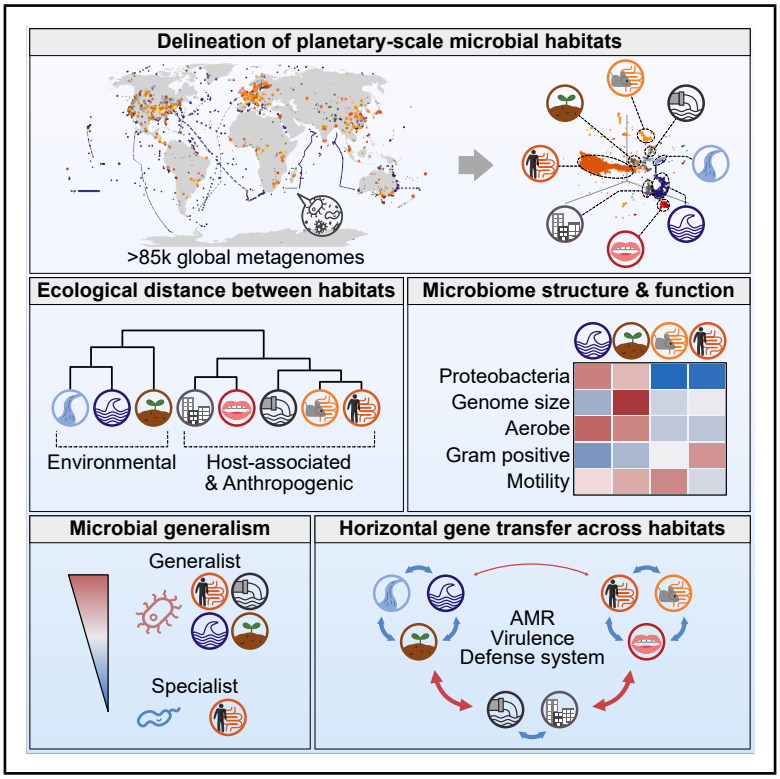
Cell

Article

Planetary microbiome structure and generalist-driven gene flow across disparate habitats

Graphical abstract



Authors

Chan Yeong Kim, Daniel Podlesny, Jonas Schiller, ..., Anastasiia Grekova, Michael Kuhn, Peer Bork

Correspondence

mkuhn@embl.de

In brief

A planetary-scale analysis of over 85,000 metagenomes establishes a framework for exploring the structure and drivers of global microbial habitats, revealing that generalist species bridge ecological boundaries to mediate gene flow across disparate habitats, including the dissemination of antimicrobial resistance.

Highlights

- Species-based clustering of >85,000 global metagenomes delineates 40 microbial habitat types
- Host and environmental filtering shape planetary microbiome structure
- Generalists facilitate horizontal gene flow across ecologically disparate habitats
- Anthropogenic habitats drive generalist-mediated resistance gene spread

Kim et al., 2026, Cell *189*, 2073–2091
April 2, 2026
https://doi.org/10.1016/j.cell.2025.12.051

CellPress

Cell

CellPress
OPEN ACCESS

Article

Planetary microbiome structure and generalist-driven gene flow across disparate habitats

Chan Yeong Kim,[1,6] Daniel Podlesny,[1,6] Jonas Schiller,[1,6] Supriya Khedkar,[1,3] Anthony Fullam,[1] Askarbek Orakov,[1,4,5] Christian Schudoma,[1] Shahriyar Mahdi Robbani,[1] Anastasiia Grekova,[1] Michael Kuhn,[1,8,*] and Peer Bork[1,2,7]

[1]European Molecular Biology Laboratory, Molecular Systems Biology Unit, 69117 Heidelberg, Germany
[2]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany
[3]Present address: BioQuant, Heidelberg University, 69120 Heidelberg, Germany
[4]Present address: School of Public Health, Harvard University, Boston, MA 02115, USA
[5]Present address: Boston Children's Hospital, Boston, MA 02115, USA
[6]These authors contributed equally
[7]Deceased
[8]Lead contact
*Correspondence: mkuhn@embl.de
https://doi.org/10.1016/j.cell.2025.12.051

SUMMARY

Microbes are ubiquitous on Earth, forming microbiomes that sustain macroscopic life and biogeochemical cycles. Microbial dispersal, driven by natural processes and human activities, interconnects microbiomes across habitats, yet most comparative studies focus on specific ecosystems. To study planetary microbiome structure, function, and inter-habitat interactions, we systematically integrated 85,604 public metagenomes spanning diverse habitats worldwide. Using species-based unsupervised clustering and parameter modeling, we delineated 40 habitat clusters and quantified their ecological similarity. Our framework identified key drivers shaping microbiome structure, such as ocean temperature and host lifestyle. Regardless of biogeography, microbiomes were structured primarily by host-associated or environmental conditions, also reflected in genomic and functional traits inferred from 2,065,975 genomes. Generalists emerged as vehicles thriving and facilitating gene flow across ecologically disparate habitat types, illustrated by generalist-mediated horizontal transfer of an antibiotic resistance island across human gut and wastewater, further dispersing to environmental habitats, exemplifying human impact on the planetary microbiome.

INTRODUCTION

Microbes are essential members of almost every ecosystem on our planet. They usually form communities, microbiomes, that drive ecosystem processes and biogeochemical cycles[1] and sustain macroscopic life, often through direct associations with their hosts.[2] Although most microbial species and genes are habitat specific,[3–5] various mechanisms facilitate their dispersal at different temporal and spatial scales, interconnecting microbiomes across diverse and geographically distant habitats.[6–9] Humans amplify this connectivity, exemplified by the rapid spread of infections and antimicrobial resistance (AMR) across the globe.[10–12] Microbial dispersal and environmental selection fundamentally shape microbiome structure (taxonomic composition),[13–15] which is composed of an evolving set of lineages that form a gradient between strictly habitat-specific species (specialists) and species that thrive in multiple and disparate habitats (generalists). However, the extent to which species traverse across habitats and the mechanisms by which they mediate gene flow at a planetary scale remain unclear.

Numerous large-scale studies of microbiome structure and function have been conducted, yet most focus on specific ecosystems (e.g., US riverine,[16] global ocean,[17] topsoil,[18] and human intestine[19]), are restricted to a national scale,[20] are relatively sparse,[21] or are limited in resolving fine-grained taxonomy.[19,22] Moreover, habitat annotations and contextual information are often incomplete or inconsistent, and human-defined ontologies of environments[23,24] may not always align with microbial habitats. Recently, databases such as SPIRE[24] or MicrobeAtlas[25] have emerged, offering uniformly processed sequencing data and annotations to facilitate planetary-scale comparative microbiome analyses. However, to fully leverage such resources, a standardized planetary-scale framework is necessary to contextualize habitat types and quantify their connectivity, disentangle the local, regional, and global drivers influencing the transmission of species and flow of genes, and better understand the

evolution of microbiome structure (specifically at the species level) and function (specifically at the gene level).

In this study, we employed an unsupervised species-based clustering and parameter modeling approach on 85,604 samples from the SPIRE database, which is the largest integration and systematic analysis of shotgun metagenomic data to date.[24] We identified 40 microbial habitat types (hereafter referred to as habitat clusters) that broadly align with existing environmental contextual data. We quantified their ecological similarity and provided a comprehensive hierarchical annotation that reflects associated functional traits and known environmental conditions. This framework enables the study of fundamental questions in microbial ecology,[26] e.g., how to trace microbial dispersal, how microbiomes are shaped, and how they are affected by human activities.

We integrated data from each of the 40 habitat clusters to compare their structural and functional composition. To systematically study gene flow on a planetary scale, we captured both horizontal gene transfer (HGT) between sympatric species as well as species transmission across habitat clusters. By devising a robust scoring scheme to quantify the gradient from prokaryotic specialism to generalism, we identified generalists as critical vehicles for gene flow between ecologically disparate habitat clusters. This is illustrated by the identification of an AMR island, widely shared among human gut specialists and a generalist species, that was able to spread across gut, wastewater (WT), lake, and estuary microbiomes. Collectively, our findings highlight the profound ecological and health implications of microbial gene flow, to which humans contribute.

RESULTS

Prokaryotic species composition reveals habitats at high resolution

To delineate microbial habitats and associated functional traits, we adopted a data-driven metagenomic approach. Leveraging the SPIRE resource,[24] we utilized a set of 85,604 quality-filtered metagenomic samples, collected across the globe. To identify metagenomic features and parameter conditions under which samples would naturally group into ecologically meaningful habitats, we systematically tested an extensive array of input profiles (taxonomic, functional, and *k*-mer-based), dimensionality reduction techniques, and unsupervised clustering methods, resulting in more than 2.06 million model parameter combinations (Figures 1A and S1A; Table S1A; STAR Methods).

We evaluated each model parameter combination using quantitative metrics that assess both cluster homogeneity and consistency (STAR Methods). Because marked differences in microbiomes among broad habitats (e.g., ocean, terrestrial, skin, and gut) are evident,[5,22,27] we imposed a penalty on their fusion. However, to allow for detection of substructures within broad habitats, we did not penalize their fission. In parallel, we assessed cluster cohesion and separation by computing the silhouette index. These complementary criteria ensured that models producing homogeneous clusters with clear boundaries were prioritized (STAR Methods). Our comprehensive parameter exploration revealed that applying Uniform Manifold Approximation and Projection (UMAP[28]) to species-level taxonomic profiles most effectively captured subtle differences among habitats (Figures S1B and S1C). Additionally, distance metrics computed on profiles transformed with logarithmic, arcsine, or centered log ratio (CLR), which emphasize lower abundance microbes, enhanced clustering performance (Figures S1D and S1E). While previous research reported that functional profiles provide lower resolution than taxonomic profiles in human microbiomes,[29] we show that this pattern extends across microbiomes from a wide range of environments: functional profiles, at the resolution of clusters of orthologous groups (COG), enabled broad distinctions such as host-associated versus environmental samples but failed to resolve finer-scale differences (Figures S2A and S2B). Similarly, *k*-mer-based approaches,[30] though more discriminative than functional profiles, did not surpass the specificity achieved by taxonomic data (Figures S2C and S2D). Based on the parameter exploration, we selected the best-performing model, which we then curated to eliminate artifacts arising from individual studies or methodological biases (Figures S2E–S2G; STAR Methods). The remaining 71,871 metagenomic samples formed 40 robust habitat clusters (Figures 1B and S2H; Tables S1B and S1C).

Host and environmental filtering structure global microbial communities

As species are mostly habitat specific,[4,5] each of the 40 clusters likely reflects a distinctive habitat condition. To construct a planetary-scale map of microbial habitats, we annotated the clusters using fine-grained and manually curated contextual data[31] where available and evaluated their ecological distances (Figure 2A). The ecological distance was quantified based on the overlap coefficient of microbial taxa, instead of relying on distances in the UMAP embedding, which prioritizes preserving local over global structure (STAR Methods). This revealed a primary structuring of all studied microbiomes into host-associated (Figure 2A, clade m) and environmental (Figure 2A, clade d) groups, regardless of the geographic sampling locations. Within these primary structures, we identified further fine-grained differentiation shaped by large-scale environmental gradients or host-specific conditions, each of which left distinct signatures in the microbiome structure. We were able to contextualize these signatures with previous literature, indicating the robustness of the framework, as illustrated with three examples: (1) human gut habitat clusters, reflecting age and lifestyle[32]; (2) animal gut habitat clusters, shaped by hosts and their diets[33,34]; and (3) marine habitat clusters, primarily structured by temperature rather than geographic proximity[17] (Figure 2; see STAR Methods for the major environmental factors and the annotation rationale for other habitat clusters).

(1) Since the human gut accounted for the largest number of samples (46,971), we achieved particularly high resolution in delineating its microbiome structure. We identified nine distinct human gut habitat clusters that separated along known age-related and lifestyle gradients. The adult fecal microbiomes were represented in human gut clusters HG1–HG5 and HG7 (for nomenclature, see Figure 2A) and clearly separated from a clade with newborn gut (NBG) and infant gut (IFG) species profiles (average

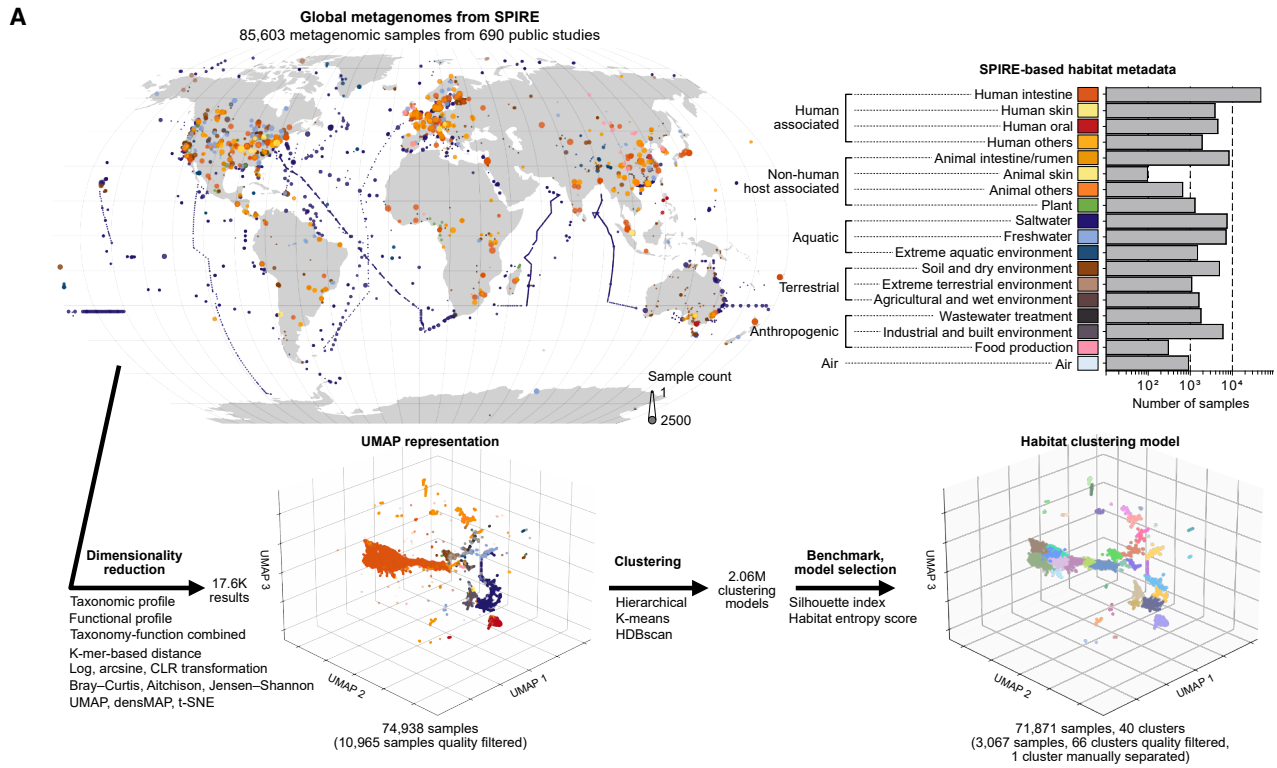


Figure 1. Planetary-scale framework for microbiome-based habitat clustering

(A) Overview of the sample collection, dimensionality reduction, clustering, and benchmarking. Upper panels show the geographical distribution of the analyzed metagenomes and the sample count per broad habitat category. 3D UMAP embedding based on Bray-Curtis dissimilarity of species compositions, used for clustering and downstream analyzes (STAR Methods).

(B) Forty clusters emerging from the best-performing habitat clustering model, showing sample count per cluster, cluster IDs from the UMAP embedding (A), and composition according to broad habitat categories (A).

See also Figures S1 and S2.

ages of 15 days and 7 months, respectively; Table S2A; Figure 2A, clade i). Following the age trajectory, the HG6 cluster (average age of 2 years) was already more similar to the adult microbiomes[35–37] (Figures 2B and 2C). In adulthood, clusters were no longer segregated by age, but lifestyle differences became increasingly prominent. Notably, HG1 was distinctly segregated from other human gut habitat clusters (Figures 2A and 2B) and was enriched in fecal samples predominantly from hunter-gatherers, indigenous Malaysian populations, paleofeces, and middle- and low-income economies[38] (two-sided Fisher's exact test, Bonferroni-corrected $p < 0.1$ and odds ratio >1; Figures S3A and S3B). This aligns with previous reports that individuals with traditional lifestyles tend to have a markedly different gut microbiota composition from those in industrialized settings.[39,40] Enrichment of samples from specific geographical regions was also evident in HG2, HG3, and HG5, which were dominated by samples from Israel, Europe, and East Asia, respectively (Figure S3A). Thus, the planetary-scale

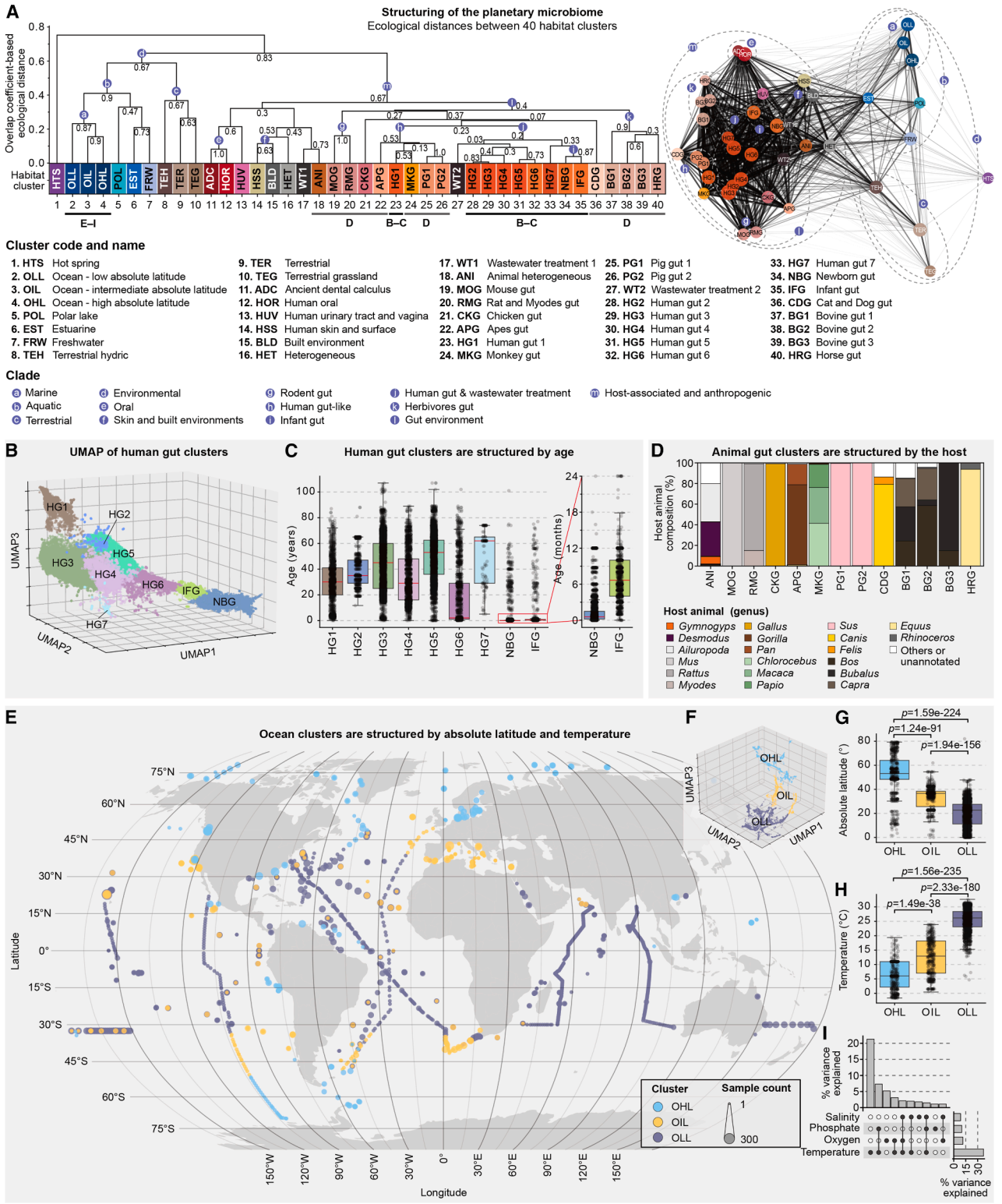


Figure 2. Annotation, ecological distances, and sample characteristics of the habitat clusters

(A) Primary structuring of the planetary microbiome into host-associated and environmental habitats, interconnected by anthropogenic microbiomes. Dendrogram (left) and network (right) showing habitat cluster ecological distances based on species profile (dis)similarity. Dendrogram: hierarchical clustering, quantifying ecological distance using the overlap coefficient of microbial taxa. Tree node values represent how consistently each clade is preserved across

(legend continued on next page)

microbiome clustering captured the major conditions underlying human gut microbiome variation.

(2) Animal gut clusters exhibited a high degree of host specificity, each comprising samples from a single animal genus or closely related genera, with the exception of a heterogeneous animal cluster (ANI) (Figure 2D; Table S2B). Primate (MKG for monkey and APG for apes) and pig (PG1 and PG2) clusters were taxonomically closer to the human gut clusters, whereas herbivores (BG1–BG3 for bovine and HRG for horse gut) and rodents (MOG for mouse and RMG for rat and myodes) formed distinct clades distant from the human gut microbiota (Figure 2A). These observations substantiate previous findings that host diet and phylogeny are strong drivers of microbial community composition.[33,34]

(3) Marine samples grouped into three habitat clusters (OHL, OIL, and OLL representing high, intermediate, and low absolute latitude, respectively) with significantly different absolute latitude ranges (two-sided Mann-Whitney U test $p < 1e{-}90$; Figures 2E–2G; Table S2C). Samples from temperate to polar zones from both hemispheres clustered together, reflecting similarities in species-level taxonomic profiles despite vast geographic distances.[41] To assess which environmental factors align best with the ocean microbiome clustering, we applied variance partitioning based on temperature, dissolved oxygen, phosphate, and salinity levels. Missing environmental measurements were imputed from the World Bank Group Ocean Atlas[38] (WOA23[42]), showing high agreement with the available *in situ* data (Spearman's $\rho = 0.98$ for temperature, 0.90 for dissolved oxygen, 0.88 for salinity, and 0.72 for phosphate; Figure S3C; STAR Methods). Temperature alone was most explanatory for the ocean microbiome clustering (adjusted $R^2 = 21\%$; Figures 2H and 2I) and explained additional shared variation jointly with phosphate (7%) and dissolved oxygen (3%) (Figure 2I), supporting temperature as the primary factor shaping global marine microbiome structure.[17]

Beyond these examples, the robustness of our habitat clustering approach enabled us to flag potential contamination and to classify clusters when habitat annotations were missing from the public data. For example, mapping of metagenomic samples lacking host annotations onto habitat clusters accurately recovered the host animal, as validated by host-DNA contamination analysis (Figure S3D; STAR Methods). Similarly, aquatic samples that clustered into the human skin cluster (HSS) (Figure S3E) exhibited significantly higher human DNA content compared with correctly clustered samples (two-sided Mann-Whitney U test $p < 1e{-}09$; Figure S3F).

Habitat conditions select for microbial functions and genomic traits

Metagenomics offers not only high taxonomic resolution but also rich functional information, providing insights into genomic and phenotypic traits of microbes. To examine how these traits shape structural differences between habitat clusters, we analyzed 2,065,975 isolate genomes and high-quality metagenome-assembled genomes (MAGs) from species within those clusters (Figures 3 and S4A; Tables S3A–S3F; STAR Methods; trait information for all taxa is accessible on the metaTraits[43] website (https://metatraits.embl.de/)).

Taxonomic composition differed markedly between habitat clusters, even at the phylum level. Whereas Firmicutes_A and Bacteroidota dominated host-associated microbiomes, Proteobacteria and Actinobacteriota were prevalent in environmental habitats (Figure 3A; Table S3A). Rarefaction analysis revealed that at equal numbers of samples, all host-associated habitats (Figure 2A, clade m) together (but not individually) harbored greater diversity of species compared with terrestrial (Figure 2A, clade c) and aquatic (Figure 2A, clade b) microbiomes. While host-associated microbiomes support a high number of closely related species[44,45] (Figure S4B), at the genus level, a broader diversity of lineages was observed across aquatic and terrestrial habitats (Figure S4C). Rarefaction analysis further indicated the ongoing discovery of taxa in terrestrial, aquatic, and animal gut microbiomes (and to a lesser extent in the human gut microbiome) (Figures S4B and S4C), suggesting numerous undiscovered taxa in these habitat types.

Functional composition, based on broad COGs, was more conserved across the habitat clusters than taxonomic composition (Figures 3A and 3B; Table S3D), which might be due to a large fraction of shared core functionalities between microbial communities.[17,46] Still, there were marked differences between host-associated (Figure 2A, clade m) and environmental (Figure 2A, clade d) habitat clusters. Environmental habitat clusters exhibited higher proportions of cellular processes (COG

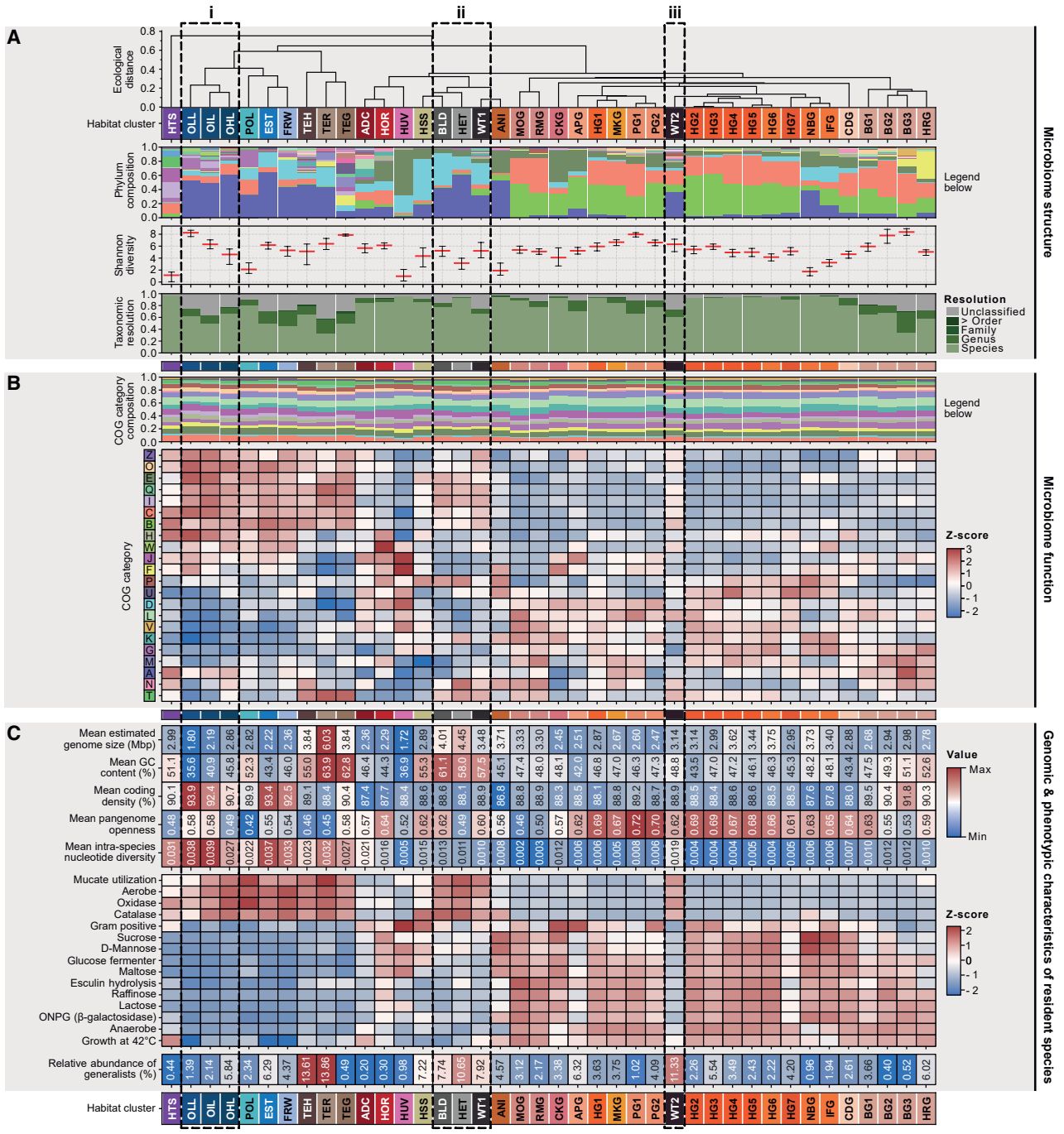
different prevalence cutoffs. Network: the positions of the habitat clusters (nodes) were determined by Prefuse Force-Directed Layout using negative logarithmic ecological distance (edges) as weights. Clusters were annotated based on the characteristics of their constituent samples (inferred from fine-grained contextual data) and are represented by three-letter abbreviations. Letters (B–I) below the cluster indices link to downstream panels.

(B and C) Human gut clusters are structured by age: (B) UMAP embedding of the nine human gut clusters shows an age-dependent gradient from newborns (NBG) to infants (IFG) and adults (HG1–HG7). (C) Age distribution of sampled hosts. The boxplot on the right zooms into the age distribution for the newborn (NBG) and infant gut (IFG) clusters. 91.22% of NBG samples were annotated as being from individuals aged 6 months or younger, and 93.63% of the IFG samples were from individuals aged 24 months or younger.

(D) Animal gut clusters are structured by the host animal: genus-level composition of the host animal phylogeny in non-human animal gut clusters. For example, the monkey gut cluster (MKG) consists of microbiome samples from *Chlorocebus*, *Macaca*, and *Papio*, which are all members of the *Cercopithecidae* family.

(E–I) Ocean samples cluster by absolute latitude: (E) geographical distribution, (F) UMAP embedding, (G) absolute latitude, and (H) temperature of samples in three ocean clusters (OHL, OIL, and OLL), and (I) variance partitioning of ocean cluster composition explained by temperature, phosphate, oxygen, and salinity. Bars show the fraction of variation (adjusted R^2) explained by each variable or their combinations. (G and H) *p* values were calculated using the two-sided Mann-Whitney U test.

See also Figure S3.



20 most abundant phyla (From the most abundant phylum, Proteobacteria, to 20th most abundant phylum, Fibrobacterota. Stacked from bottom)

Proteobacteria, Actinobacteriota, Cyanobacteria, Chloroflexota, Acidobacteriota, Campylobacterota, Gemmatimonadota, Firmicutes_A, Firmicutes, Firmicutes_C, Patescibacteria, Fusobacteriota, Thermoplasmatota*, Fibrobacterota, Bacteroidota, Verrucomicrobiota, Thermoproteota*, Methanobacteriota*, Spirochaetota, Deinococcota, (* Archaeal phylum)

COG category

- Z Cytoskeleton
- O Post-translational modification, protein turnover, chaperones
- E Amino acid transport and metabolism
- Q Secondary metabolites biosynthesis, transport, & catabolism
- I Lipid transport and metabolism
- C Energy production and conversion
- B Chromatin structure and dynamics
- H Coenzyme transport and metabolism
- W Extracellular structures
- J Translation, ribosomal structure & biogenesis
- F Nucleotide transport and metabolism
- P Inorganic ion transport and metabolism
- U Intracellular trafficking, secretion, and vesicular transport
- D Cell cycle control, cell division, chromosome partitioning
- L Replication, recombination and repair
- V Defense mechanisms
- K Transcription
- G Carbohydrate transport and metabolism
- M Cell wall/membrane/envelope biogenesis
- A RNA processing and modification
- N Cell motility
- T Signal transduction mechanisms

(legend on next page)

categories B, O, and W) and metabolic functions (COG categories C, E, H, I, and Q), pointing to high metabolic versatility (Figure 3B). They were also enriched in traits related to aerobic conditions, whereas human and animal gut clusters were dominated by species with anaerobic lifestyles and higher potential to utilize sugars like sucrose, D-mannose, glucose, maltose, raffinose, and lactose (Figure 3C).

Microbes in each habitat optimized their genetic repertoires differently (Tables S3E and S3F): environmental microbiomes showed high intra-species nucleotide diversity but closed intra-species gene pools (pangenomes), indicating more frequent adaptation through point mutations than HGT. Host-associated microbiomes, conversely, exhibited low intra-species nucleotide diversity and often open pangenomes, especially in the gut of omnivorous animals (humans, pigs, apes, and monkeys) (Figure 3C), suggesting that they instead rely more on HGT for functional innovation, which requires spatial proximity.[47] Genome sizes of host-associated microbiomes were constrained to an upper limit of around 3.5 Mbp, with the human urinary tract and vagina cluster (HUV) displaying the smallest average genome size (1.72 Mbp). Aquatic clusters had compact genomes (<2.86 Mbp) with high coding density (90%–94%) and low GC content (<46%), whereas terrestrial clusters featured less streamlined genomes with larger genomes (>3.84 Mbp), lower coding density (88%–90%), and high GC content (>55%) (Figure 3C). This is consistent with previous reports[48–50] and may reflect a trade-off from balancing the energetic costs of genome maintenance with the need for adaptive potential under fluctuating habitat conditions. Since these patterns are visible at the level of habitat clusters, habitat-specific conditions may play a dominant role in shaping these optimization strategies and thereby genomic characteristics.[51]

We observed three ocean clusters (OIL, OLL, and OHL) separated along distinct gradient patterns. The latitude segregation correlated with taxonomic Shannon diversity, peaking in low-latitude tropical and subtropical regions (OLL) and declining toward intermediate latitudes beyond the subtropics (OIL) and high latitudes (OHL) (Figure 3A, dashed rectangle i). This transition was accompanied by an increase in genome size and GC content of the microbes inhabiting these communities, consistent with previous findings[52] (Figure 3C). Additionally, the proportion of aerobic microbes increased toward high latitudes, coinciding with higher dissolved oxygen levels in colder waters (Figure 3C). Microbiomes from OLL possessed a high abundance of amino acid transporter genes[53] (COG E), while motility-related functions (COG N) were more frequent in microbiomes of high latitudes (OHL), potentially facilitating attachment to particulate organic matter, which is abundant in (sub)polar oceans[54] (Figure 3B). These results highlight the impact of gradient-dependent environmental filtering on the function and diversity of the ocean microbiome.

Together, host and environmental filtering of global microbiome structure was reflected by functional composition and genomic adaptation processes (e.g., pangenome openness, intra-species nucleotide diversity, and genome streamlining). Therefore, microbes in each habitat cluster employed fundamentally different strategies to adapt to their environment, likely influenced by abiotic (e.g., nutrient availability, temperature, and oxygen) and biotic (e.g., species interactions and competition) factors.

Metabolic versatility and oxygen utilization are key traits associated with prokaryotic generalism

From the microbes' perspective, habitats define the boundaries of where species can persist. Across habitat conditions, distinct adaptive strategies emerge, from specialists restricted to narrow niches to generalists that tolerate a broad range of environmental conditions. To quantify this continuum, we used our framework capturing microbiome structure to develop a generalism score that integrates relative abundances of taxa with the ecological distances between habitat clusters they were observed in. At the species level, the generalism score ranges from 0 for species predominantly abundant in a single habitat cluster (indicative of specialization[55]) to 0.512 for *Afipia sp000497575*, a species thriving across disparate habitat clusters at comparable relative abundances (Figures 4A, 4B, S5A, and S5B; Table S4A; STAR Methods).

Compared with simply counting the number of habitat clusters in which a species was detected,[56] our scoring method yielded lower estimates of generalism (Fisher's moment coefficient = 6.004 versus 3.357). This aligns with previous reports that most microbes are specialists, while only a small fraction are generalists capable of colonizing highly disparate habitats.[4,5] For simplicity, we operationally define species with a generalism score below 0.1 as specialists, which includes 95.4% of all species, and those with a score of 0.1 or above as generalists (Figure 4C). Identifying and quantifying generalism can be biased by factors such as human-derived contamination, where oral and skin microbes are detected in environmental samples and may be misinterpreted as evidence of broad adaptability (for which we find indications; Figures S2E and S2F). However, by taking the species' relative abundance into account, our scoring scheme appeared robust against such biases. For instance, *Cutibacterium acnes*, a known skin commensal and a

Figure 3. Microbiome structure, function, and characteristics of species within habitat clusters

Planetary microbiome structure is reflected in the habitat clusters' functional composition and in resident species' genomic and phenotypic traits.

(A) Taxonomic properties of habitat clusters. Dendrogram of ecological distances, phylum-level composition, Shannon diversity (median with interquartile range), and taxonomic classification resolution, from top to bottom. Host-associated microbiomes were predominantly composed of Firmicutes_A and Bacteroidota, whereas environmental habitats were enriched in Proteobacteria and Actinobacteriota.

(B) COG-level functional profiles show contrasting patterns in host-associated and environmental microbiomes. COG composition raw values are shown as a stacked bar plot (top), and normalized *Z* scores as a heatmap (bottom).

(C) Genomic (upper panel) and phenotypic traits (center panel) of species residing within habitat clusters. The top 15 phenotypic traits with the highest variability across clusters are shown (Figure S4A for the complete trait set). These traits mirror habitat conditions: for example, frequently oxygenated environments, such as aquatic and terrestrial habitats, show a high proportion of aerotolerant microbes. Proportion of generalist species (bottom panel, generalism score $\geq$ 0.1) per habitat cluster (Figures 4A and S5A for a detailed introduction into the generalism score).

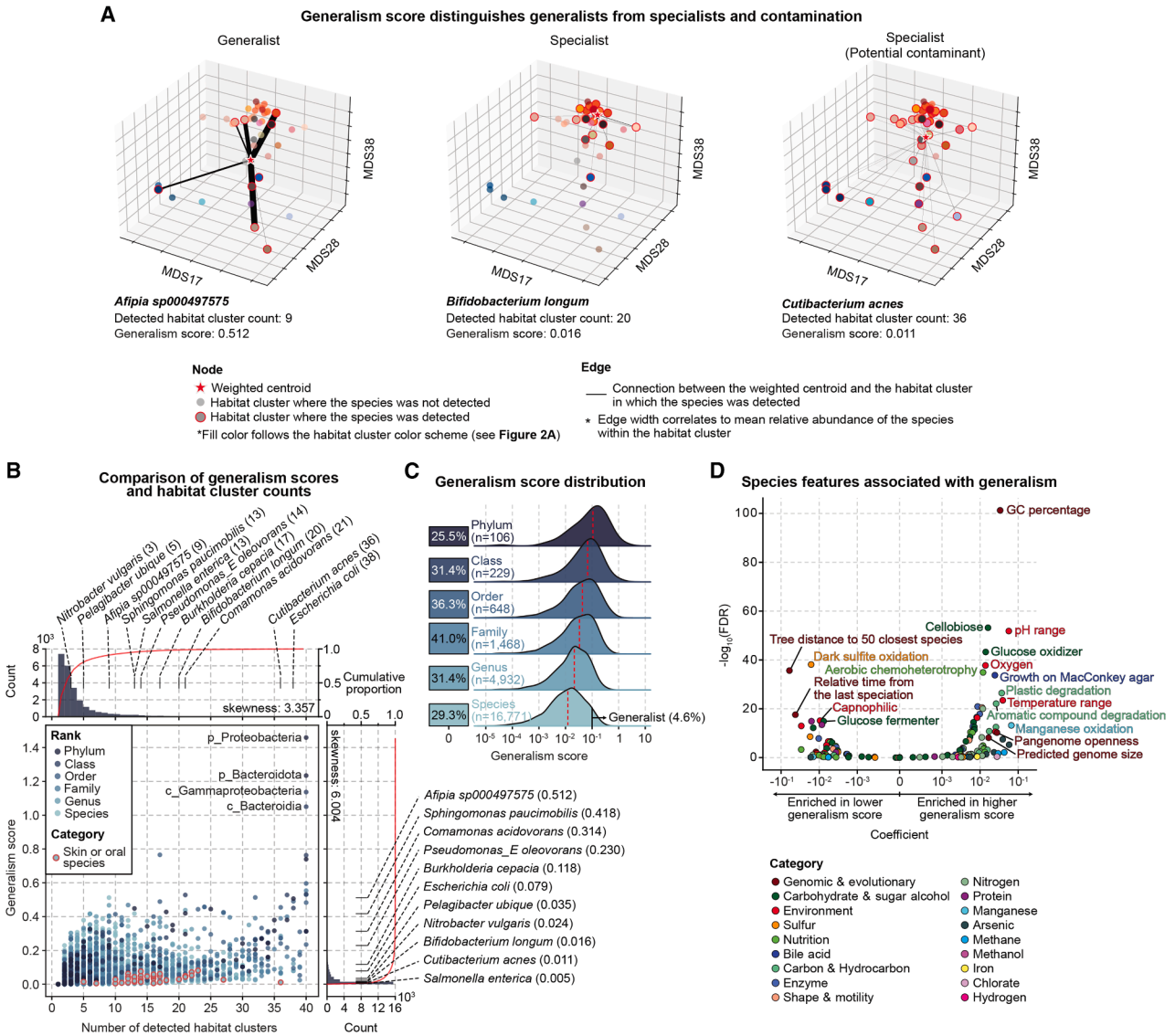


Figure 4. Prokaryotic generalism score and traits enriched in generalists

(A) Generalism scores for *Bifidobacterium longum*, *Afipia sp000497575*, and *Cutibacterium acnes*. *B. longum* was detected in 20 animal gut clusters that were ecologically similar, resulting in a low generalism score (0.016). *Afipia sp000497575* was detected in fewer habitat clusters (9) but showed consistent abundance across ecologically distinct habitats (including aquatic, terrestrial, and bovine gut clusters), yielding a higher generalism score of 0.512. A potential contaminant species, *C. acnes*, was detected in 36 habitat clusters but had disproportionately high abundance in human skin-associated (HSS) clusters, resulting in a low generalism score of 0.011.

(B) Comparison of generalism scores and habitat cluster counts with histograms and cumulative proportions (red lines) in the outer panels. Species frequently (>50%) found in skin and oral samples are highlighted in red. Raw habitat cluster counts can inflate generalism predictions when taxa are frequent sample contaminants (e.g., skin or oral species). Skewness was calculated using Fisher's moment coefficient.

(C) Distribution of generalism scores for each taxonomic rank. The bar indicates the proportion of taxa with a score of 0, while the histogram area shows nonzero scores. The red vertical line marks the median for each rank. At the species level, the black line demarcates the 0.1 cutoff used to classify generalists and specialists.

(D) Species traits associated with prokaryotic generalism. Positive coefficients indicate traits enriched in generalist species, while negative coefficients represent traits enriched in specialists. GC content is the trait most positively associated with prokaryotic generalism. FDR: Benjamini-Hochberg correction of logistic-regression *p* values.

See also Figure S5.

contamination indicator of metagenomic samples,[57] was detected in 36 out of 40 habitat clusters. Due to its abundance pattern with disproportionate preference for the HSS cluster, *Cutibacterium acnes* was not misinterpreted as a generalist (generalism score = 0.011; Figures 4A and 4B).

We identified genomic and phenotypic traits and pathways associated with generalism using linear regression analysis, which allowed us to control for technical confounders (genome quality) and phylogeny (Table S4B; STAR Methods). The generalism score showed strong positive associations with a broader tolerance range for both temperature and pH, which suggests environmental flexibility (linear regression false discovery rate [FDR]-corrected $p\ (q) < 1e{-}20$) (Figure 4D). These associations were more pronounced than those based on simple habitat cluster counts (Figures S5C and S5D). We also observed a marked positive correlation between generalism and GC content (linear regression $q < 1e{-}100$) (Figures 4D and S5E), consistent with previous findings linking higher GC content to resource opportunism, a key factor in generalism.[58] This correlation is reflected in the high fraction of generalists in terrestrial and anthropogenic habitat clusters, whose microbiomes exhibited high GC contents (Figure 3C). Notably, generalism was associated with higher speciation rates (Figure 4D; specialists showed larger phylogenetic distances to their closest species and longer intervals since their last speciation, linear regression $q < 1e{-}15$), consistent with previous reports.[56,59]

Traits associated with oxygen utilization (e.g., aerobic chemoheterotrophy, glucose oxidation, and aerobic respiration) were positively correlated with generalism (linear regression $q < 1e{-}20$). By contrast, traits linked to oxygen-deficient or high-CO_2 environments (e.g., capnophilic metabolism, glucose fermentation, and dark sulfite oxidation) were negatively correlated with the generalism score (linear regression $q < 1e{-}10$) (Figure 4D). This suggests that aerobic respiration supports generalism, aligning with the need to adapt to the oxygenated nature of most habitats. Furthermore, traits and pathways involved in degrading plastics and aromatic compounds with frequently low bioavailability[60,61] were enriched (Figures 4D and S5F), highlighting the importance of resource opportunism for generalists. Specialists, on the other hand, were enriched for pathways associated with antibiotics and polyketide biosynthesis, secondary metabolites commonly used for antibiotic production (Figures 4D and S5F). This implies that antibiotic production may increase fitness for specialists by competing with sympatric species,[62] consistent with the competition-dispersal trade-off concept in ecology.[63]

Generalists mediate HGT between ecologically disparate habitats

Due to their adaptability to diverse environmental conditions and resources, generalist microbes thrive across a broad range of habitats. Through dispersal, generalists can transport both their own genetic material and horizontally acquired genes from sympatric species into new habitats (Figure 5A). Because species co-occurrence increases the likelihood of HGT,[64] and generalists can be members of microbiomes from different habitats, generalists have the potential to be key facilitators of planetary gene flow, particularly across ecologically disparate habitats. To quantify and validate this hypothesis, we screened ~2 million genomes and MAGs for high-confidence HGT events, adapting methods from Khedkar et al.[65] (species from different families sharing adjacent genes, including a recombinase; STAR Methods). We integrated these HGT events with our habitat cluster and generalism framework, normalizing for sampling biases among habitat clusters (Table S5A; STAR Methods).

Most of the 84,049 HGT events detected this way occurred between ecologically similar-habitat clusters, such as within host-associated or ocean clusters (Figures 5B and 5C). Specifically, at an ecological distance threshold of 0.5 as an operational separation of similar and disparate habitat clusters (differentiating aquatic, terrestrial, and host-associated habitats; Figure 2A), similar-habitat HGT accounted for 97.62% (Figure 5C) of all HGT events. Frequent HGT between similar habitats likely reflects microbes' high phylogenetic relatedness and ample opportunities for them to physically interact. Within these similar-habitat transfers, most events occurred among specialists (64.4%), followed by those between specialists and generalists (35.5%), and only 0.1% among generalists. By contrast, transfers spanning disparate habitats (2.38%) mostly occurred between specialists and generalists (74.8%), while 25.2% occurred among specialists, and none among generalists exclusively (Figure 5D). Independent of the operational ecological distance threshold, the fraction of generalist-involved HGT increased significantly with greater ecological distance between the sampled environments (Pearson's $r = 0.589$, $p = 1.82e{-}35$; Figure S6A). Thus, despite being rare, HGT between disparate habitat clusters was disproportionately mediated by generalist taxa. Since generalists can tolerate a wide range of conditions, encountering diverse microbial lineages, this may increase their chances of participating in HGT across habitat boundaries. Notably, species from only three prokaryotic families, *Pseudomonadaceae*, *Burkholderiaceae*, and *Alteromonadaceae*, account for the majority (63%) of disparate-habitat HGT (Figure 5E). This suggests that targeted measures aimed at interrupting gene flow between disparate gene reservoirs (e.g., antibiotic resistance) could effectively focus on these key taxa.

These findings uncover fundamental principles of planetary-scale gene flow, highlighting generalists as key vehicles capable of overcoming ecological and geographical boundaries, thereby facilitating HGT across ecologically distinct microbial habitats.

Generalists at anthropogenic interfaces facilitate gene flow between host-associated and environmental microbiomes

Having identified generalist microbes as major vehicles of HGT across ecologically disparate habitats, we next explored the role of anthropogenic microbiomes within the planetary microbiome landscape. At the interface of host-associated and environmental habitats, anthropogenic clusters such as WT (WT1–2), built environments (BLD), and heterogeneous surfaces frequently exposed to human contact (HET) exhibit a unique pattern: taxonomically, they resemble host-associated microbiomes, yet functionally they share traits with environmental microbiomes (Figure 3, dashed rectangle ii and iii; Figure 6A). Importantly, they harbor a high relative abundance of generalist species (Figure 3C; higher than all habitat clusters except for

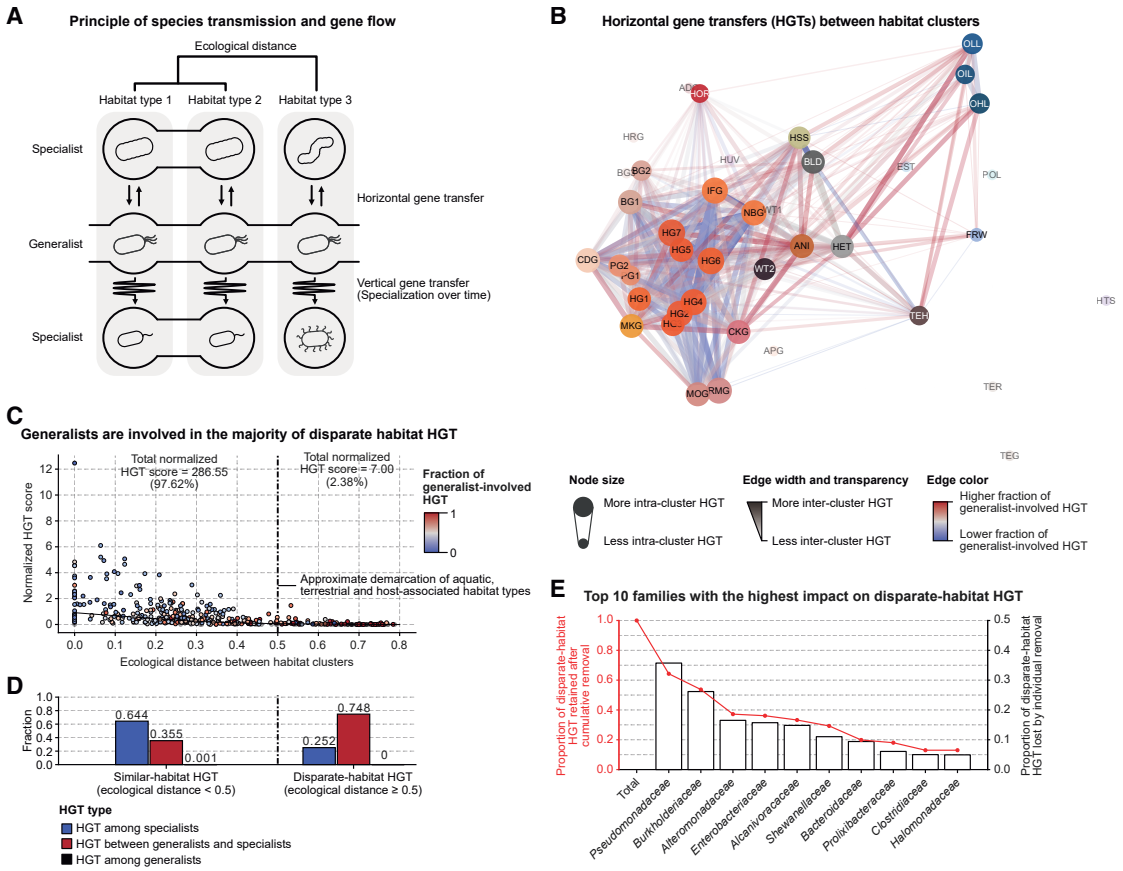


Figure 5. Generalists as a vehicle for HGT across ecologically disparate habitats

(A) Conceptual visualization of possible gene flow between habitats, by microbial dissemination or horizontal gene transfer (HGT) between microbes. Generalists can transport their own or horizontally acquired genes into new, disparate habitats.

(B–D) Generalists are involved in the majority of HGT events between disparate habitat clusters.

(B) HGT network between habitat clusters. The position of the nodes (habitat clusters) was determined based on shared taxa (Figure 2A). Edges represent HGT events between microbes detected in different habitat clusters. Edge width and transparency scale with the count of HGT events between habitat clusters (normalized by the count of MAGs per cluster). Edge colors indicate the fraction of generalists that were involved in the HGT events. Habitat clusters with <1,000 MAGs with >50% completeness were not considered for the HGT analysis (transparent nodes).

(C) Normalized HGT score across ecological distances. The point color (same as edges in B) scales with the fraction of generalists that were involved in the HGT events. The vertical dashed line marks an ecological distance of 0.5 as an operational definition of disparate habitat clusters. The smooth line represents a LOWESS local regression.

(D) Proportion of HGT events between specialists and generalists within short ecological distances (<0.5) or across disparate habitat clusters (>0.5 ecological distance).

(E) Families involved in most disparate-habitat HGT events. The top three families are jointly involved in 63% of HGT events across disparate habitat clusters. See also Figure S6.

terrestrial [TER] and terrestrial hydric [TEH]), suggesting an advantage for generalists in these environments. We therefore hypothesized that adaptive generalist taxa from human microbiomes disseminate and integrate into anthropogenic microbiomes, structurally shaping the local microbial communities and facilitating genetic connectivity between disparate habitats.

To test this, we specifically focused on WT1–2, given their high and continuous influx of host-associated microbiota and their well-established role as hubs for selection[66,67] and horizontal transfer of antibiotic resistance genes (ARGs).[68] A substantial fraction of species (60.95%, 2,283 out of 3,746) identified in the WT clusters were also present in animal (including human) gut clusters (HG1–7, NBG, IFG, MOG, RMG, CKG, APG, MKG, PG1–2, BG1–3, HRG, and cat and dog gut [CDG]), in line with previous research.[69,70] The vast majority of these shared species (81.95%, 1,871 out of 2,283) exhibited lower relative abundance in WT compared with their respective gut clusters (Figure 6B), whereas a minority of species showed a higher relative abundance in WT compared with gut clusters. We found a significant positive correlation between relative abundance fold change from gut to WT habitats and generalism scores (Pearson's r = 0.331, p = 1.42e−59; Figure 6C), supporting our hypothesis of a selective enrichment of human-derived generalists in WT, where they contribute to shaping the local microbial communities. Species exhibiting higher relative abundance in WT demonstrated a significant enrichment of traits associated

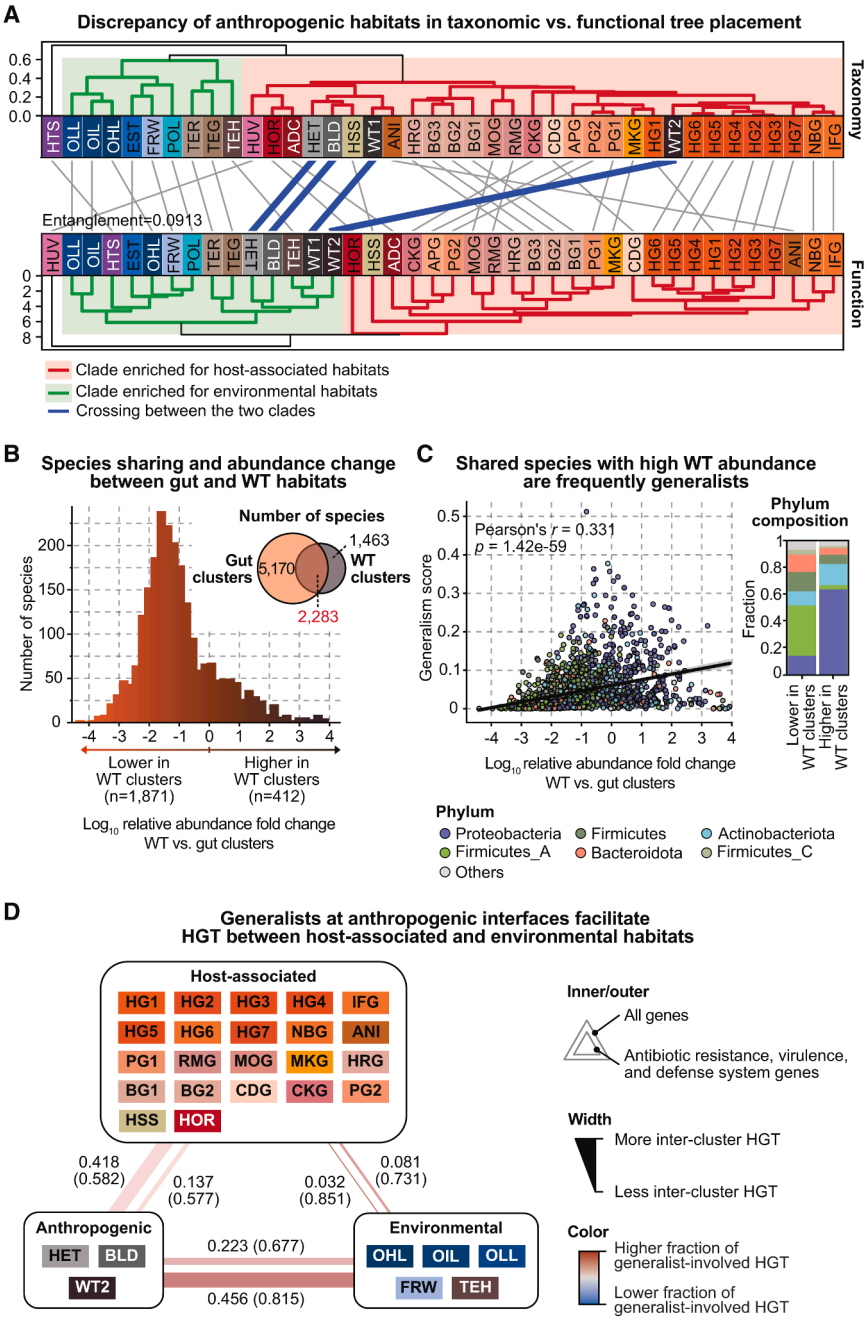


Figure 6. Anthropogenic habitats interconnect host-associated and environmental microbiomes

(A) Tanglegram comparing the placement of habitat clusters in the taxonomic versus functional microbiome landscape (hierarchical clustering). Lines between the dendrograms connect the same habitat clusters. Bold blue lines highlight habitat clusters that are taxonomically placed in the host-associated branch (red) but functionally in the environmental branch (green), which exclusively marks anthropogenic habitats.

(B) Venn diagram showing the number of species detected in gut and wastewater (WT) clusters (top-right). Distribution of the relative abundance fold change for 2,283 overlapping gut/WT species. Positive abundance change indicates higher relative abundance in WT compared with gut clusters. The majority of taxa detected in WT clusters were also found in gut microbiomes (60.95%), but frequently (81.95% of cases) at a lower relative abundance.

(C) Generalism scores are positively correlated with the relative abundance change of taxa in WT (compared with gut clusters). Data points are colored by phylum. The barplot represents the phylum composition with increased or decreased relative abundance in WT compared with gut clusters. Proteobacteria species exhibited increased relative abundance in WT clusters compared with gut clusters, whereas Firmicutes_A species showed a decrease.

(D) HGT among clusters from host-associated, anthropogenic, and environmental microbiomes. Outer links represent normalized between-group HGT scores for all transferred genes, and inner links report scores restricted to antibiotic resistance, virulence, and defense system genes. Edge width scales with the normalized score (values shown), and edge color indicates the fraction of generalist-involved HGT (proportion in parentheses).

See also Figure S6.

with adaptation to aerobic conditions (e.g., aerobe, oxidase activity, and glucose oxidation; logistic-regression $q < 1e{-}30$; Figure S6B; STAR Methods).

Because anthropogenic habitats link host-associated and environmental microbiomes and harbor a high abundance of generalists, we examined their contribution to horizontal gene dissemination across ecologically disparate habitats. HGT between host-associated and environmental habitats was markedly rarer (normalized HGT count = 0.081; STAR Methods) than transfers between anthropogenic and host-associated (0.418) or environmental (0.456) habitats (Figure 6D). Consistent with the involvement of generalists in mediating disparate HGT (Figures 5B–5D), most HGT events between host-associated and anthropogenic (58.2%) and between anthropogenic and environmental (81.5%) habitats involved generalists (Figure 6D), underscoring their importance in connecting host-associated and environmental microbiomes.

Putative multidrug resistance island disseminates through generalist-mediated HGT across habitats

To examine potential implications for human and planetary health, we next focused on the subset of HGT events carrying

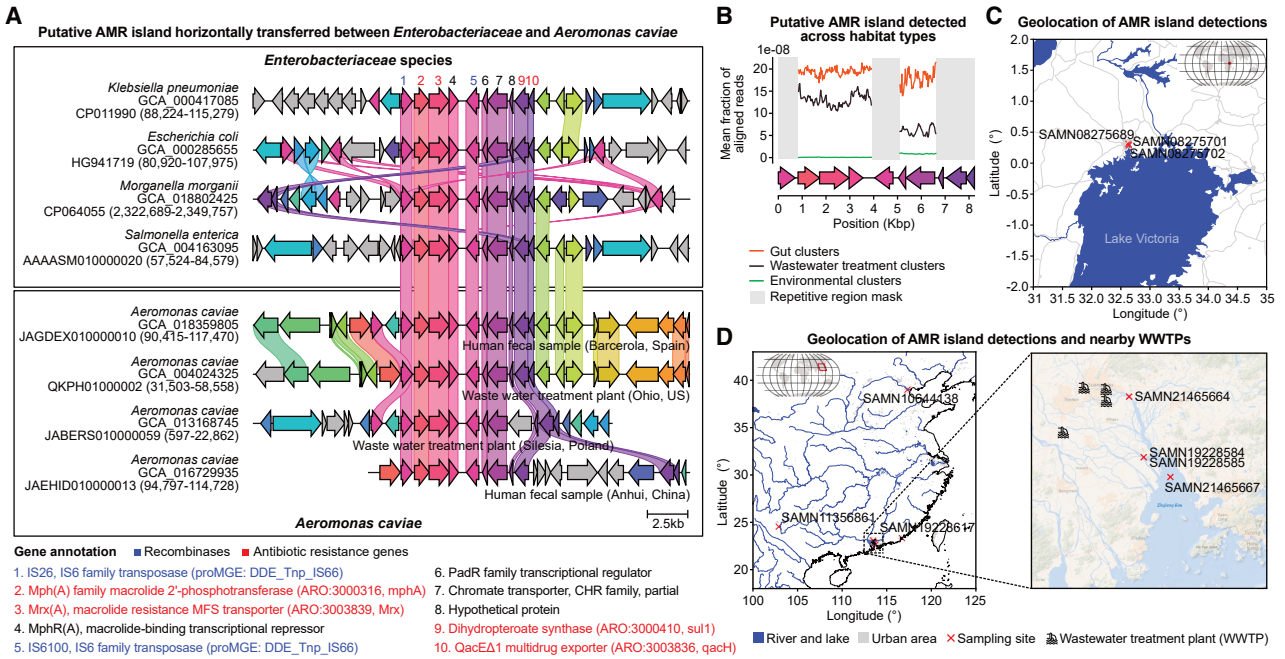


Figure 7. Tracing a generalist-mediated HGT of ARG across habitats

(A) Evidence of a putative antimicrobial resistance (AMR) island consisting of 10 genes, shared between gut specialist *Enterobacteriaceae* species (showing 4 out of 36 genomes) and the generalist *Aeromonas caviae* (showing 4 out of 13 genomes). Genes with 100% sequence identity are shown in identical colors. The annotations for these genes are provided below, with blue text indicating insertion sequences and red text indicating ARGs.

(B) Vertical coverage of the putative AMR island (A) by metagenomic reads summarized by habitats. Genomic regions found repeatedly across prokaryotic genomes (Figure S7E) are masked. The island was mostly detected in gut and WT habitat clusters and rarely in environmental samples.

(C–E) Geographical locations (red crosses) of environmental samples where the putative AMR island was detected: (C) Lake Victoria near Kampala, (D) sites in China, and (E) a zoom-in of the Guangdong-Hong Kong-Macao Greater Bay Area together with nearby WT treatment plants.

See also Figure S7.

defense, antibiotic resistance, or virulence genes (n = 36,830). These transfers mirrored the overall pattern, with generalists in anthropogenic habitats acting as major mediators of gene exchange between host-associated and environmental microbiomes (Figure 6D). Such transfers occurred globally rather than being confined to specific regions (Figure S6C). Since the emergence and spread of resistance to last-resort antibiotics is a critical global health concern, we further stratified ARG-carrying HGT using the World Health Organization (WHO) AWaRe classification,[71,72] specifically focusing on resistance to "Watch" and "Reserve" class antibiotics (STAR Methods). These high-priority ARGs were exchanged most frequently within gut microbiomes, but substantial transfer also occurred across disparate habitats, including WT-gut, skin-gut, and terrestrial-gut axes (Figure S7A). To reveal key microbial drivers behind this global gene dissemination, we identified the 15 generalists involved in most disparate-habitat HGT and quantified their ARG dissemination (Figure S7B). These include potentially pathogenic species for which *in vitro* studies have demonstrated capability to horizontally transfer ARG, including *Aeromonas caviae*,[73–75] *Vibrio cholerae*,[76] and *Vibrio parahaemolyticus*.[77]

Given the increasing public health threat posed by multidrug resistance, we specifically ranked transfer events containing multiple ARGs shared between host-associated and anthropogenic environments. This revealed an AMR island with ten genes, including four putative ARGs (*mphA* [ARO:3000316], *Mrx* [ARO:3003839], *sul1* [ARO:3000410], and *qacH* [ARO:3003836]) that confer resistance to macrolide antibiotics (ARO:0000000), sulfonamide antibiotics (ARO:3000282), and disinfecting agents and antiseptics (ARO:3005386). The AMR island was shared between the generalist *A. caviae* and multiple *Enterobacteriaceae* gut specialists (Figures 7A and S7C; STAR Methods). *A. caviae* is recognized for its robustness to WT treatment procedures[78] and is classified by the US Environmental Protection Agency as an opportunistic pathogen and potential water contaminant.[79,80] Within shared coding regions, the identified island was identical in sequence across all carrier species, indicating recent horizontal transfer, coinciding with the emergence of WT treatment plants as hubs of such gene exchange.[68,81]

The resistance island was most abundant in gut and WT samples (Figures 7B, S7D, and S7E), but notably also appeared in estuary samples downstream from WT facilities (Figure S7D). We mapped reads from *A. caviae*-positive metagenomes to the resistance island (STAR Methods) and detected it in 10 environmental samples. These environmental samples were likely impacted by anthropogenic activities: Lake Victoria near Kampala, where untreated WT was discharged into the lake[82]; biofilms on plants irrigated by Haihe River water; Fuxian Lake near Kunming; and estuarine sites downstream of WT treatment plants (Figures 7C and 7D) within the densely populated

metropolitan region of the Guangdong-Hong Kong-Macao Greater Bay Area. Although directionality and causality still need to be shown, our results suggest that generalists such as *A. caviae* can engage in ARG-containing HGT with host-associated microbes and introduce these genes into environmental microbiomes, further supported by a comparable generalist-mediated ARG transfer event involving *V. cholerae* (Figure S7F). Our habitat and generalism framework thus provides a powerful approach for tracing HGT (including that of AMR) across habitats at a planetary scale.

DISCUSSION

In this study, we leveraged a planetary-scale metagenomic dataset[24] to develop and apply a robust and fine-grained microbiome-based habitat clustering framework. Our framework harmonizes dispersed knowledge from various ecosystem-specific studies, allowing us to quantify the ecological drivers, functional traits, and adaptation strategies of microbiomes that emerge across their ecological distances. By developing a scoring scheme to characterize the continuum from prokaryotic specialists to generalists, quantifying their dispersal between habitat clusters, and integrating HGT events detected in our dataset, we identified a fundamental principle that underlies the growing importance of generalists as mediators of gene flow with increasing ecological habitat distance.

Our study establishes a model that captures the planetary-scale microbial metacommunity,[83] thereby extending ecological frameworks previously well-established for animals[84] and plants[85,86] to prokaryotes. The metacommunity view is a prerequisite for understanding how the planetary microbiome's structure and function are shaped and how they evolve: our findings reinforce the classical microbiological tenet formulated by Baas Becking that "everything is everywhere, but the environment selects."[87,88] This is evident through the clustering of metagenomes by habitat conditions rather than geography, best illustrated by the similarity of ocean microbiomes toward polar regions (Figure 2E) and by previous reports on terrestrial and lacustrine microbiomes from the Arctic and Antarctica.[41] We note that the Baas Becking tenet primarily applies to specialists, whereas generalist species deviate because their pronounced adaptability allows them to transcend ecological constraints. Furthermore, we propose specific mechanisms underlying gene flow across disparate habitats, notably mediated by HGT events between specialists and generalist species, as exemplified by the spread of ARGs between gut, WT, and other aquatic habitats.

Future research could build upon our habitat clustering framework to monitor and quantify microbial gene flow on our planet. Beyond antibiotic resistance spread, the framework can be utilized to elucidate HGT of other adaptive traits, including metabolic pathways or biosynthetic gene clusters, in response to changes in environmental and anthropogenic pressures. As datasets continue to grow, this foundational work paves the way to predict microbiome structure from habitat descriptors and to better understand and manage human impacts on the global dissemination of microbial genes with relevance for planetary health.

Limitations of the study

While our study synthesizes various concepts from molecular ecology and evolution and is supported by established knowledge, our approach has limitations.

First, sampling imbalances in public research are biasing our dataset. For example, human gut microbiomes have the largest number of public samples, have been most intensively studied, and most taxa could be classified at the species level. For these reasons, we identified the largest number of clusters (nine) across human gut microbiomes. However, this does not imply that the human gut microbiome is inherently more variable or more complex than other microbiomes. To minimize data imbalances in downstream analyses, we consistently applied strategies to adjust for uneven sampling efforts (STAR Methods).

Second, our analysis focused exclusively on prokaryotes, omitting eukaryotic microorganisms and viruses, despite their well-documented importance in shaping microbial communities.[89,90] Including these, although technically challenging, could redefine cluster boundaries and uncover finer-scale ecological patterns.

Third, the habitat cluster model provides a simplified framework to capture major structures within the planetary microbiome landscape but may not reflect the full metadata resolution and can occasionally mislabel samples due to environmental complexity. Such discrepancies may result from mixed or transitional habitats, contamination (e.g., human skin particles in air; aquatic samples in Figures S3E and S3F), environmental gradients (e.g., land-sea transitions), or technical and labeling errors. The model is therefore intended to reveal structural patterns rather than classify every sample with fine-grained accuracy.

Lastly, as a cross-sectional study, our research does not capture evolutionary processes such as speciation or long-term adaptation, which would require paleogenomic data spanning evolutionary timescales[91] and is beyond the scope of this work.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Michael Kuhn (mkuhn@embl.de).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

The underlying sequence data are available through the proGenomes (http://progenomes.embl.de/) and SPIRE websites (https://spire.embl.de/). Trait information and generalism scores for all taxa are accessible and can be interactively explored on the metaTraits (https://metatraits.embl.de/) and SPIRE websites, respectively. The Python scripts for the calculation of the generalism score and the HGT analysis are available at https://github.com/grp-bork/PlanetaryMicrobiomeHabitat_Kim_2025.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (no. RS-2023-00240807); by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 460129525 (NFDI4Microbiota); by the Ministry of Science, Research, and the Arts, Baden-Württemberg (M.K.) within the framework of LIBIS/de.NBI; and by the Federal Ministry of Research, Technology,

and Space in the frame of de.NBI & ELIXIR-DE (W-de.NBI-014). This work used EMBL IT Services HPC resources (https://doi.org/10.5281/zenodo.12785829). The authors thank the Bork group, the von Mering group at the University of Zürich (João Frederico Matias Rodrigues, Janko Tackmann, and Christian von Mering), and T.S.B. Schmidt for discussions and critical reading of the manuscript. The authors are deeply grateful to P.B. for his invaluable guidance, vision, and intellectual contributions to this work. He passed away after the acceptance of this manuscript.

AUTHOR CONTRIBUTIONS

Overall bioinformatics analysis: C.Y.K., D.P., and J.S.; conceptualization: C.Y.K., D.P., and A.O.; metadata curation: M.K.; data processing: A.F., M.K., C.S., and S.M.R.; HGT analysis: S.K. and A.G.; supervision: P.B.; writing manuscript: C.Y.K., D.P., J.S., and P.B.

DECLARATION OF INTERESTS

P.B. was a member of the advisory board of *Cell*.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT (OpenAI) to improve language and clarity. After the usage of this tool, the authors reviewed and edited the contents as needed and take full responsibility for the content of the article.

STAR★METHODS

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

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Taxonomic, functional, and *k*-mer-based distance between metagenomic samples
  - Dimensionality reduction and clustering of metagenomic samples
  - Benchmarking of clustering results and selection of habitat clusters
  - Annotation of HTS, EST, POL, and FRW clusters: aquatic samples collected from transitional environments and extreme temperatures form distinct clusters
  - Annotation of TEH, TER, and TEG clusters: the hydric status affects terrestrial microbiome composition
  - Annotation of HOR and ADC clusters: differentiation of ancient and modern oral microbiomes
  - Annotation of HSS and BLD clusters: human skin microbiomes are a major source of anthropogenic surface microbiomes
  - Annotation of HET and ANI clusters: clusters consist of heterogeneous samples
  - Ecological distance between habitat clusters
  - Network visualization
  - Variance partitioning of the ocean clusters
  - Analysis of host and human DNA contents
  - Cluster-level summary of species' genomic and phenotypic features
  - Weighted centroid-based generalism score
  - Comparison between dendrograms
  - Identifying horizontal gene transfer in the context of generalism and ecological distance
  - Habitat group stratification and analysis of ARGs, virulence, and defense genes
  - Identification of putative resistance islands and measuring the abundance from metagenomic samples
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Species features associated with the generalism score and abundance change in the WT cluster
  - Additional statistical analyses used in this study

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.12.051.

Received: July 18, 2025
Revised: November 19, 2025
Accepted: December 25, 2025
Published: February 9, 2026

REFERENCES

1. Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008). The microbial engines that drive Earth's biogeochemical cycles. Science *320*, 1034–1039. https://doi.org/10.1126/science.1153213.
2. Schmidt, T.S.B., Raes, J., and Bork, P. (2018). The Human Gut Microbiome: From Association to Modulation. Cell *172*, 1198–1215. https://doi.org/10.1016/j.cell.2018.02.044.
3. Coelho, L.P., Alves, R., Del Río, Á.R., Myers, P.N., Cantalapiedra, C.P., Giner-Lamia, J., Schmidt, T.S., Mende, D.R., Orakov, A., Letunic, I., et al. (2022). Towards the biogeography of prokaryotic genes. Nature *601*, 252–256. https://doi.org/10.1038/s41586-021-04233-4.
4. von Mering, C., Hugenholtz, P., Raes, J., Tringe, S.G., Doerks, T., Jensen, L.J., Ward, N., and Bork, P. (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. Science *315*, 1126–1130. https://doi.org/10.1126/science.1133420.
5. von Meijenfeldt, F.A.B., Hogeweg, P., and Dutilh, B.E. (2023). A social niche breadth score reveals niche range strategies of generalists and specialists. Nat. Ecol. Evol. *7*, 768–781. https://doi.org/10.1038/s41559-023-02027-7.
6. Hooper, S.D., Raes, J., Foerstner, K.U., Harrington, E.D., Dalevi, D., and Bork, P. (2008). A molecular study of microbe transfer between distant environments. PLoS One *3*, e2607. https://doi.org/10.1371/journal.pone.0002607.
7. Custer, G.F., Bresciani, L., and Dini-Andreote, F. (2022). Ecological and Evolutionary Implications of Microbial Dispersal. Front. Microbiol. *13*, 855859. https://doi.org/10.3389/fmicb.2022.855859.
8. Barbour, K.M., Barrón-Sandoval, A., Walters, K.E., and Martiny, J.B.H. (2023). Towards quantifying microbial dispersal in the environment. Environ. Microbiol. *25*, 137–142. https://doi.org/10.1111/1462-2920.16270.
9. Choudoir, M.J., and DeAngelis, K.M. (2022). A framework for integrating microbial dispersal modes into soil ecosystem ecology. iScience *25*, 103887. https://doi.org/10.1016/j.isci.2022.103887.
10. Czatzkowska, M., Wolak, I., Harnisz, M., and Korzeniewska, E. (2022). Impact of Anthropogenic Activities on the Dissemination of ARGs in the Environment-A Review. Int. J. Environ. Res. Public Health *19*, 12853. https://doi.org/10.3390/ijerph191912853.
11. Karkman, A., Pärnänen, K., and Larsson, D.G.J. (2019). Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. Nat. Commun. *10*, 80. https://doi.org/10.1038/s41467-018-07992-3.
12. von Wintersdorff, C.J.H., Penders, J., van Niekerk, J.M., Mills, N.D., Majumder, S., van Alphen, L.B., Savelkoul, P.H.M., and Wolffs, P.F.G. (2016). Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Front. Microbiol. *7*, 173. https://doi.org/10.3389/fmicb.2016.00173.
13. Li, H., Zhou, H., Yang, S., and Dai, X. (2023). Stochastic and Deterministic Assembly Processes in Seamount Microbial Communities. Appl. Environ. Microbiol. *89*, e0070123. https://doi.org/10.1128/aem.00701-23.
14. Wang, J., Shen, J., Wu, Y., Tu, C., Soininen, J., Stegen, J.C., He, J., Liu, X., Zhang, L., and Zhang, E. (2013). Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. ISME J. *7*, 1310–1321. https://doi.org/10.1038/ismej.2013.30.

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

16. Borton, M.A., McGivern, B.B., Willi, K.R., Woodcroft, B.J., Mosier, A.C., Singleton, D.M., Bambakidis, T., Pelly, A., Daly, R.A., Liu, F., et al. (2025). A functional microbiome catalogue crowdsourced from North American rivers. Nature *637*, 103–112. https://doi.org/10.1038/s41586-024-08240-z.

17. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. Science *348*, 1261359. https://doi.org/10.1126/science.1261359.

18. Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M., Bengtsson-Palme, J., Anslan, S., Coelho, L.P., Harend, H., et al. (2018). Structure and function of the global topsoil microbiome. Nature *560*, 233–237. https://doi.org/10.1038/s41586-018-0386-6.

19. Abdill, R.J., Graham, S.P., Rubinetti, V., Ahmadian, M., Hicks, P., Chetty, A., McDonald, D., Ferretti, P., Gibbons, E., Rossi, M., et al. (2025). Integration of 168,000 samples reveals global patterns of the human gut microbiome. Cell *188*, 1100–1118.e17. https://doi.org/10.1016/j.cell.2024.12.017.

20. Singleton, C.M., Jensen, T.B.N., Delogu, F., Knudsen, K.S., Sørensen, E.A., Jørgensen, V.R., Karst, S.M., Yang, Y., Sereika, M., Petriglieri, F., et al. (2025). The Microflora Danica atlas of Danish environmental microbiomes. Nature. https://doi.org/10.1038/s41586-025-09794-2.

21. Shaffer, J.P., Nothias, L.-F., Thompson, L.R., Sanders, J.G., Salido, R.A., Couvillion, S.P., Brejnrod, A.D., Lejzerowicz, F., Haiminen, N., Huang, S., et al. (2022). Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. Nat. Microbiol. *7*, 2128–2150. https://doi.org/10.1038/s41564-022-01266-x.

22. Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature *551*, 457–463. https://doi.org/10.1038/nature24621.

23. Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., and Lewis, S.E.; ENVO Consortium (2013). The environment ontology: contextualising biological and biomedical entities. J. Biomed. Semant. *4*, 43. https://doi.org/10.1186/2041-1480-4-43.

24. Schmidt, T.S.B., Fullam, A., Ferretti, P., Orakov, A., Maistrenko, O.M., Ruscheweyh, H.-J., Letunic, I., Duan, Y., Van Rossum, T., Sunagawa, S., et al. (2024). SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. Nucleic Acids Res. *52*, D777–D783. https://doi.org/10.1093/nar/gkad943.

25. Rodrigues, J.F.M., Tackmann, J., Malfertheiner, L., Patsch, D., Perez-Molphe-Montoya, E., Näpflin, N., Gaio, D., Rot, G., Danaila, M., Peluso, M.E., et al. (2025). The MicrobeAtlas database: Global trends and insights into Earth's microbial ecosystems. Preprint at bioRxiv. https://doi.org/10.1101/2025.07.18.665519.

26. Antwis, R.E., Griffiths, S.M., Harrison, X.A., Aranega-Bou, P., Arce, A., Bettridge, A.S., Brailsford, F.L., de Menezes, A., Devaynes, A., Forbes, K.M., et al. (2017). Fifty important research questions in microbial ecology. FEMS Microbiol. Ecol. *93*. https://doi.org/10.1093/femsec/fix044.

27. Jing, G., Zhang, Y., Liu, L., Wang, Z., Sun, Z., Knight, R., Su, X., and Xu, J. (2021). A Scale-Free, Fully Connected Global Transition Network Underlies Known Microbiome Diversity. mSystems *6*, e0039421. https://doi.org/10.1128/msystems.00394-21.

28. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. *3*, 861. https://doi.org/10.21105/joss.00861.

29. Xu, Z., Malmer, D., Langille, M.G.I., Way, S.F., and Knight, R. (2014). Which is more important for classifying microbial communities: who's there or what they can do? ISME J. *8*, 2357–2359. https://doi.org/10.1038/ismej.2014.157.

30. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

31. Kuhn, M., Schmidt, T.S.B., Ferretti, P., Głazek, A., Robbani, S.M., Akanni, W., Fullam, A., Schudoma, C., Cetin, E., Hassan, M., et al. (2025). Metalog: curated and harmonised contextual data for global metagenomics samples. Nucleic Acids Res. gkaf1118. https://doi.org/10.1093/nar/gkaf1118.

32. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227. https://doi.org/10.1038/nature11053.

33. Youngblut, N.D., Reischer, G.H., Walters, W., Schuster, N., Walzer, C., Stalder, G., Ley, R.E., and Farnleitner, A.H. (2019). Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. Nat. Commun. *10*, 2200. https://doi.org/10.1038/s41467-019-10191-3.

34. Groussin, M., Mazel, F., Sanders, J.G., Smillie, C.S., Lavergne, S., Thuiller, W., and Alm, E.J. (2017). Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat. Commun. *8*, 14319. https://doi.org/10.1038/ncomms14319.

35. Stewart, C.J., Ajami, N.J., O'Brien, J.L., Hutchinson, D.S., Smith, D.P., Wong, M.C., Ross, M.C., Lloyd, R.E., Doddapaneni, H., Metcalf, G.A., et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature *562*, 583–588. https://doi.org/10.1038/s41586-018-0617-x.

36. Arrieta, M.-C., Stiemsma, L.T., Amenyogbe, N., Brown, E.M., and Finlay, B. (2014). The intestinal microbiome in early life: health and disease. Front. Immunol. *5*, 427. https://doi.org/10.3389/fimmu.2014.00427.

37. Wernroth, M.-L., Peura, S., Hedman, A.M., Hetty, S., Vicenzi, S., Kennedy, B., Fall, K., Svennblad, B., Andolf, E., Pershagen, G., et al. (2022). Development of gut microbiota during the first 2 years of life. Sci. Rep. *12*, 9080. https://doi.org/10.1038/s41598-022-13009-3.

38. World Bank Group. (2024). The World by Income and Region. https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html.

39. Schnorr, S.L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. Nat. Commun. *5*, 3654. https://doi.org/10.1038/ncomms4654.

40. Conteville, L.C., Oliveira-Ferreira, J., and Vicente, A.C.P. (2019). Gut Microbiome Biomarkers and Functional Diversity Within an Amazonian Semi-Nomadic Hunter-Gatherer Group. Front. Microbiol. *10*, 1743. https://doi.org/10.3389/fmicb.2019.01743.

41. Kleinteich, J., Hildebrand, F., Bahram, M., Voigt, A.Y., Wood, S.A., Jungblut, A.D., Küpper, F.C., Quesada, A., Camacho, A., Pearce, D.A., et al. (2017). Pole-to-Pole Connections: Similarities between Arctic and Antarctic Microbiomes and Their Vulnerability to Environmental Change. Front. Ecol. Evol. *5*, 300617. https://doi.org/10.3389/fevo.2017.00137.

42. Reagan, J.R., Boyer, T.P., García, H.E., Locarnini, R.A., Baranova, O.K., Bouchard, C., Cross, S.L., Mishonov, A.V., Paver, C.R., Seidov, D., et al. (2023). World Ocean Atlas. (National Centers for Environmental Information (NCEI)). https://doi.org/10.25921/va26-hv25.

43. Podlesny, D., Kim, C.Y., Robbani, S.M., Schudoma, C., Fullam, A., Reimer, L.C., Koblitz, J., Schober, I., Iyappan, A., Van Rossum, T., et al. (2025). metaTraits: a large-scale integration of microbial phenotypic trait information. Nucleic Acids Res. gkaf1241. https://doi.org/10.1093/nar/gkaf1241.

44. de Jonge, N., Carlsen, B., Christensen, M.H., Pertoldi, C., and Nielsen, J.L. (2022). The Gut Microbiome of 54 Mammalian Species. Front. Microbiol. *13*, 886252. https://doi.org/10.3389/fmicb.2022.886252.

45. Kim, N., Kim, C.Y., Ma, J., Yang, S., Park, D.J., Ha, S.-J., Belenky, P., and Lee, I. (2024). MRGM: an enhanced catalog of mouse gut microbial genomes substantially broadening taxonomic and functional landscapes. Gut Microbes *16*, 2393791. https://doi.org/10.1080/19490976.2024.2393791.

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

47. Gophna, U., and Altman-Price, N. (2022). Horizontal Gene Transfer in Archaea-From Mechanisms to Genome Evolution. Annu. Rev. Microbiol. *76*, 481–502. https://doi.org/10.1146/annurev-micro-040820-124627.

48. Raes, J., Korbel, J.O., Lercher, M.J., von Mering, C., and Bork, P. (2007). Prediction of effective genome size in metagenomic samples. Genome Biol. *8*, R10. https://doi.org/10.1186/gb-2007-8-1-r10.

49. Korth, N., Yang, Q., Van Haute, M.J., Tross, M.C., Peng, B., Shrestha, N., Zwiener-Malcom, M., Mural, R.V., Schnable, J.C., and Benson, A.K. (2024). Genomic co-localization of variation affecting agronomic and human gut microbiome traits in a meta-analysis of diverse sorghum. G3 (Bethesda) *14*, jkae145. https://doi.org/10.1093/g3journal/jkae145.

50. Rodríguez-Gijón, A., Nuy, J.K., Mehrshad, M., Buck, M., Schulz, F., Woyke, T., and Garcia, S.L. (2021). A Genomic Perspective Across Earth's Microbiomes Reveals That Genome Size in Archaea and Bacteria Is Linked to Ecosystem Type and Trophic Strategy. Front. Microbiol. *12*, 761869. https://doi.org/10.3389/fmicb.2021.761869.

51. Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S., Rodrigues, J.F.M., von Mering, C., Pedro Coelho, L., Huerta-Cepas, J., et al. (2020). Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J. *14*, 1247–1259. https://doi.org/10.1038/s41396-020-0600-z.

52. Ngugi, D.K., Acinas, S.G., Sánchez, P., Gasol, J.M., Agusti, S., Karl, D.M., and Duarte, C.M. (2023). Abiotic selection of microbial genome size in the global ocean. Nat. Commun. *14*, 1384. https://doi.org/10.1038/s41467-023-36988-x.

53. Dlugosch, L., Poehlein, A., Wemheuer, B., Pfeiffer, B., Badewien, T.H., Daniel, R., and Simon, M. (2022). Significance of gene variants for the functional biogeography of the near-surface Atlantic Ocean microbiome. Nat. Commun. *13*, 456. https://doi.org/10.1038/s41467-022-28128-8.

54. Wu, H., Cui, L., Wang, L., Sun, R., and Zheng, Z. (2023). A method for estimating particulate organic carbon at the sea surface based on geodetector and machine learning. Front. Mar. Sci. *10*, 1295874. https://doi.org/10.3389/fmars.2023.1295874.

55. Malfertheiner, L., Tackmann, J., Rodrigues, J.F.M., and von Mering, C. (2024). Community conservatism is widespread throughout microbial phyla and environments. Preprint at bioRxiv. https://doi.org/10.1101/2024.11.06.622256.

56. Sriswasdi, S., Yang, C.-C., and Iwasaki, W. (2017). Generalist species drive microbial dispersion and evolution. Nat. Commun. *8*, 1162. https://doi.org/10.1038/s41467-017-01265-1.

57. Lou, Y.C., Hoff, J., Olm, M.R., West-Roberts, J., Diamond, S., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2023). Using strain-resolved analysis to identify contamination in metagenomics data. Microbiome *11*, 36. https://doi.org/10.1186/s40168-023-01477-2.

58. Aliperti, L., Aptekmann, A.A., Farfañuk, G., Couso, L.L., Soler-Bistué, A., and Sánchez, I.E. (2023). r/K selection of GC content in prokaryotes. Environ. Microbiol. *25*, 3255–3268. https://doi.org/10.1111/1462-2920.16511.

59. Xu, Q., Vandenkoornhuyse, P., Li, L., Guo, J., Zhu, C., Guo, S., Ling, N., and Shen, Q. (2022). Microbial generalists and specialists differently contribute to the community diversity in farmland soils. J. Adv. Res. *40*, 17–27. https://doi.org/10.1016/j.jare.2021.12.003.

60. Kalbitz, K., Schwesig, D., Schmerwitz, J., Kaiser, K., Haumaier, L., Glaser, B., Ellerbrock, R., and Leinweber, P. (2003). Changes in properties of soil-derived dissolved organic matter induced by biodegradation. Soil Biol. Biochem. *35*, 1129–1142. https://doi.org/10.1016/S0038-0717(03)00165-2.

61. Obrador-Viel, T., Zadjelovic, V., Nogales, B., Bosch, R., and Christie-Oleza, J.A. (2024). Assessing microbial plastic degradation requires robust methods. Microb. Biotechnol. *17*, e14457. https://doi.org/10.1111/1751-7915.14457.

62. Kinkel, L.L., Schlatter, D.C., Xiao, K., and Baines, A.D. (2013). Sympatric inhibition and niche differentiation suggest alternative coevolutionary trajectories among Streptomycetes. ISME J. *8*, 249–256. https://doi.org/10.1038/ismej.2013.175.

63. Yawata, Y., Cordero, O.X., Menolascina, F., Hehemann, J.-H., Polz, M.F., and Stocker, R. (2014). Competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. Proc. Natl. Acad. Sci. USA *111*, 5622–5627. https://doi.org/10.1073/pnas.1318943111.

64. Dmitrijeva, M., Tackmann, J., Matias Rodrigues, J.F., Huerta-Cepas, J., Coelho, L.P., and von Mering, C. (2024). A global survey of prokaryotic genomes reveals the eco-evolutionary pressures driving horizontal gene transfer. Nat. Ecol. Evol. *8*, 986–998. https://doi.org/10.1038/s41559-024-02357-0.

65. Khedkar, S., Smyshlyaev, G., Letunic, I., Maistrenko, O.M., Coelho, L.P., Orakov, A., Forslund, S.K., Hildebrand, F., Luetge, M., Schmidt, T.S.B., et al. (2022). Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. Nucleic Acids Res. *50*, 3155–3168. https://doi.org/10.1093/nar/gkac163.

66. Berglund, B., Fick, J., and Lindgren, P.-E. (2015). Urban wastewater effluent increases antibiotic resistance gene concentrations in a receiving northern European river. Environ. Toxicol. Chem. *34*, 192–196. https://doi.org/10.1002/etc.2784.

67. Mukherjee, M., Laird, E., Gentry, T.J., Brooks, J.P., and Karthikeyan, R. (2021). Increased Antimicrobial and Multidrug Resistance Downstream of Wastewater Treatment Plants in an Urban Watershed. Front. Microbiol. *12*, 657353. https://doi.org/10.3389/fmicb.2021.657353.

68. Brown, C.L., Maile-Moskowitz, A., Lopatkin, A.J., Xia, K., Logan, L.K., Davis, B.C., Zhang, L., Vikesland, P.J., and Pruden, A. (2024). Selection and horizontal gene transfer underlie microdiversity-level heterogeneity in resistance gene fate during wastewater treatment. Nat. Commun. *15*, 5412. https://doi.org/10.1038/s41467-024-49742-8.

69. Cai, L., Ju, F., and Zhang, T. (2013). Tracking human sewage microbiome in a municipal wastewater treatment plant. Appl. Microbiol. Biotechnol. *98*, 3317–3326. https://doi.org/10.1007/s00253-013-5402-z.

70. García-Aljaro, C., Blanch, A.R., Campos, C., Jofre, J., and Lucena, F. (2019). Pathogens, faecal indicators and human-specific microbial source-tracking markers in sewage. J. Appl. Microbiol. *126*, 701–717. https://doi.org/10.1111/jam.14112.

71. World Health Organization (2025). The Selection and Use of Essential Medicines, 2025 (World Health Organization). https://doi.org/10.2471/B09489.

72. Moja, L., Zanichelli, V., Mertz, D., Gandra, S., Cappello, B., Cooke, G.S., Chuki, P., Harbarth, S., Pulcini, C., Mendelson, M., et al. (2024). WHO's essential medicines and AWaRe: recommendations on first- and second-choice antibiotics for empiric treatment of clinical infections. Clin. Microbiol. Infect. *30*, S1–S51. https://doi.org/10.1016/j.cmi.2024.02.003.

73. Walsh, T.R., Weeks, J., Livermore, D.M., and Toleman, M.A. (2011). Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study. Lancet Infect. Dis. *11*, 355–362. https://doi.org/10.1016/S1473-3099(11)70059-7.

74. Drk, S., Puljko, A., Dželalija, M., and Udiković-Kolić, N. (2023). Characterization of Third Generation Cephalosporin- and Carbapenem-Resistant Aeromonas Isolates from Municipal and Hospital Wastewater. Antibiotics (Basel) *12*, 513. https://doi.org/10.3390/antibiotics12030513.

75. Rhodes, G., Huys, G., Swings, J., McGann, P., Hiney, M., Smith, P., and Pickup, R.W. (2000). Distribution of oxytetracycline resistance plasmids between aeromonads in hospital and aquaculture environments: implication of Tn1721 in dissemination of the tetracycline resistance determinant Tet A. Appl. Appl. Environ. Microbiol. *66*, 3883–3890. https://doi.org/10.1128/AEM.66.9.3883-3890.2000.

76. Verma, J., Bag, S., Saha, B., Kumar, P., Ghosh, T.S., Dayal, M., Senapati, T., Mehra, S., Dey, P., Desigamani, A., et al. (2019). Genomic plasticity associated with antimicrobial resistance in Vibrio cholerae. Proc. Natl. Acad. Sci. USA *116*, 6226–6231. https://doi.org/10.1073/pnas.1900141116.

77. Song, Y., Yu, P., Li, B., Pan, Y., Zhang, X., Cong, J., Zhao, Y., Wang, H., and Chen, L. (2013). The mosaic accessory gene structures of the SXT/R391-like integrative and conjugative elements derived from Vibrio spp. isolated from aquatic products and environment in the Yangtze River Estuary, China. BMC Microbiol. *13*, 214. https://doi.org/10.1186/1471-2180-13-214.

78. Varela, A.R., Nunes, O.C., and Manaia, C.M. (2016). Quinolone resistant Aeromonas spp. as carriers and potential tracers of acquired antibiotic resistance in hospital and municipal wastewater. Sci. Total Environ. *542*, 665–671. https://doi.org/10.1016/j.scitotenv.2015.10.124.

79. Carusi, J., Kabuki, D.Y., de Seixas Pereira, P.M., and Cabral, L. (2024). Aeromonas spp. in drinking water and food: Occurrence, virulence potential and antimicrobial resistance. Food Res. Int. *175*, 113710. https://doi.org/10.1016/j.foodres.2023.113710.

80. United States Environmental Protection Agency (U.S. EPA). (2006). Aeromonas: Human Health Criteria Document. https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1004WS5.TXT.

81. Karkman, A., Do, T.T., Walsh, F., and Virta, M.P.J. (2018). Antibiotic-Resistance Genes in Waste Water. Trends Microbiol. *26*, 220–228. https://doi.org/10.1016/j.tim.2017.09.005.

82. Schneeberger, P.H.H., Fuhrimann, S., Becker, S.L., Pothier, J.F., Duffy, B., Beuret, C., Frey, J.E., and Utzinger, J. (2019). Qualitative microbiome profiling along a wastewater system in Kampala, Uganda. Sci. Rep. *9*, 17334. https://doi.org/10.1038/s41598-019-53569-5.

83. Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., et al. (2004). The metacommunity concept: a framework for multi-scale community ecology. Ecol. Lett. *7*, 601–613. https://doi.org/10.1111/j.1461-0248.2004.00608.x.

84. Parris, K.M. (2006). Urban amphibian assemblages as metacommunities. J. Anim. Ecol. *75*, 757–764. https://doi.org/10.1111/j.1365-2656.2006.01096.x.

85. Alexander, H.M., Foster, B.L., Ballantyne, F., IV, Collins, C.D., Antonovics, J., and Holt, R.D. (2012). Metapopulations and metacommunities: combining spatial and temporal perspectives in plant ecology. J. Ecol. *100*, 88–103. https://doi.org/10.1111/j.1365-2745.2011.01917.x.

86. Spasojevic, M.J., Copeland, S., and Suding, K.N. (2014). Using functional diversity patterns to explore metacommunity dynamics: a framework for understanding local and regional influences on community structure. Ecography *37*, 939–949. https://doi.org/10.1111/ecog.00711.

87. Fondi, M., Karkman, A., Tamminen, M.V., Bosi, E., Virta, M., Fani, R., Alm, E., and McInerney, J.O. (2016). "Every Gene Is Everywhere but the Environment Selects": Global Geolocalization of Gene Sharing in Environmental Samples through Network Analysis. Genome Biol. Evol. *8*, 1388–1400. https://doi.org/10.1093/gbe/evw077.

88. Baas Becking, L.G.M. (1934). Geobiologie of Inleiding tot de Milieukunde. (W.P. Van Stockum & Zoon).

89. de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., et al. (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science *348*, 1261605. https://doi.org/10.1126/science.1261605.

90. Breitbart, M., and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? Trends Microbiol. *13*, 278–284. https://doi.org/10.1016/j.tim.2005.04.003.

91. Shapiro, B., and Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. Science *343*, 1236573. https://doi.org/10.1126/science.1236573.

92. Ruscheweyh, H.-J., Milanese, A., Paoli, L., Karcher, N., Clayssen, Q., Keller, M.I., Wirbel, J., Bork, P., Mende, D.R., Zeller, G., et al. (2022). Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. Microbiome *10*, 212. https://doi.org/10.1186/s40168-022-01410-z.

93. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

94. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

95. van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

96. Narayan, A., Berger, B., and Cho, H. (2021). Assessing single-cell transcriptomic variability through density-preserving data visualization. Nat. Biotechnol. *39*, 765–774. https://doi.org/10.1038/s41587-020-00801-7.

97. McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. J. Open Source Softw. *2*, 205. https://doi.org/10.21105/joss.00205.

98. Müllner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. J. Stat. Softw. *53*, 1–18. https://doi.org/10.18637/jss.v053.i09.

99. Parks, D.H. (2022). PhyloRank. GitHub. https://github.com/donovan-h-parks/PhyloRank.

100. Katz, K.S., Shutov, O., Lapoint, R., Kimelman, M., Brister, J.R., and O'Sullivan, C. (2021). STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. Genome Biol. *22*, 270. https://doi.org/10.1186/s13059-021-02490-0.

101. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

102. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

103. Heer, J., Card, S.K., and Landay, J.A. (2005). Prefuse: A Toolkit for Interactive Information Visualization. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 421–430. https://doi.org/10.1145/1054972.1055031.

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

105. Bizzotto, E., Fraulini, S., Zampieri, G., Orellana, E., Treu, L., and Campanaro, S. (2024). MICROPHERRET: MICRObial PHEnotypic tRait ClassifieR using Machine lEarning Techniques. Environ. Microbiome *19*, 58. https://doi.org/10.1186/s40793-024-00600-6.

106. Barnum, T.P., Crits-Christoph, A., Molla, M., Carini, P., Lee, H.H., and Ostrov, N. (2024). Predicting microbial growth conditions from amino acid composition. Preprint at bioRxiv. https://doi.org/10.1101/2024.03.22.586313.

107. Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat. Methods *20*, 1203–1212. https://doi.org/10.1038/s41592-023-01940-w.

108. Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S.B., and Bork, P. (2021). GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biol. *22*, 178. https://doi.org/10.1186/s13059-021-02393-0.

109. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

110. Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. Nat. Commun. *9*, 2542. https://doi.org/10.1038/s41467-018-04964-5.

111. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

112. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

113. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L.S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome *6*, 23. https://doi.org/10.1186/s40168-018-0401-z.

114. Nguyen, N., Chawshin, K., Berg, C.F., and Varagnolo, D. (2022). Shuffle & untangle: novel untangle methods for solving the tanglegram layout problem. Bioinform. Adv. *2*, vbac014. https://doi.org/10.1093/bioadv/vbac014.

115. Gilchrist, C.L.M., and Chooi, Y.-H. (2021). clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics *37*, 2473–2475. https://doi.org/10.1093/bioinformatics/btab007.

116. Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., et al. (2025). vegan: an R package for community ecologists. GitHub. https://github.com/vegandevs/vegan.

117. Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., and Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat. Commun. *13*, 2561. https://doi.org/10.1038/s41467-022-30269-9.

118. Seemann, T. (2020). ABRicate. GitHub. https://github.com/tseemann/abricate.

119. Ugarcina Perovic, S., Ramji, V., Chong, H., Duan, Y., Maguire, F., and Coelho, L.P. (2025). argNorm: normalization of antibiotic resistance gene annotations to the Antibiotic Resistance Ontology (ARO). Bioinformatics *41*, btaf173. https://doi.org/10.1093/bioinformatics/btaf173.

120. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794. https://doi.org/10.1093/nar/gkab776.

121. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics *38*, 5315–5316. https://doi.org/10.1093/bioinformatics/btac672.

122. Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

123. Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

124. Granehäll, L., Huang, K.D., Tett, A., Manghi, P., Paladin, A., O'Sullivan, N., Rota-Stabelli, O., Segata, N., Zink, A., and Maixner, F. (2021). Metagenomic analysis of ancient dental calculus reveals unexplored diversity of oral archaeal Methanobrevibacter. Microbiome *9*, 197. https://doi.org/10.1186/s40168-021-01132-8.

125. Pilliol, V., Mahmoud Abdelwadoud, B., Aïcha, H., Lucille, T., Gérard, A., Hervé, T., Michel, D., Ghiles, G., and Elodie, T. (2024). Methanobrevibacter oralis: a comprehensive review. J. Oral Microbiol. *16*, 2415734. https://doi.org/10.1080/20002297.2024.2415734.

126. Danko, D., Bezdan, D., Afshin, E.E., Ahsanuddin, S., Bhattacharya, C., Butler, D.J., Chng, K.R., Donnellan, D., Hecht, J., Jackson, K., et al. (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell *184*, 3376–3393.e17. https://doi.org/10.1016/j.cell.2021.05.002.

127. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004. https://doi.org/10.1038/nbt.4229.

128. Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J.R., and O'Sullivan, C. (2022). The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. *50*, D387–D390. https://doi.org/10.1093/nar/gkab1053.

129. Brito, I.L., Gurry, T., Zhao, S., Huang, K., Young, S.K., Shea, T.P., Naisilisili, W., Jenkins, A.P., Jupiter, S.D., Gevers, D., et al. (2019). Transmission of human-associated microbiota along family and social networks. Nat. Microbiol. *4*, 964–971. https://doi.org/10.1038/s41564-019-0409-6.

130. Mende, D.R., Letunic, I., Huerta-Cepas, J., Li, S.S., Forslund, K., Sunagawa, S., and Bork, P. (2017). proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. Nucleic Acids Res. *45*, D529–D534. https://doi.org/10.1093/nar/gkw989.

131. Hyun, J.C., Monk, J.M., and Palsson, B.O. (2022). Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. BMC Genomics *23*, 7. https://doi.org/10.1186/s12864-021-08223-8.

132. Manghi, P., Blanco-Míguez, A., Manara, S., NabiNejad, A., Cumbo, F., Beghini, F., Armanini, F., Golzato, D., Huang, K.D., Thomas, A.M., et al. (2023). MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. Cell Rep. *42*, 112464. https://doi.org/10.1016/j.celrep.2023.112464.

133. Podlesny, D., Arze, C., Dörner, E., Verma, S., Dutta, S., Walter, J., and Fricke, W.F. (2022). Metagenomic strain detection with SameStr: identification of a persisting core gut microbiota transferable by fecal transplantation. Microbiome *10*, 53. https://doi.org/10.1186/s40168-022-01251-w.

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

135. Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J.G., Haendiges, J., Haft, D.H., Hoffmann, M., Pettengill, J.B., Prasad, A.B., Tillman, G.E., et al. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci. Rep. *11*, 12728. https://doi.org/10.1038/s41598-021-91456-0.

136. Florensa, A.F., Kaas, R.S., Clausen, P.T.L.C., Aytan-Aktug, D., and Aarestrup, F.M. (2022). ResFinder - an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microb. Genom. *8*, 000748. https://doi.org/10.1099/mgen.0.000748.

137. Bonin, N., Doster, E., Worley, H., Pinnell, L.J., Bravo, J.E., Ferm, P., Marini, S., Prosperi, M., Noyes, N., Morley, P.S., et al. (2023). MEGARes and

AMR++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencingv3.0. Nucleic Acids Res. *51*, D744–D752. https://doi.org/10.1093/nar/gkac1047.

138. Gupta, S.K., Padmanabhan, B.R., Diene, S.M., Lopez-Rojas, R., Kempf, M., Landraud, L., and Rolain, J.-M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob. Agents Chemother. *58*, 212–220. https://doi.org/10.1128/AAC.01310-13.

139. Alcock, B.P., Huynh, W., Chalil, R., Smith, K.W., Raphenya, A.R., Wlodarski, M.A., Edalatmand, A., Petkau, A., Syed, S.A., Tsang, K.K., et al. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res. *51*, D690–D699. https://doi.org/10.1093/nar/gkac920.

140. Zhou, S., Liu, B., Zheng, D., Chen, L., and Yang, J. (2025). VFDB 2025: an integrated resource for exploring anti-virulence compounds. Nucleic Acids Res. *53*, D871–D877. https://doi.org/10.1093/nar/gkae968.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
SPIRE	Schmidt et al.[24]	https://spire.embl.de/
Metalog	Kuhn et al.[31]	https://metalog.embl.de/
metaTraits	Podlesny et al.[43]	http://metatraits.embl.de/
World Bank Group Ocean Atlas[38]	Reagan et al.[42]	https://www.ncei.noaa.gov/access/world-ocean-atlas-2023/
Software and algorithms		
mOTUs (v3.0.1)	Ruscheweyh et al.[92]	https://github.com/motu-tool/mOTUs
eggNOG-mapper (v2.1.3)	Cantalapiedra et al.[93]	https://github.com/eggnogdb/eggnog-mapper
Prodigal (v2.6.3)	Hyatt et al.[94]	https://github.com/hyattpd/Prodigal
Mash (v2.3)	Ondov et al.[30]	https://github.com/marbl/mash
UMAP	McInnes et al.[28]	https://github.com/lmcinnes/umap
t-SNE	Maaten et al.[95]	https://github.com/lvdmaaten/bhtsne
densMAP	Narayan et al.[96]	https://github.com/hhcho/densvis
HDBSCAN	Campello et al.[97]	https://pypi.org/project/hdbscan/
Fastcluster (v1.2.6)	Müllner[98]	https://github.com/fastcluster/fastcluster
Phylorank (v0.1.12)	Parks[99]	https://github.com/donovan-h-parks/PhyloRank
SRA Taxonomic Analysis Tool	Katz et al.[100]	https://github.com/ncbi/ngs-tools/tree/tax/tools/tax
Cytoscape (v3.10.2)	Shannon et al.[101]	https://cytoscape.org/download.html
Scikit-learn (v1.3.0)	Pedregosa et al.[102]	https://scikit-learn.org/stable/install.html
Prefuse Force-Directed Layout	Heer et al.[103]	N/A
Traitar (v1.1.2)	Weimann et al.[104]	https://github.com/hzi-bifo/traitar
MICROPHERRET (hash: fd21931)	Bizzotto et al.[105]	https://github.com/BizzoTL/MICROPHERRET/
GenomeSPOT (v1.0.1)	Barnum et al.[106]	https://github.com/cultivarium/GenomeSPOT
CheckM2 (v1.0.1)	Chklovski et al.[107]	https://github.com/chklovski/CheckM2
GUNC (v1.0.5)	Orakov et al.[108]	https://github.com/grp-bork/gunc
MMSeq2 (v13.45111)	Steinegger and Söding[109]	https://github.com/soedinglab/MMseqs2
Linclust	Steinegger and Söding[110]	https://github.com/soedinglab/MMseqs2
Bowtie2 (v2.5.2)	Langmead and Salzberg[111]	https://github.com/BenLangmead/bowtie2
Samtools (v1.18)	Li et al.[112]	https://github.com/samtools/samtools
deepARG (v1.0)	Arango-Argoty et al.[113]	https://github.com/gaarangoa/deeparg
tanglegram (v0.2.0)	Galili[114]	https://github.com/schlegelp/tanglegram
Clinker (v0.0.31)	Gilchrist and Chooi[115]	https://github.com/gamcil/clinker
vegan (v2.6.10)	Oksanen et al.[116]	https://github.com/vegandevs/vegan
DefenseFinder (v1.0.9)	Tesson et al.[117]	https://github.com/mdmparis/defense-finder-models
ABRicate (v1.0.1)	Seemann[118]	https://github.com/tseemann/abricate
argNORM (v1.1.0)	Perovic et al.[119]	https://github.com/BigDataBiology/argNorm

METHOD DETAILS

Taxonomic, functional, and *k*-mer-based distance between metagenomic samples

We compiled mOTU v3.0.1[92]-based taxonomic profiles for 85,604 metagenomic samples collected from 690 studies worldwide using SPIRE.[24] Samples with < 100 mapped reads or more than 70% of their profile unclassified (indicated as '-1') were excluded to maintain data integrity. We removed the unclassified feature from the profiles and renormalized each profile so that the sum of sample values equaled one. The mOTUs profiles were converted to the GTDB r207 taxonomy,[120,121] and we generated profiles at different phylogenetic ranks, from family to species. Additionally, we created concatenated profiles spanning from domain to species. We applied arcsine, log, and centered log-ratio (CLR) transformations to the relative abundance profiles. The arcsine transformation

involved taking the square root of the profile values and then calculating the arcsine. For the log transformation, we added a pseudocount of half the global minimum value, took the natural logarithm, and adjusted the transformed values to range between 0 and 1.

For each sample, functional profiles were generated on different functional levels. From assemblies with a length of > 1,000 bp, open reading frames (ORFs) were predicted using Prodigal[94] (v2.6.3) and functionally annotated with eggNOG-mapper[93] (v2.1.3). To approximate ORF abundances, each ORF was multiplied by the average sequencing coverage of the contig it was found on. Then, the counts were summed on the single-letter Cluster of Orthologous Groups (COG) category level and on the root-level COG annotation. A sample-wise compositional transformation was applied, yielding relative abundance values per functional category. To generate a functional profile on the level of single-letter COG categories per habitat cluster, proportions from all samples from the same habitat cluster were summed and transformed to cluster-wise relative abundances.

For the taxonomic (mOTUs level) and functional (root-level COG) profiles, we calculated pairwise Bray–Curtis dissimilarity (on relative abundance, arcsine, and log transformed profiles), Jensen–Shannon divergence (on relative abundance), and Euclidean distances (on CLR transformed profile namely Aitchison distance) for all sample pairs. Additionally, we combined the functional profile-based distance matrix with the taxonomic profile-based distance matrices derived from mOTUs, GTDB-species, and GTDB-domain-to-species. The combined distance matrices were generated using varying ratios, ranging from 1:9 to 9:1, in increments of 1. We also calculated 21-mer overlap based distance between raw sequencing reads of the metagenomic samples using Mash[30] (v2.3). All the profiles, transformations, and distance metrics we used in this study are summarized in Figure S1A; Table S1A.

Dimensionality reduction and clustering of metagenomic samples

For each distance matrix, we applied dimensionality reduction techniques using UMAP,[28] DensMAP,[96] and t-SNE.[95] For UMAP and densMAP, we experimented with both 3-dimensional and 2-dimensional configurations, varying the *'n_neighbors'* parameters (16 levels: 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 80, 90, 100) and *'min_dist'* parameters (4 levels: 0, 0.01, 0.05, 0.1). For t-SNE, we also used both 3-dimensional and 2-dimensional configurations, adjusting *'early_exaggeration'* (4 levels: 12, 24, 36, 48, 100) and *'perplexity'* parameters (9 levels: 5, 10, 15, 20, 30, 40, 50, 75, 100).

Utilizing the coordinates obtained from each dimensionality reduction space, we then clustered the samples using HDBSCAN,[97] *k*-means, and hierarchical clustering methods. For HDBSCAN, we tested various *'min_cluster_size'* (5 levels: 10, 50, 100, 150, 200) and *'min_samples'* (7 levels: 1, 5, 10, 20, 50, 75, 100) parameters. Hierarchical clustering was performed using Euclidean distance-based average linkage. For both *k*-means and hierarchical clustering, we tried multiple *'n_clusters'* (41 levels: every 5 from 10 to 200, 250, and 300) parameters. We observed that t-SNE underperforms compared to UMAP and densMAP for our habitat clustering purpose. Therefore, we omitted t-SNE for the function-taxonomy combined distance matrix to reduce the computational burden.

As a result, a total of 2,063,646 clustering models were created by combining different profile pre-processing methods, dimensionality reduction algorithms, and clustering methods with many of their parameters (Figure S1A; Table S1A).

Benchmarking of clustering results and selection of habitat clusters

For each clustering model, we calculated the Shannon entropy[122] for each cluster based on broad habitat categories, defined based on the SPIRE microntology[24] (Figure 1A). Clusters composed exclusively of samples originating from a single broad habitat term had an entropy value of zero, whereas clusters comprising samples from heterogeneous terms exhibited higher entropy values. We employed the entropy score to penalize the fusion of metagenomes with fundamentally different broad habitat annotation but did not impose a penalty for their fission. We used the mean entropy score as the representative entropy value for each clustering model. Additionally, we calculated the silhouette index[123] based on Euclidean distances in the UMAP embedding for each clustering model. We utilized the silhouette index to quantify the robustness of the clusters and to offset the tendency toward excessive fission of highly similar samples when relying solely on the entropy score. We conducted a comprehensive comparison across various features (taxonomy, function, combined taxonomy-function, *k*-mer), transformation methods (raw, log, arcsine), dimensionality reduction methods (UMAP, densMAP, t-SNE), and clustering methods (HDBSCAN, hierarchical, *k*-means), exploring a range of dimensionality reduction and clustering parameters. Based on this parameter exploration, we observed that the combination of Bray–Curtis dissimilarity on log-transformed profile, three-dimensional UMAP, and *k*-means or hierarchical clustering resulted in consistently lower entropy and higher silhouette index values compared to other dimensionality reduction methods. Therefore, among the clustering models based on the combination, we selected the model with the lowest mean entropy that also satisfied the following criteria: silhouette index in the top 10% (> 0.511), mean entropy in the bottom 10% (< 0.306), < 100 clusters, and > 95% of samples included in clusters with a size of $\geq$ 100 (Best performing model: log-transformed mOTUs profiles, Bray–Curtis dissimilarity, 3D UMAP with n_neighbors=25 and min_dist=0.01, hierarchical clustering with n_clusters=105, Figure S2H).

If a cluster consists primarily of samples from a single study, it is difficult to determine whether it represents a unique study protocol or the environmental characteristics of the habitat where the samples were collected. Therefore, to select clusters that genuinely represent a habitat type rather than study-specific bias, we analyzed the associations among the number of samples, the number of unique studies from which the samples were derived, and the fraction of samples derived from the most dominant study. As expected, smaller cluster size, fewer unique studies, and higher dominance by a single study were correlated. Therefore, we excluded

clusters that had 100 or fewer samples, clusters derived from two or fewer studies, and clusters in which more than 95% of samples were derived from a single study. This process retained 39 out of the 105 clusters (Figure S2E).

Among the 39 clusters obtained, we manually divided a cluster, APG-HRG, into two subclusters, APG and HRG. This decision was prompted by the observation that cluster 13 was clearly separated into two distinct groups in the UMAP embedding, with any pairwise distance between samples from different subclusters being greater than any pairwise distance within the same subcluster. Additionally, we confirmed that the distribution of Bray–Curtis distances significantly decreased after dividing into subclusters, indicating increased homogeneity within each group (Figure S2F). Furthermore, both subclusters comprised animal gut microbiome samples; however, subcluster APG was predominantly composed of samples from apes (*Gorilla* and *Pan* species), while subcluster HRG was mainly composed of horse samples (Figure S3G). Although these subclusters were not defined by the initial systematically benchmarked clustering parameters and their sizes were < 100 samples, we concluded that dividing and analyzing this cluster based on their clearly distinguishable characteristics would better define clusters that reflect microbiome habitats. Through this process, we ultimately defined 40 habitat clusters consisting of a total of 71,871 samples (Table S1C).

Annotation of HTS, EST, POL, and FRW clusters: aquatic samples collected from transitional environments and extreme temperatures form distinct clusters

Except for the groundwater samples, the majority of lentic and lotic freshwater samples clustered into the freshwater (FRW) cluster. However, samples collected from estuarine environments, such as the Savannah River Estuary in eastern USA, the Mississippi Sound in southern USA, and the Pearl River Estuary in southern China, formed a distinct cluster labeled EST (Estuary). Additionally, samples from hot springs in geographically diverse regions, including Yellowstone (USA), Hawaii (USA), Kootenay Rockies (Canada), Tshipise hot spring (South Africa), and Tengchong (China), clustered together in a separate designated Hot spring (HTS) cluster. Samples collected from Arctic and Antarctic lakes or ice sources also formed an independent cluster, Polar lake (POL). Although each of these clusters (EST, HTS, and POL) contains fewer than 200 metagenomic samples, the clustering of samples from geographically distant locations based on environmental similarities suggests that estuarine environments, as transitional zones, and extreme temperature environments such as hot springs and polar lakes support distinct microbial communities.

Annotation of TEH, TER, and TEG clusters: the hydric status affects terrestrial microbiome composition

While limited metadata on soil hydric conditions make it difficult to draw definitive conclusions, an examination of the environmental metadata for clustered soil samples suggests that soil samples may primarily be distinguished by the degree of hydric conditions. Most soil samples were classified into either the terrestrial hydric (TEH) or terrestrial (TER) clusters. The TEH cluster, in particular, included soil samples from boundary environments with aquatic characteristics, such as groundwater, aquifers, and sediment, which were annotated as aquatic (SPIRE) but are taxonomically close to terrestrial environments. Additionally, this cluster contained a large proportion of soil samples from wetlands, subsurface environments, and some sludge samples from wastewater treatment facilities. In contrast, the TER cluster was predominantly composed of samples from environments that could be considered less hydric, though not necessarily dry, such as forests, rhizospheres, arid regions, and tundra. The terrestrial grassland (TEG) cluster contained 180 out of 182 samples sourced from grassland environments. Although these samples were collected across three different studies (PRJNA405447, PRJNA365981, PRJNA449266) and over different years, they all originated from the same specific location: the Angelo Coast Range Reserve in California, USA. This unique clustering warrants careful interpretation, with the caveat that it is unclear whether it reflects general grassland environments, the specific soil characteristics of this region, or some other shared environmental variable.

Annotation of HOR and ADC clusters: differentiation of ancient and modern oral microbiomes

Samples collected from the human oral cavity, primarily saliva or dental plaque, were mostly classified into the Human oral (HOR) cluster. Interestingly, samples from multiple studies analyzing dental plaque of ancient humans formed a distinct ancient dental calculus (ADC) cluster. The microbial species found in the ADC cluster were a subset of those observed in the HOR cluster, with 96.6% of species prevalent in more than 5% in the ADC cluster were also prevalent in the HOR cluster (Figure S3G). However, the prevalence and relative abundance of detected taxa was substantially different between the clusters (Figures 3A and S3G). A notable difference between ADC and HOR was an archaeal species, *Methanobrevibacter_A oralis*, which was detected in 87% of samples in the ADC cluster but only in 0.6% of samples in the HOR cluster (Figure S3G), consistent with previous studies.[124,125] As suggested by the studies, this reflects both shifts in oral microbiota composition over time between ancient and modern humans, as well as changes due to taphonomic processes that influenced microbial survival and abundance, contributing to the observed separation of the two clusters. Nevertheless, the HOR and ADC clusters remained the ecologically closest to each other, forming a robust clade (Figure 2A-clade e).

Annotation of HSS and BLD clusters: human skin microbiomes are a major source of anthropogenic surface microbiomes

The Human skin and surface (HSS) cluster contains most human skin samples as well as a substantial number of metagenomic samples collected from built environment surfaces, such as hospital surfaces (PRJEB13117) and subways (PRJNA413474) (Figure S3E). Its composition was most similar to that of the Built Environment (BLD) cluster, which consists almost exclusively of samples from

anthropogenic surfaces (Figures 1B and 2A). This finding is consistent with previous reports indicating that urban microbiomes closely resemble human skin microbiomes.[126]

Annotation of HET and ANI clusters: clusters consist of heterogeneous samples

Although most animal gut clusters exhibited high host specificity, the ANI cluster notably comprised gut samples from diverse hosts, including *Desmodus*, *Gymnogyps*, and *Ailuropoda* (Figure 2D). Considering that these animals are evolutionarily distant, inhabit distinct environments, and have markedly different dietary habits, the ANI cluster likely does not represent a specific animal or dietary pattern. Similarly, although approximately 70% of the samples in the HET cluster originated from anthropogenic environments, such as built-environment surfaces and wastewater, it also included samples with unclear commonalities, such as terrestrial samples and animal intestine microbiomes. It is possible that habitat types with limited sampling, lower sequencing quality, or particularly sparse microbiomes did not provide sufficient variability to robustly differentiate each habitat cluster via UMAP and subsequent clustering. Nevertheless, the ANI cluster was analyzed downstream as representative of animal gut environments, while the HET cluster, consisting predominantly (70%) of anthropogenic samples, was analyzed as representative of anthropogenic environments.

Ecological distance between habitat clusters

The dimensionality reduction algorithms employed in this study (UMAP, densMAP, and tSNE) prioritize the preservation of local structure over global structure. Therefore, the distances between clusters were calculated based on the distance from the original taxonomic profile, not from dimensionality reduction space.

For each habitat cluster C, we defined the set of detected taxa T as the taxa whose prevalence exceeds a specified cutoff value. The overlap coefficient distance D between two clusters $\boldsymbol{c_i}$ and $\boldsymbol{c_j}$ is defined as:

$$D(c_i, c_j) = \frac{|T(c_i) \cap T(c_j)|}{min(|T(c_i)|, |T(c_j)|)}$$

where $T(c_i)$ and $T(c_j)$ represent the sets of detected taxa for clusters c_i and c_j, respectively. Different prevalence cutoffs have distinct implications. A higher prevalence cutoff emphasizes the overlap of core taxa between habitat clusters, while a lower prevalence cutoff accounts for the overlap of taxa that are sporadically present in the clusters. Rather than using a single prevalence cutoff, we calculated D across a range from 0.01 to 0.3, increasing by 0.01 at each step, and calculated the average. We evaluated the robustness of clades in the hierarchical dendrogram generated from these distances by assessing how consistently each clade is recovered across the different cutoffs. The average distance was computed at various taxonomic ranks: from species to phylum. Since differences at higher taxonomic ranks indicate more substantial divergence, we assigned weights to the rank-specific distances inversely proportional to the median GTDB Relative Evolutionary Divergence (RED) score[127] at each rank (phylum:class:order:family:genus:species = 3.03:2.22:1.61:1.32:1.09:1). The final ecological distances between the habitat clusters were determined as the weighted average of these rank-specific distances.

Network visualization

The network was visualized using Cytoscape v3.10.2.[101] In the network shown in Figure 2A, the Prefuse Force-Directed Layout[103] was used. The weight for layout computation was set as the negative logarithm of the ecological distance. The layout parameters were set as follows: number of iterations = 50,000, default spring coefficient = 1, default spring length = 50, and default node mass = 5. The width and transparency of the edges were determined by the ecological distance between the connected habitat clusters (shorter distances resulted in thicker and more opaque edges). The node size was determined by the logarithm of the number of samples belonging to each habitat cluster.

In the network shown in Figure 5B, the layout determined in Figure 2A was preserved (maintaining the relative positions of habitat clusters based on ecological distance), and horizontal gene transfer (HGT) information was overlaid. In this network, node size represents the normalized score of intra-cluster HGT, while the edge width and transparency represent the normalized score of inter-cluster HGT. The edge color was determined by the fraction of generalist-involved HGTs among all HGT events between the two clusters. A total of 11 clusters were excluded (transparent nodes) from the network visualization because the small number of samples and resulting low number of assembled MAGs could introduce distortions in the calculation of normalized HGT scores.

Variance partitioning of the ocean clusters

We compiled environmental measurements for samples assigned to the ocean clusters (OHL, OIL, and OLL) focusing on temperature, dissolved oxygen, phosphate, and salinity. We preferentially used *in situ* data if available and imputed missing measurements from the World Bank Group Ocean Atlas[38] (WOA23[42]), matched to the closest coordinates, sampling month, and depth. For samples with both sources, we confirmed agreement between the *in situ* and imputed measurements (Figure S3C).

We then quantified the unique and shared contributions of the four environmental variables to cluster identity using variance partitioning using the varpart function in the R vegan package. Cluster identity was represented as binary (one-hot encoded) variables to allow multivariate analysis. To ensure approximately normal distribution of all environmental variables, phosphate concentrations were log-transformed.

Analysis of host and human DNA contents

To detect human or host DNA contamination in metagenomic samples, we utilized the NCBI SRA Taxonomic Analysis Tool (STAT).[100,128] We focused on samples from the FijiCOMP[129] project that lacked host annotations; samples annotated as "ocean" or "freshwater" in the SPIRE microntology and Metalog[31] but classified within the HSS cluster; and randomly selected samples from OIL, OLL, OHL, and FRW. For these samples, we compiled all associated run accessions and used the BigQuery UI to examine the taxa to which the reads were mapped. To identify the missing hosts in the FijiCOMP project samples, we investigated the most dominant vertebrate genus among those with more than 1,000 mapped reads. For assessing human DNA contamination, we calculated the proportion of reads mapped to the genus *Homo*.

Cluster-level summary of species' genomic and phenotypic features

From the GTDB metadata, we selected up to 50 genomes with the highest completeness for each species. These genomes were chosen among isolated genomes and metagenome-assembled genomes (MAGs) with a CheckM2[107] completeness > 50% and contamination < 10%. The estimated genome size for each genome was calculated as genome size × (100 / completeness). For each species, we used the mean GC content, mean coding density, and mean estimated genome size as representative values.

For pangenome openness, we compiled metagenome-assembled genomes (MAGs) and isolate-derived genomes with > 80% completeness and < 5% contamination, as assessed by CheckM2,[107] from the SPIRE[24] and Progenomes3[130] databases. For each GTDB species, we selected up to 500 genomes evenly from both databases. Species with < 10 genomes were excluded from pangenome openness calculations. Genes were predicted using Prodigal and clustered at 95% nucleotide identity using MMseqs2 linclust[109,110] (–cov-mode 0, -c 0.95). Genomes were added incrementally, and unique gene clusters were counted. This process was repeated 50 times with randomized genome order, and the resulting rarefaction curve was fit to a power law equation ($y = kx^{\alpha}$). The power law exponent, α, from this model was used to quantify pangenome openness. A higher α indicates a more open pangenome, where new genes are continually discovered as more genomes are sampled.[131] The power law equation fits the data well across most species, with a median R-squared value > 0.997.

For evolutionary features, we utilized the phylogenetic tree provided by GTDB (release 207) to calculate the tree distance to the 50 closest species for each species. Additionally, we computed the Relative Evolutionary Divergence (RED) score for all species using PhyloRank (v0.1.12). For each species (leaf node), the RED score of its immediate parent node was used as the relative time since the last speciation event.

To quantify intra-species strain diversity, quality controlled metagenomic reads were aligned to species-specific markers (mpa_vJun23_CHOCOPhlAnSGB_202307) with MetaPhlAn.[132] The resulting alignments were processed with SameStr[133] (v1.2024.02) to calculate the fraction of polymorphic sites for each species-level alignment, which served as a metric for intra-species population diversity as previously described.[134]

Phenotypic trait predictions for species in GTDB (r207) were obtained using MICROPHERRET,[105] Traitar,[104] GenomeSPOT.[106] We ran MICROPHERRET and Traitar on MAGs and genomes from SPIRE[24] and proGenomes3.[130] For this, we compiled 1,158,587 MAGs (> 50% CheckM2[111]-based completeness > 50%, contamination < 5%, and GUNC[108] passed) from SPIRE and 907,388 genomes (completeness > 90%, contamination < 5%, GUNC passed) from Progenomes. We then binarized traits on a per-species level based on a 75% trait prevalence cutoff. GenomeSPOT predictions were taken directly from the manuscript[106] and mapped from GTDB release r214 to r207 using GTDB's metadata files.

We summarized all species features at the cluster level by calculating weighted averages using each species' mean relative abundance within its habitat cluster as the weighting factor.

Weighted centroid-based generalism score

We computed the coordinates of each habitat cluster in a 40-component Multidimensional Scaling (MDS) space based on the ecological distances between habitat clusters. We confirmed that the Euclidean distances between habitat clusters in the MDS space exhibited a significantly high linear correlation with the phylogenetic distances (Pearson's $r > 0.97$, $p <$ 1e-300, Figure S5B), indicating that the MDS representation accurately reflects the phylogenetic relationships among habitat clusters.

For each taxon t, we defined the set of habitat clusters, H, where the taxon is detected with the 1% prevalence cutoff. For each taxon t, we calculated the weight $w_t(h)$, of the taxon for each habitat cluster h, in which it was present. The weight was defined as the ratio of the taxon's mean relative abundance, A, in the habitat cluster, to its mean abundance in the most abundant habitat cluster:

$$w_t(h) = \frac{A_t(h)}{A_t(h_{max})}$$

Where h_{max} is the habitat cluster where the taxon t has the highest mean relative abundance. Therefore, the weight $w_t(h)$ is proportional to the mean abundance of the taxon in the habitat cluster, h, and is constrained to a maximum value of 1. We computed the coordinates of weighted centroid C_t for each taxon in the MDS space as the weighted average of the coordinates of the habitat clusters where the taxon is present, weighted by w_t:

$$C_t = \frac{\sum_{h \in H} w_t(h) \cdot X_h}{\sum_{h \in H} w_t(h)}$$

Where X_h is the coordinate vector of habitat cluster h in the MDS space. The generalism score S, for each taxon t was calculated as the weighted sum of the squared Euclidean distances between the weighted centroid C_t and the coordinates of the habitat clusters X_h where the taxon is present:

$$S_t = \sum_{h \in H} w_t(h) \cdot d(C_t, X_h)^2$$

Where d represents the Euclidean distance function between the two input vectors. This score reflects both the distribution and abundance of the taxon across different habitat clusters. The scoring scheme is conceptually equivalent to weighted variance: the species' weighted centroid corresponds to the mean, the clusters in which the species is present correspond to the observations, the Euclidean distance corresponds to the difference, and the relative abundance corresponds to the weight. According to our definition, taxa receive higher generalism scores if they are found in clusters that are more ecologically distant from each other, and if the taxon's abundance is relatively uniform across the habitat clusters (Figure S5A).

Comparison between dendrograms

Taxonomy-based dendrograms were constructed by performing average linkage hierarchical clustering using the distance between clusters calculated based on the overlap coefficient as described above. We generated functional and trait-based dendrograms using average linkage hierarchical clustering based on the Euclidean distance between summarized root-level COG profiles (for functional) and resident species trait profiles (for trait-based) within each cluster. The entanglement between the different dendrograms was analyzed using the Python tanglegram package with an untangling process, which aims to find the pair of trees with minimized entanglement by flipping internal nodes while preserving the tree structures.[114]

Identifying horizontal gene transfer in the context of generalism and ecological distance

Horizontal gene transfer (HGT) was detected using the method described by Khedkar et al. (2022)[65] on SPIRE[24] MAGs and Progenomes3[130] isolated genomes. We used the same set of 20.7 million MAGs and genomes previously employed for phenotypic traits prediction. We clustered their genes with the mmseqs linclust[109] using a 95% nucleotide identity cutoff with > 95% mutual alignment coverage (–search-type 3 –min-seq-id 0.95 –cov-mode 0 -c 0.95). If five or more consecutive genes appear in multiple species from different taxonomic families, and at least one of these genes encodes a recombinase, we define this as an HGT event. We used a threshold of five consecutive genes because the median length of the smallest mobile genetic element type (IS_Tn, insertion sequences, and transposons) is approximately 5 kbp, which corresponds to roughly five prokaryotic genes. To confirm that our identity-based criterion effectively excludes vertically inherited genes, we analyzed ten highly conserved single-copy housekeeping genes (Table S5B) that are broadly shared across prokaryotes and quantified their nucleotide identity as a function of phylogenetic distance, grouping species pairs by the rank of their lowest common ancestor (Figure S6D). If the species pair is from different families (LCA rank $\geq$ order), only 0.000031% exceed 95% identity across the conserved genes, indicating that observing five or more consecutive genes with at least 95 percent identity between species from different families is extremely unlikely to reflect vertical inheritance. Requiring at least one recombinase within the region further increases the stringency of the call.

To interpret HGT events in the context of prokaryotic generalism and ecological distances, we determined the ecological distance of each HGT event based on the habitat cluster information of the genomes or MAGs in which the event was detected. For MAGs derived from the SPIRE database, we used the habitat cluster of the sample from which the MAG was assembled. For genomes from the Progenome database, we identified MAGs with $\geq$ 99% genome ANI and used the habitat cluster information of the sample from which the matched MAG originated. If the HGT event occurred between a pair of habitat clusters with an ecological distance > 0.5, we classified this event as a disparate-habitat HGT.

Additionally, we classified each HGT event based on the generalism scores of the species involved. If all species involved in the event had a generalism score below 0.1, the event was classified as a HGT event among specialists. If at least one species had a generalism score of 0.1 or higher, the event was classified as a generalist-involved HGT. Among these, if all species had generalism scores $\geq$ 0.1, the event was further classified as a HGT event among generalists; otherwise, it was classified as a generalist–specialist HGT.

The number of detected horizontal gene transfer (HGT) events between habitat clusters increased with the number of assembled MAGs in each habitat cluster, as a higher MAG count provides more opportunities for HGT detection. Due to limitations caused by sampling bias, the number of samples and the number of assembled MAGs differed across habitat clusters. To address this bias, the HGT count between pairs of habitat clusters was normalized by dividing by the product of the number of MAGs assembled in the two

clusters. Similarly, for intra-habitat cluster HGT, the count was normalized by the square of the number of assembled MAGs within the cluster. To minimize errors introduced by normalization, 11 habitat clusters (ADC, APG, BG3, EST, HRG, HTS, HUV, POL, TEG, TER, and WT1) with < 1,000 MAGs were excluded from the analysis.

Habitat group stratification and analysis of ARGs, virulence, and defense genes

To simplify the habitat clusters into the three groups shown in Figure 6D, we grouped habitat clusters within Figure 2-clade d, as the environmental group and within Figure 2-clade m, as the host-associated group. Exceptionally, the HET, BLD, WT1, and WT2 clusters were grouped into the anthropogenic group, according to their distinct patterns shown in Figure 3-dashed rectangles ii and iii, and Figure 6A. After assigning clusters to the three groups, we averaged the HGT score and the involvement of generalists across habitat clusters. Similar to the individual cluster-level HGT analysis, clusters containing fewer than 1,000 MAGs were excluded.

To focus on genes related to planetary health, we annotated horizontally transferred genes as antibiotic resistance genes (ARGs), virulence genes, and defense system genes. Antibiotic resistance genes were identified using ABRicate with the NCBI AMRFinderPlus,[135] ResFinder,[136] MEGARES,[137] and ARG-ANNOT[138] databases, and were additionally predicted with deepARG.[113] The resulting annotations were collated and normalized to the Antibiotic Resistance Ontology (ARO)[139] using argNorm.[119] To evaluate the clinical relevance of each resistance gene, we mapped the associated antibiotic to its corresponding WHO AWaRe category[71,72] (Table S5C). These categories progress from Access to Watch to Reserve, and reflect increasing potential for antimicrobial resistance and decreasing availability of therapeutic alternatives. If a resistance island contained genes assigned to multiple AWaRe categories, the island was classified according to the highest category present. Virulence genes and defence system genes were annotated with ABRicate using the VFDB[140] database and with DefenseFinder,[117] respectively. We defined co-transfer regions as regions containing at least one gene belonging to one of these three categories (ARG, virulence gene, and defense system genes), and we used these regions to quantify the amount of HGT and the involvement of generalists.

Identification of putative resistance islands and measuring the abundance from metagenomic samples

We focused on the horizontal gene transfer between *Aeromonas caviae*, a generalist species with a high proportion of disparate-habitat HGT and the highest number of putative AMR islands, and gut bacteria belonging to *Enterobacteriaceae*. The putative resistance island is visualized with Clinker.[115]

To quantify the abundance of the putative AMR island across habitat clusters, we randomly selected 50 samples from each habitat cluster, for a total of 2,000 samples. Raw sequencing reads from these 2,000 samples were aligned to the resistance island region using Bowtie2[111] (v2.5.2) with default parameters, and the number of aligned reads per DNA locus was determined using samtools[112] (v1.18) depth function. Vertical coverage was calculated as the number of aligned reads divided by the total number of reads, while horizontal coverage was calculated as the length of the aligned region divided by the total length of the resistance island. We further aligned the reads from environmental samples in which *Aeromonas caviae* was detected (mOTUs relative abundance > 1e−05) using Bowtie2 as described above. A horizontal-coverage threshold of 90% was applied to designate the resistance island as detected.

We observed that the regions from positions 0–850 bp, 3900–5100 bp, and 6600–8233 bp exhibited exceptionally high vertical coverage compared to the other regions (Figure S7E). In particular, the 0–850 bp, 3900–5100 bp regions are aligned with insertion sequences, which are repetitive in the genome (Figure 7A). Therefore, we excluded these regions when comparing the relative abundance of the resistance island (Figure 7B) between different habitat clusters.

QUANTIFICATION AND STATISTICAL ANALYSIS

Species features associated with the generalism score and abundance change in the WT cluster

We summarized predicted traits (from Traitar,[104] MICROPHERRET,[105] and GenomeSPOT[106]), the presence and absence of KEGG Orthologous groups (predicted by EggNOG-mapper[93]), and other genomic features, such as genome length, pangenome openness, and tree distances to the 50 closest species, for each species (see "cluster-level summary of species' genomic and phenotypic features" for details). All feature values were min-max normalized across species to standardize their ranges for linear regression. To prevent bias in the correlation analysis due to genome completeness and contamination, these factors were measured by CheckM2[107] and included as covariates in the regression model. Since the range of generalism scores varies substantially across phyla, we accounted for phylogenetic effects by including binary covariates representing phylum membership in the linear regression model below:

$$y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot Completeness_i + \beta_3 \cdot Contamination_i + \sum_{p \in P} \beta_p \cdot p_i$$

Where y_i is generalism score for species i, X_i is min-max normalized feature value for species i, and p_i is binary value for phylum membership of species i. For the phylum-specific analysis, a similar linear regression model was applied, replacing phylum membership with class membership within the respective phylum.

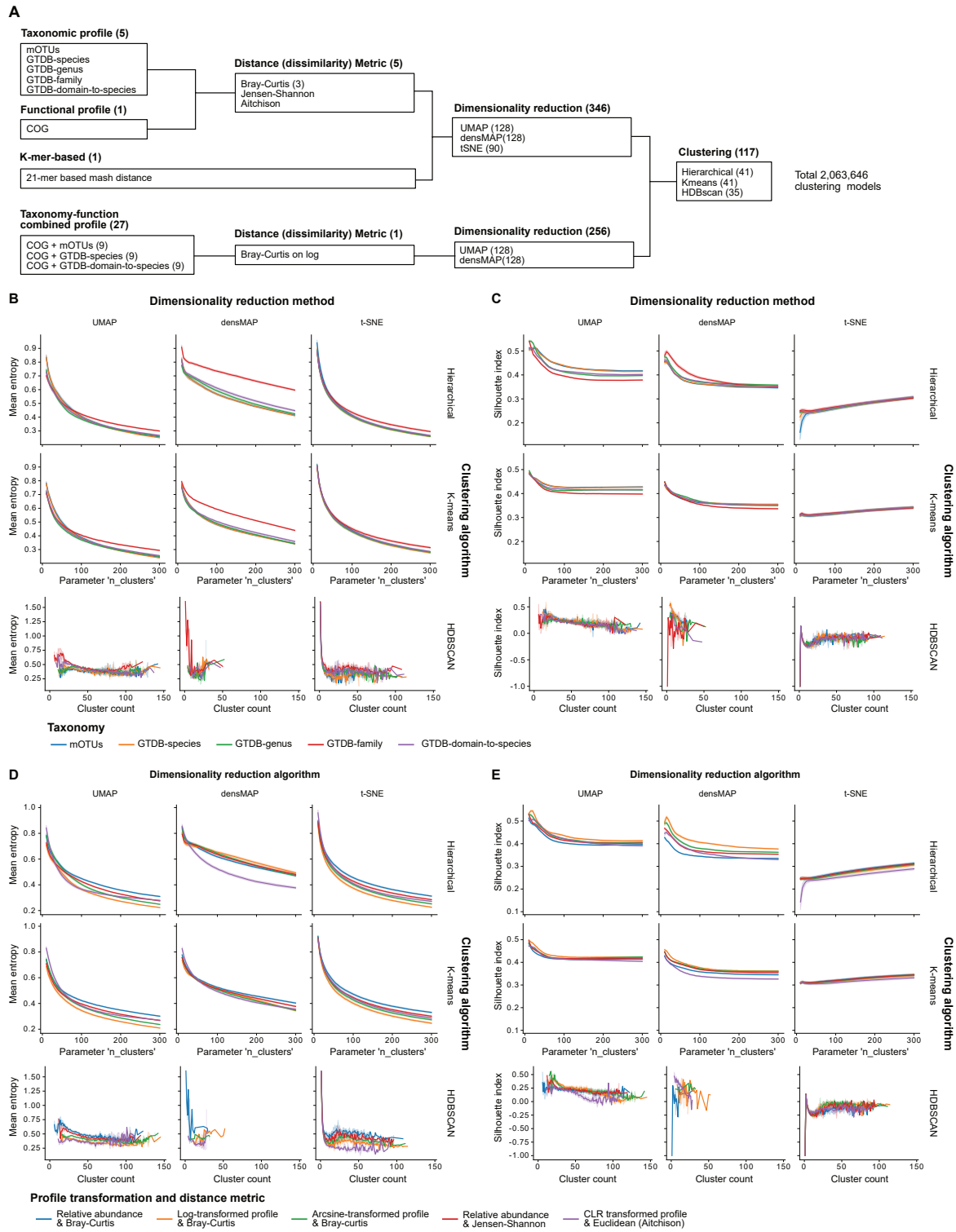
To investigate the association between species traits and the abundance change in the WT cluster, we only used predicted traits from Traitar for simplicity. Logistic regression was performed to analyze the association between the predicted traits and abundance change in the WT cluster.

Additional statistical analyses used in this study

Two-tailed Mann–Whitney U tests were performed to compare ocean latitude (Figure 2G), ocean temperature (Figure 2H), human DNA contents (Figure S3F), temperature range (Figure S5C), pH range (Figure S5D), and GC percentage (Figure S5E). For pairwise Mann–Whitney U tests, multiple testing correction was performed using the Holm–Bonferroni method (Figures S5C–S5E). For comparisons of temperature range, pH range, and GC percentage across generalism, we binned generalism scores in 0.05 increments; values $\geq$ 0.15 were collapsed into a single bin due to the low number of species in higher ranges. habitat cluster count bins were set to maximize similarity in sample size to minimize the influence of N on statistical power (Figures S5C–S5E). For the regression analyses (see "species features associated with the generalism score and abundance change in WT cluster") multiple testing correction was performed using the Benjamini–Hochberg procedure (Figures 4D, S5F, and S6B). Two-sided Fisher's exact tests, followed by Bonferroni adjustment for multiple comparisons, were performed to compare country enrichment across the human gut clusters (Figure S3A).

CellPress
OPEN ACCESS

Supplemental figures



(legend on next page)

Figure S1. Workflow and performance of taxonomic profile-based clustering models, related to Figure 1

(A) Analysis framework for the planetary-scale microbiome-based habitat clustering. The diagram shows the parameters and number of parameter settings (in parentheses) at each step (Table S1A), yielding 2,063,646 clustering models.

(B and C) Comparison of clustering models generated using different taxonomic resolutions. (B) Habitat entropy (lower scores represent better alignment with known broad habitat categories) and (C) silhouette indexes (higher scores indicate better cluster separation) are compared across models based on species-level mOTUs, GTDB species, genus, family, and domain-to-species-level profiles.

(D and E) Comparison of clustering models generated using different data transformations. (D) Habitat entropy and (E) silhouette indexes are compared across models using Bray-Curtis dissimilarity based on raw, log-transformed, and arcsine-transformed profiles; Jensen-Shannon divergence based on raw profiles; and the compositional Aitchison distance.

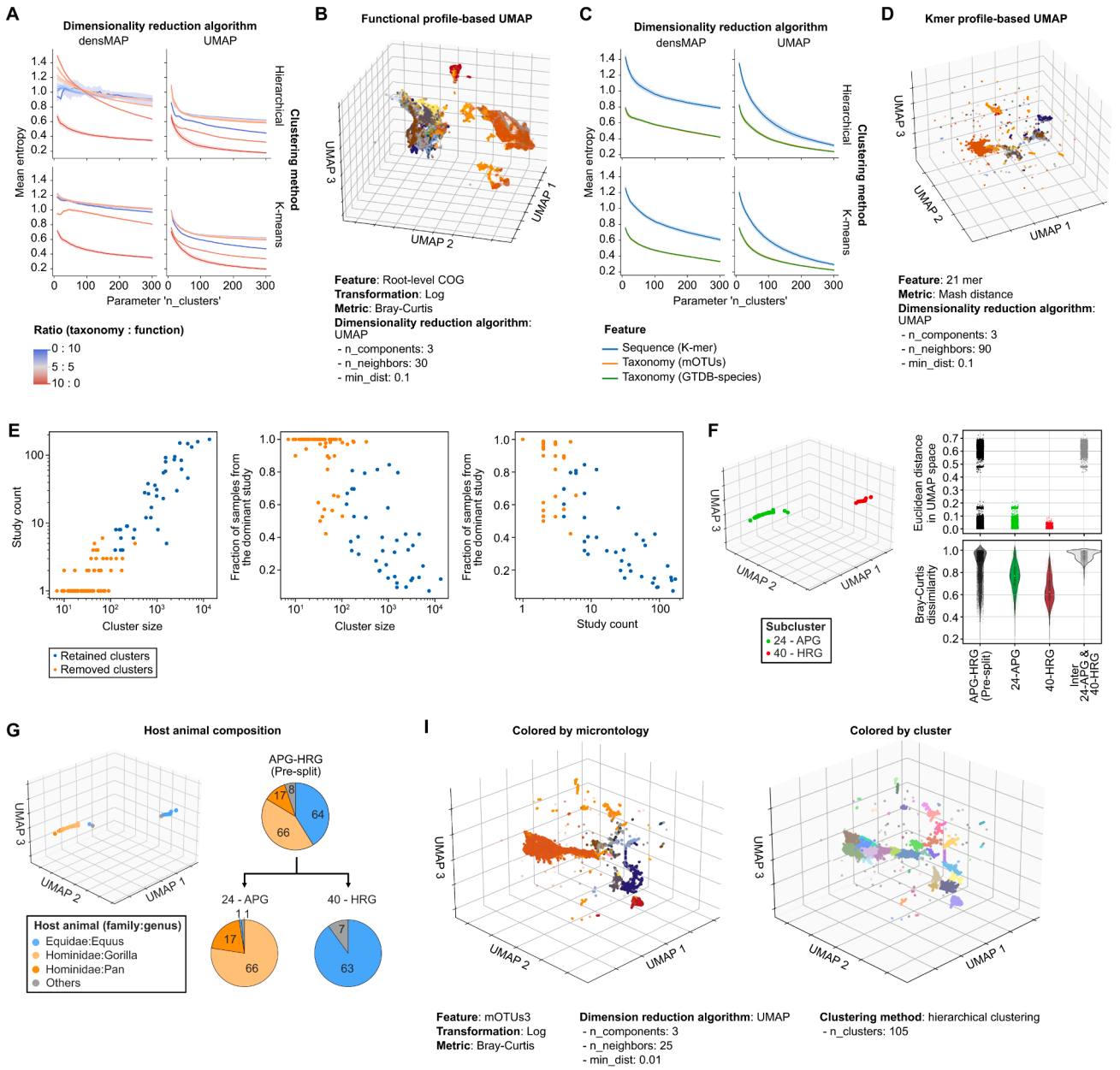


Figure S2. Benchmarking and post hoc refinement of the clustering models, related to Figure 1

(A) Comparison of habitat entropy across clustering models based on functional profiles (blue), taxonomic profiles (red), and various ratios of combined functional-taxonomic profiles (blue-to-red gradient; STAR Methods).

(B) 3D UMAP embedding of the clustering model built using the functional root-level COG profile.

(C) Comparison of habitat entropy for clustering models based on mash distances and taxonomic profiles.

(D) 3D UMAP embedding of the clustering model built using the mash distance profiles. Datapoint colors in (B) and (D) represent the broad habitat category per sample, matching the color scheme in Figure 1.

(E) Correlations of cluster size (numbers of samples within each cluster), study count (number of unique studies from which the samples were derived), and the fraction of samples from the dominant study for all clusters. The 39 retained clusters are highlighted as blue data points included in the final habitat clustering model.

(F) 3D UMAP embedding of the pre-split 24-APG and 40-HRG clusters and comparison of inter- and intra-cluster distances for the pre- and post-split of the 24-APG and 40-HRG clusters. (Upper-right) Euclidean distances calculated in UMAP embedding, and (lower-right) Bray-Curtis dissimilarity calculated from the original taxonomic profiles.

(legend continued on next page)

(G) UMAP embedding of 24-APG and 40-HRG clusters, visualized by host animal (family: genus) and host animal composition of 24-APG and 40-HRG clusters before and after the split.
(H) Three-dimensional UMAP embedding of the final habitat clustering model, colored by (left) broad habitat category and (right) habitat cluster. See Figure 1A for the color scheme. The parameters used to build the final model are specified.

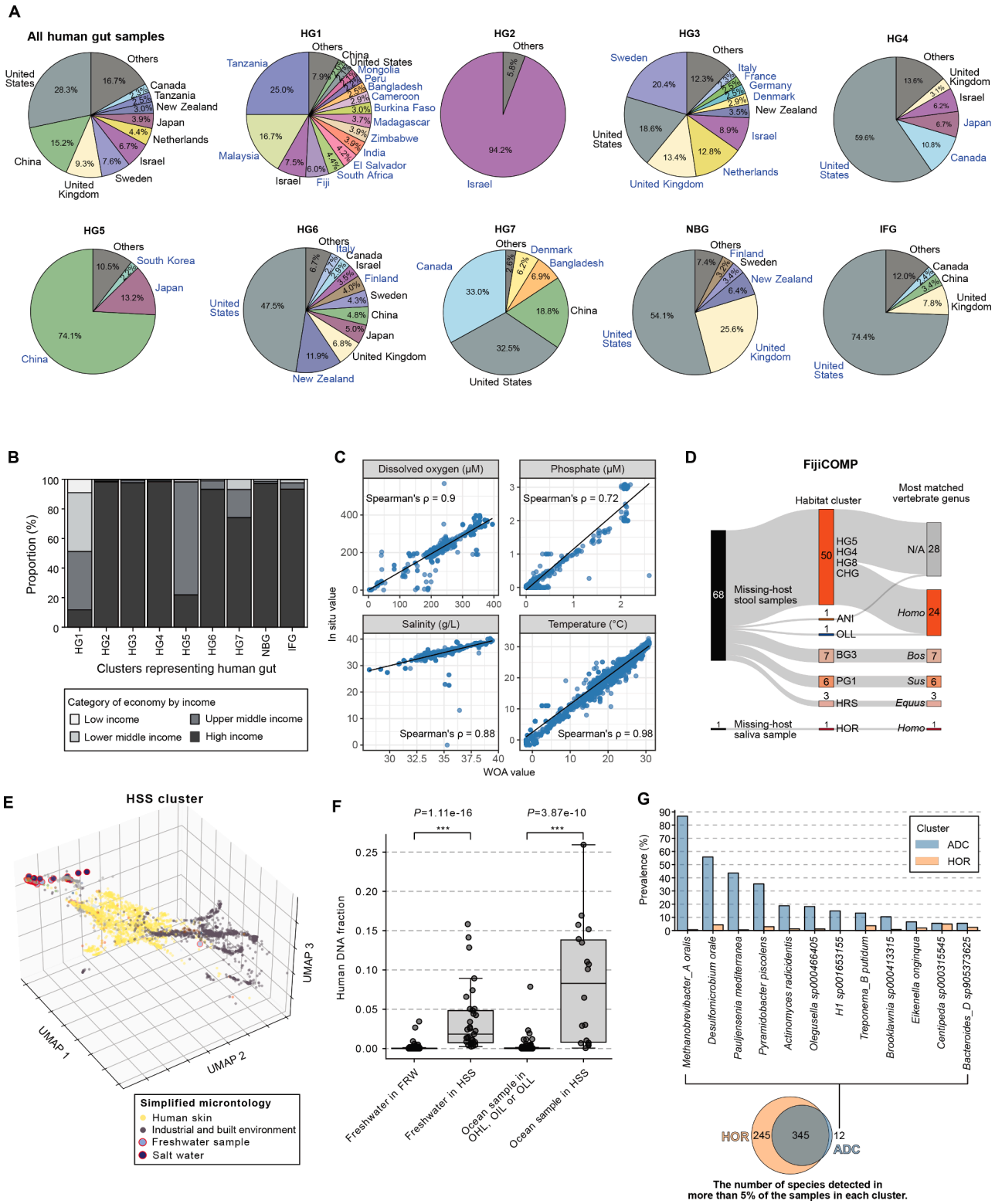


Figure S3. Contextual data and QC for the habitat clusters, related to Figure 2

(A) Country composition of samples in each human gut cluster. Countries labeled in blue contributed a significantly higher proportion of samples to that cluster (two-sided Fisher's exact test, Bonferroni-corrected $p < 0.1$, odds ratio > 1).

(B) Income categories of countries (The World Bank, 2024) by human gut clusters.

(legend continued on next page)

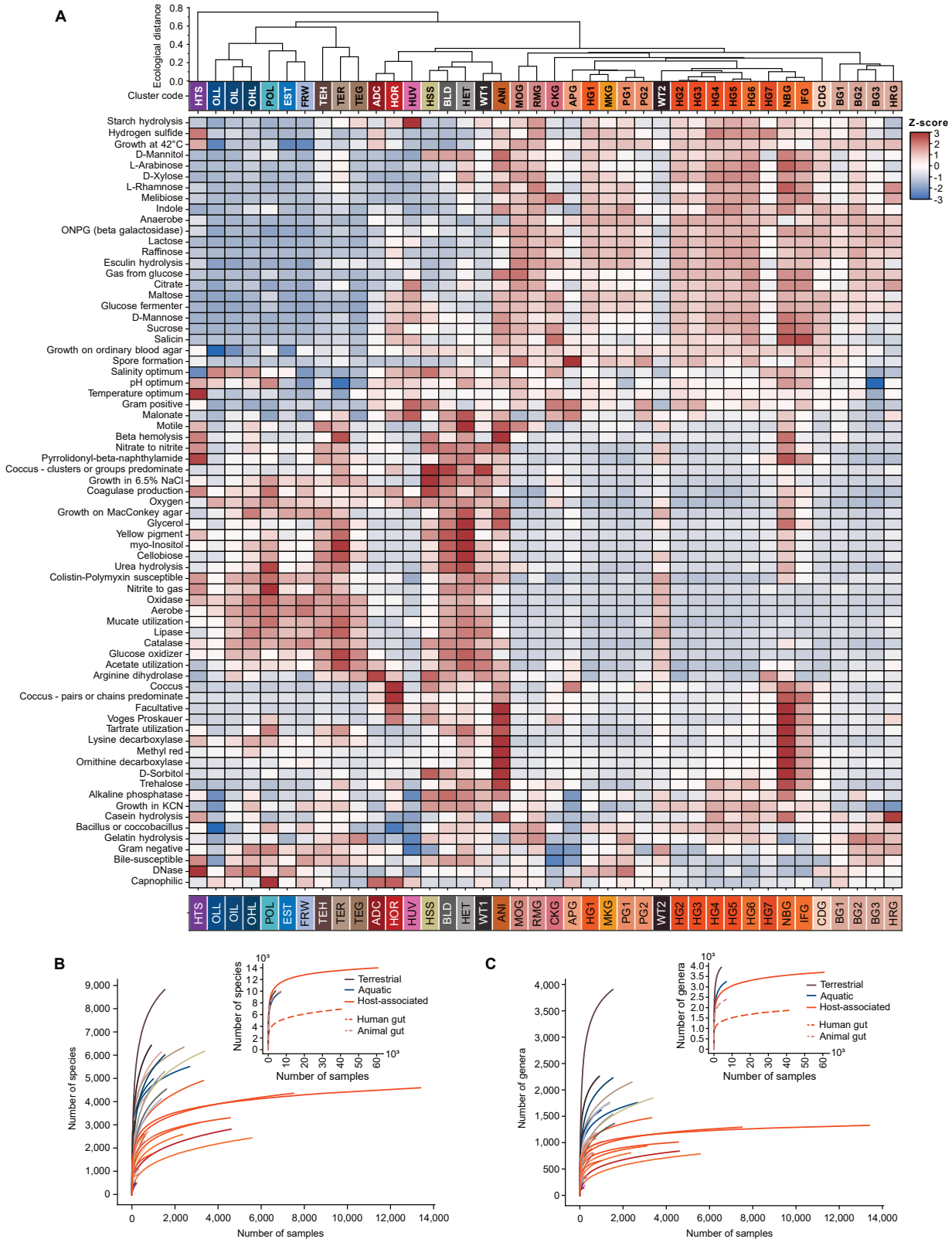
(C) Comparison of *in situ* metadata with World Ocean Atlas (WOA) values for dissolved oxygen, phosphate, salinity, and temperature across samples from the three ocean clusters. Each sample was matched to the nearest WOA grid point at a similar depth. Solid lines indicate the linear regression, and statistics are Spearman's ρ.

(D) Habitat classification and host-DNA contamination analysis of samples lacking host annotations from the FijiCOMP dataset. These stool samples were mapped to habitat clusters, indicating their best-matching vertebrate genus based on host-DNA contamination in the Sankey diagram.

(E) UMAP embedding of the human skin and surface cluster (HSS). Samples with metadata indicating freshwater or ocean environments are outlined in red.

(F) Comparison of human DNA content between freshwater and ocean samples classified into HSS and those assigned to clusters matching their habitat metadata (freshwater [FRW], OIL, OLL, or OHL). Statistical significance was evaluated using a two-tailed Mann-Whitney U test.

(G) Venn diagram showing the species overlap between human oral (HOR) and ancient dental calculus (ADC) clusters, based on species that were detected at >5% prevalence in each cluster, and a bar chart representing the prevalence of species that were detected with >5% prevalence only in the ADC cluster.

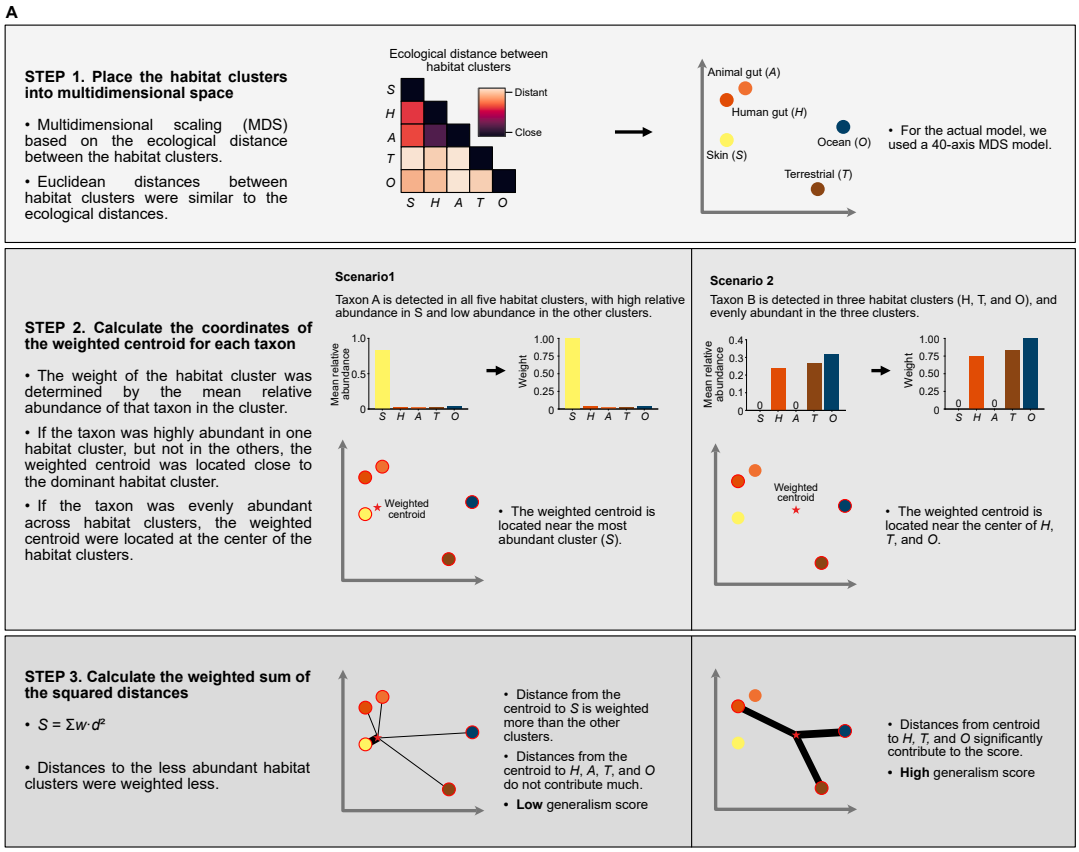


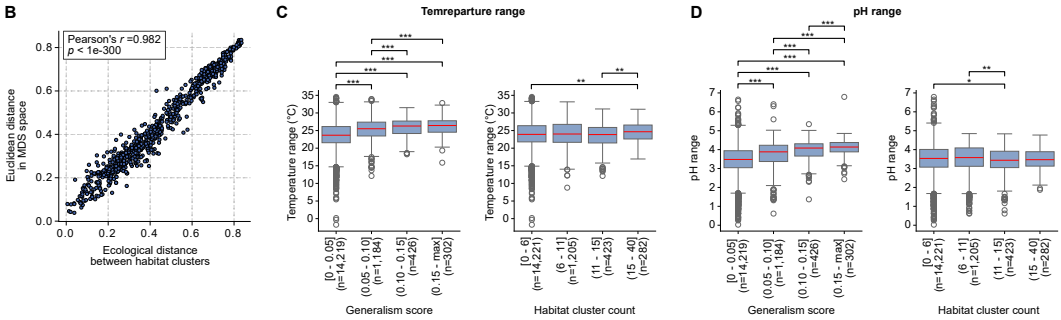
(legend on next page)

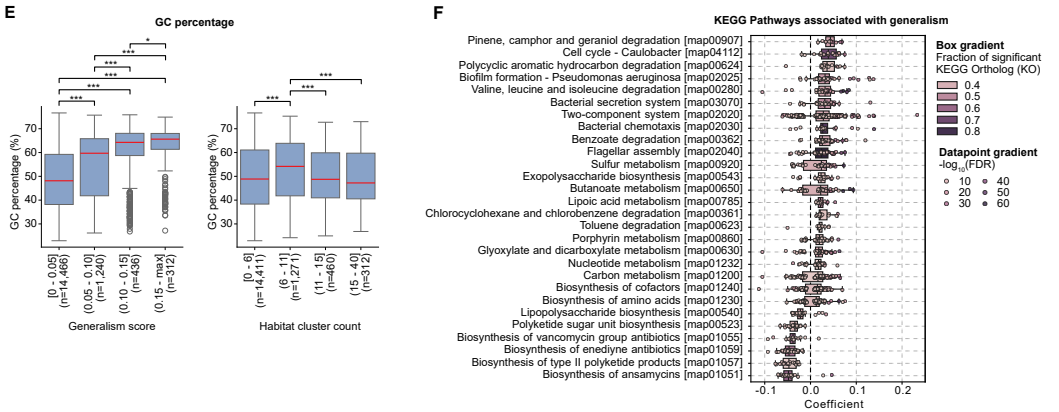
Figure S4. Phenotypic traits and rarefaction analysis, related to Figure 3

(A) Heatmap illustrating the predicted phenotypic traits of species residing within habitat clusters. Genome-based predictions of microbial phenotypic traits were quantified at the habitat level by multiplying each species' trait prevalence by its habitat-specific mean relative abundance.

(B and C) Assessment of species-level (B) and genus-level (C) richness across simulated sampling efforts (rarefaction analysis). Insets show rarefaction curves after aggregating habitat clusters into terrestrial (Figure 2A, clade c), aquatic (Figure 2A, clade b), host-associated (Figure 2A, clade m), human gut (HG1–7, IFG, and NBG), and animal gut (ANI, MOG, RMG, CKG, APG, MKG, PG1–2, CDG, BG1–3, and HRG) categories. Human gut and animal gut are part of the host-associated category but are shown separately due to their relevance in microbiome research.







(legend on next page)

Figure S5. Generalism score calculation and genomic and phenotypic traits associated with generalism, related to Figure 4

(A) Schematic of the centroid-based generalism score calculation, which considers the ecological distance between a microbe's habitat(s) and how evenly the microbe is distributed among them.

(B) The correlation between the overlap coefficient-based phylogenetic distance and the Euclidean distance in multidimensional scaling (MDS) space was evaluated using Pearson's correlation coefficient.

(C–E) Boxplots of (C) temperature range (maximum minus minimum across occurrences), (D) pH range, and (E) GC percentage as a function of generalism score (left panel of each pair) and the number of habitat clusters occupied (right panel). Sample sizes per bin are indicated on the *x* axis. Pairwise differences between adjacent bins were assessed with Mann-Whitney U tests and Holm correction (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

(F) KEGG pathways (KPs) associated with prokaryotic generalism. Data points represent KEGG orthologs (KOs) within each pathway. Only KPs with more than 40% of their KOs significantly (FDR < 1e−10) enriched in either generalist or specialist species are shown.

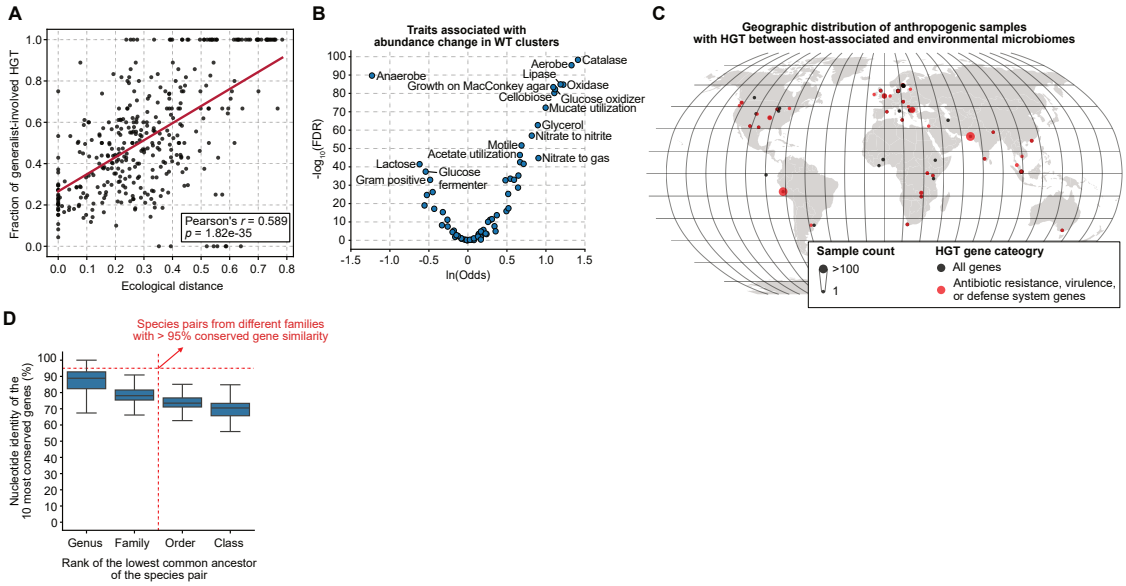


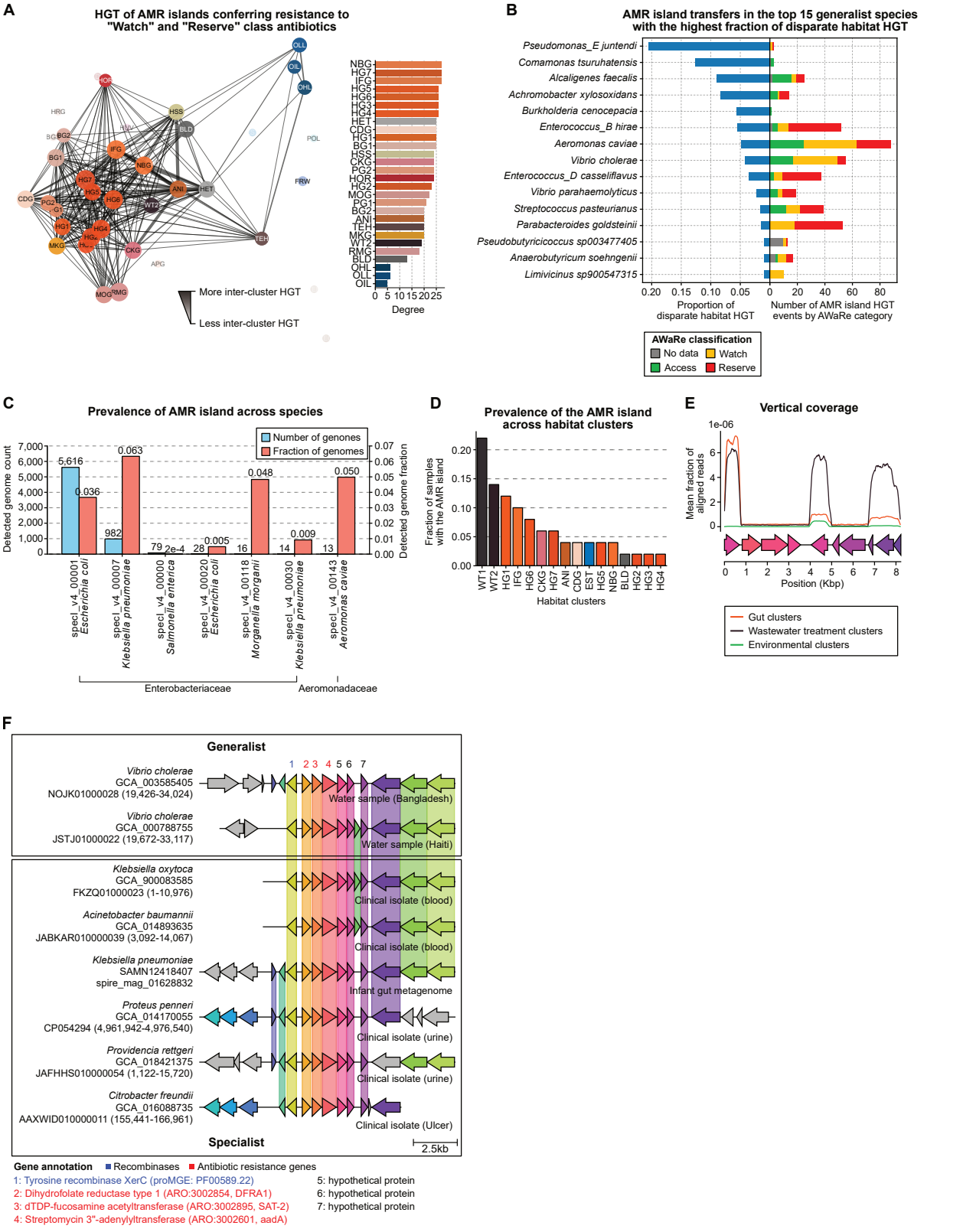
Figure S6. HGT across habitat types, related to Figures 5 and 6

(A) Linear relationship (Pearson's $r = 0.589$, $p = 1.82e{-}35$) between the ecological distance of habitat cluster pairs (>1,000 MAGs with >50% completeness) and the generalist (generalism score ≥ 0.1) involvement in HGT.

(B) Traits associated with changes in abundance from gut to wastewater (WT) clusters. Positive odds indicate traits enriched in species with increased dominance in WT clusters.

(C) Geographic distribution of anthropogenic samples in which HGT with host-associated or environmental samples was observed. Red points mark locations where HGT involved antibiotic resistance, virulence, or defense system genes.

(D) Boxplots show the nucleotide identity of the ten most conserved single-copy genes for species pairs grouped by the rank of their lowest common ancestor (genus, family, order, and class). The horizontal red line marks 95% sequence identity, and the vertical line highlights cross-family comparisons. Species pairs from different families rarely exceed 95% sequence identity across these conserved genes.



(legend on next page)

Figure S7. Case studies of generalist-mediated ARG transfer across habitats, related to Figure 7

(A) HGT network of AMR islands conferring resistance to "watch" and "reserve" class antibiotics by AWaRe classification. The bar chart on the right represents the network degree for each habitat cluster.

(B) For each of the 15 generalist species with the highest fraction of disparate-habitat HGT (between habitat cluster pairs with ecological distance > 0.5), the proportion of disparate-habitat HGT events and the number of AMR island HGT events by AWaRe category are shown.

(C) Number and fraction of genomes containing the resistance island, organized by specI clusters. SpecI clusters with <10 genomes containing the resistance island were excluded.

(D) Proportion of samples containing the resistance island (horizontal coverage >90%) across habitat clusters.

(E) Average vertical coverage of the resistance island by metagenomic reads per environment.

(F) Evidence for horizontal transfer of a putative AMR island between environmental and host-associated contexts. Genomes of the generalist, potentially pathogenic species *Vibrio cholerae* harbor the same AMR island found in diverse clinical isolates of specialist species. Genes with 100% nucleotide identity are shown in identical colors. Gene annotations are listed below, with blue text indicating recombinase enzymes and red text indicating ARGs.