritis, obesity and hypertension among others (Extended Data Fig. 5).

To delve deeper into differences between sites, we investigated microbial diversity and abundance. Prokaryotic diversity differs significantly between sites (Kruskal–Wallis $P \leq 2.2 \times 10^{-16}$) (Fig. 1c), mirroring the site gradient observed for the second principal coordinate, with

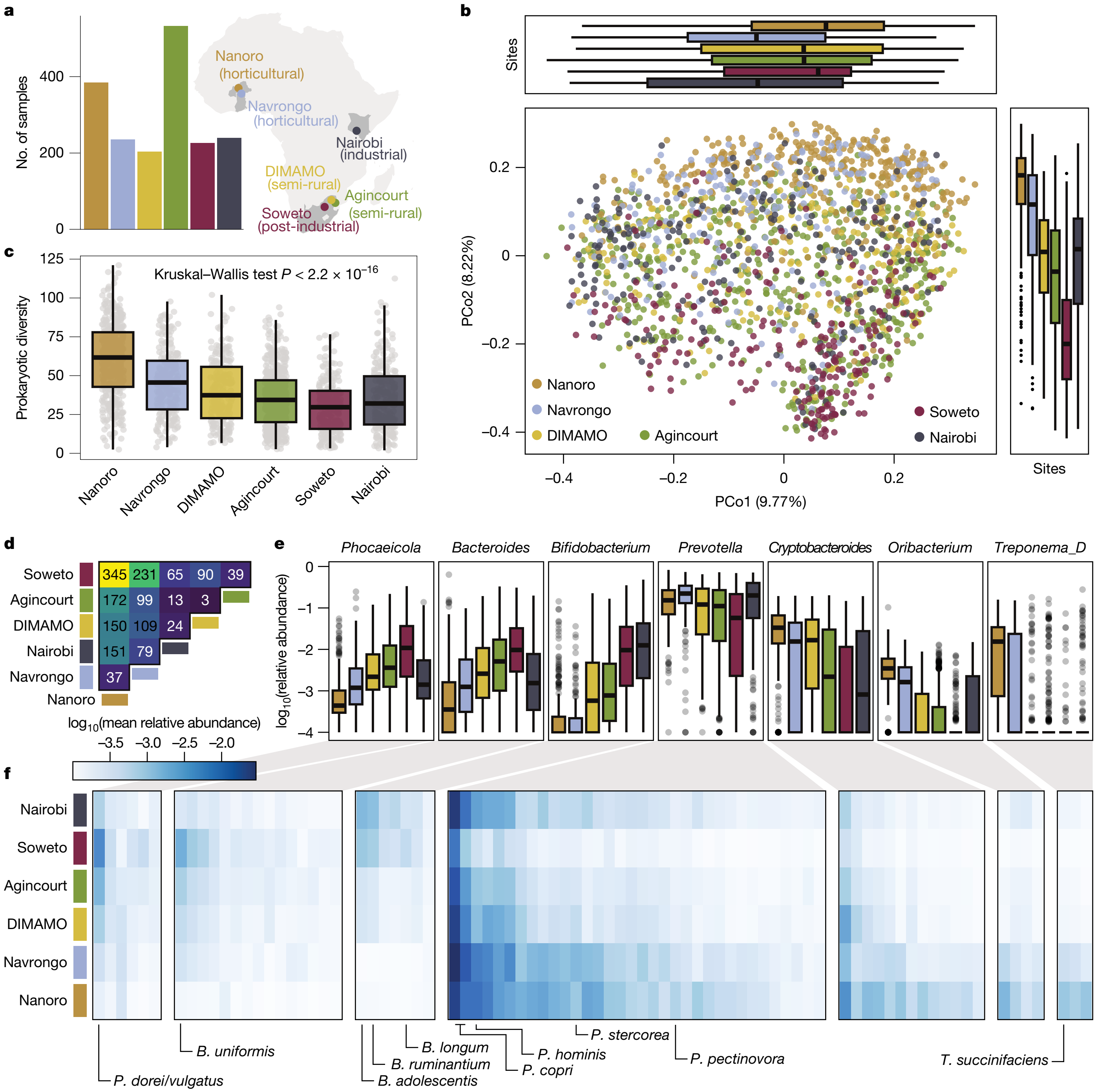


Fig. 1 | Microbiome composition and diversity in the AWI-Gen 2 cohort. **a**, Sample number and location of each study site. Countries containing sites are dark grey. **b**, Principal coordinate analysis of all samples on the basis of Bray–Curtis distance on species-level prokaryotic profiles. Study site is colour-coded and the boxplots show the samples per site projected onto the first and second principal coordinate. **c**, Prokaryotic diversity (inverse Simpson's index after rarefaction) per site (Kruskal–Wallis test, $P < 2 \times 10^{-16}$, $n = 1{,}796$ after quality control and removing data from male individuals). **d**, Heatmap showing the number of prokaryotic species with high generalized fold change between sites; sites are clustered on the basis of this number of species. **e**, The $\log_{10}$(relative abundance) of genera with the highest variance in fold change and median across sites. **f**, The $\log_{10}$ of the mean relative abundance per site is shown for all species within the genera shown in **e**. For *Prevotella*, *Oribacterium*, *Cryptobacteroides* and *Treponema*, all species with scientific names are highlighted; only the top abundant species with scientific names are indicated for the other genera. All panels represent data from $n = 1{,}796$ biologically independent samples. Boxplot boxes denote the interquartile range (IQR), thick black lines indicate the median, and whiskers indicate the most extreme points within 1.5-fold IQR. Supplementary Methods contain photographs and further information for each site.

Nairobi again falling out of sequence. Many taxa have distinct relative abundance and prevalence patterns between study sites (Supplementary Data 3). Few taxa are differentially abundant between sites with similar population densities and subsistence strategies (Agincourt and DIMAMO, 3 species; Nanoro and Navrongo, 37 species) and several are differentially abundant between sites with distinct characteristics (Soweto and Nanoro, 345 species) (Fig. 1d and Supplementary Data 4). Among the bacterial genera that varied most across study sites (Methods), some have relative abundance that is positively correlated with the study site gradient observed previously, such as *Phocaeicola*,

Table 1 | Site-level health and demographic summary statistics

	Nanoro, Burkina Faso	Navrongo, Ghana	DIMAMO, South Africa	Agincourt, South Africa	Soweto, South Africa[a]	Nairobi, Kenya
HDSS catchment area, km^2	594	1,675	545	420	200	6
HDSS catchment area population	63,000	156,000	36,000	115,000	1,200,000	75,000
Population density per km^2	105	91	113	274	6,357	14,833
Map coordinates	12.68N, 2.19W	10.89N, 1.09W	23.72S, 29.78E	24.82S, 31.26E	26.24S, 27.84E	1.25S, 36.89E 1.31S, 36.87E
Altitude, m	313	196	1,250	400–600	1,632	1,790
Household size, median (IQR)	12 (7, 20)	6 (5, 7)	5 (3, 7)	5 (3, 7)	4 (2, 6)	4 (2, 6)
Electricity (%)	3.9	33.9	98.0	99.6	98.2	82.9
Age, median (IQR)	50 (48, 53)	55 (52, 61)	53 (50, 57)	61 (54, 68)	54 (50, 58)	50 (48, 53)
Prevalence of HIV in this study						
Seronegative, *n* (%)	–	–	–	341 (80.6)	164 (75.6)	216 (90.4)
Seropositive, +ART, *n* (%)	–	–	–	60 (14.2)	50 (23.0)	19 (7.9)
Seropositive, −ART, *n* (%)	–	–	–	22 (5.2)	3 (1.4)	4 (1.7)
Body mass index, mean ± s.d.	21.3 ± 3.7	21.1 ± 4.0	33.7 ± 8.0	29.8 ± 6.8	32.8 ± 6.7	29.6 ± 6.1

Geographic and population density statistics summarize the catchment area of the sites' surveillance area. Note that the catchment area of the site does not always correspond exactly to administrative units for which census data are available. See the Supplementary Methods for a full description of the sites and inclusion criteria. Household size, electricity, age, HIV and body mass index statistics are specific to the AWI-Gen 2 Microbiome Project study populations. All HIV+ participants were women; body mass index is shown for women only.
For normally distributed parameters, values represent the mean ± s.d.; for not normally distributed parameters, values represent the median with IQR (P25, P75). Categorical variables are represented by counts and percentages. GPS coordinates (decimal degrees) are listed twice when catchment area surveys from two distinct areas.
+ART, participant reports currently receiving antiretroviral therapy; −ART, participant does not report currently receiving antiretroviral therapy.
[a] The Soweto study area is managed by the SAMRC/Wits Developmental Pathways to Health Research Unit, not an HDSS.

Bacteroides and *Bifidobacterium*, and others correspond to the inverse gradient, such as *Prevotella*, *Cryptobacteroides*, *Oribacterium* and *Treponema_D* (Fig. 1e). These abundance gradients are often shared among the bacterial species within each genus (Fig. 1f), and are also reflected in overall species prevalence (Extended Data Fig. 6a), where some taxa are ubiquitous and others have increasing or decreasing prevalence across the site gradient.

The comparison between Nairobi and Soweto, which are both industrial or post-industrial urban sites, is particularly interesting. Despite very high population density in Nairobi, individuals from Nairobi often have more similar microbiome composition to individuals from Agincourt and DIMAMO than to individuals from Soweto (Fig. 1d–f and Extended Data Fig. 6a,b). The relative abundance of *Phocaeicola*, *Bacteroides, Prevotella* and *Treponema_D* species is somewhat similar between individuals in Nairobi and individuals from semi-rural sites. By contrast, *Bifidobacterium* species have high relative abundance in both Soweto and Nairobi populations. Crucially, these findings illustrate that 'urbanization' or 'industrialization', which are commonly cited variables that impact the microbiome in previous research including our own, cannot adequately capture lifestyle and environmental differences between different urban areas and between different rural areas.

Altogether, site is the dominant factor explaining overall microbiome composition. Disease and medication variables explain smaller amounts of microbiome composition, and HIV status represents one of the largest subsequent contributors to microbiome variation after site. These findings indicate that study sites represent varied subsistence strategies, industrialization levels, health-care access and overall adversity that together affect microbiome composition. These findings also suggest a model for microbiome transition dynamics. Taxa such as *Treponema_D* are less abundant in sites practising large-scale agriculture or industry, whereas *Bacteroides* and *Phocaeicola* species gradually expand in abundance and *Prevotella* species gradually decrease. Despite reports in other cohorts[41], we do not observe mutual exclusion between *Prevotella* and *Bacteroides* (Extended Data Fig. 6c). We observe interesting taxonomic profiles in participants from Nairobi, many of whom probably migrated to Nairobi from rural parts of Kenya[38,42] into the informal settlements: high *Prevotella* abundance may reflect a microbial signature retained from participants' former rural residences, whereas the high abundance of *Bifidobacterium* species and low abundance of *Cryptobacteroides* species may reflect taxa that are strongly influenced by changing environments.

Novel prokaryotic genomes

African gut microbes are under-represented in public reference collections, and when present, are often sourced from relatively isolated populations with lifestyle practices that are not representative of the African continent. To identify previously unknown taxa in the AWI-Gen 2 sample collection, we performed metagenomic assembly and binned contigs into metagenome-assembled genomes (MAGs), yielding a total of 69,539 genomes, of which 34,215 genomes are high quality (more than 90% complete and less than 5% contaminated) and 26,660 are medium quality (more than or equal to 50% complete, less than 10% contaminated). To condense redundant genomes, we dereplicated all MAGs with a minimum genome completeness of 50% and maximum genome contamination of 5% at 95% average nucleotide identity (ANI). The resulting 2,613 MAGs span 19 bacterial phyla (Fig. 2a and Supplementary Data 5). We constructed a protein catalogue from all medium- and high-quality MAGs clustered at 95% amino acid identity, yielding 63.8M unique proteins.

We compared our prokaryotic genome and protein catalogues to the Unified Human Gastrointestinal Genome (UHGG) catalogue of 4,744 prokaryotic species representatives and the Unified Human Gastrointestinal Protein 95 (UHGP95) catalogue of 20.5M proteins. The AWI-Gen 2 dataset includes 1,005 new prokaryotic MAGs relative to UHGG (Fig. 2b,c), and 7.6M new proteins relative to UHGP95 (Extended Data Fig. 7a). Most new bacterial MAGs fall under the phyla Bacillota A, Actinomycetota and Bacillota. We also observe 29 unique MAGs from the archaeal phyla Methanobacteriota, Thermoplasmatota and Halobacteriota (Fig. 2d), and nine are not found in the UHGG, indicating that the AWI-Gen 2 population contains substantial archaeal novelty. Most individual samples yielded several previously unknown prokaryotic genomes and tens of thousands of new proteins relative to reference

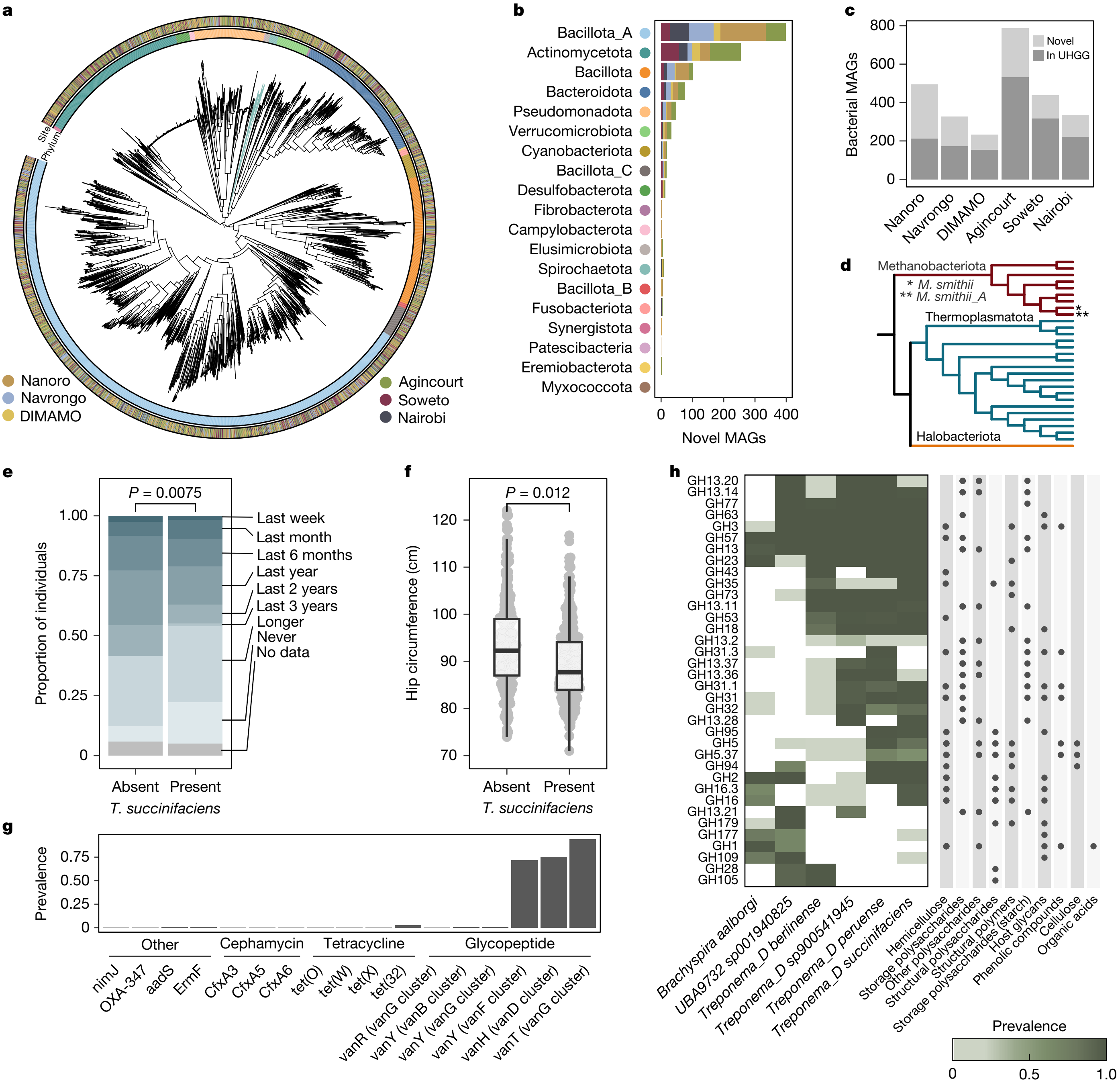


Fig. 2 | Prokaryotic novelty and features. a, Phylogenetic tree of 2,584 dereplicated bacterial MAGs. Outer ring indicates study site of origin, inner ring indicates GTDB phylum and teal branches indicate Spirochaetota. **b**,**c**, Total number of previously unknown bacterial genomes by phylum (**b**) and new and existing bacterial genomes in the AWI-Gen assemblies (**c**), relative to the UHGG collection. Only representative genomes are shown. **d**, Archaeal phyla and species found in the AWI-Gen 2 genome collection. **e**,**f**, Self-reported antibiotic use (**e**) and hip circumference in centimetres (**f**) of $n = 617$ individuals from Nanoro, Burkina Faso and Navrongo, Ghana with and without *T. succinifaciens* present in the gut microbiome. Differences tested with a linear model that adjusted accounting for site (**e**) and for site and antibiotic history (**f**) as random effects. Boxplot boxes denote the IQR, thick black lines indicate the median and whiskers indicate the most extreme points within 1.5-fold IQR. **g**, Prevalence of antibiotic resistance genes in $n = 244$ *T. succinifaciens* MAGs, ordered by drug class. **h**, Prevalence of glycoside hydrolase genes among the six species of Spirochaetota with the largest number of MAGs in the AWI-Gen 2 genome collection. Only glycoside hydrolases present in at least 5% of the genomes are shown.

collections (Extended Data Fig. 7b,c), with individuals in Nanoro contributing the most new genomes per individual.

Relative to recent large gut microbiome metagenomic datasets from other populations around the globe[43–47], 598 MAGs are unique to the AWI-Gen 2 study, a much larger number than those contributed by other studies from outside the African continent (Extended Data Fig. 7d). Further, rarefaction analysis of the prokaryotic genomes (Extended Data Fig. 7e) and proteins (Extended Data Fig. 7f) generated from samples across each of the six study sites indicates that no feature has reached saturation. Crucially, these results imply that further measurement of gut microbiomes in these communities will continue to show new microbiome diversity.

The extensive AWI-Gen 2 microbial genome catalogue enables investigation of taxa that cannot be studied using standard microbiological techniques. One example is *T. succinifaciens*, a commensal anaerobic bacterial species in the phylum Spirochaetota that is non-spore forming, consumes a wide range of sugars and produces short-chain fatty acids and succinate[48]. *T. succinifaciens* is thought to be present in rural hunter-gatherer, pastoralist and agriculturalist populations, and lost in urban populations[17,40]. The first phase of AWI-Gen identified that *T. succinifaciens* is indeed present in urban populations[33]; however, we observe that *T. succinifaciens* abundance is inversely correlated with population density (Fig. 1e). Despite the emerging interest in this gut commensal, *T. succinifaciens* dispersal and acquisition are poorly understood.

Treponema are difficult to culture, making MAG catalogues an invaluable resource for understanding their biology. One complete genome from a cultured *T. succinifaciens* is available[48], and this type strain was isolated from the swine gut[49]. Only 71 high-quality *T. succinifaciens* MAGs exist in UHGG, predominantly from human gut samples from populations in Madagascar, Peru and Fiji. Our genome catalogue includes 244 high-quality MAGs for *T. succinifaciens*, primarily from sites with low population density (Extended Data Fig. 8a and Supplementary Data 6). *T. succinifaciens* genomic features are characteristic of host-adapted microbes: regardless of site of origin, genome lengths are relatively small (2.52 ± 0.15 megabases; Extended Data Fig. 8b), and their shared core genome of 1,589 genes constitute most of each genome (68.59% ± 3.95%; Extended Data Fig. 8c). *T. succinifaciens* genomes from this study and others[43,50] cluster by geographic origin, demonstrating a strong phylogeographic signal (Extended Data Fig. 8d; $P = 0.001$), further supporting a hypothesis of limited environmental dispersal.

We sought to further explore *T. succinifaciens* presence and absence in the AWI-Gen 2 population. In Nanoro, Burkina Faso and Navrongo, Ghana, where *T. succinifaciens* prevalence is high, individuals with *T. succinifaciens* reported less recent antibiotic use ($P = 0.0075$; Fig. 2e), and had a lower hip circumference ($P = 0.012$; Fig. 2f). Consistent with *T. succinifaciens* absence among individuals with recent antibiotic use, *T. succinifaciens* genomes have low prevalence of antibiotic resistance genes, excepting three genes related to vancomycin resistance (Fig. 2g). As vancomycin inhibits cell-wall synthesis in Gram-positive bacteria and *T. succinifaciens* is a Gram-negative bacterium, these genes may not be directly related to vancomycin resistance, but instead related to cell-wall synthesis and modification. Although this analysis does not capture antibiotic resistance genes on mobile genetic elements, it suggests limited antimicrobial resistance within the core genome. The association between *T. succinifaciens* presence and lower hip circumference may imply a connection between diet, nutrient availability and *T. succinifaciens* persistence. CAZyme profiling of high-quality MAGs within the Spirochaetota phylum demonstrates a broad glycoside hydrolase repertoire across species, and that *T. succinifaciens* has capacity to degrade hemicellulose and starch (Fig. 2h), indicating potential for *T. succinifaciens* loss during shifts to low fibre diets. We also observe that 58% of *T. succinifaciens* glycoside hydrolases are found in 95% of the *T. succinifaciens* genomes, implying that these genes are part of the core genome and probably vertically inherited rather than horizontally acquired. Together, these findings connect human exposure and phenotype data to *T. succinifaciens* persistence, and demonstrate how MAG catalogues built from populations in understudied areas can be used to investigate bacterial biology.

Viral fraction across sites

Although most gut microbiome research focuses on prokaryotes, this sample collection also represents a source of viral diversity. We generated a viral genome catalogue from all metagenomic assemblies and clustered all genomes at 95% ANI (Supplementary Data 7). Of 44,506 viral genomes, 381 are present at the assembly level in at least 18 individuals (about 1% prevalence) (Fig. 3a), and 2,701 (4.65% of the catalogue) are observed in at least 1% of participants in at least one study site. Similar to bacterial taxa, some genomes have higher prevalence in South African sites and others have higher prevalence in Nanoro and Navrongo. We compared the viral catalogue with the Metagenomic Gut Virus (MGV) catalogue of 54,118 viral operational taxonomic units, and find 40,135 new viruses relative to MGV (Fig. 3b). On average, each individual microbiome yielded dozens of new genomes relative to MGV (Extended Data Fig. 9a) and relative to a more recent viral catalogue[51] (Extended Data Fig. 9b). Rarefaction indicates that this catalogue has not saturated viral discovery among the AWI-Gen populations (Extended Data Fig. 9c). Viral richness, measured as the number of assembled genomes per sample, does not follow the same population density site gradient as prokaryotic richness (Fig. 3c), even though prokaryotic and viral richness correlate well (Extended Data Fig. 9d). These findings were independent of sequencing depth (Extended Data Fig. 3b) and richness trends are similar with reference-based phage profiling (Extended Data Fig. 9e).

Crassvirales, an order of abundant gut dsDNA bacteriophage, is prevalent across all sites (Fig. 3d). This finding is consistent with previous descriptions of crAss-like genera having 77% global prevalence[52]. P-crAssphage, the first discovered representative of the clade, is more prevalent in Soweto relative to other sites, again consistent with previous findings of low p-crAssphage prevalence in populations residing outside urban and highly industrial contexts[53]. We also identified several jumbophages (phages with genomes larger than 200 kilobases (kb)) and identified nine previously unknown jumbophages, with stringent thresholds of maximum alignment length of less than 10% to any genome in MGV and presence at the assembly level in at least five individuals (Fig. 3e). These jumbophage genomes largely consist of genes with no predicted functional annotation, but contain several features that relate to persistence in the host, including CRISPR arrays, sporulation regulators, addiction module toxins and large suites of tRNAs (Supplementary Data 8). These phages are prevalent in the studied cohort, with all reported new jumbophages reaching a prevalence of at least 5% in at least one site (Extended Data Fig. 9f). Further, one phage (phage A) shows evidence of integration into a *Clostridium* sp., and both phages A and G have unidirectional gene orientation. These results indicate the existence of several highly prevalent jumbophages that evaded previous discovery due to the narrow scope of previous population studies.

Taxonomic associations with HIV status

Finally, we investigated the relationship between the gut microbiome and HIV status. HIV represents one of the biggest public health concerns in the Kenyan and South African AWI-Gen study populations, especially as a rapidly increasing number of people aged 50 and older are living with HIV because of high antiretroviral therapy (ART) uptake: HIV prevalence was 17.2% among individuals aged 50–64 years in South Africa in 2017[54] and 9.1% among individuals aged 45–54 years in Kenya[55]. Despite advances in ART that have reduced population viral load and transmission, viral suppression is not sufficient to control HIV-related mortality and morbidity[56], thus motivating deeper investigation into the gut microbiome as a possible mediator. Gut microbiota and their metabolites have been implicated in HIV-related inflammation and immune activation: gut-associated lymphoid tissue serves as a main reservoir for HIV[57], and gut microbial metabolites can promote HIV transcription[58]. In turn, HIV infection can diminish epithelial barrier integrity[59], allowing for microbial translocation that promotes immune activation and chronic inflammation[60]. Moreover, obesity has become a notable problem for individuals on the latest generation of ART such as dolutegravir[61]. In HICs, microbiome dysbiosis in people living with HIV (PLWH) is characterized by an enrichment of Pseudomonadota and depletion of Bacteroidota[60,62]. Perhaps confounded by a Prevotella-enrichment signature often observed in men who have sex with men[63], it has even been described that the microbiomes of PLWH seem more similar to

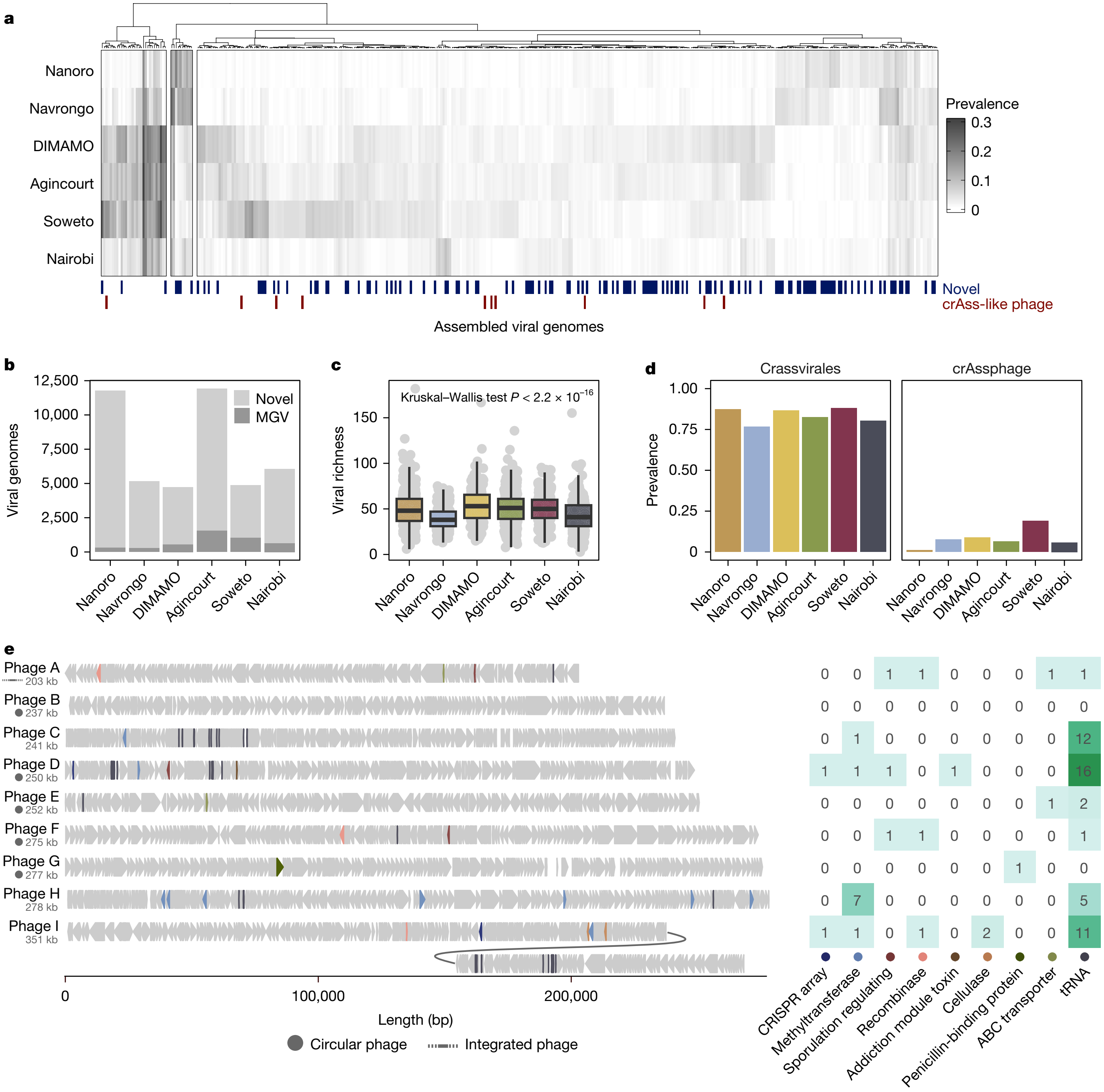


Fig. 3 | Viral novelty and diversity in the AWI-Gen 2 cohort. a, Prevalence of viral genomes found in at least 18 individuals (approximately 1% of the AWI-Gen 2 population). Prevalence is measured as the proportion of the population in a given site that yielded an assembled viral genome that shares 95% ANI with the representative viral genome. Viral genomes that are new relative to the MGV catalogue and viral genomes that fall under the Crassvirales order are highlighted. **b**, Total number of new and existing viral genomes relative to MGV. **c**, Phage richness (number of phage species clusters present in each sample) per site (Kruskal–Wallis test $P < 2 \times 10^{-16}$, $n = 1{,}796$). Boxplot boxes denote the IQR, thick black lines indicate the median and whiskers indicate the most extreme points within 1.5-fold IQR. **d**, Prevalence of Crassvirales viruses and prototypical crAssphage by site, determined by read-level abundance. **e**, Genome maps of nine previously unknown jumbophages with genome annotations and length in kb, and count of notable genetic features in each genome.

seronegative individuals in agrarian populations[64]. Few studies have measured associations between the gut microbiome and HIV status in African populations[14,65–67], where baseline microbiome composition and disease profiles are distinct from those observed in HICs. Instead, studies have largely focused on men in HICs and have often been confounded by sexual practice.

We compared microbiome composition between women living with HIV on ART and seronegative (HIV−) women in Agincourt (PLWH $n = 60$; HIV− $n = 341$), Soweto (PLWH $n = 50$; HIV− $n = 164$) and Nairobi (PLWH $n = 19$; HIV− $n = 214$) (Fig. 4a). HIV status was not assessed in Nanoro and Navrongo because of low population prevalence, and DIMAMO was excluded from this analysis because only six individuals were found to be living with HIV. Our dataset also included participants with positive HIV status, but not self-reporting as receiving ART (Supplementary Data 9), possibly because they learned of their HIV diagnosis in the course of participation within AWI-Gen 2. Because of the low number

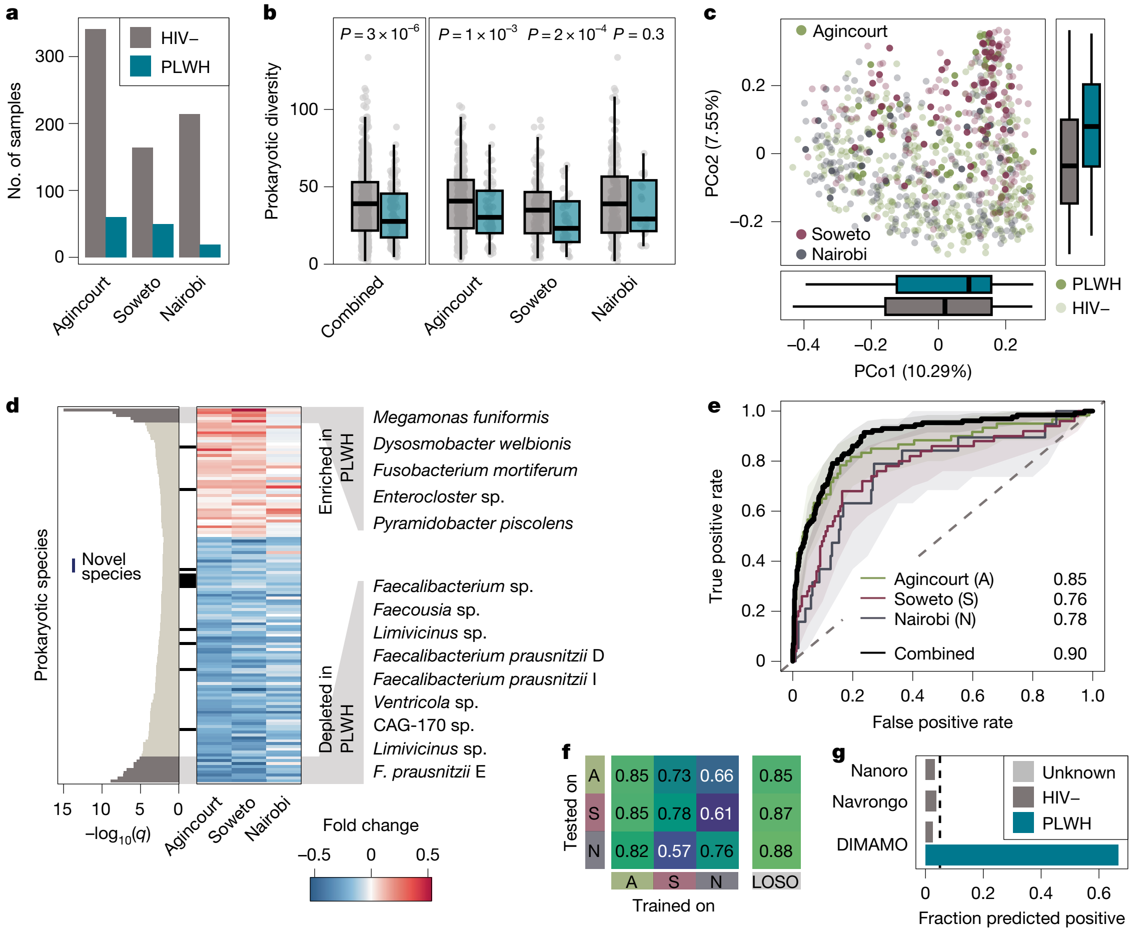


Fig. 4 | Microbial composition and diversity in PLWH. a, Number of seronegative individuals (HIV−) and PLWH on antiretroviral treatment. **b**, Prokaryotic diversity (inverse Simpson's index on rarefied counts) by site and HIV status. Points represent individual samples. Differences in alpha-diversity for each individual site were tested with two-sided analysis of variance and for all sites combined with a linear mixed effect model accounting for site as a random effect. **c**, Principal coordinate analysis of species-level Bray–Curtis distance. Points represent individual samples, coloured by site, and PLWH are shaded. Boxplots show samples by HIV status projected onto the principal coordinates. **d**, Differentially abundant species (q-value < 0.01) determined by a linear mixed effect model. Species with q-value $<1 \times 10^{-5}$ are annotated. Shading indicates abundance fold change between seronegative individuals and PLWH. Black bars indicate previously unknown taxa from this study. **e**, ROC for machine learning models trained to distinguish HIV status on samples from each site or for all samples. Shading indicates the 95% confidence interval and numbers indicate AU-ROC. **f**, AU-ROC for machine learning model evaluation. Models trained on each site were applied to the other sites and external predictions were evaluated by means of AU-ROC. In LOSO validation, models were trained on two sites and validated on the left-out site. **g**, Fraction of samples from other sites predicted to be positive calibrated at a 5% false positive rate (indicated by dashed black line). For DIMAMO, HIV status is known and false positive rate and true positive rate can be evaluated. Serostatus is unknown for individuals in Nanoro and Navrongo but is expected to be below 2%. Boxplot boxes denote the IQR, thick black lines indicate the median and whiskers indicate the most extreme points within 1.5-fold IQR.

of ART-naive PLWH in our dataset, we focused our analysis on HIV− and ART+ PLWH participants (see Extended Data Fig. 10 for comparison between ART+ and ART− PLWH). Prevalence of tuberculosis co-infection was low among PLWH ($n = 6$ individuals reported being on active treatment for tuberculosis) (Supplementary Data 9). Beyond co-infection, we do not expect many demographic differences between the HIV− and PLWH populations, but observe that PLWH are younger than HIV− individuals in Agincourt and Soweto, and have lower cholesterol in Nairobi and slightly lower waist-to-hip ratio in Agincourt (Extended Data Fig. 10).

First, we compared the microbiome composition of PLWH and HIV− individuals using alpha- and beta-diversity metrics. Consistent with previous descriptions[14], alpha-diversity is lower in PLWH overall and within each study site (Fig. 4b). In terms of the beta-diversity, HIV status significantly varies over the second principal coordinate axis (PCo1 $P = 0.10$, PCo2 $P = 6.2 \times 10^{-12}$, Wilcoxon test) but is again outweighed by site differences (PCo1 $P = 1.4 \times 10^{-9}$, PCo2 $P = 2.3 \times 10^{-28}$, Kruskal–Wallis test) (Fig. 4c).

To identify specific taxa that vary with HIV status, we performed differential abundance testing with a linear mixed effect model that accounted for confounders such as site, antibiotic treatment or recency of diarrhoea. After correcting for multiple testing with the Benjamini–Hochberg procedure reported with a q-value, 131 prokaryotic species had significant differences with HIV status (q-value < 0.01) (Fig. 4d and Supplementary Data 10), most of which were lower in PLWH, agreeing with the finding of lower prokaryotic diversity in PLWH. Overall, the effect sizes generally agree across sites, with Nairobi exhibiting smaller effect sizes because of a lower number of PLWH. Some of the most significantly associated (q-value $< 1 \times 10^{-5}$) (Fig. 4d) taxa have been associated with HIV status in other cohorts: *Faecalibacterium prausnitzii* is a known butyrate producer that has been associated with reduced inflammatory biomarkers, and has been negatively associated with HIV status previously, probably because of increased oxygen levels in the gut during HIV infection[68,69]. By contrast, the genus *Fusobacterium* has well-described pro-inflammatory associations with HIV and other diseases[70], and has been associated with poor immune recovery following ART administration[71] and shown potential to reactivate latent HIV[58]. Interestingly, other taxa that are negatively associated with HIV, including *Ventricola* sp., *Faecousia* sp. and *Limivicinus* sp., are better represented in metagenomic studies focused on livestock[72]. Other taxa that are positively associated with HIV have conflicting associations with inflammatory disorders: *Dysosmobacter welbionis* is a recently isolated gut bacterium that has not been characterized in the context of HIV, but has been shown to counteract diet-induced obesity and improve glucose tolerance[73,74]. *Megamonas uniformis* has been described as enriched in PLWH[75], but has discordant associations with inflammation and obesity[76,77]. These results highlight the value of investigating microbiome and disease associations in broader cohorts, as we observe several taxa that are not present in studies conducted in HICs.

To explore which microbial features can differentiate HIV status within and across sites, we trained machine learning models for each site individually and for all data combined (Fig. 4e,f). The goal of these models is to identify features that are strongly associated with HIV status in one or several populations as targets for future study; it is not proposed as a diagnostic. The models achieved accurate distinction between HIV− and PLWH individuals in all sites, yielding area under the receiver operating characteristics (AU-ROC) of more than 0.75 in all cases. When transferred across sites, only the model trained on data from Agincourt maintains high classification accuracy (Fig. 4f), perhaps because of the larger dataset available for model training. In line with this hypothesis, when data from two sites were combined in a leave-one-site-out (LOSO) validation, samples from the left-out site are accurately classified (AU-ROC ≥ 0.85 in all cases), even when samples from Agincourt are not used for training. As another test for generalization, we calibrated the model trained on all data to an internal 5% false positive rate and applied it to samples from Nanoro, Navrongo and DIMAMO (Fig. 4g). Even though information about HIV status in Nanoro and Navrongo is not available, we can assess the fraction of samples predicted to be HIV positive, which we would not expect to exceed 5% given the population HIV prevalence in these sites and our model calibration. Indeed, the model predicts very few samples to be HIV positive (3.1% in Nanoro, 3.7% in Navrongo), highlighting its specificity, and correctly classifies two thirds of the HIV-positive samples from DIMAMO (Fig. 4g).

We also considered the viral fraction of the microbiome of PLWH. Phage richness is lower in PLWH on ART compared with seronegative individuals, and is not significantly different between PLWH who are receiving ART and who are ART-naive (Extended Data Fig. 10). We observe 89 phages that show significant differences in abundance with HIV status (q-value < 0.01) (Supplementary Data 11), but find that machine learning models trained on the viral features achieve less accurate distinction between seronegative individuals and PLWH than models trained on prokaryotic features (Extended Data Fig. 10). Future work may identify which viral features are independent of their host bacterial abundance and have independent associations with HIV status.

Overall, HIV-associated microbiome differences are consistent across study settings, despite the strong effect of study context on overall microbiome composition. Some strongly associated taxa have been described in the context of HIV previously, but we also identify taxa that have not been well-described in human gut microbiomes or in microbial associations with disease. Although single-time-point sampling cannot determine causality or whether microbiome changes precede or follow HIV infection and ART administration, these results lend insight into possible targets for gut remediation.

Discussion

In 2007, the Human Microbiome Project set the goal of measuring the human microbiome and its contribution to disease[6]. Subsequent studies have built upon this goal, studying the human microbiome in large cohorts in HICs[8,9]. Here, 17 years later, the AWI-Gen 2 Microbiome Project is a landmark collaborative research effort that extends these goals to diverse LMIC populations in Africa: rural and predominantly horticultural areas (Nanoro, Burkina Faso and Navrongo, Ghana), rapidly transitioning rural areas (Agincourt and DIMAMO, South Africa), urban industrial informal settlements (Viwandani and Korogocho settlements in Nairobi, Kenya) and an urban post-industrial settlement (Soweto, South Africa). This is the largest cross-sectional and population-based survey of gut microbiome composition in relation to human health, environment and disease in low- and middle-income settings, and will prove invaluable in future microbiome discovery research.

This study enables comparison between populations that span a range of subsistence strategies and resource access. Site has a strong effect on microbiome variation, with alpha-diversity and taxon prevalence correlating with gradients in population density and resource access. Unexpectedly, we observe differences between sites that have similar subsistence strategies and industrialization levels. For example, *Bifidobacterium* and *Cryptobacteroides* have similar abundance in Nairobi and Soweto, whereas abundances of *Prevotella*, *Bacteroides* and *Phocaeicola* in Nairobi are more typical of the rural and semi-rural sites. HDSS data can contextualize these findings: Nairobi informal settlements have high in- and out-migration rates, and extensive circular migration[38,42]. Paired phenotype and metagenomic data also provide insights into taxonomic composition. For example, we identify genomic and host phenotype features that define a diet-related nutrient niche that supports *T. succinifaciens* persistence, and predict sensitivity to antibiotics that may drive *T. succinifaciens* loss. Here and previously[33], we have found *Treponema* in the guts of urban individuals, perhaps because of low rates of antibiotic exposure or high rates of circular migration between urban and rural areas. These findings underscore the complex interplay between subsistence, industrialization and lifestyle factors in shaping the gut microbiome.

Our shotgun sequencing approach yielded assembly-based discovery of 40,135 previously unknown viral genomes. We identify nine new and prevalent jumbophages, including a putative integrated jumbophage, and jumbophages with intriguing unidirectional gene organization. To our knowledge, unidirectional gene organization is undescribed in jumbophages, and may be indicative of phage integration into a host with leading strand-biased gene distribution; indeed, the flanking regions of putative integrated Phage A map to a *Clostridium* species, and Bacillota have been previously described to have strong strand-biased gene distribution[78]. These findings provide an exciting opportunity for deeper study of previously unknown phages and their hosts.

AWI-Gen's population-representative enrollment paired with clinical and lifestyle information also support the capture of population-level disease associations. HIV is prevalent in South Africa and Kenya, and improved viral load management with ART is not sufficient to protect against HIV-associated comorbidities[56]. This study represents one of the largest microbiome studies of women with HIV so far, and provides unique insights into microbiome–HIV associations in LMICs. Several taxa are enriched in PLWH receiving ART relative to seronegative participants, including taxa that have not been well-described in the context of HIV and inflammation. We cannot conclude whether these taxa are enriched in response to HIV infection, HIV-related comorbidities or antiretroviral medication, or whether taxa pre-existed HIV infection because of lifestyle or exposure differences in at-risk individuals. These results further demonstrate that existing disease associations are probably not broadly portable across global populations, and more research is necessary to disentangle the effects of HIV infection and other confounding variables on microbiome composition.

We emphasize that the AWI-Gen 2 Microbiome Project does not exhaustively represent any country or region. There is tremendous diversity within LMIC populations, and population density alone is not a sufficient indicator of microbiome composition or population lifestyle. Rather, the microbiome field needs to improve representation of LMIC populations to maximize sampling diversity and ensure portability of study findings. Even within this study, we focus on older adult women. Although sex differences have less impact on microbiome composition than other factors[79], these studies may not capture health and lifestyle differences that may exist between sexes in LMICs. We also specifically highlight key variables that explain the greatest amount of microbial variation, leaving several disease and lifestyle variables open for future investigation. For example, we do not consider diet, which varies greatly across the AWI-Gen study populations and across LMIC contexts. We also acknowledge the limitations of identifying causality in microbiome associations with disease within this study design. We anticipate that longitudinal sampling of these populations and others will improve our understanding of the timing of microbiome changes in relation to HIV infection and other diseases. Further studies can also incorporate other microbiome measurements, such as eukaryotic profiling, to explore the complex relationship between eukaryote infection and immune activation, or total microbial concentration quantification to shed light on whether the taxonomic shifts observed in this study are due to blooms or losses of specific taxa. Future work can leverage the extensive AWI-Gen 2 participant data and more quantification methods to investigate the interplay between the microbiome and host genetics, environmental exposures, health status and participant demographics.

We strove to conduct the AWI-Gen 2 Microbiome Project ethically and equitably, taking into account recommendations for ethical research partnerships[26,27,80]. AWI-Gen hires field workers locally through the community-embedded infrastructures, and study staff host community advisory group discussions before study onset and return results to participants upon study completion. Through community discussions, the research team can identify pressing health issues within each study centre and ensure that research questions prioritize community needs. This study represents a strong scientific partnership between Stanford University in the USA and University of the Witwatersrand in South Africa. Trainees and faculty from both groups contributed to study design and data analysis, and a trainee from each institution has participated in a one-year research exchange with mentorship from both institutions. Further, the team has led three microbiome and bioinformatic training workshops to support genomics research capacity in South Africa. Altogether, this study illustrates that equitable research and impactful science do not represent a ‘zero-sum’ trade-off, but in fact lead to more robust research with benefit-sharing among all stakeholders.

The AWI-Gen 2 Microbiome Project contributes to advancing the investigation of human gut microbiomes from diverse populations around the globe. The study provides extensive opportunities for continued exploration, including identifying microbiome and disease associations, measuring human genetic contributions to microbiome composition, and defining lifestyle factors that shape microbial community assembly. Future studies can leverage these data, along with the foundation for community-engaged and equitable research described herein, to close the gap in global representation in microbiome research. Moreover, there is every reason to anticipate that the platforms and findings that emerge will enhance disease management, health and wellbeing among communities living in a diversity of contexts.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-08485-8.

1. World Bank Open Data. *Population, total – Low & middle income, High income*. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=XO-XD (2023).
2. Brewster, R. et al. Surveying gut microbiome research in Africans: toward improved diversity and representation. *Trends Microbiol.* **27**, 824–835 (2019).
3. Allali, I. et al. Human microbiota research in Africa: a systematic review reveals gaps and priorities for future research. *Microbiome* **10**, 10 (2022).
4. Abdill, R. J., Adamowicz, E. M. & Blekhman, R. Public human microbiome data are dominated by highly developed countries. *PLoS Biol.* **20**, e3001536 (2022).
5. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
6. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
7. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
8. Gacesa, R. et al. Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732–739 (2022).
9. Salosensaari, A. et al. Taxonomic signatures of cause-specific mortality risk in human gut microbiome. *Nat. Commun.* **12**, 2671 (2021).
10. Zeng, Q. et al. Discrepant gut microbiota markers for the classification of obesity-related metabolic abnormalities. *Sci. Rep.* **9**, 13424 (2019).
11. Jie, Z. et al. A transomic cohort as a reference point for promoting a healthy human gut microbiome. *Med. Microecol.* **8**, 100039 (2021).
12. Lu, J. et al. Chinese gut microbiota and its associations with staple food type, ethnicity, and urbanization. *NPJ Biofilms Microbiomes* **7**, 71 (2021).
13. Yooseph, S. et al. Stool microbiota composition is associated with the prospective risk of *Plasmodium falciparum* infection. *BMC Genomics* **16**, 631 (2015).
14. Parbie, P. K. et al. Dysbiotic fecal microbiome in HIV-1 infected individuals in Ghana. *Front. Cell. Infect. Microbiol.* **11**, 646467 (2021).
15. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
16. Schnorr, S. L. et al. Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, 3654 (2014).
17. Obregon-Tito, A. J. et al. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* **6**, 6505 (2015).
18. Vangay, P. et al. US immigration westernizes the human gut microbiome. *Cell* **175**, 962–972.e10 (2018).
19. Ecklu-Mensah, G. et al. Gut microbiota and fecal short chain fatty acids differ with adiposity and country of origin: the METS-microbiome study. *Nat. Commun.* **14**, 5160 (2023).
20. Reyes, A. et al. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc. Natl Acad. Sci. USA* **112**, 11941–11946 (2015).
21. Rouhani, S. et al. Diarrhea as a potential cause and consequence of reduced gut microbial diversity among undernourished children in Peru. *Clin. Infect. Dis.* **71**, 989–999 (2020).
22. Vatanen, T. et al. A distinct clade of *Bifidobacterium longum* in the gut of Bangladeshi children thrives during weaning. *Cell* **185**, 4280–4297.e12 (2022).
23. Robertson, R. C. et al. The gut microbiome and early-life growth in a population with high prevalence of stunting. *Nat. Commun.* **14**, 654 (2023).
24. Hibberd, M. C. et al. Bioactive glycans in a microbiome-directed food for children with malnutrition. *Nature* **625**, 157–165 (2024).
25. De Wolfe, T. J., Arefin, M. R., Benezra, A. & Rebolleda Gómez, M. Chasing ghosts: race, racism, and the future of microbiome research. *mSystems* **6**, e0060421 (2021).
26. Oduaran, O. H. & Bhatt, A. S. Equitable partnerships and the path to inclusive, innovative and impactful human microbiome research. *Nat. Rev. Gastroenterol. Hepatol.* **19**, 683–684 (2022).
27. Mangola, S. M., Lund, J. R., Schnorr, S. L. & Crittenden, A. N. Ethical microbiome research with Indigenous communities. *Nat. Microbiol.* **7**, 749–756 (2022).
28. Ramsay, M. et al. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob. Health Epidemiol. Genom.* **1**, e20 (2016).
29. Mashinya, F., Alberts, M., Mashaba, R. G. & Tindana, P. O. Community engagement in genomics research; Challenges and lessons learnt in the AWI-Gen study at Dikgale Health and Demographic Surveillance System (HDSS) site, South Africa. *AAS Open Res.* https://doi.org/10.12688/aasopenres.13076.1 (2020).
30. Agongo, G. et al. Community engagement and feedback of results in the H3Africa AWI-Gen project: experiences from the Navrongo Demographic and Health Surveillance site in Northern Ghana. *AAS Open Res.* **4**, 15 (2021).
31. Ali, S. A. et al. Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for phase 1 of the AWI-Gen population cross-sectional study. *Glob. Health Action* **11**, 1507133 (2018).
32. Oduaran, O. H. et al. Gut microbiome profiling of a rural and urban South African cohort reveals biomarkers of a population in lifestyle transition. *BMC Microbiol.* **20**, 330 (2020).
33. Tamburini, F. B. et al. Short- and long-read metagenomics of urban and rural South African gut microbiomes reveal a transitional composition and undescribed taxa. *Nat. Commun.* **13**, 926 (2022).
34. Derra, K. et al. Profile: Nanoro Health and Demographic Surveillance System. *Int. J. Epidemiol.* **41**, 1293–1301 (2012).
35. Oduro, A. R. et al. Profile of the Navrongo Health and Demographic Surveillance System. *Int. J. Epidemiol.* **41**, 968–976 (2012).
36. Kahn, K. et al. Profile: Agincourt health and socio-demographic surveillance system. *Int. J. Epidemiol.* **41**, 988–1001 (2012).
37. Alberts, M. et al. Health & demographic surveillance system profile: the Dikgale Health and Demographic Surveillance System. *Int. J. Epidemiol.* **44**, 1565–1571 (2015).
38. Beguy, D. et al. Health & demographic surveillance system profile: the Nairobi Urban Health and Demographic Surveillance System (NUHDSS). *Int. J. Epidemiol.* **44**, 462–471 (2015).
39. Ware, L. J. et al. Social vulnerability, parity and food insecurity in urban South African young women: the healthy life trajectories initiative (HeLTI) study. *J. Public Health Policy* **42**, 373–389 (2021).
40. Angelakis, E. et al. *Treponema* species enrich the gut microbiota of traditional rural populations but are absent from urban individuals. *New Microbes New Infect.* **27**, 14–21 (2019).

41. Costea, P. I. et al. Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2018).
42. Beguy, D., Bocquier, P. & Zulu, E. M. Circular migration patterns and determinants in Nairobi slum settlements. *Demogr. Res.* **23**, 549–586 (2010).
43. Carter, M. M. et al. Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell* **186**, 3111–3124.e13 (2023).
44. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
45. Franzosa, E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
46. Schirmer, M. et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1125–1136.e8 (2016).
47. Ni Lochlainn, M. et al. Effect of gut microbiome modulation on muscle function and cognition: the PROMOTe randomised controlled trial. *Nat. Commun.* **15**, 1859 (2024).
48. Han, C. et al. Complete genome sequence of *Treponema succinifaciens* type strain (6091T). *Stand. Genomic Sci.* **4**, 361–370 (2011).
49. Harris, D. L., Kinyon, J. M., Mullin, M. T. & Glock, R. D. Isolation and propagation of spirochetes from the colon of swine dysentery affected pigs. *Can. J. Comp. Med.* **36**, 74–76 (1972).
50. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
51. Zolfo, M. et al. Discovering and exploring the hidden diversity of human gut viruses using highly enriched virome samples. Preprint at *bioRxiv* https://doi.org/10.1101/2024.02.19.580813 (2024).
52. Guerin, E. et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).
53. Edwards, R. A. et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol.* **4**, 1727–1736 (2019).
54. Simbayi, L. C. et al. *South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, 2017* (HSRC, 2019).
55. National AIDS and STI Control Programme, Ministry of Health, Kenya. *Kenya AIDS Indicator Survey 2012: Preliminary Report* (2013).
56. Hunt, P. W., Lee, S. A. & Siedner, M. J. Immunologic biomarkers, morbidity, and mortality in treated HIV infection. *J. Infect. Dis.* **214**, S44–S50 (2016).
57. Chun, T.-W. et al. Persistence of HIV in gut-associated lymphoid tissue despite long-term antiretroviral therapy. *J. Infect. Dis.* **197**, 714–720 (2008).
58. Imai, K., Yamada, K., Tamura, M., Ochiai, K. & Okamoto, T. Reactivation of latent HIV-1 by a wide variety of butyric acid-producing bacteria. *Cell. Mol. Life Sci.* **69**, 2583–2592 (2012).
59. Nazli, A. et al. Exposure to HIV-1 directly impairs mucosal epithelial barrier integrity allowing microbial translocation. *PLoS Pathog.* **6**, e1000852 (2010).
60. Dinh, D. M. et al. Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *J. Infect. Dis.* **211**, 19–27 (2015).
61. Chandiwana, N. C. et al. Weight gain after HIV therapy initiation: pathophysiology and implications. *J. Clin. Endocrinol. Metab.* **109**, e478–e487 (2024).
62. Vujkovic-Cvijin, I. et al. Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Sci. Transl. Med.* **5**, 193ra91 (2013).
63. Noguera-Julian, M. et al. Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine* **5**, 135–146 (2016).
64. Lozupone, C. A. et al. Alterations in the gut microbiota associated with HIV-1 infection. *Cell Host Microbe* **14**, 329–339 (2013).
65. Jackson, C. L. et al. Evolution of the gut microbiome in HIV-exposed uninfected and unexposed infants during the first year of life. *mBio* **13**, e0122922 (2022).
66. Rocafort, M. et al. Evolution of the gut microbiome following acute HIV-1 infection. *Microbiome* **7**, 73 (2019).
67. Monaco, C. L. et al. Altered virome and bacterial microbiome in human immunodeficiency virus-associated acquired immunodeficiency syndrome. *Cell Host Microbe* **19**, 311–322 (2016).
68. Serrano-Villar, S. et al. The effects of prebiotics on microbial dysbiosis, butyrate production and immunity in HIV-infected subjects. *Mucosal Immunol.* **10**, 1279–1293 (2017).
69. Dubourg, G. et al. Gut microbiota associated with HIV infection is significantly enriched in bacteria tolerant to oxygen. *BMJ Open Gastroenterol.* **3**, e000080 (2016).
70. Bashir, A., Miskeen, A. Y., Hazari, Y. M., Asrafuzzaman, S. & Fazili, K. M. *Fusobacterium nucleatum*, inflammation, and immunity: the fire within human gut. *Tumour Biol.* **37**, 2805–2810 (2016).
71. Lee, S. C. et al. Enrichment of gut-derived *Fusobacterium* is associated with suboptimal immune recovery in HIV-infected individuals. *Sci. Rep.* **8**, 14277 (2018).
72. Gilroy, R. et al. Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture. *PeerJ* **9**, e10941 (2021).
73. Le Roy, T. et al. *Dysosmobacter welbionis* is a newly isolated human commensal bacterium preventing diet-induced obesity and metabolic disorders in mice. *Gut* **71**, 534–543 (2022).
74. Moens de Hase, E. et al. *Dysosmobacter welbionis* effects on glucose, lipid, and energy metabolism are associated with specific bioactive lipids. *J. Lipid Res.* **64**, 100437 (2023).
75. Ako, S. E. et al. Unique community of gut bacterial microbiome as indicator for HIV infection and progression. *Int. J. Trop. Dis. Health* **44**, 33–45 (2023).
76. Wan, Y. et al. Habitual animal fat consumption in shaping gut microbiota and microbial metabolites. *Food Funct.* **10**, 7973–7982 (2019).
77. Duan, M. et al. Characteristics of gut microbiota in people with obesity. *PLoS ONE* **16**, e0255446 (2021).
78. Tomasch, J., Kopejtka, K., Shivaramu, S., Mujakić, I. & Koblížek, M. On the evolution of chromosomal regions with high gene strand bias in bacteria. *mBio* **15**, e0060224 (2024).
79. Kim, Y. S., Unno, T., Kim, B. Y. & Park, M. S. Sex differences in gut microbiota. *World J. Mens Health* **38**, 48–60 (2020).
80. Haelewaters, D., Hofmann, T. A. & Romero-Olivares, A. L. Ten simple rules for Global North researchers to stop perpetuating helicopter research in the Global South. *PLoS Comput. Biol.* **17**, e1009277 (2021).

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

the AWI-Gen 2 Collaborative Centre

Ovokeraye H. Oduaran[1,15], Luicer A. Ingasia Olubayo[1], Natalie Smyth[1], Theophilous Mathema[1], Godfred Agongo[5,6], Palwendé R. Boua[1,7], Solomon S. R. Choma[8], F. Xavier Gómez-Olivé[9], Isaac Kisiangani[10], Given R. Mashaba[8], Lisa Micklesfield[11], Shukri F. Mohamed[10], Engelbert A. Nonterah[6], Shane Norris[11,12], Hermann Sorgho[7], Stephen Tollman[9], Floidy Wafawanaka[9], Furahini Tluway[1], Michèle Ramsay[1] & Scott Hazelhurst[1,14,16]

Methods

Ethics approval

Human subjects research approval was obtained (University of the Witwatersrand Human Research Ethics (Medical) Committee Clearance Certificate No. M170880, M2210108), and ethics approvals were also obtained at each study centre. Informed consent was obtained from participants for all samples collected. Every participant was provided with an information sheet and consent documents, either in English or translated into the local language. Participants had opportunities to discuss concerns with the interviewer, and participants who could not read or write had documents read aloud with a witness[31]. The Stanford University Institutional Review Board deemed that the de-identified data transferred to Stanford University do not constitute human subjects research and thus did not require a further ethics approval beyond the human subjects research approval obtained at the University of the Witwatersrand.

Community engagement

Each study centre conducted pre-study engagement before recruitment during both AWI-Gen 1 and AWI-Gen 2, adapting to the local contexts to engage with community members and discuss feedback and concerns related to the study. For example, in DIMAMO, South Africa, pre-study engagement involved meeting with tribal leaders, the community advisory team and community representatives. In Navrongo, Ghana, the community engagement team visited chiefs and elders of the various study communities and informed them of the proposed study, and followed up with a community sensitization gathering before AWI-Gen 1 with a larger audience of chiefs, elders and people of the study communities. The community durbar was excluded from AWI-Gen 2, because of the continuing COVID-19 outbreak. In Nairobi, Kenya, the community engagement team held several consultative meetings with members of the community advisory committee, village elders, community health volunteers and AWI-Gen study participants before, during and after the study. The village elders and community health volunteers were crucial in mobilizing study participants who could not be reached by telephone. Questions from participants were related to how blood and stool samples would be used and why the study was focused on women. If during recruitment and sample collection there were notable health concerns[28] (for example, hypertension), the participants are referred into their clinical health-care service infrastructures. These mechanisms and processes varied from country to country and for sites within a country, depending on resources and local context.

Study design and cohort selection

Inclusion criteria included previous participation in the AWI-Gen 1 study[28] and continued participation in the AWI-Gen 2 study. This AWI-Gen 2 microbiome study is a companion study to an AWI-Gen 2 menopause study, and so only participants self-identifying as female were surveyed for the microbiome sub-study. A small number of men were recruited owing to a fieldwork mix-up. Given the understudied nature of these populations, we did not fully exclude samples from men in downstream analyses; rather, samples from men were excluded from site comparisons and disease associations, but included when cataloguing genomic novelty. Participants were chosen semi-randomly from the overall AWI-Gen 2 participant pool, with extra measures taken to ensure a cross-section of individuals with respect to menopause status and hypertension. See Supplementary Methods for extended recruitment details.

A harmonized approach for stool sample collection was implemented in all study sites to ensure equal temperature exposures and handling of all samples. In Soweto, Nairobi and Nanoro, participants came to central locations for interviews and biomarker collection. Participants were given stool sample collection kits that were either collected the same day or collected from their homes or at a central location in the following days. At the Navrongo, DIMAMO and Agincourt study centres, participants were visited in their homes for interviews and biomarker collection. Participant phenotype data and survey information were stored in REDCap servers based in South Africa, Burkina Faso, and Ghana (v.9 to v.13, regularly updated through the course of the study). Participants were given a stool sample collection kit to use at their home, which was collected by fieldworkers within 24 h.

Each participant self-collected a single stool samples using an OMNIGene GUT OMR-200 Collection Kit (DNA Genotek). This preservation kit maintains DNA integrity and taxonomic composition across a wide range of ambient temperatures[81], including the temperatures that are experienced year-round at each of the study sites. Samples were immediately frozen at study centres and then collectively shipped frozen to a central laboratory in Johannesburg, South Africa, where they were thawed, aliquoted into cryovials and stored at −80 °C. After obtaining necessary exportation and importation permits, all samples were shipped on dry ice in a single shipment to the United States for downstream processing. Samples were thawed once more to retrieve aliquots for DNA extraction. We previously conducted analysis to ensure that storage and shipping conditions would not significantly affect measured microbial composition[81]. Altogether, this approach minimized any technical confounders that would have coincided with study site, and we do not anticipate any other site-level methodological variation that would affect sample composition. Participant metadata, including age, demographic information, health history and blood biomarkers were collected as part of the larger AWI-Gen 2 project, with methods similar to those used in AWI-Gen 1 (ref. 31).

DNA extraction and metagenomic sequencing

All stool samples were extracted at the same time, in the same facility to minimize batch effect. DNA was extracted from samples using the QIAamp PowerFecal Pro DNA Kit (Qiagen, catalogue no. 51804) from 300 µl of stool sample according to manufacturer's instructions. Bead beating was performed for 10 min at 30 Hz, followed by rotation of the adapter and an extra 10 min of bead beating using a TissueLyser II (Qiagen, catalogue no. 85300) using a 2-ml Tube Holder Set (Qiagen, catalogue no. 11993), and DNA extractions were eluted with C6 Elution Buffer in a final volume of 80 µl. DNA concentration was quantified by spectrophotometer using the DropSense 96 platform (Trinean, catalogue no. 10100096). Every extraction batch of 96 samples included one water blank as a negative control and one mock community aliquot (Zymo Research, catalogue no. D6300) as a positive control.

All libraries were prepared concurrently at the same facility and sequenced at the same time across several flow cells. Samples were evaluated for concentration, integrity and purity before library preparation using the 5400 Fragment Analyzer System (Agilent, catalogue no. M5312AA). Metagenomic libraries were prepared using the NEB Ultra II kit (NEB, catalogue no. E7645L) according to the manufacturer's instructions. Library concentration was quantified using quantitative polymerase chain reaction and fragment length distribution was analysed using a 2100 Bioanalyzer (Agilent, catalogue no. G2939BA). Libraries were pooled and 2 × 150-base-pair reads were generated using the NovaSeq 6000 platform (Illumina, catalogue no. 20012850).

Metagenomic read preprocessing and taxonomy profiling

Metagenomic reads were deduplicated using HTStream SuperDeduper v.1.3.3 with default parameters, trimmed using TrimGalore v.0.6.7 with a minimum quality score of 30 and a minimum read length of 60. Reads aligning to version hg38 of the human genome were removed using BWA v.0.7.17 (ref. 82). Metagenomic reads were taxonomically profiled using mOTUs v.3.0.3 (ref. 83) and counts were distributed to GTDB[84] species using the GTDB_v207 mapping file available as part of the mOTUs database.

Given the number of previously unknown bacterial taxa observed in our assembly approach (see below), we aimed to better characterize the taxonomic composition by including our assembled bacterial genomes into the mOTUs database. To do so, we extended the mOTUs database with the scripts available under https://github.com/motu-tool/mOTUs-extender/. In brief, marker genes were identified in all high-quality assembled genomes using fetchMG v.1.2 (ref. 85). Those genes were then clustered together with the genes in the mOTUs database v.3.0.3. The resulting extended database contained 662 new genome clusters and reduced the fraction of unassigned reads for nearly all samples. Particularly in samples from Nanoro and Navrongo, the new genome clusters carried a large part of the relative abundance (Extended Data Fig. 3). For GTDB-level profiling, the GTDB-tk classification of our assembled genomes were added to the GTDB_v207 mapping file. Unless indicated otherwise, all analyses shown here are based on the extended mOTUs database.

All samples were used for metagenome assembly and new feature discovery ($n = 1,820$). Samples from males, one sample with a potential label mismatch, and samples with high percentages of human reads (percentage of human reads more than or equal to 70%, $n = 4$ samples) were excluded from classification-based analyses and site comparisons, leaving 1,796 samples for other analyses.

Participant covariate processing

Extensive participant data were collected as part of the AWI-Gen study, including demographic, ethnolinguistic, family composition, pregnancy, cognition, frailty, household amenity, substance use, general health, diet, infection history, cardiometabolic disease and physical activity information. Participants also gave blood, urine and stool samples, and underwent ultrasound, blood pressure, blood and urine testing for various metrics. Not all data were available for every participant, and some participants gave stool samples for microbiome analysis but did not complete other testing or questionnaires. At the time of analysis for the microbiome study, not all participant data had gone through quality control. In total, 59 variables were available to use as covariates in the microbiome study.

Before using covariate data in microbiome analysis, we first collapsed the covariate dataset to only those variables that we expected to be most meaningful to avoid unnecessary multiple-hypothesis testing and measuring associations between dependent variables. First, we removed variables that had overwhelmingly missing data, excluding those that had entries for 100 or fewer participants (for example, several ultrasound measurements). Second, we filtered variables with not enough unique values (such as sex, which had only one group). Lastly, we excluded variables with an entropy (calculated with the infotheo package v.1.2.0.1 (ref. 86) in R) of less than 0.2 to avoid variables that were too uniform in the participant set to power comparisons (for example, breast cancer or cervical cancer status with only 10 and 12 cases, respectively).

To calculate correlation between covariates and associations between covariates and microbiome composition, we transformed non-numerical covariates into numerical values on the basis of ordered factor levels. For example, values for the *Menopause* covariate were changed from Pre-menopausal to 1, from Peri-menopausal to 2 and from Post-menopausal to 3. Most covariates were binary (for example, Probiotics could contain either the value Yes or No) and were converted to 1 (for Yes) and 2 (for No) in this process. The full list of binary variables is: Arthritis, Diabetes status, Diabetes treatment, Hypertension status, Hypertension treatment, Pesticides, Vigorous work, Weekend work, HIV medication, HIV status, Cattle, Other livestock, Potable water, Poultry, Refrigerator, Toilet, Deworming treatment, Probiotics, Chew tobacco and Smokeless tobacco. The variables describing time (Deworming period, Probiotics period, Antibiotics and Diarrhoea last) were ordered according to recency with the order WithinLastWeek < WithinLastMonth < WithinLastSixMonths < WithinLastYear < WithinLastTwoYears < WithinLastThreeYears < Longer < Never. Employment was ordered as Self-Employed < FormalFull-time < FormalPart-time < Informal < Unemployed. Site density was ordered as Nanoro < Navrongo < DIMAMO < Agincourt < Soweto < Nairobi.

Microbial diversity, composition and site differences

To measure prokaryotic alpha-diversity, species counts were rarefied to 5,000 using the rrarefy function available through the vegan R package v.2.6-4 (ref. 87). Alpha-diversity was measured as inverse Simpson index after rarefaction, and prokaryotic richness was measured as number of species with relative abundance greater than or equal to 1×10^{-4} after rarefaction).

Beta-diversity was calculated on the Bray–Curtis distance using the vegdist function from vegan[87] and the pco function from the labdsv R package v.2.1-0 (ref. 88). To assess the amount of variance explained by covariates, we undertook distance-based redundancy analysis with the dbrda function from vegan. In an iterative manner, the covariate explaining the highest amount of variance was added to the model formula. To reduce redundancy of highly correlated covariates, all available covariates were transformed into numerical values (using ordinal factors, whenever applicable) and the Pearson correlation between covariates was calculated. In cases of highly correlated covariates (Pearson's $r \geq 0.8$), the covariate that explained the higher amount of variance in the prokaryotic composition was chosen for the iterative model (Extended Data Fig. 5).

Prokaryotic species prevalence was defined as the fraction of individuals in a study site in which a given species is found at a relative abundance of more than or equal to 1×10^{-4}. The difference between sites for individual taxa was calculated using a generalized fold change[89]. In short, instead of comparing the median (the 50% quantile) between distributions, the generalized fold change is the mean of the differences between two distributions at several quantiles and can therefore resolve differences also in low-prevalence taxa. Figure 3d shows the number of taxa for which the generalized fold change between sites exceeds the 90% quantile of all pairwise site comparisons across all prokaryotic species.

The number of samples for these analyses (Fig. 1 and associated supplements) was distributed across the different sites as follows: Nanoro, $n = 382$; Navrongo, $n = 218$; DIMAMO, $n = 201$; Agincourt, $n = 532$; Soweto, $n = 226$; Nairobi, $n = 237$.

Metagenome assembly and external dataset comparison

All samples ($n = 1,820$), including samples for male participants, were included in metagenomic assembly analyses (Nanoro, $n = 384$; Navrongo, $n = 235$; DIMAMO, $n = 203$; Agincourt, $n = 533$; Soweto, $n = 226$; Nairobi, $n = 239$). Metagenomic reads were assembled using megahit v.1.2.9 (ref. 90) and assembly quality was assessed using QUAST v.5.2.0 (ref. 91). Metagenomic assemblies were binned into draft genomes using MetaBAT v.2.5 (ref. 92), CONCOCT v.1.1.0 (ref. 93) and MaxBin v.2.2.7 (ref. 94), and subsequently dereplicated and aggregated on a per-sample basis using DAS Tool v.1.1.6 (ref. 95). Bin quality was assessed using CheckM v.1.2.2 (ref. 96). To create a dereplicated genome set, MAGs were dereplicated using dRep v.3.4.3 (ref. 97), filtering to only include genomes with a minimum CheckM completeness of 50% and maximum CheckM contamination of 5%. In dereplication, we implemented a primary clustering threshold (-pa) of 0.9 and secondary alignment threshold (-sa) of 0.95, requiring minimum overlap between genomes (-nc) of 0.3, using multiround primary clustering (--multiround_primary_clustering) and greedy secondary clustering with fastANI v.1.33 (ref. 98) (--greedy_secondary_clustering, --S_algorithm fastANI) to reduce the computational complexity of dereplicating a large genome set. For dereplication, cluster representatives were chosen using scoring criteria that included a completion weight (-comW) of 1, contamination weight (-conW) of 5, N50 weight (-N50W) of 0.5, size weight (-sizeW) of 0, and centrality weight (-centW) of 0. Genome filters and scoring were consistent with standards used in the UHGG[50]. The final genome set was taxonomically classified and placed in a tree with GTDB-tk v.2.3.0 (ref. 99) using the GTDB r214 catalogue and default parameters. Phylogenetic trees were visualized with iTOL v.6 (ref. 100). The dereplicated prokaryotic genome set was compared against the UHGG v.2.0.1 species representatives using dRep v.3.4.3 with the same parameters as previously stated above.

Protein-coding genes were predicted from each medium-quality and high-quality prokaryotic genome, before genome dereplication, using prodigal v.2.6.3 (ref. 101) with parameters -c -p meta to exclude partial genes. Putative proteins were clustered successively using mmseqs v.14.7e284 (ref. 102) linclust command with alignment coverage (-c) of 0.8 in target coverage mode (--cov-mode 1) and greedy secondary clustering (--cluster-mode 2) at 100% and 95% amino acid identity (--min-seq-id). The 95% identity protein set was compared against the UHGP95 v.2.0.1 proteins using mmseqs v.14.7e284, and proteins sharing 95% amino acid identity over 80% of the UHGG protein were considered to match the UHGP set.

Modelled accumulation of previously unknown prokaryotic genomes and proteins with further participant sampling was determined by randomly subsetting the full participant set or site-specific participant sets to a range of individuals (1–1,500) in 100 iterations and counting the number of prokaryotic genome cluster or protein clusters represented by the participant subset.

Comparison with external metagenomic studies (Extended Data Fig. 7) used the same pipelines for read preprocessing, assembly and binning, with the exception of Carter et al.[43] who published a MAG catalogue. All genomes from the UHGG, AWI-Gen 2 and external metagenomic studies were dereplicated using the same parameters as described above.

Treponema succinifaciens core genome analysis and functional profiling

We evaluated the complete set of *T. succinifaciens* MAGs in our genome catalogue before dereplication. To identify *T. succinifaciens* genomes, we selected all genomes with completeness of more than 90% and contamination less than 5% that fell into a secondary cluster with genomes classified as *Treponema_D succinifaciens* by GTDB-tk in our dereplicated genome catalogue ($n = 244$). Coding sequences were annotated with bakta v.1.8.2 (ref. 103). Core genes, defined here as genes present in at least 80% of genomes, were identified with roary v.3.12.0 (ref. 104).

Public *T. succinifaciens* genomes with completeness of more than 90% and contamination less than 5% were downloaded from the UHGG[50], Carter et al.[43] and National Center for Biotechnology Information (NCBI). To build a global phylogenetic tree, core genes were identified and incorporated into a core gene multiple sequence alignment using roary v.3.12.0 (ref. 104) and MAFFT v.7.407 (ref. 105). The core gene multiple sequence alignment was used as input to FastTree v.2.1.11 (ref. 106), and the resulting phylogenetic trees were visualized in iTOL v.6 (ref. 100). Phylogeographic signal was statistically quantified using the same method as Hildebrand et al.[107]: we calculated pairwise phylogenetic distance between all genomes on the basis of branch length using DendroPy[108], and implement a permuted multivariate analysis of variance test with 1,000 permutations with adonis2 (ref. 87) to evaluate whether phylogenetic distances within countries are smaller than phylogenetic differences between countries.

Associations between *T. succinifaciens* presence and host phenotype were performed for all participants from Nanoro, Burkina Faso and Navrongo, Ghana. Host phenotype measurements included antibiotic history, anthropometric measurements, livestock ownership, hypertension status and all biomarkers. Associations were tested using a linear model that adjusted for site and for antibiotic history, excepting the association with antibiotic history, which only adjusted for site. Correction for multiple-hypothesis testing was performed with the Benjamini–Hochberg procedure. Antimicrobial resistance profiling was performed with the Resistance Gene Identifier[109], and 'Loose' matches were omitted. Carbohydrate-active enzyme (CAZyme) annotation was performed on all high-quality MAGs from AWI-Gen 2 using dbCAN3 v.4.1.4 (ref. 110) with the prok parameter for conservative, high-confidence annotations. Substrate annotation was performed at the CAZyme family level using the high-level substrate annotations from the dbCAN3 substrate mapping table, and substrates were grouped according to biological origin.

Viral fraction characterization

Phage genomes were annotated from metagenomic assemblies with VIBRANT v.1.2.1 (ref. 111), and genome quality was determined with checkV v.1.0.1 (ref. 112). Redundant genomes from each sample were removed by clustering medium- and high-quality genomes using a database built with BLAST 2.14.0 (ref. 113), clustering at a minimum of 95% ANI and 85% alignment fraction using checkV supporting scripts with default parameters. Phage richness was measured as the number of assembled phage genomes per sample after removal of duplicate genomes. A unified catalogue of phage genomes was built by clustering the representative phages from each individual using the same clustering parameters, and this catalogue was compared against the MGV v.1.0 (ref. 114) vOTU representative phage genomes using the same BLAST clustering approach and parameters. Alternate phage profiling using read-based classification (Extended Data Fig. 9) was performed with Phanta v.1.1.0 (ref. 115) using the combination of MGV and UHGG as the reference database. Phage richness measured with Phanta was defined as the number of phage species clusters present at greater than or equal to 10^{-5}% relative abundance. Differences between alpha-diversity metrics across sites were tested with a linear model, using the anova function from base R to estimate the significance of the difference. Modelled accumulation of previously unknown phage genomes with further sampling was performed using the same methods as described above for modelled prokaryotic genome and protein accumulation.

Crassvirales and crAssphage prevalence was defined as the fraction of individuals with taxon relative abundance of greater than or equal to 10^{-5}%. Previously unknown jumbophages were defined as viral genomes in the dereplicated genome catalogue with length greater than 200 kb that did not cluster with an MGV vOTU representative. We further filtered the new jumbophages to highlight only jumbophages with evidence supporting prevalence and novelty, by including only those with assembled genomes present in at least five individuals, and with alignment fractions less than 10% against any MGV vOTU representative. Jumbophage genes were annotated with bakta v.1.8.2 (ref. 103). Read-level presence of jumbophages was defined at greater than 0.1 coverage threshold as measured using CoverM v.0.7.0 (ref. 116).

Association between microbiome features and HIV status

Participants from Agincourt, South Africa, Soweto, South Africa and Nairobi, Kenya were included in this analysis. Participants from DIMAMO, South Africa were excluded because of the low number of PLWH ($n = 6$). Participants from Nanoro, Burkina Faso and Navrongo, Ghana were excluded because HIV status was not measured in these populations owing to a low national prevalence of HIV. A total of 848 participants were included in this analysis, capturing 129 PLWH and 719 seronegative individuals (Table 1 and Supplementary Data 9). The rest of the samples from those sites were either HIV positive, but reported not to take ART ($n = 28$, $n = 22$ in Agincourt, $n = 3$ in Soweto, $n = 3$ in Nairobi). Male individuals and individuals with missing/discrepant HIV/ART data or with low read counts were excluded. Prokaryotic alpha- and beta-diversity were calculated as described above.

We undertook differential abundance analysis using a linear mixed effect model implemented in the lmerTest R package v.3.1-3 (ref. 117), including site, exposure to antibiotics and self-reported recency of diarrhoea as random effects, because those factors had shown to be related to microbiome composition in the previous analyses. Overall effect size was estimated through the lmerTest package as well and generalized fold change within each site was calculated as described above.

For the machine learning analysis, we trained statistical models using the SIAMCAT R package v.2.5.0 (ref. 89) for both all data combined and for each site separately. In short, relative abundances were normalized using the log.std method in SIAMCAT. Samples were split for five-times repeated fivefold cross-validation (20% of samples were retained for testing and not included in model training) and for each split, an L1-regularized logistic regression model was trained on the training folds, using standard parameters. Model evaluation was performed within the cross-validation (for example, within a site) by applying each model to the respective left-out test fold. The predictions for each sample were averaged across repeats and AU-ROC was calculated with the pROC package v.1.18.2 (ref. 118). For cross-site evaluation, the external data was normalized with the recorded normalization parameters (frozen normalization), all models from the cross-validation were applied to the normalized data, and predictions were averaged again for AU-ROC analysis. For the LOSO analysis, models were trained as described on data from two sites combined (for example, Agincourt and Soweto) and were then applied on the data from the left-out site (Nairobi).

To test the fraction of positive prediction in other sites, we calibrated the model prediction to an internal 5% false positive rate; that is, recorded at which prediction threshold 5% of HIV− samples were incorrectly classified as PLWH. The model trained on all data combined was then applied to the data from Nanoro, Navrongo and DIMAMO to quantify the number of samples that resulted in a prediction above the threshold value.

Viral feature comparison between seronegative individuals, PLWH who are ART-naive, and PLWH who are ART+ was performed using phage relative abundance profiles generated by Phanta. Phage richness was calculated as the number of phage species present at greater than or equal to 10^{-5}% abundance in Phanta profiles, as opposed to using total count of assembled phages, because Phanta abundance profiles have features with sufficient prevalence for differential abundance analysis and machine learning models. Differential feature analysis and machine learning models were performed using the same methods as the prokaryotic analysis above.

Statistical analysis

Statistical analyses were performed with R v.4.1.2 using the statistical test specified in the respective Methods section. Correction for multiple-hypothesis testing was performed with the Benjamini–Hochberg procedure[119] as implemented in the p.adjust function in base R in all analyses where several tests were performed. Plots were generated in R using the packages ggplot2 v.3.4.2 (ref. 120), cowplot v.1.1.1 (ref. 121), pheatmap v.1.0.12 (ref. 122) and tidyverse v.2.0.0 (ref. 123).

Ethics and inclusion statement

All authors of this study fulfilled criteria for authorship inclusion, and researchers from each study centre are represented as authors. Researchers from all institutions were involved throughout the study process. Study centre staff facilitated community engagement sessions, which identified specific community concerns and determined that this study is locally relevant. Roles and responsibilities were agreed upon amongst collaborators before conducting the research. Authors of this study have led formal capacity-building genomics workshops for local scientists during the course of the study (Extended Data Fig. 1), along with further informal training. This study has been approved by local ethics review committees (Methods). Research pertinent to the study centres and led by local researchers has been taken into account in the citations.

Reporting summary

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

Data availability

To maximize public availability of our data while protecting participants, we have split our data into two sets: sequences that may contain human reads, which is available from the European Genome-Phenome Archive (EGA) as EGAD00001015463 on application to the H3Africa Data and Biospecimens Access Committee; and an open dataset available from the NCBI Sequence Read Archive as PRJNA115737 (Supplementary Methods). The dereplicated genome sets are available at Zenodo (https://doi.org/10.5281/zenodo.13761309)[124]. Participant phenotype is stored at the EGA under accession EGAD00001015440. Participant phenotype data are under restricted access due to ethics requirements of the AWI-Gen 2 study. Applications must be made to the independent Human Heredity and Health in Africa Data and Biospecimen Access Committee by registering and applying at https://catalog.h3africa.org/. The H3Africa Data Sharing Policy can be found at https://h3africa.org/wp-content/uploads/2020/06/H3Africa-Consortium-Data-Access-Release-Policy-April-2020.pdf. Decisions on requests made by the third week of the month should be made by the end of the subsequent month. Source data for figures is available. Classification tables, genome summary statistics, taxon prevalence and differential feature tables are available as supplementary data. Reference data used in this study are available from the Unified Human Gastrointestinal Genome collection in the European Nucleotide Archive under project accession PRJEB33885, the Metagenomic Gut Virus catalogue at https://portal.nersc.gov/MGV and the Genome Taxonomy Database at https://data.gtdb.ecogenomic.org/releases/. The hg38 human reference genome is available at NCBI Genome under accession number GCF_000001405.26.

Code availability

Source code for analysis and figure generation is publicly available at Zenodo (https://doi.org/10.5281/zenodo.14231329)[125] and on GitHub at https://github.com/bhattlab/AWIGen2Microbiome.

81. Maghini, D. G. et al. Quantifying bias introduced by sample collection in relative and absolute microbiome measurements. *Nat. Biotechnol.* **42**, 328–338 (2024).
82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
83. Ruscheweyh, H.-J. et al. Reference genome-independent taxonomic profiling of microbiomes with mOTUs3. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.20.440600 (2022).
84. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
85. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
86. Meyer, P. E. *Information-Theoretic Variable Selection and Network Inference from Microarray Data*. PhD thesis, Universite Libre de Bruxelles, Brussels, Belgium (2008).
87. Oksanen, J. et al. vegan: Community ecology package v2.6-4 https://CRAN.R-project.org/package=vegan (2022).
88. Roberts, D. W. labdsv: Ordination and multivariate analysis for ecology. Comprehensive R Archive Network (CRAN) v2.1-0 https://cran.r-project.org/web/packages/labdsv/index.html (2023).
89. Wirbel, J. et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* **22**, 93 (2021).
90. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
91. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
92. Kang, D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
93. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
94. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
95. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
96. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
97. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
98. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
99. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).

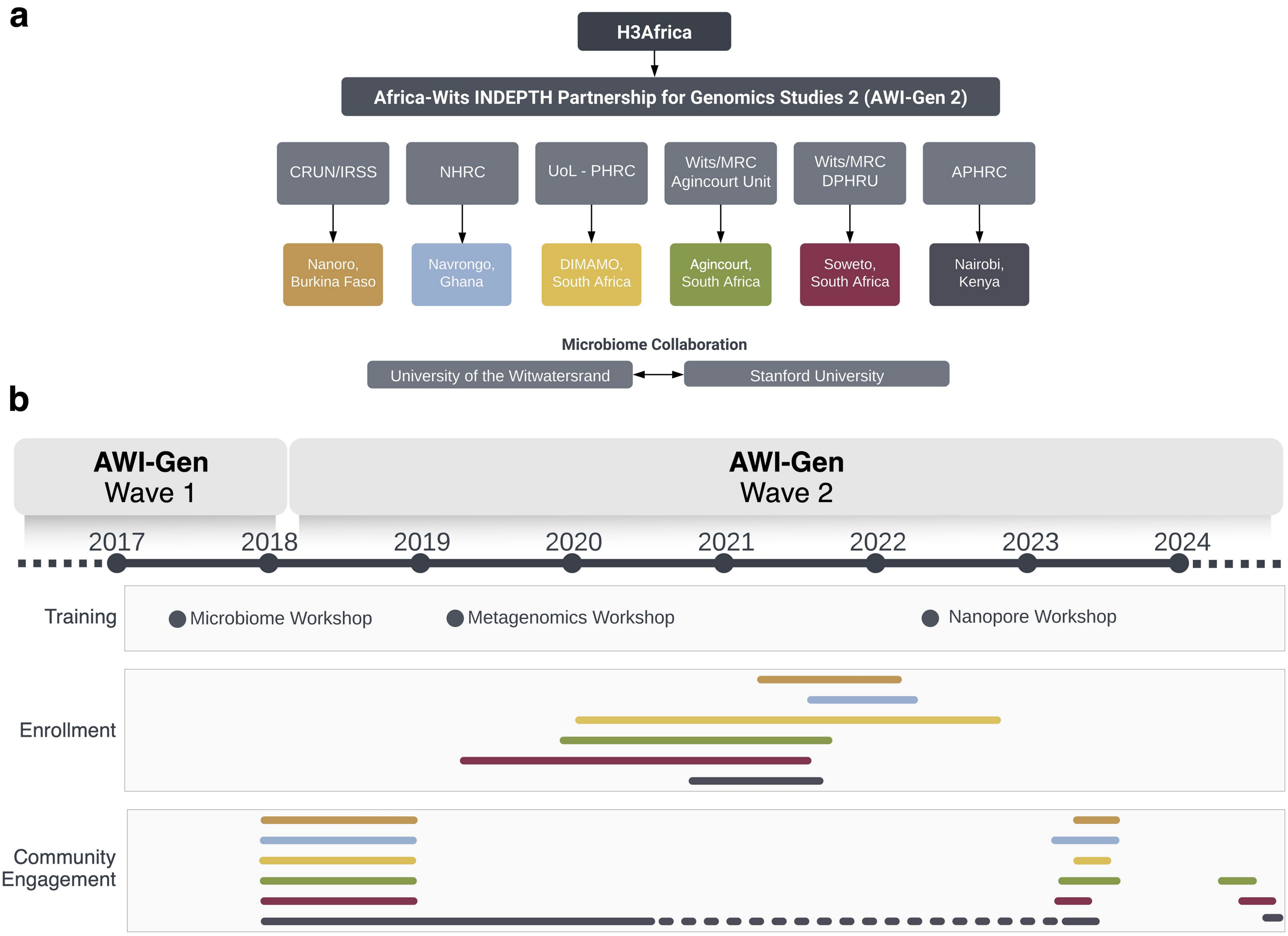
100. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
101. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
102. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
103. Schwengers, O. et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genom.* **7**, 000685 (2021).
104. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
105. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
106. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
107. Hildebrand, F. et al. Dispersal strategies shape persistence and evolution of human gut bacteria. *Cell Host Microbe* **29**, 1167–1176.e9 (2021).
108. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).
109. Alcock, B. P. et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699 (2023).
110. Zheng, J. et al. dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res.* **51**, W115–W121 (2023).
111. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
112. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
113. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
114. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
115. Pinto, Y., Chakraborty, M., Jain, N. & Bhatt, A. S. Phage-inclusive profiling of human gut microbiomes with Phanta. *Nat. Biotechnol.* **42**, 651–662 (2024).
116. Aroney, S. T. N. et al. CoverM: read coverage calculator for metagenomics (v0.7.0). *Zenodo* https://doi.org/10.5281/zenodo.10531253 (2024).
117. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
118. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
119. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
120. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer Science & Business Media, 2009).
121. Wilke, C. O. cowplot: streamlined plot theme and plot annotations for 'ggplot2' v1.1.1 https://CRAN.R-project.org/package=cowplot (2020).
122. Kolde, R. Pheatmap: pretty heatmaps. R package version 1.0.12, https://github.com/raivokolde/pheatmap (2018).
123. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
124. Maghini, D. & Wirbel, J. AWI-Gen 2 Microbiome Project MAGs. *Zenodo* https://doi.org/10.5281/zenodo.13761309 (2024).
125. Maghini, D. & Wirbel, J. Figure generation and analysis source code: Expanding the human gut microbiome atlas of Africa. *Zenodo* https://doi.org/10.5281/zenodo.14231329 (2024).

Acknowledgements We thank the study participants for participating in this research project, along with each community advisory group for their recommendations on study design; the fieldworkers and staff at each study centre for participant enrollment and sample collection; Navrongo centre principal investigator A. Oduro and Navrongo Health Research Centre Director P. Ansah, as well as C. Santuah and R. Afiya for supporting sample collection; the Soweto study and data collection team: V. Baloyi, S. Sekwati, T. Ngobeni, O. Godongwana and M. Ritsuri; C. Soo for her role as biobank manager through 2021 and for assistance with standard operating procedure development and sample shipping; the Stanford Research Computing Center for computational infrastructure; E. Brooks, G. Reynolds and C. Chen for assistance in sample processing; S. Nayfach for advice regarding phylogeography calculations and phage catalogues; M. Barry and S. Luby for guidance regarding the project; the eLwazi Open Data Science Platform (N. Mulder, Z. Parker and S. Aron) for support with data management. This work was supported by the National Human Genome Research Institute, the Eunice Kennedy Shriver National Institute of Child Health & Human Development and Office of the Director of the National Institutes of Health (NIH) of the USA under award number U54HG006938 for the AWI-Gen study, and its supplements, as part of the H3Africa Consortium. Further funding was granted by the Department of Science and Technology (now Department of Science and Innovation), South Africa (award no. DST/CON 0056/2014). This work was also supported, in part, by the Stanford Center for Innovation in Global Health. This work used supercomputing resources provided by the Stanford Genetics Bioinformatics Service Center, supported by NIH S10 Instrumentation Grant S10OD023452. D.G.M. was supported by the Stanford Gerald J. Lieberman Fellowship and the NIH Fogarty Global Health Equity Scholars Program (NIH FIC D43TW010540). O.H.O. was partially supported by the NIH Fogarty Global Health Equity Scholars Program (NIH FIC D43TW010540). L.A.I.O. was supported by the Sydney Brenner Charitable Trust (SBCT) for her postdoctoral fellowship. M.R. is the South African Research Chair in Genomics and Bioinformatics of African Populations hosted by the University of the Witwatersrand, funded by the South African Department of Science and Innovation, and administered by the National Research Foundation. J.W. is a Damon Runyon Quantitative Biology Fellow supported by the Damon Runyon Cancer Research Foundation (grant no. DRQ-22-24) and additionally acknowledges support from the School of Medicine Dean's Postdoctoral Fellowship. The Bhatt lab is supported by NIH R01AI148623 and R01AI143757, and a Stand Up 2 Cancer Grant. A.S.B. is supported by the Allen Distinguished Investigator Award. S.H. was partially supported by the SA National Research Foundation (CPRR160421162721). This paper describes the views of the authors and does not necessarily represent the official views of the National Institutes of Health (USA), the South African Department of Science and Innovation or the National Research Foundation (South Africa) who funded this research.

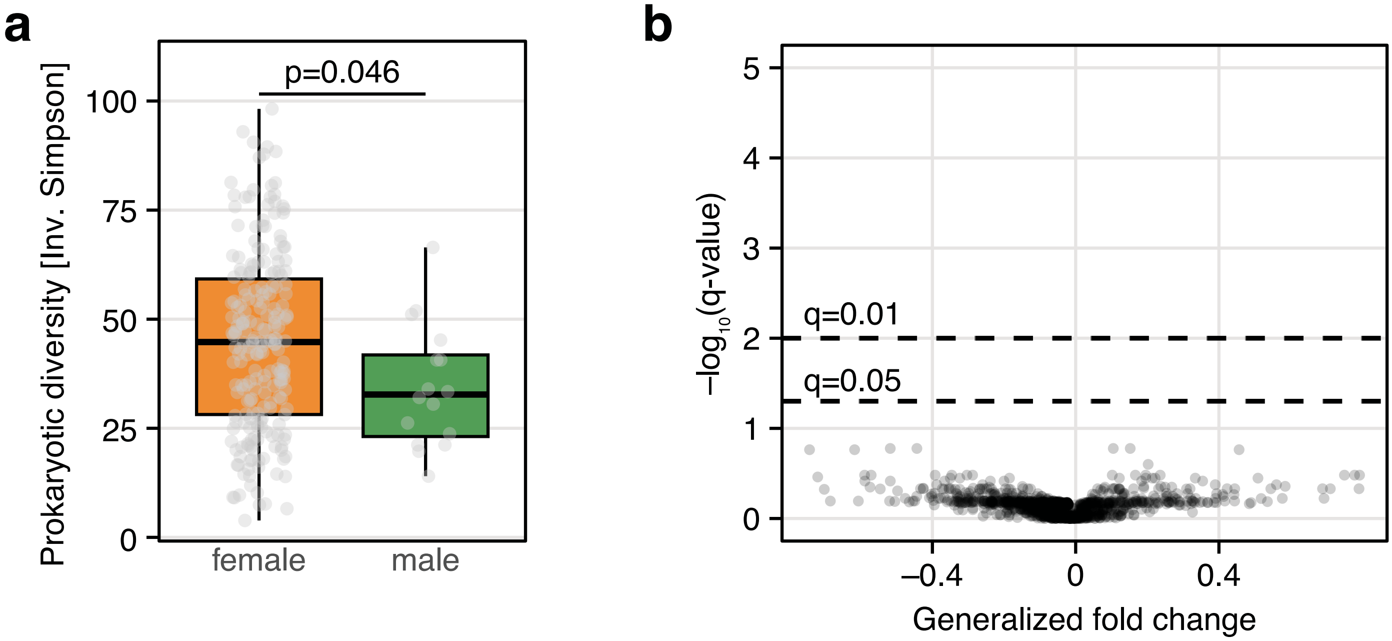
Author contributions A.S.B., S.H. and M.R. conceived of the study, secured funding and obtained necessary ethics approvals. S.H., O.H.O., D.G.M., L.A.I.O., M.R., F.T., G.A., P.R.B., S.S.R.C., F.X.G.-O., I.K., G.R.M., L.M., S.F.M., E.A.N., S.N., H.S., S.T. and F.W. organized study logistics, community engagement, participant enrollment and data collection. N.S., O.H.O. and D.G.M. coordinated sample preparation, shipping and sequencing. T.M. performed metadata cleaning and management. D.G.M. and J.W. performed compositional and assembly-based analysis. D.G.M., L.A.I.O. and J.W. calculated associations between HIV status, microbiome composition and other phenotype variables. D.G.M., J.A.C. and J.W. performed the CAZyme analysis. D.G.M., L.A.I.O., J.A.C., C.W.B. and J.W. performed the literature review. D.G.M., O.H.O., L.A.I.O., J.W., A.S.B. and S.H. wrote and edited the manuscript. A.S.B. and S.H. supervised the work and oversaw the collaboration, contributing equally. All authors have read and approved the manuscript.

Competing interests The authors declare no competing interests.

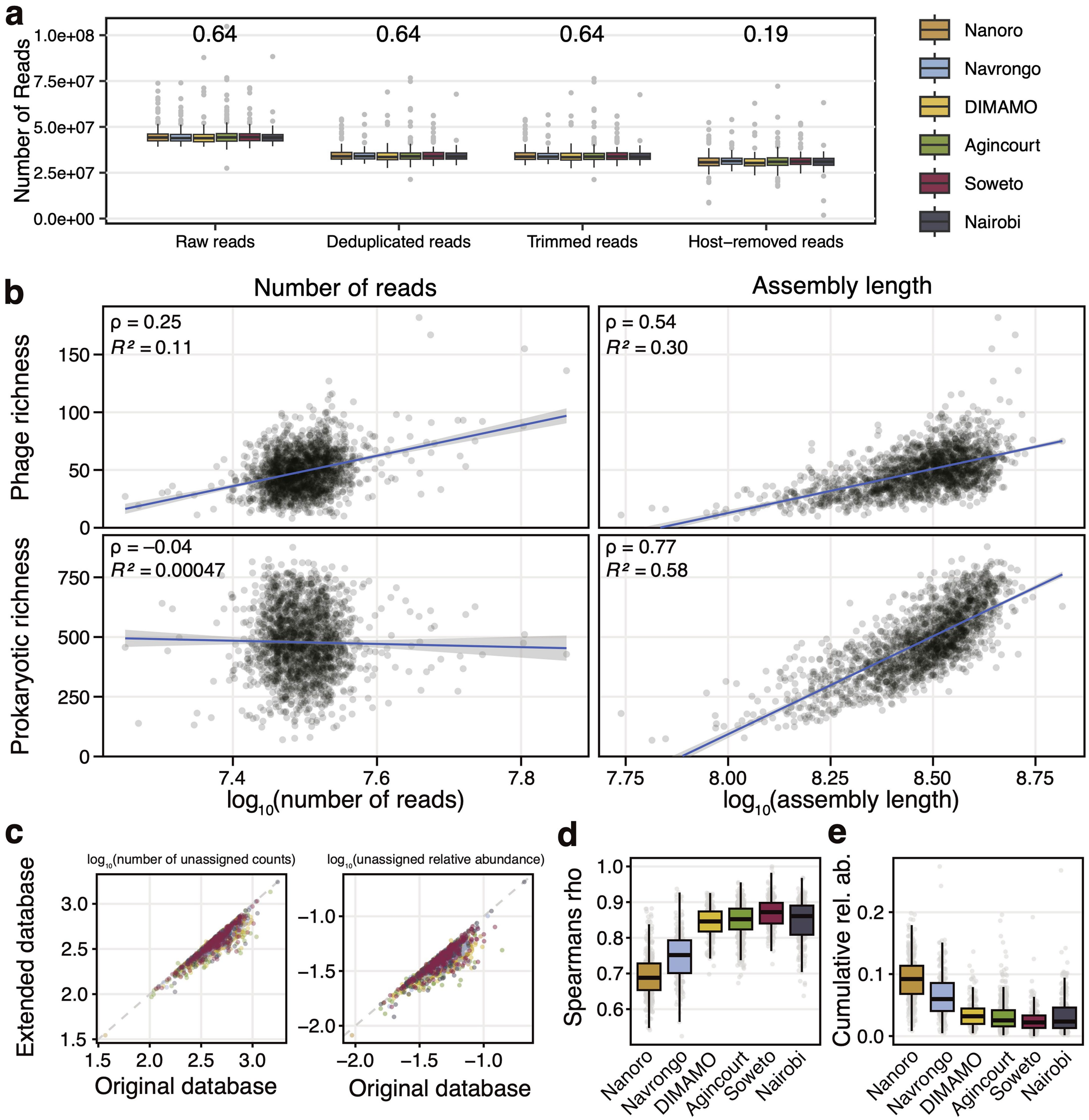
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08485-8.
Correspondence and requests for materials should be addressed to Ami S. Bhatt or Scott Hazelhurst.
Peer review information *Nature* thanks Hilary Browne and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.



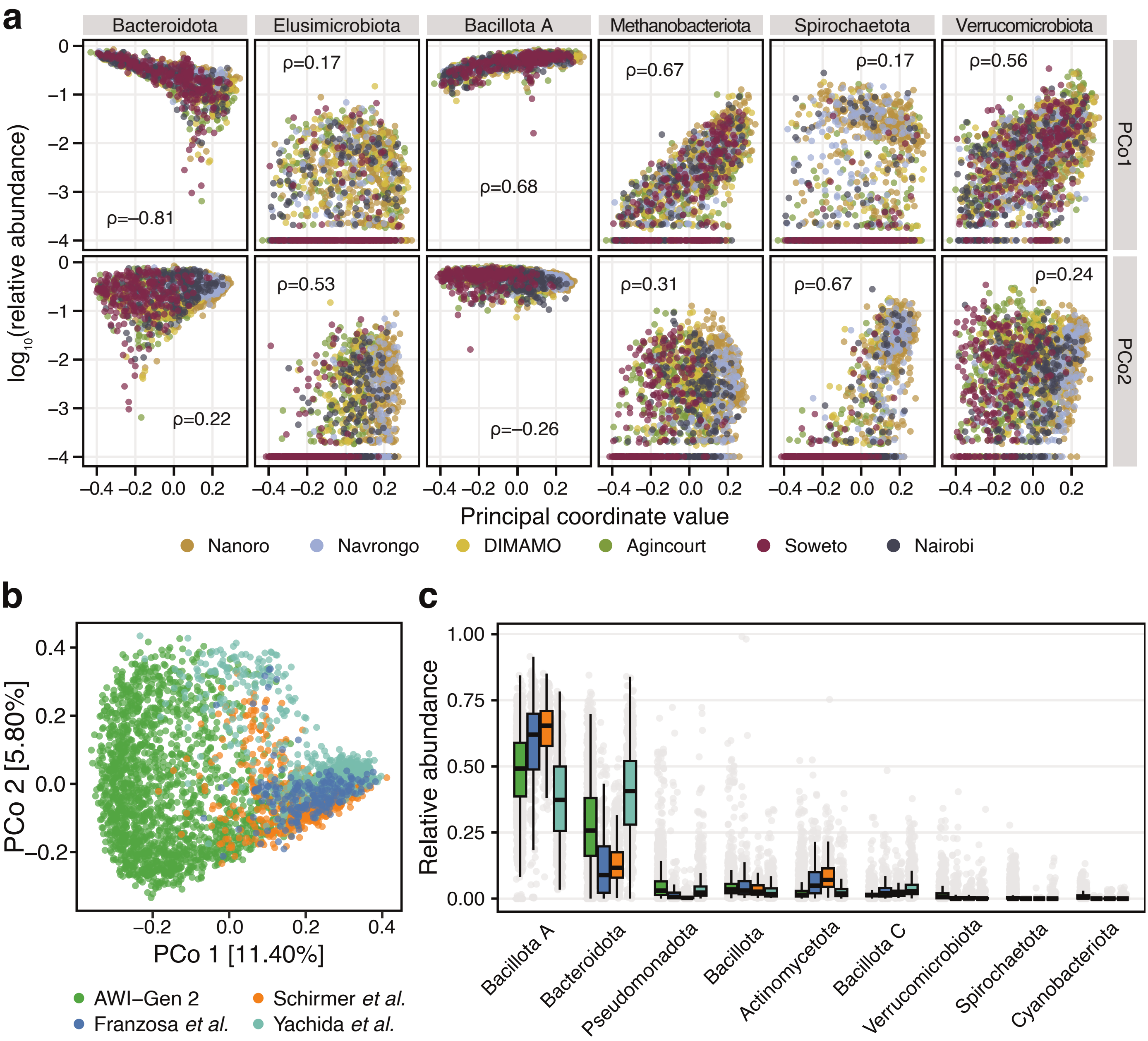
Extended Data Fig. 1 | Overview of the AWI-Gen 2 Microbiome study. **a**) Organizational chart of the AWI-Gen 2 project. The partnership, funded by the National Institutes of Health under the umbrella of the Human Heredity and Health in Africa consortium (H3Africa), includes five Health and Demographic Surveillance Sites (HDSSs) and the Soweto MRC/Wits Developmental Pathways for Health Research Unit (DPHRU). The HDSSs and DPHRU are managed by the Clinical Research Unit of Nanoro Institut de Recherche en Sciences de la Santé (CRUN/IRSS), Navrongo Health Research Centre (NHRC), University of Limpopo Population Health Research Centre (UoL–PHRC), University of the Witwatersrand and the South African Medical Research Council (Wits/MRC), and African Population Health and Research Center (APHRC). Researchers from Stanford University and the University of the Witwatersrand led the microbiome analysis. **b**) Timeline of the AWI-Gen 2 microbiome study research activities, including study administration, sample collection, and community engagement. During both AWI-Gen phases, researchers led microbiome and bioinformatic workshops for local researchers. Community engagement preceded sample collection at all sites, and participants with concerning health-related results were referred to their local healthcare facilities in accordance with site-specific protocols. Community engagement in Nairobi continued intermittently throughout sample collection to accommodate roadblocks during the COVID-19 pandemic. Post-study engagement was conducted at all sites, and microbiome-specific return of results is complete at three study sites.



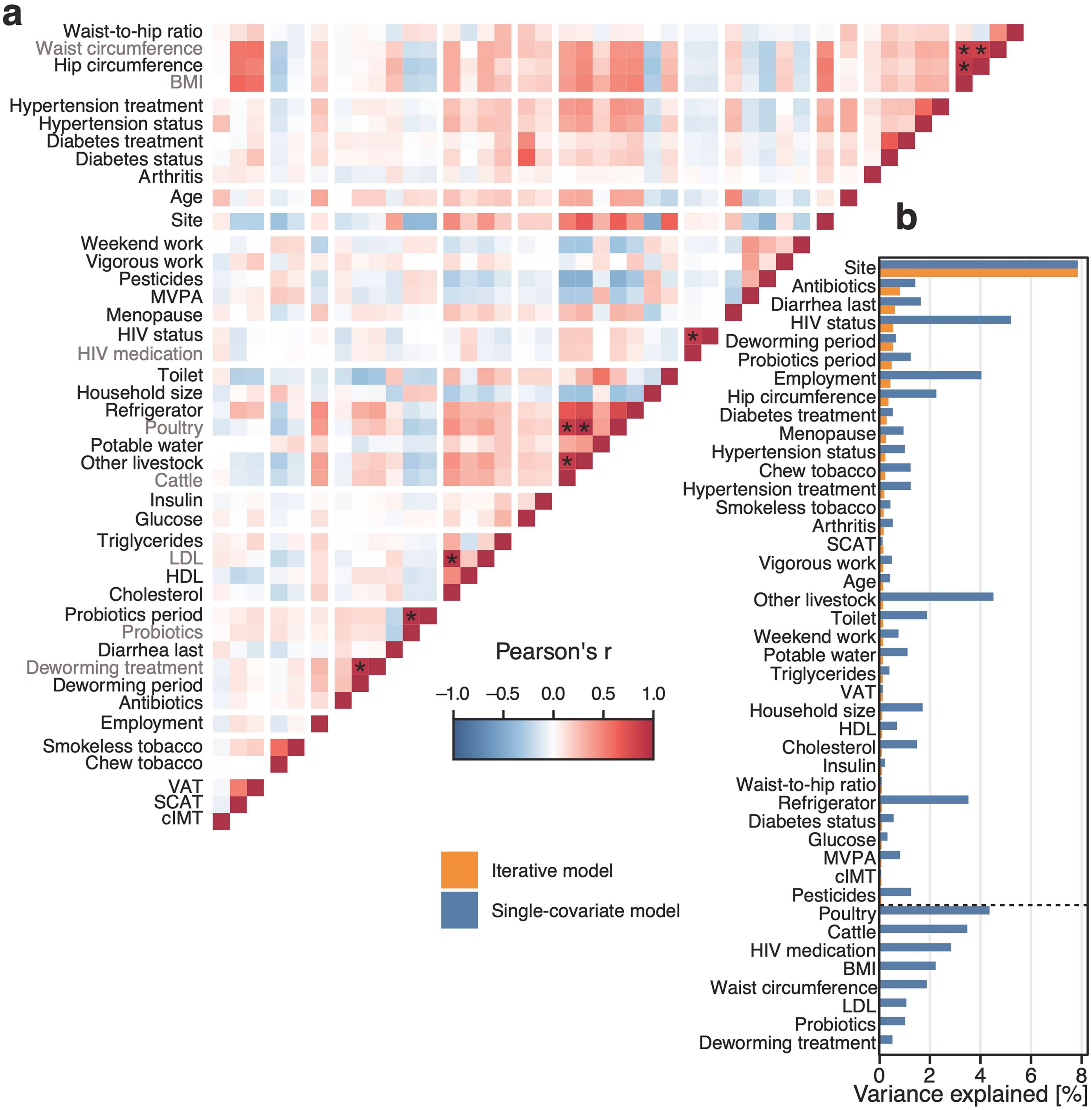
Extended Data Fig. 2 | Microbiome composition of male and female participants in Navrongo, Ghana. a) Prokaryotic richness (number of prokaryotic species present at ≥1 × 10^{-4}% relative abundance after rarefaction, see Methods) in $n = 16$ males and $n = 218$ females in Navrongo, Ghana (Wilcoxon test, $P = 0.027$). Points indicate individual samples. (In total, 19 samples from male participants were sequenced). **b**) Generalized fold change between male and female participants for all species with a prevalence higher than 5% in Navrongo is plotted against the negative log10-transformed q-value (Benjamini-Hochberg corrected p-value). Positive values correspond to higher relative abundance in males, whereas negative fold change values indicate higher relative abundance in female participants. No species meet the threshold of significance after correction for multiple testing. For all boxplots, boxes denote the interquartile range (IQR) with the median as a thick black line and the whiskers extending up to the most extreme points within 1.5-fold IQR.



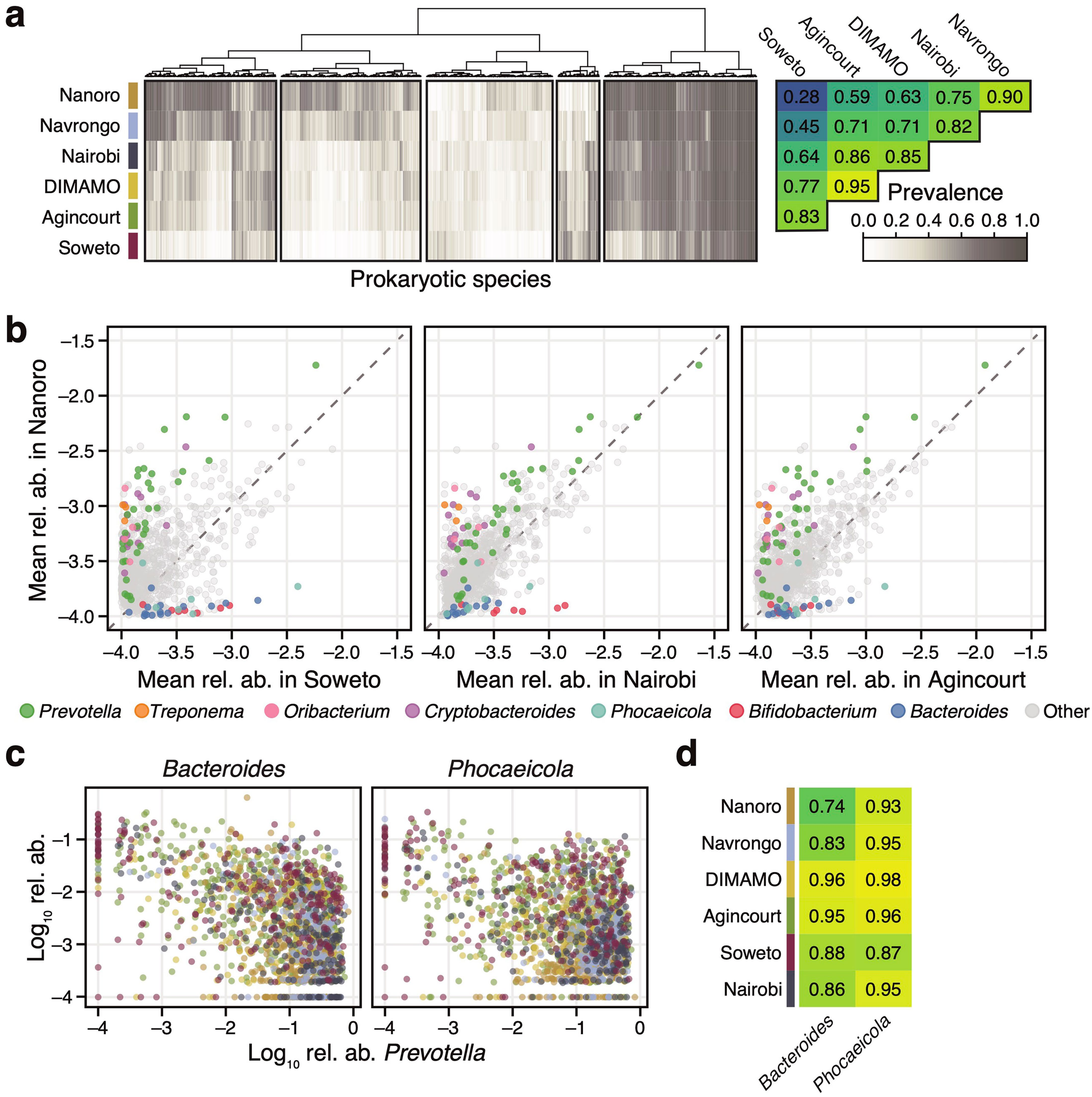
Extended Data Fig. 3 | Sequencing depth and database effects on taxonomic classification. a) Reads per sample throughout quality control, including original read count, and reads remaining after deduplication, end trimming, and removal of host reads. *P*-values indicate Kruskal-Wallis tests with Benjamini Hochberg multiple testing correction. **b**) Spearman correlation coefficient (Spearman's ρ) and R^2 for a linear model between phage richness (number of assembled phages) or prokaryotic richness (number of prokaryotic species present at ≥1 × 10^{-4}% relative abundance after rarefaction) and total read count and total assembly length (length of the total assembly in base pairs). Points represent individual samples. Blue line indicates a linear association model with 95% confidence intervals shown as shaded areas. **c**) Count and relative abundance of unassigned reads per sample, as estimated by the mOTUs profiler using the original database (v3.0.3) or extended database. **d**) Spearman's ρ between the original and extended database for each sample, separated by study site. Prokaryotic species with abundance of zero in both the original and extended database were removed on a per-sample basis. **e**) The cumulative abundance of the genomes added to the database for profiling are shown for each sample, separated by study site. Figures represent data from *n* = 1,796 samples. For all boxplots, boxes denote the interquartile range (IQR) with the median as a thick black line and the whiskers extending up to the most extreme points within 1.5-fold IQR.



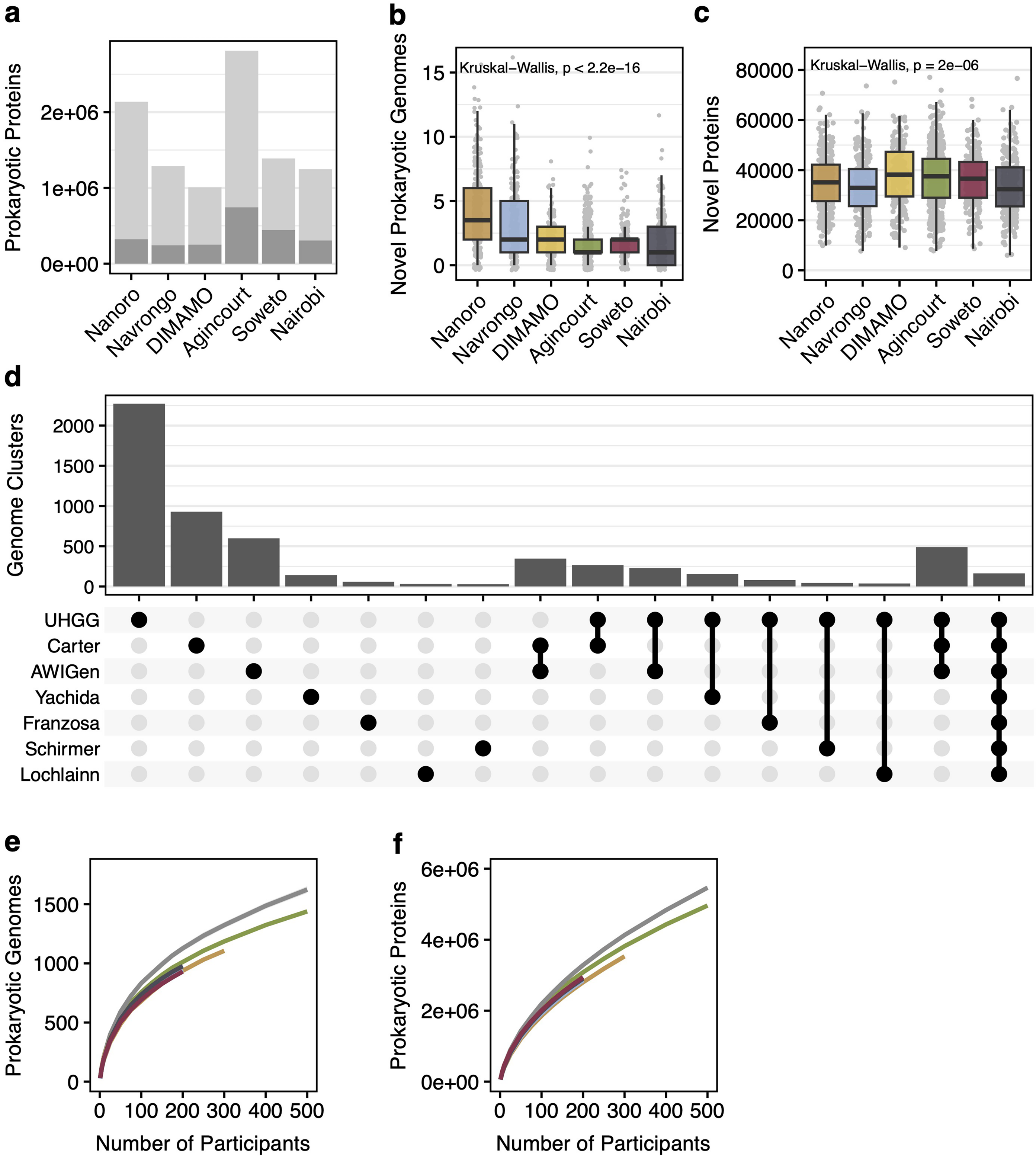
Extended Data Fig. 4 | Phylum-level differences between AWI-Gen sites and in external datasets. a) Spearman correlation coefficient (Spearman's ρ) between principal coordinate values and the relative abundance of selected prokaryotic phyla. Phyla with an absolute correlation coefficient higher than 0.5 for either of the first two principal coordinates are shown (see Fig. 1 in the main text). Points represent individual samples and are coloured by site. **b**) Principal coordinate analysis of all AWI-Gen 2 samples based on Bray-Curtis distance on species-level prokaryotic profiles together with other large datasets, color-coded by study. Franzosa et al. and Schirmer et al. are datasets collected in the USA and the Netherlands, focusing on patients with inflammatory bowel disease and healthy controls, respectively. Yachida et al. is a dataset from Japan for the study of colorectal cancer. **c**) Relative abundance of the most abundant phyla across the different datasets. Phyla are ordered by mean abundance across all included samples. Figures represent data from *n* = 1,796 AWI-Gen, *n* = 220 Franzosa et al., *n* = 471 Schirmer et al., and *n* = 645 Yachida et al. samples. For all boxplots, boxes denote the interquartile range (IQR) with the median as a thick black line and the whiskers extending up to the most extreme points within 1.5-fold IQR.



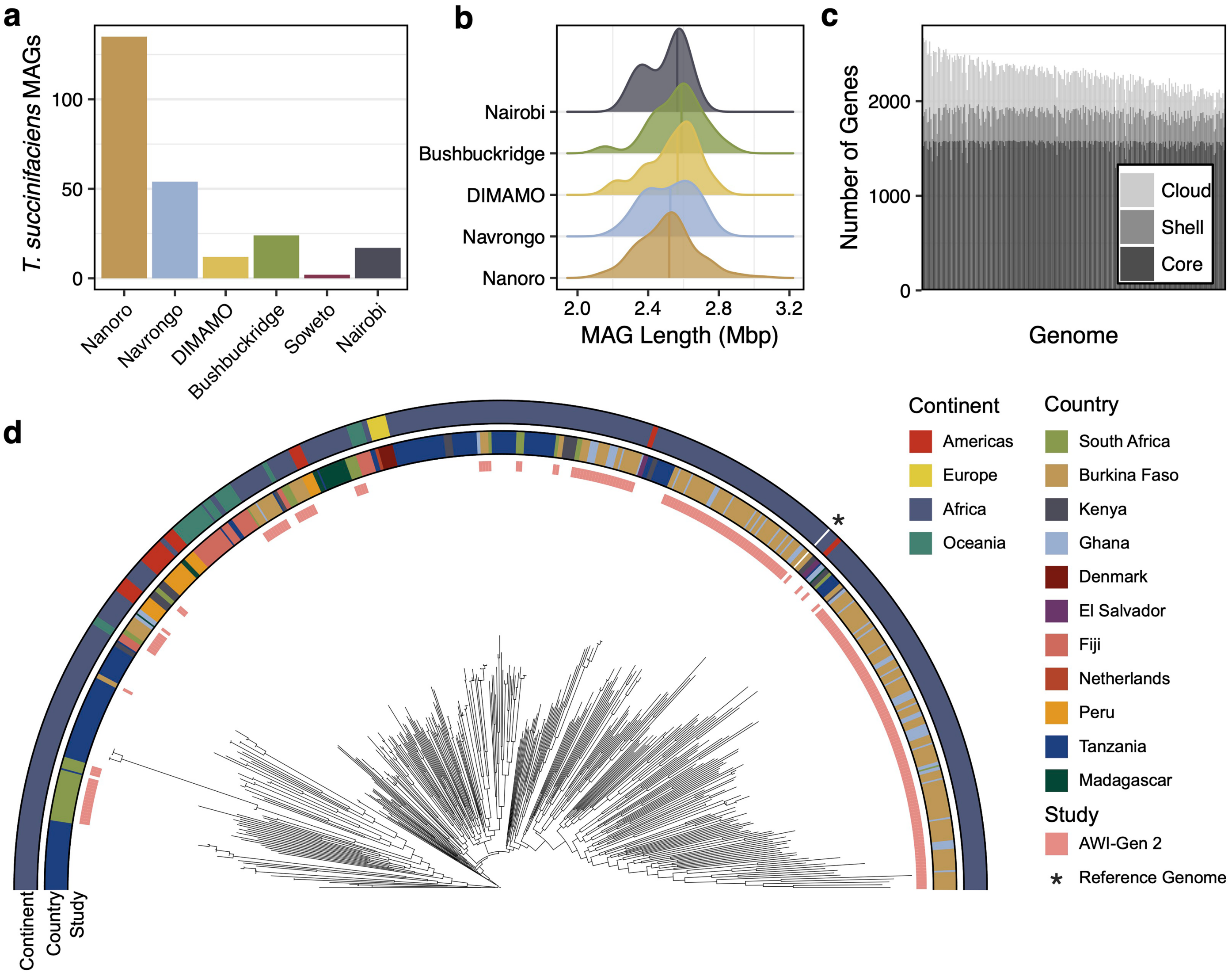
Extended Data Fig. 5 | Metadata correlation and distance-based redundancy analysis. a) Pearson correlation coefficient (Pearson's *r*) between available participant covariates, calculated on all participants included in the site comparison (*n* = 1,796). Non-numerical covariates were transformed into numerical values based on ordered factor levels (see Supplementary Methods). Asterisks indicate highly correlated covariates (Pearson's *r* ≥ 0.8). In those cases, the covariate that explained the higher amount of variance in the prokaryotic composition (see panel b) was selected (redundant variables are indicated by grey labels). **b**) The amount of variance in the prokaryotic composition that is explained by covariates in distance-based redundancy analysis. Blue bars indicate single-covariate models (each covariate associated with prokaryotic composition individually), whereas orange bars show the amount of variance explained in the iterative model in which the variable explaining the most additional variation is added iteratively to a multi-covariate model (see Methods). Covariates below the dashed line were removed before the iterative modelling since they were highly correlated with other covariates. BMI: body mass index, MVPA: moderate to vigorous physical activity, LDL: low-density lipoproteins, HDL: high-density lipoproteins, VAT: visceral adipose tissue, SCAT: subcutaneous adipose tissue, cIMT: carotid intima-media thickness.



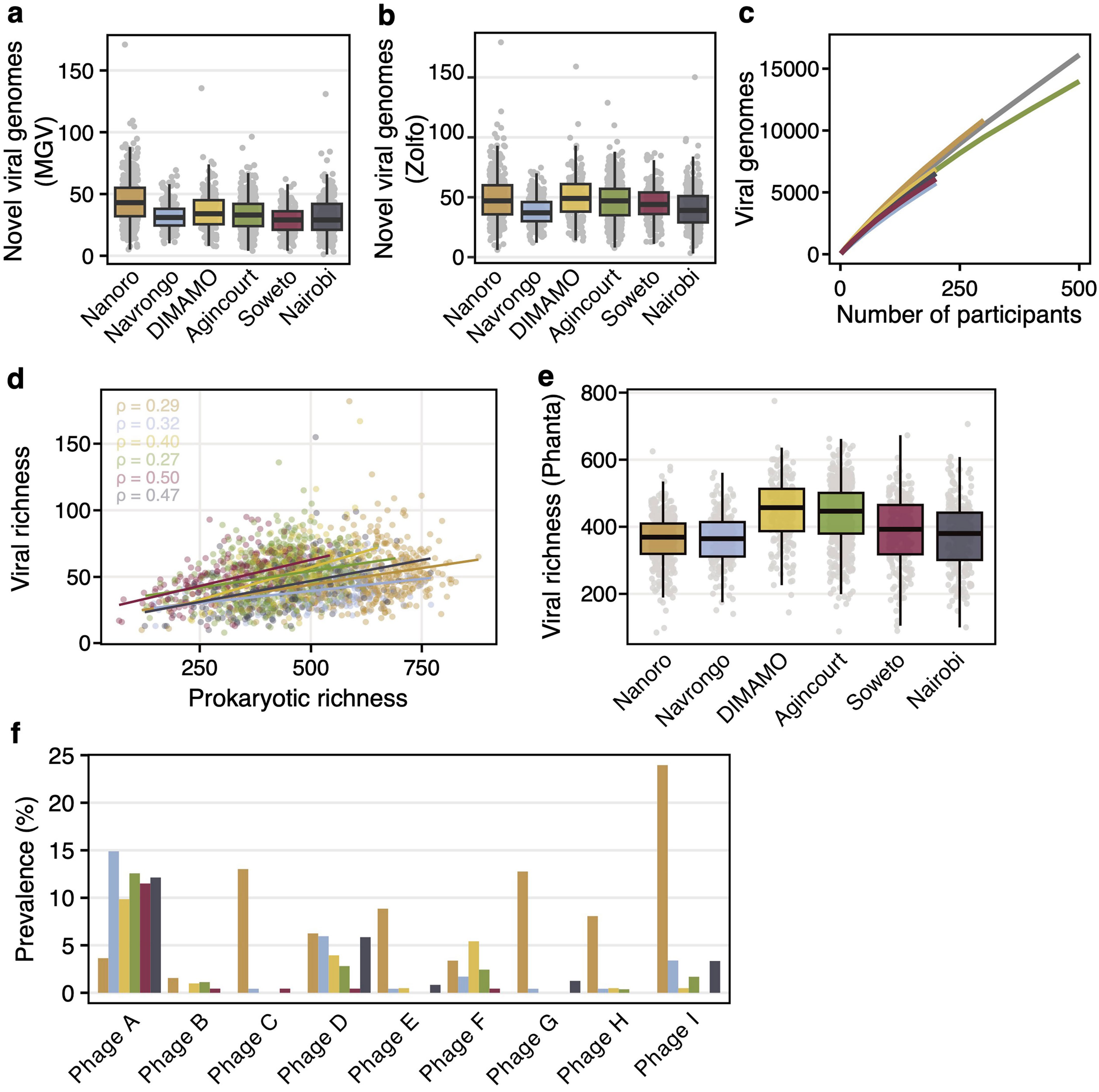
Extended Data Fig. 6 | Site-level prevalence and differential abundance of microbial taxa. a) The prevalence per site is shown for all prokaryotic species with prevalence higher than 5% in at least 2 sites ($n = 1{,}071$ species), clustered using the Ward algorithm as implemented in the R stats v4.2.2 package. Spearman correlation between sites is shown on the right. Population prevalence is calculated for each study site, where prevalence of zero indicates that the species is absent in all individuals in a site, and prevalence of one indicates that the species is present in all individuals in a site. **b**) The mean log10-transformed abundance of the same prokaryotic species as in a). Species that belong to the genera with the highest variance in fold change across all sites are highlighted by colours. **c**) The log10-transformed relative abundance of the genus *Prevotella* plotted against the relative abundance for the genera *Bacteroides* and *Phocaeicola*. Points represent $n = 1{,}796$ individual samples, coloured by site. **d**) The fraction of samples in which both *Prevotella* and either *Bacteroides* or *Phocaeicola* are present (relative abundance $\geq 1 \times 10^{-4}$) is shown across sites, indicating that these genera co-exist in most samples.



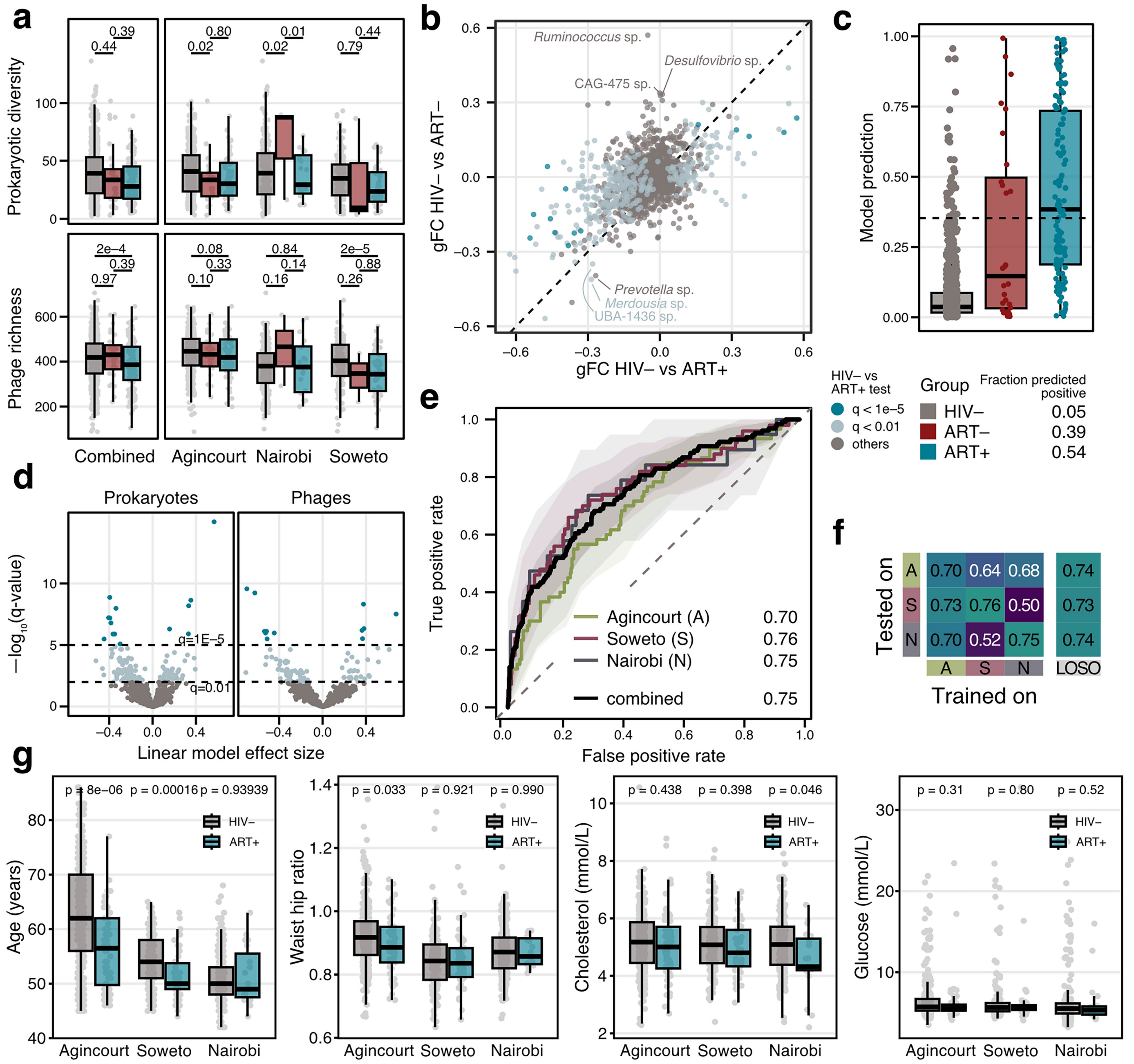
Extended Data Fig. 7 | Prokaryotic novelty in the AWI-Gen 2 cohort. a) Total number of novel and existing prokaryotic proteins in the AWI-Gen assemblies, relative to UHGP. Only representative proteins after feature clustering are represented. Number of novel **b**) prokaryotic genomes relative to the UHGG and **c**) prokaryotic proteins relative to the UHGP95 present in each sample. Points indicate the number of genome or protein clusters present per sample (n = 1,820 total samples) that are not found in respective feature databases. For all boxplots, boxes denote the interquartile range (IQR) with the median as a thick black line and the whiskers extending up to the most extreme points within 1.5-fold IQR. **d**) Comparison of number of representative genomes contributed by several metagenomic gut microbiome studies, including the UHGG[50] (global), Carter et al.[43] (Tanzania), Yachida et al.[44] (Japan), Franzosa et al.[45] (USA, Netherlands), Schirmer et al.[46] (western Europe), and Lochlainn et al.[47] (United Kingdom). The UpSet plot shows the number of genomes that are shared between or unique to each study. Note that Carter et al. performed ultra-deep metagenomic sequencing, leading to a high number of MAGs generated per individual sample. Rarefaction curves of the number of **e**) prokaryotic genomes and **f**) prokaryotic proteins detected as a function of the number of individuals sampled, by study site or from the full AWI-Gen sample set (grey). Each random subset was repeated a hundred times, and lines represent the mean feature count and standard deviation.



Extended Data Fig. 8 | Features of *Treponema succinifaciens* metagenome-assembled genomes (MAGs). a) Number of high-quality *T. succinifaciens* metagenome-assembled genomes by study site. **b**) Distribution of the length, in megabase pairs (Mbp), of each *T. succinifaciens* MAG. MAGs from Soweto are not pictured, as Soweto samples only contained two MAGs. **c**) Number of genes in each MAG that were classified as core (≥ 80% prevalence), shell (25 ≤ prevalence < 80%), or cloud genes (< 25% prevalence) in the complete MAG set. **d**) Midpoint-rooted phylogenetic tree of *T. succinifaciens* MAGs from this study (noted in pink inner ring) and public data sets ($n = 430$ total genomes). Middle ring indicates the country of origin, and outer ring indicates the continent of origin. White line and asterisk indicate the *T. succinifaciens* DSM 2489 type strain reference genome. PERMANOVA test indicates significant difference in phylogenetic distance by country of origin ($P = 0.001$).



Extended Data Fig. 9 | Additional characterization of viral novelty and diversity in the AWI-Gen 2 cohort. Number of novel viral genomes relative to the MGV (**a**) and the Zolfo et al. viral catalogue (**b**) present in each sample ($n = 1{,}820$ total samples). Points indicate the number of genome clusters present per sample that are not found in respective feature databases. **c**) Rarefaction curves of the number of viral genomes detected as a function of the number of individuals sampled, by study site or from the full AWI-Gen sample set (grey). Each random subset was repeated a hundred times, and lines represent the mean feature count. **d**) Spearman correlation coefficient (Spearman's ρ) between prokaryotic richness and viral richess. Points represent individual samples. **e**) Viral richness per sample, based on Phanta profiles (number of phage species clusters present ≥ 10^{-5}% relative abundance). **f**) Prevalence of jumbophages across sites, where prevalence indicates the percent of individuals at each site with 0.1× coverage of the indicated jumbophage genome, as measured by CoverM (see Methods). All colors indicate site, using colour-code in panel a. For all boxplots, boxes denote the interquartile range (IQR) with the median as a thick black line and the whiskers extending up to the most extreme points within 1.5-fold IQR.



Extended Data Fig. 10 | Phage, prokaryotic, and phenotypic differences in ART+ and ART- PLWH. a) Prokaryotic diversity (inverse Simspon's index after rarefaction) and phage richness (species present at ≥10^{-5}% abundance) by HIV and antiretroviral therapy status. Points represent individual samples. Differences in diversity by site were tested with ANOVA and across sites with a linear mixed effect model accounting for site as a random effect. **b**) Generalized fold change (gFC) for all species in HIV+ ART+ relative to HIV− individuals and for HIV+ ART− relative to HIV− individuals. Species are coloured by *q*-value in HIV− vs HIV+ ART+ comparison. Species with an absolute gFC ≥ 0.3 in the HIV− vs HIV+ ART− comparison (that do not exhibit a gFC ≥ 0.3 in the HIV− vs HIV+ ART+ comparison) are annotated. **c**) Prediction from machine learning model trained prokaryotic data from HIV− and HIV+ ART+ participants and applied to HIV+ ART- participants. Sample fraction predicted to be positive at a 5% internal false positive rate (dashed line) is listed below. **d**) HIV-associated effect size for prokaryotic and phage species. Species are colored by *q*-value. **e**) Receiver-operating characteristic (ROC) for models trained to distinguish HIV status using phage composition. Shading indicates 95% confidence intervals and numbers show area under the ROC curve (AU-ROC). **f**) AU-ROC for models trained on participants from each site (panel e) and applied to other sites. Models were trained on two sites and validated on the left-out site for leave-one-site-out (LOSO) validation. **g**) Statistics for age, waist-to-hip ratio, cholesterol, and glucose for individuals who are HIV seronegative and seropositive on ART. All *p*-values result from Wilcox rank sum test. For all panels, *n* = 129 HIV+ ART+, *n* = 28 HIV+ ART−, *n* = 719 HIV−. For all boxplots, boxes denote the interquartile range (IQR) with the median as a thick black line and the whiskers extending up to the most extreme points within 1.5-fold IQR.

nature portfolio

Corresponding author(s): Ami S. Bhatt, Scott Hazelhurst

Last updated by author(s): November 28, 2024

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	REDCap v14.7.3 was used for collection of participant data.
Data analysis	Workflows for read quality control, classification, assembly, and binning are available at https://github.com/bhattlab/AWIGen2Microbiome/. Tools used for data analysis include: SuperDeduper v1.3.3, TrimGalore v0.6.7, BWA v0.7.17, mOTUs v3.0.3, Phanta v1.1.0, vegan v2.6-4, labdsv v2.1-0, megahit v1.2.9, QUAST v5.2.0, prodigal v2.6.3, MetaBAT v2.5, CONCOCT v1.1.0, MaxBin v2.2.7, DAS Tool v1.1.6, CheckM v1.2.2, dRep v3.4.3, GTDB-tk v2.3.0, iTOL v6, VIBRANT v1.2.1, BLAST v2.14.0, roary v3.12.0, MAFFT v7.407, FastTree v2.1.11, R v4.1.2, ggplot2 v3.4.2, cowplot v1.1.1, tidyverse v2.0.0, mmseqs v14.7e284, checkV v1.0.1, lmerTest v3.1-3, SIAMCAT v2.5.0, pROC v1.18.2., bakta v1.8.2, DendroPy v4.6.1, CoverM v0.7.0, fetchMG v1.2, dbCAN3 v4.1.4, stats v4.2.2

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

To maximise public availability our data while protecting participants, we have split our data in two sets: (a) sequences that may contain human reads — this is available from EGA as EGADXXXXXXX on application to the H3Africa Data and Biospecimens Access Committee; and (b) an open data set available from the NCBI Sequence Read Archive as PRJNA115737 (see Supplementary Methods). The dereplicated genome sets are available at 10.5281/zenodo.13761309. Participant phenotype is stored at the European Genome-Phenome Archive under accession EGAD00001015440. Participant phenotype data are under restricted access due to ethics requirements of the AWI-Gen 2 study. Applications must be made to the independent Human Heredity and Health in Africa Data and Biospecimen Access Committee by registering and applying at https://catalog.h3africa.org/. The H3Africa Data Sharing Policy can be found at https://h3africa.org/wp-content/uploads/2020/06/H3Africa-Consortium-Data-Access-Release-Policy-April-2020.pdf. Decisions on requests made by the third week of the month should be made by the end of the subsequent month. Source data, including classification tables, genome summary statistics, taxon prevalence, and differential feature tables, are available as source data within the manuscript.
Reference data used in this study are available from the Unified Human Gastrointestinal Genome collection in the European Nucleotide Archive under project accession PRJEB33885, the Metagenomic Gut Virus catalogue at https://portal.nersc.gov/MGV, and the Genome Taxonomy Database at https://data.gtdb.ecogenomic.org/releases/. The hg38 human reference genome is available at NCBI Genome under accession number GCF_000001405.26. .

Research involving human participants, their data, or biological material

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

Reporting on sex and gender

We aimed at recruiting self-identified females primarily because menopause and hormonal transition is another arm of the AWI-Gen 2 study and we want to explore the link in a subsequent paper (see Supplementary Methods). A small number of self-identified males were recruited. We recruited 1,803 females and 17 males. Males were excluded from all comparative analyses to avoid confounding. However, a comparison of microbiome composition between male and female participants is included in the Supplementary Figures.

Reporting on race, ethnicity, or other socially relevant groupings

Race, ethnicity, and other socially relevant groupings were not reported on within this study. We controlled for confounding variables by conducting cross-sectional recruitment at each study site. That said, each study site has a different ancestry background, which is accounted for by adjusting for 'site' as a confounder in comparative analyses.

Population characteristics

Participants are 42-86 years old. Individuals were recruited cross-sectionally, and therefore have a range of health conditions that are typical of the larger populations at each study site, including hypertension, obesity, HIV, and others.

Recruitment

The AWI-Gen 2 study performed cross-sectional recruitment based on population records from the participating Health and Demographic Surveillance Sites. The participants in the microbiome sub-study were semi-randomly chosen from the larger AWI-Gen 2 participant pool, with a randomization approach that favored balancing participants by menopause and hypertension status. Approximately 35% of the overall AWI-Gen participants were lost to follow up in the second study phase. Participants were cross-sectionally and randomly sampled across each site catchment area, and therefore we do not expect a strong self-selection bias.

Ethics oversight

Human subjects research approval was obtained from the University of the Witwatersrand Human Research Ethics Committee, Clearance Certificate No. M170880, M2210108.

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size

Microbiome data was generated for 1,820 participants. No sample size calculation was performed. Sample sizes were chosen based on study feasibility and budget, which allowed for target enrollment of at least 200 individuals from each site, with greater enrollment at Agincour and Nanoro to support orthogonal research efforts. These sample sizes are sufficient to capture a range of microbiome diversity and to detect microbial taxa with significantly different prevalence and abundance across groups, based on sample sizes used in previous literature and our pilot studies in Agincourt and Soweto.

Data exclusions

Samples that did not meet DNA yield thresholds for shotgun sequencing were excluded from these analyses. 1,820 participants had sequencing data that met inclusion thresholds, which were determined prior to analysis. Samples from men and samples with high proportions of human reads were excluded from analyses comparing microbiome composition between sites, to avoid confounding.

Replication

As this study compares samples from six populations, we did not replicate the full study with an additional set of sample collection. The analysis in this study is reproducible through the availability of our computational pipelines (see Data and Code Availability).

Randomization

Participants/samples were not randomized into experimental groups for this study as we compared the microbiomes of six distinct communities. Samples were randomized across plates for DNA extraction and sequencing to avoid batch effects between groups. Covariates between groups were not controlled for, as the purpose of this study was to understand how geography and lifestyle covariates relate to microbiome composition.

Blinding

Investigators were not blinded to group during data collection. Blinding was not possible as participants were surveyed at their respective locales. Blinding was not relevant to our study as all data were processed through the same computational pipelines.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Plants

Seed stocks

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

Novel plant genotypes

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

Authentication

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