Title

A metagenomic perspective on the microbial prokaryotic genome census

Journal

Science Advances, 11(3)

ISSN

2375-2548

Authors

Wu, Dongying
Seshadri, Rekha
Kyrpides, Nikos C
et al.

Publication Date

2025-01-17

DOI

10.1126/sciadv.adq2166

Peer reviewed

MICROBIOLOGY

A metagenomic perspective on the microbial prokaryotic genome census

Dongying Wu, Rekha Seshadri, Nikos C. Kyrpides, Natalia N. Ivanova*

Following 30 years of sequencing, we assessed the phylogenetic diversity (PD) of >1.5 million microbial genomes in public databases, including metagenome-assembled genomes (MAGs) of uncultivated microbes. As compared to the vast diversity uncovered by metagenomic sequences, cultivated taxa account for a modest portion of the overall diversity, 9.73% in bacteria and 6.55% in archaea, while MAGs contribute 48.54% and 57.05%, respectively. Therefore, a substantial fraction of bacterial (41.73%) and archaeal PD (36.39%) still lacks any genomic representation. This unrepresented diversity manifests primarily at lower taxonomic ranks, exemplified by 134,966 species identified in 18,087 metagenomic samples. Our study exposes diversity hotspots in freshwater, marine subsurface, sediment, soil, and other environments, whereas human samples yielded minimal novelty within the context of existing datasets. These results offer a roadmap for future genome recovery efforts, delineating uncaptured taxa in underexplored environments and underscoring the necessity for renewed isolation and sequencing.

INTRODUCTION

The biodiversity contained within the bacterial and archaeal domains is vast, and these microorganisms play critical roles in sustaining the biosphere and supporting the health of our planet. Whole-genome sequences serve as essential tools for studying these organisms, and, as of August 2023, the National Center for Biotechnology Information (NCBI) GenBank houses more than 1,316,387 genomes of prokaryotic isolates. Note that, in the past, genome collections showed a bias toward certain taxonomic groups of organisms and specific environments (*1*). To address this limitation and promote a more comprehensive understanding of prokaryotic life, initiatives such as the Tree of Life and the Genomic Encyclopedia of Bacteria and Archaea (*2*) were implemented. Despite these efforts, there remains a conspicuous amount of redundancy in the available genome data, especially (and understandably) for clinically important pathogens.

More recently, advancements in metagenomics have allowed scientists to study genetic material directly from environmental samples, without the need for laboratory culture. This has led to the recovery of numerous uncultivated genome equivalents known as metagenome-assembled genomes (MAGs) from a diverse range of environments (*3*–*8*). These MAGs have provided unprecedented insights into previously inaccessible and unknown taxa (microbial "dark matter") (*9*, *10*), greatly enriching our understanding of microbial diversity and challenging long-held precepts gleaned from studying isolated or domesticated species (*9*, *11*, *12*). The numbers of MAGs submitted to GenBank and other databases are rapidly mounting and rivaling those of isolate genomes. While MAGs represent a welcome advancement, the recovery of high-quality (HQ) MAGs (*13*) from individual metagenomes is relatively low.

Given the bounty of sequenced genomes from both cultivated and uncultivated sources, we aimed to conduct a comprehensive census of *Bacteria* and *Archaea* and gauge the biodiversity represented by these genome sequences. Through this analysis, we can pinpoint the missing taxa and their samples for further exploration, aiming to build a comprehensive, phylogenetically diverse compendium of microbial genomes and cultures to advance our understanding of the microbial planet.

DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
*Corresponding author. Email: nnivanova@lbl.gov

RESULTS

Net phylogenetic diversity of *Bacteria* and *Archaea*

Derived from summing the branch lengths connecting taxa on a phylogenetic tree, we calculated phylogenetic diversity (PD), a simple yet effective proxy measure of biodiversity (*2*, *14*). For tree building, we tested five universally conserved single-copy marker genes: Profile Hidden Markov Models (HMMs) for each gene were used to retrieve markers from a comprehensive dataset of 1,889,638 genomes including both isolates and MAGs collected from NCBI and the Integrated Microbial Genomes (IMG) databases, as well as 43,965 metagenomes from IMG. We chose to use protein-coding genes instead of the 16*S* ribosomal RNA (rRNA) gene, primarily because the 16*S* gene is poorly recovered in metagenome assemblies because of its highly conserved, repetitive nature and low complexity regions (in addition to other known issues) (*15*, *16*). Trees were constructed on the basis of aligned protein sequences of marker genes arising from 137,658 dereplicated isolate genomes, 28,269 HQ IMG MAGs, 68,655 medium-quality (MQ) IMG MAGs, 209,568 NCBI MAGs (unknown quality), and 34,068 metagenomes from diverse environmental samples (Fig. 1 and Methods).

Our analysis included independent evaluation of results and inferences from five distinct marker gene trees, namely, ribosomal protein L1 (COG0081), tRNA A37 threonyl carbamoyltransferase TsaD (COG0533), signal recognition particle guanosine triphosphatase (COG0541), alanyl-tRNA synthetase (COG0013), and RNA polymerase beta prime subunit RpoC (COG0086) (Table 1). While similar numbers of genes were retrieved for isolates and MAGs for each marker, we observed considerable variability in the numbers of genes retrieved from the unbinned portion of metagenomes. This variability exhibited an inverse relationship with the length of the marker gene, with the shortest (COG0081) retrieving the most genes and the longest (COG0086) retrieving the fewest (Table 1). This is due to the increased likelihood of fragmentation in metagenomic assemblies for longer markers, leading to potential undercounting. This variation in gene

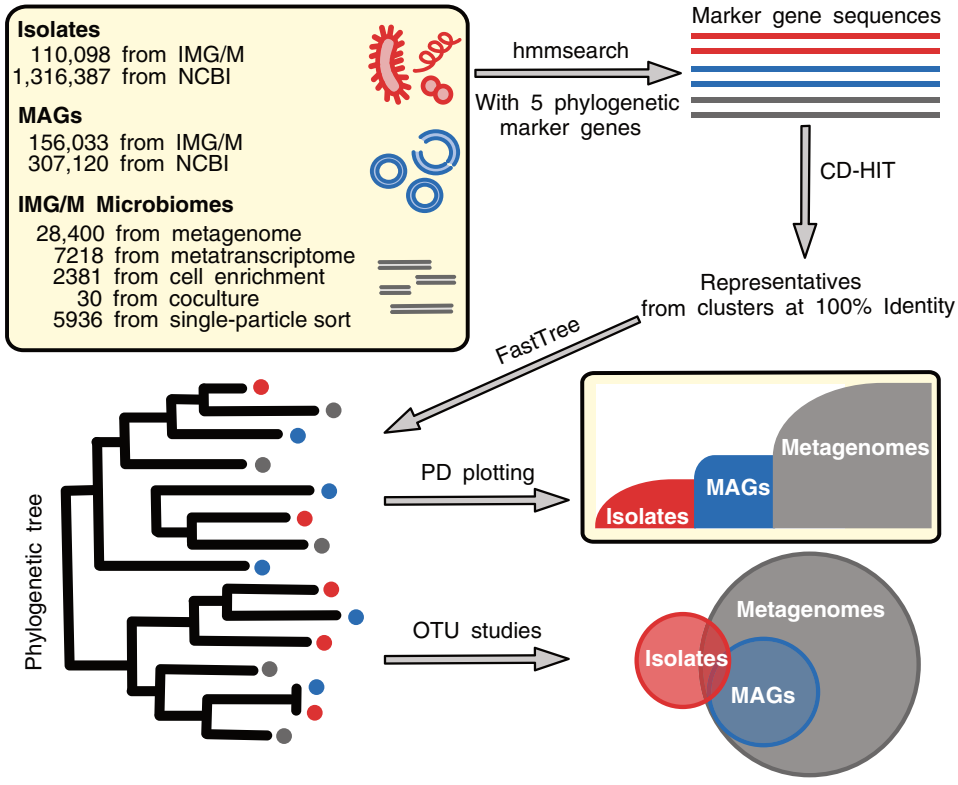

Fig. 1. Schematic overview of the workflow used in this study. The amino acid sequences of phylogenetic marker genes were obtained from isolates, MAGs, and metagenomic datasets using hmmsearch with the hmm models. Redundancy reduction was performed using CD-HIT for each marker, followed by alignment building and phylogenetic tree construction. The resulting trees were used to evaluate the phylogenetic diversity (PD) contribution of sequences derived from isolates, MAGs, and metagenomes. Operational taxonomic units (OTUs) were generated at various taxonomy levels based on these trees, and the distribution of OTUs containing genes from different categories was analyzed at each taxonomy level.

Table 1. Protein counts of five phylogenetic markers from different databases. Only genes with ≥60% non-gap positions of the alignment length are included. Split marker genes were excluded. GTPase, guanosine triphosphatase.

COG	DESCRIPTION	HMM	Length	Proteins from IMG/M isolates	Proteins from NCBI isolates	Proteins from IMG/M MAGs	Proteins from NCBI MAGs	Proteins from IMG/M metagenomes
COG0081	Ribosomal protein L1	COG0081	230	96,796	1,317,582	112,686	228,678	2,037,918
COG0533	tRNA A37 threonylcarbamoyltransferase TsaD	COG0533	342	97,118	1,319,320	128,471	264,130	1,169,060
COG0541	Signal recognition particle GTPase	COG0541	452	95,049	1,319,907	119,735	245,018	948,009
COG0013	Alanyl-tRNA synthetase	Pfam01411 Pfam07973 pfam02272	734	96,549	1,315,914	124,701	252,468	625,171
COG0086	RpoC	Pfam04997 Pfam00623 Pfam04983 Pfam05000 pfam04998	1010	96,942	1,313,029	105,330	207,437	505,068

recovery notably affected the proportional contributions to PD for each category, especially for bacterial PD. Applying different length cutoffs to the metagenome scaffolds used for recovery of the shortest marker gene (COG0081), similar variations in the PD contributions were measured for the bacterial domain (fig. S1C). COG0081 PD contribution distribution was most similar to those of COG0533 and COG0541 at a scaffold length cutoff of 1 kb [Pearson's correlation coefficient (r) of 0.9998 and 0.9992, respectively]. COG0013 was the closest match at a scaffold length cutoff of 2 kb (Pearson's r of 0.9993), while COG0086 was a match at a scaffold length cutoff of 3 kb (Pearson's r of 0.9922). We observe similar variation in archaeal PD contributions for different phylogenetic markers with the exception of COG0086 RpoC (Fig. 2, C and D). COG0086 genes are broken into two genes in many archaeal lineages (e.g., *Methanobacteriaceae* and *Halorubraceae*). These fragmented genes were excluded to avoid inclusion of pseudogenes caused by bad sequence/assembly quality in metagenomes at the cost of introducing errors in the archaeal PD analysis.

However, when the five marker genes were used to reconstruct phylogenies including all genomes, they resulted in largely similar "topologies" as measured by adjusted mutual information (AMI) between the best matched operational taxonomic units (OTUs) and Genome Taxonomy Database (GTDB) taxonomic groups (table S1) (see Methods). Briefly, a relative evolutionary divergence (RED) analysis was performed (*17*) with various cutoffs to organize markers into OTUs and gain an approximate hierarchical classification by comparing against GTDB-based taxonomy of reference genomes. These RED OTUs generated from the five distinct phylogenetic markers exhibited a high degree of comparability across each taxonomic level for both bacterial and archaeal domains. Of the 120 pairs of OTUs at the same taxonomic level within the same domain, 107 pairs demonstrate pairwise distances estimated by AMI above 0.8. The remaining 13 pairs exhibit pairwise distances above 0.7, with 10 of these occurring at the phylum level and 3 at the class level (fig. S2A).

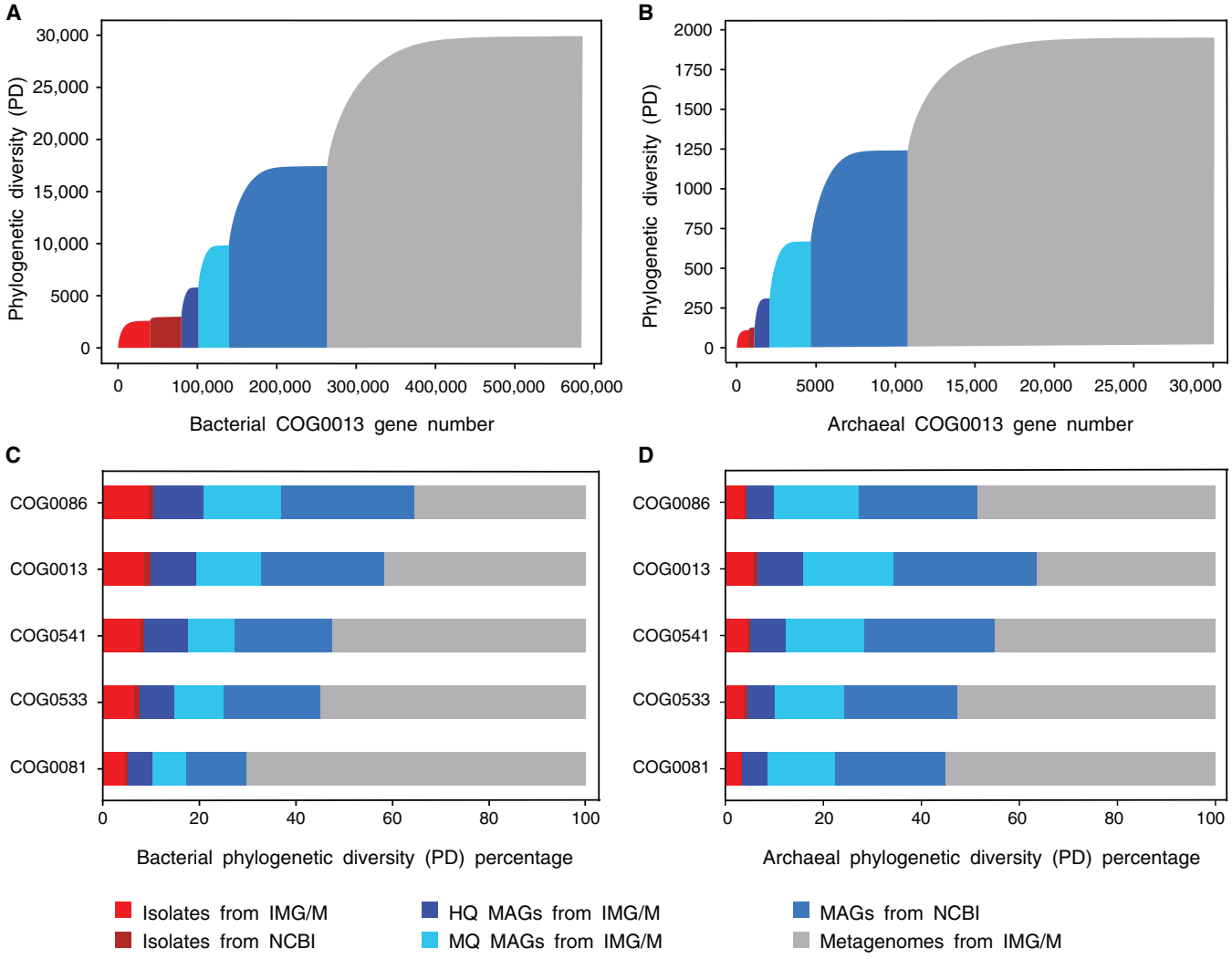

Fig. 2. Estimates of total bacterial and archaeal PD. (**A**) Bacterial phylogenetic diversity (PD) accumulation curve depicting incremental increase in PD inferred from computed branch lengths of alanyl-tRNA synthetase (COG0013) tree. The units on the x axis represent individual taxa or their equivalents (from metagenomes) ordered by genome category as the "accumulation units." PD score based on summed branch lengths is shown on the y axis. (**B**) Archaeal PD accumulation curve. (**C**) Marker-based variability in relative PD contributions for bacteria. (**D**) Marker-based variability in relative PD contributions for archaea.

The variations in OTU composition across different phylogenetic markers are most evident at the species and genus levels. Shorter genes yield a higher count of metagenome-only OTUs (mOTUs). For instance, COG0081 (230 amino acid long) identifies 480,821 species-level and 84,676 genus-level OTUs for bacteria. By contrast, COG0086 (1010 amino acid long) yields 152,349 and 41,444 OTUs, respectively. This discrepancy is less pronounced at higher taxonomic levels. We also used five different markers to overcome potential complications afflicting any single marker. For example, we observed gene duplications for COG0081, potential horizontal gene transfer of COG0013 from bacteria to archaea, and spanning of one marker over multiple genes in COG0086. We have selected COG0013 to present further findings (applying a minimum 2-kb metagenome scaffold size cutoff to exclude lower-quality assemblies and gene calls) because COG0013 OTUs consistently align more closely with the GTDB taxonomies across most taxonomic levels (table S1) compared to other markers.

Using the alanyl-tRNA synthetase marker gene (COG0013), we found that bacterial isolate genomes represent 9.73% of the total estimated diversity in the domain *Bacteria* and 6.55% in *Archaea* (Fig. 2, A and B). This estimate contrasts the oft-cited "great plate count anomaly experiment," which suggested that <1% of microorganisms observed under a microscope were cultivated (*18*). IMG MAGs including both HQ and MQ MAGs recovered from diverse metagenome datasets account for 22.95% of total bacterial PD. In addition, NCBI MAGs (of unknown quality and provenance) contribute an additional 25.59%. It is worth noting the inherent limitations of MAGs such as higher rates of incompleteness (down to 50% for MQ MAGs), genome fragmentation, absence of mobile genetic elements, their population-level composite structure (collapsing many closely related strains), lack of appropriate standards, and possible chimeric nature (*19*–*21*). The remaining 41.73% of bacterial diversity remains without any type of genomic representation and is identifiable only from the current metagenomic marker gene data. This figure represents a rather conservative estimate: Uncaptured bacterial PD ranges from 35.56 to 70.22%, depending on the length of the marker used for estimations as discussed above (Fig. 2C). For domain *Archaea*, only 6.55% of total estimated PD is captured by isolates, while HQ, MQ, and NCBI MAGs genomes boost this coverage by 9.33, 18.37, and 29.35%, respectively (Fig. 2B). This leaves 36.39% of total archaeal PD without any genomic representation.

Narrowing in from domain to phylum, we focused on phyla with the largest numbers of sequenced isolate genomes: *Pseudomonadota* (formerly *Proteobacteria*), *Bacillota* (formerly *Firmicutes*), *Actinomycetota* (formerly *Actinobacteria*), and *Bacteroidota* (formerly *Bacteroidetes*). At this taxonomic level, the PD captured by isolates is slightly higher (ranging from 9.57 to 17.06 %) compared to 9.78 % within domain *Bacteria*. Similar to before (*22*), we observed that 52.67% of actinobacterial diversity remains uncaptured by genome sequences, followed by *Pseudomonadota* with 42.35% and *Bacteroidota* with 32.59% outstanding PD (fig. S1, A and B). By contrast, for *Bacillota*, only 18.31% remains uncaptured, with isolate genomes representing 18.94% of PD and MAGs contributing the majority 62.74%. This suggests that the genomic landscape of *Bacillota* may be nearing saturation in the now sampled environments. This may partially reflect the relative ease of recovering *Bacillota* (*Firmicutes*) MAGs due to their smaller genome sizes compared to other lineages (*23*) as well as focused efforts such as the Human Microbiome Project (HMP) (*24*) and other genome recovery attempts from the human gastrointestinal tract, where *Bacillota* predominate (*4*). However, there is a small possibility that *Bacillota* estimates could be affected by sequencing bias against low GC sequences as previously noted (*25*). *Bacteroidota* display the lowest PD contribution by isolates (11.15%) among these phyla. While the reason for this observation is not entirely clear, one possible explanation may be more challenging growth requirements for members of *Bacteroidota* compared to other phyla.

Taxonomy of the uncaptured taxa

We assessed the taxonomic breadth of the uncaptured portion of the bacterial and archaeal PD to determine whether metagenome-only marker genes (not present in a MAG) represent as yet undiscovered phyla or distant taxonomic relatives of extant genomes. A RED analysis was performed (*17*), and various cutoffs were applied to group markers into OTUs to achieve a provisional hierarchical classification by comparing against GTDB-based taxonomy of reference genomes. Examining mOTUs arising from 18,087 samples, a total of 134,858 "species" (127,766 bacterial and 7200 archaeal) were uncaptured in genomic data. This amounts to 53.22% of the 253,400 total species-level OTUs (240,059 bacterial and 13,341 archaeal) generated by the entire dataset (Fig. 3A). An important caveat is that these figures are influenced by sample complexity and the taxa captured during the sampling, sequencing and assembly processes. Taxa may be underrepresented or entirely excluded from the analysis for a variety of reasons (discussed below).

Most of the added diversity represented by mOTUs is clearly at the species level (Fig. 3A and fig. S2B) and predominantly assigned to existing taxa within the phyla *Pseudomonadota*, *Acidobacteriota*, *Actinomycetota*, *Patescibacteria*, *Chloroflexota*, followed by *Planctomycetota*, *Bacteroidota*, *Myxococcota*, *Verrucomicrobiota*, and others. At the genus level, 18,942 mOTUs (18,276 bacterial and 666 archaeal) represent about 36.24% of the total OTUs. However, these proportions decrease at higher taxonomic ranks, e.g., only 1599 family-level mOTUs (1490 bacteria and 109 archaea), representing 18.88% of total, were discovered. At the phylum level, only six bacterial and two archaeal mOTUs (with a requirement of a minimum of three genes) were discovered, predominantly originating from various marine or freshwater samples. The only exceptions are an archaeal mOTU arising from hydrothermal vent samples and a bacterial mOTU from soil or termite-gut samples.

Most of the archaeal species-level OTUs are classified into the phyla *Thermoproteota*, *Thermoplasmatota*, *Halobacteriota*, *Nanoarchaeota*, and others. We focused on well-established sub-clades of archaeal methanogens (such as *Methanobacteriales*, *Methanococcales*, *Methanomicrobiales*, *Methanosarcinales*, *Methanocellales*, and *Methanopyrales*) to perform a thorough census of these important taxonomic groups. Discovering and characterizing previously unknown methanogenic taxa from various ecosystems is critical to refining our understanding of their ecological roles and methane production processes. This is increasingly important as methane emissions from both natural and human-driven sources are major contributors to global warming, and addressing methane emissions could help mitigate climate change (*26*). Moreover, methanogens are increasingly being exploited for their ability to degrade organic matter in wastewater treatment processes and for capturing emitted methane as a fuel source. Our analysis revealed that few mOTUs related to these clades were identified (at least one MQ MAG was

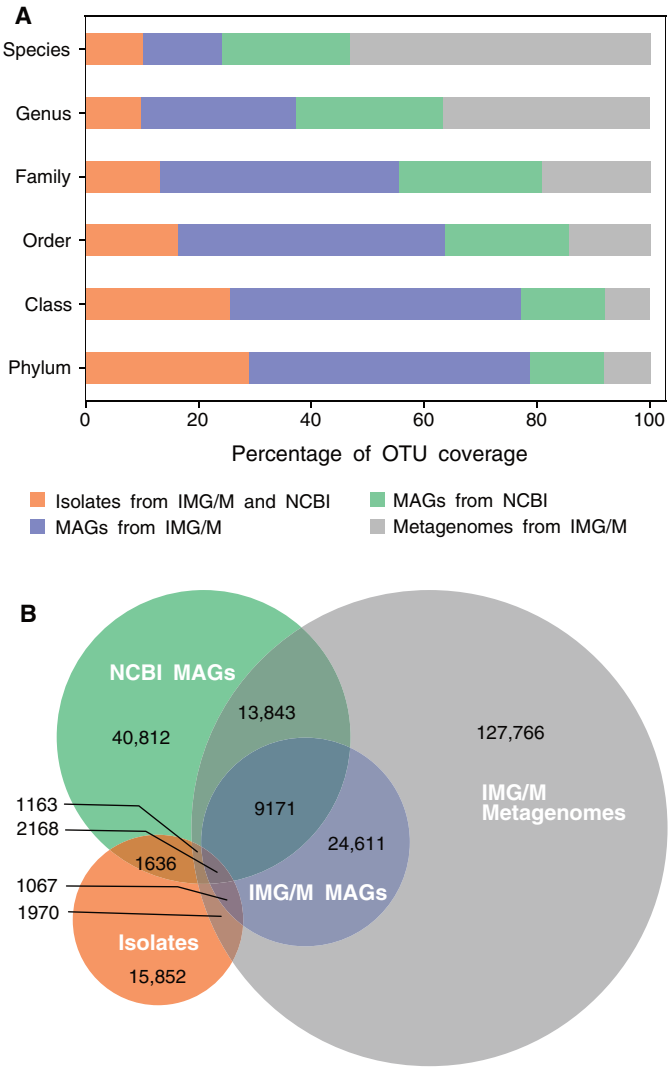

Fig. 3. Distribution of bacterial operational taxonomic units (OTUs) based on alanyl-tRNA synthetase (COG0013) gene sequences. (**A**) Proportion of total OTUs contributed at each taxonomic level by genome categories. (**B**) Venn diagram of shared and unique species-level OTUs obtained from various genome categories.

available in the vast majority of cases), suggesting that these clades are already well represented by genomes, particularly MAGs (*27*–*29*). A few examples of unrepresented methanogenic subclades include those related to isolated *Methanobrevibacter* spp. from termite gut and rumen samples or a potential unknown species of *Methanobacterium* from coal/oil affected environments (fig. S3). Many other large clades of metagenomic sequences with only a single MQ MAG are worthy of in-depth exploration, particularly those latent in permafrost or polar samples (*30*).

Isolate genomes "in the wild"

Examining species-level OTUs composed of sequences from both isolates and metagenomes (including MAGs), we can infer that about one-third of sequenced isolates are constituent members of an environmental sample, potentially playing key roles in these biomes (Fig. 3B). Specifically, of the 24,672 isolate-containing OTUs (23,856 bacterial and 816 archaeal), 8237 contain at least one metagenomic sequence. Of these, 8004 are bacterial OTUs encompassing 1,169,854 genomes and 233 are archaeal OTUs encompassing 1045 genomes. Conversely, 16,435 isolate-containing OTUs (66.61% of total) lack any type of metagenomic sequence. Of those, 15,852 are bacterial encompassing 234,117 genomes and 583 are archaeal encompassing 1380 genomes. This indicates that most of the isolated species are undetected in environmental samples. Some of these undetected isolates could be part of the rare biosphere of the sampled environments or occupy niches that are under-sampled or unsequenced such as those of obligate symbionts or pathogens or organisms that are sequestered in specialized niches (e.g., endoliths) or sessile or adherent organisms (e.g., biofilms and mucilage) that traditional sampling approaches might miss. For example, many model intracellular pathogens are recognized in this set such as *Bartonella* spp., *Salmonella* spp., *Yersinia* spp., *Blattabacterium* spp., *Brucella* spp., *Chlamydia* spp., *Rickettsia* spp., and *Helicobacter* spp., as well as select agents like *Coxiella burnetii*, *Francisella tularensis*, and *Rickettsia prowazekii*. Other instances include *Demequina* spp., *Microbulbifer* spp., *Xenorhabdus* spp., and *Nesterenkonia* spp., typically isolated from invertebrate hosts or high-salt environments that tend to be under-sampled or possibly under-sequenced (especially high-complexity soil communities).

Habitat of the uncaptured taxa

mOTUs are considered unique because no isolates or MAGs have been identified in these groups. These mOTUs arise from various environmental samples at different taxonomic levels (table S2). mOTUs (genus level and higher ranks) are primarily found in freshwater, marine, and terrestrial environments (Fig. 4 and tables S2 and S3). Individual metagenome samples with a preponderance of mOTUs are predominantly from the marine subsurface and sediment. Notable sources include various hydrothermal vents or oxygen minimum zones, as well as soil and termite gut. These environments are proposed as priority targets for in silico and other recovery efforts. The "most-wanted" taxa from these environments are summarized in Fig. 4B. In terrestrial environments, genus-level mOTUs within the *Acidobacteriota* and *Actinomycetota* are predominantly uncaptured, while, in aquatic samples, *Patescibacteria*, *Pseudomonadota*, and *Chloroflexota* are. By contrast, very few mOTUs are identified in the subset of human samples in this study (Fig. 4A), possibly owing to concentrated efforts over the past decades to sample and recover these genomes, as well as their relative lower species richness (*4*, *31*, *32*). However, we acknowledge that these samples are biased toward Western populations and diets, and sequencing microbiomes of more diverse populations could provide critical insights into human health and wellbeing.

Other host-associated environments may be similarly tractable for genome recovery due to lower complexity (species richness) compared to other environmental samples. As an example, we assessed OTUs from rumen samples and observed a high number of species-level mOTUs per sample (Fig. 5A). We chose to examine the rumen environment that was subject to extensive cultivation and sequencing efforts similar to the HMP (*26*). Recomputing the bacterial PD plot using only tree branches or leaves containing at least one rumen isolate or rumen metagenome sequence, we observed that isolates account for less than a quarter of total PD (17.19%). While MAGs represent a larger portion (37.91%), this still leaves 44.90% of total PD in rumen samples without genomic

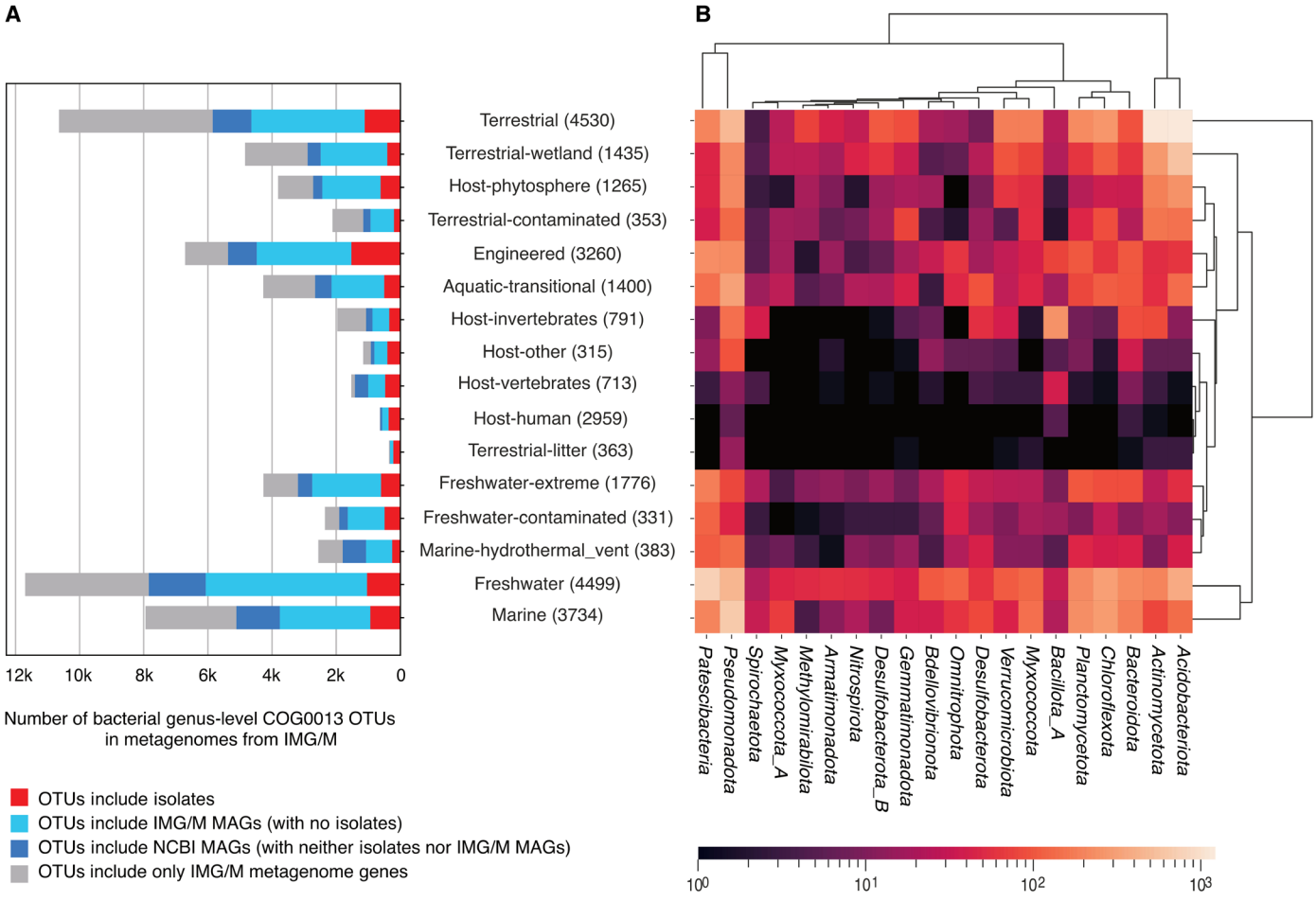

Fig. 4. Habitat distribution and taxonomy of genus-level metagenome-only OTUs (mOTUs) for COG0013. (**A**) Numbers of mOTUs by habitat: Relative contribution of mOTUs (gray) can be compared with other OTUs including isolates (red) or metagenome assembled genomes (MAGs) (blues) in each habitat. Total numbers of samples in each habitat are shown in parentheses. (**B**) Number of genus-level mOTU in each habitat for each phylum. Only the phylum with ≥150 genus-level mOTU are included in the heatmap (table S3). The colors represent the log scale of the number of mOTUs for each phylum in each habitat, as indicated by the color bar. Habitats and phyla are clustered by average linkage clustering, based on the Euclidean distances of genus-level mOTU counts.

representation (Fig. 5B). Archaeal PD is very low in this environment as expected (total PD is only 16) (*26*, *33*) and well covered by genome representatives, especially MAGs (47.32%) and isolates (36.97%), leaving only 15.72% of the PD unrepresented. Examining mOTUs containing sequences of rumen origin suggests that there are no new phyla or classes therein (with the same limitations of sampling as stated above for human samples). Instead, these mOTUs primarily represent undiscovered species and genera belonging to extant lineages.

While targeted isolation and sequencing efforts like the Hungate 1000 project (*27*) have made obvious contributions to our understanding of the rumen microbiome, more efforts are clearly needed (*34*, *35*). A saturated genome collection will be key in addressing food security and global warming challenges because ruminants are among the leading emitters of anthropogenic methane (*35*). We also examined the overall taxonomic differences between bacterial isolates and the uncultivated genomes (MAGs) to highlight phyla with few to no isolate genomes like *Spirochaeta*, *Verrucomicrobiota*, *Patescibacteria*, or *Elusimicrobiota* (Fig. 5C). Members of *Spirochaeta*, *Verrucomicrobiota*, and *Planctomycetota* are most likely to encode functions important to various aspects of rumen function. The need for cultivation of previously uncultivated taxa within the *Bacteroidota* and *Bacillota* (*Firmicutes*) that dominate the rumen microbiome was also previously recognized in a 16*S* rRNA-based census (*36*). Reconstructed metabolic pathways from genomes of uncultivated lineages could potentially inform culture criteria in the pursuit of gaining a saturated isolate reference collection for the rumen environment.

Last, we examined the environmental distribution of all the isolate-containing species-level OTUs to determine whether there are any potentially cosmopolitan species. While microbial species tend to be niche-adapted and consequently restricted to specific environments, multi-specialists with wider niche breadth are also probable (*37*). Excluding known contaminants such as *Stenotrophomonas maltophila*, *Ralstonia pickettii*, or *Cutibacterium acnes* (*38*), a handful of species were detected in a variety of environments. For example, *Marinobacter* sp. DSM 26671 (*39*) was detected in a range of aquatic and terrestrial samples, and *Priestia megaterium* DSM 319 (*40*) was found in diverse terrestrial and rhizosphere settings. These findings suggest that just a few species have remarkable adaptability,

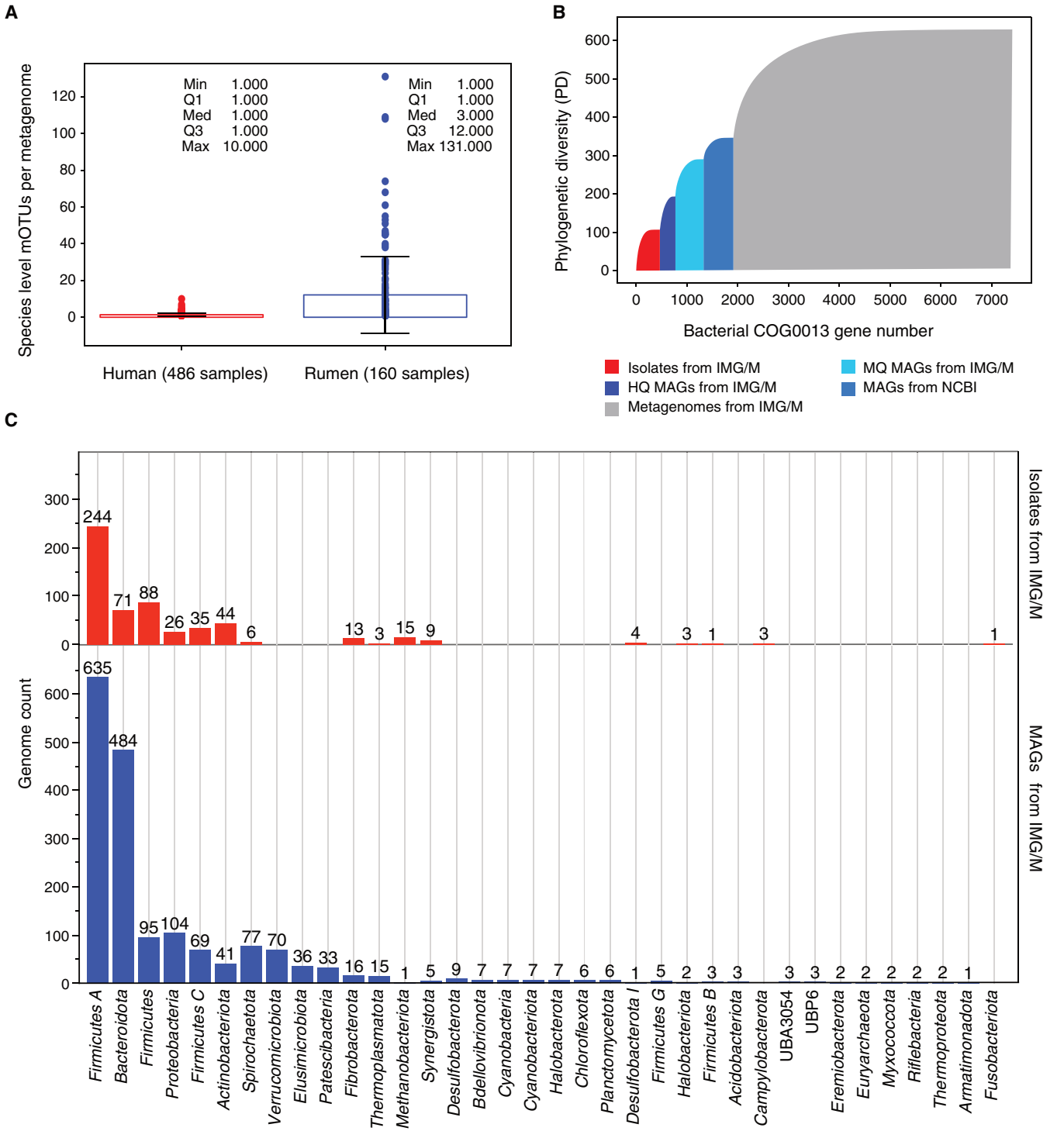

Fig. 5. Gaps in data for the rumen microbiome. (**A**) Comparing per sample metagenome-only OTUs (mOTUs) arising from the human large intestine versus the rumen environment. (**B**) Bacterial phylogenetic diversity (PD) accumulation curve for rumen sequences (based on COG0013 tree). (**C**) Counts of available isolate genomes versus uncultivated metagenome-assembled genomes (MAGs) for rumen taxa.

enabling them to inhabit a wide range of ecological niches, a property that could render them as great candidate model organisms for microbiology, and with the added advantage of proven genetic tractability (*41*, *42*).

DISCUSSION

This study reveals that, after three decades of microbial genome sequencing, cultivated taxa account for a rather small portion of the overall microbial diversity estimated from metagenomic sequences, contributing 9.73% in bacteria and 6.55% in archaea. While they remain the gold-standard resource in the laboratory supporting an array of experiments, including the development of microbial model systems and analyses of biotechnologically relevant pathways, their absence from metagenomic samples (two-thirds of isolates undetected) is noteworthy. This may reflect a historic bias in cultivation toward pathogens or organisms with desired metabolic capabilities, which may not accurately reflect "the natural world" and resulting in the isolation of taxa that are not numerically abundant in natural ecosystems, i.e., part of the "rare biosphere" (falling below the limits of detection at current sequencing depths). Others may be sequestered in specialized environments that are under-sampled or inaccessible by routine sampling approaches.

MAGs represent a much larger proportion of the total diversity, contributing 48.54% in bacteria and 57.05% in archaea, and, while they have yielded many fresh insights into ecology and evolution, a widely perceived problem is that they are characterized by relatively low quality (see above). By our conservative accounting, an outstanding fraction of PD (41.73% in bacteria and 36.39% in archaea) remains without any genomic representation, primarily at lower taxonomic ranks. Note that these figures do not fully represent the extensive PD present in global microbiomes as estimated by other methods (*43*) but rather reflect only what has been captured in available metagenomic datasets. Furthermore, these metagenome assemblies provide a mere snapshot of the most abundant and active microorganisms in a given sample and do not (and cannot) capture global diversity. Some of the less-abundant or unsampled members may well represent important keystone species within the ecosystem (*44*).

These results highlight the need for more targeted efforts to capture the full range of microbial diversity using a multifaceted approach involving renewed cultivation, deeper sequencing, and advanced genome reconstruction techniques. For the latter, redoubling sequencing of select samples leveraging recent innovations in long-read sequencing (*45*–*47*), co-assemblies (*48*–*50*), and improved binning strategies (*51*–*54*) could help. However, their success will undoubtedly vary by environment influenced by determining factors like species richness and diversity, genome size, and the complexity of resident genomes (e.g., large repetitive structures). For high-complexity samples like soil where metagenomes are often massively under-sequenced, Hi-C sequencing, co-assembly, and improved binning could help narrow the gap, but it is likely that less abundant components or members of the rare biosphere may remain elusive.

Recovery efforts should obviously include renewed isolation approaches because the ultimate goal is to recover cultures, thereby addressing the limitations of MAGs (*20*, *55*). With notable exceptions like Hungate1000 (*26*) or the mouse gut microbial biobank (*56*), culture-based efforts have been scarce or narrowly focused on specific metabolic guilds (like methanogens and lignocellulose-degrading bacteria). Therefore, we hope the "hotspots" highlighted in this manuscript reinvigorate cultivation efforts on a global scale, including extensive sampling and the capture of the full range of microbial diversity via renewed emphasis on classical microbiology and physiology, in addition to leveraging modern approaches such as subtractive enrichment of key guilds and/or taxa, co-culture, and diffusion chambers (*57*, *58*). Insights gained from -omics data can also help inform cultivation efforts as previously demonstrated (*58*–*63*).

We also propose that uncaptured taxa include some that are so not due to culturing difficulties but rather due to limited effort or access to samples. Therefore, even traditional culturing approaches may yield gains when applied to such samples (*10*, *64*). While it might not be feasible to recover all bacterial species, especially those from highly specialized environments with fastidious growth conditions or those with unknown dependencies, establishing a comprehensive and phylogenetically diverse collection of microbial cultures spanning various habitats, ecosystems, and geographical regions remains essential (*65*–*67*). Targeting species from specific environmental samples such as those with a preponderance of uncaptured taxa (as identified in this study) or with potential for biotechnological applications like high temperature, ionic liquids, hypersaline, or acidic environments is recommended. Prioritization based on diversity, phylogenetic distance from extant isolates, and ecological relevance can guide these efforts, but collaborative initiatives with pooled resources and expertise from varied researchers and institutions can play a pivotal role.

METHODS

Delineating datasets for analysis and marker selection

A total of 50,286 HQ bacterial and 705 archaeal isolate genomes (checkM completeness ≥ 95% and contamination < 5%) from IMG were used to assess universally distributed phylogenetic marker genes identified using COGs (Clusters of Orthologous Groups) and Pfams. Alignments of proteins from these reference genomes were constructed using hmmsearch (option: -cut_ga -A) from the hmmer3.3.2 package with related hmm profiles. To qualify as reliable phylogenetic markers, each COG needed to contain single-copy genes in at least 97% of the reference isolate genomes, and each gene was required to have non-gapped positions accounting for at least 60% of the total minimum alignment length of ≥200 amino acids. Using these thresholds, five COGs were identified as reliable phylogenetic marker genes to proceed with further analysis: COG0081, COG0533, COG0541, COG0013, and COG0086.

To obtain a comprehensive census, we queried datasets from both the Integrated Microbial Genomes and Microbiomes (IMG/M) database (*68*) and NCBI on 4 August 2023 (Fig. 1). IMG/M includes more than 50,000 metagenomes from a wide range of environmental samples including freshwater, marine, various terrestrial (e.g., forest, desert, cave, and coal bed), and host associated (e.g., human, insects, mammals, and plants).

For NCBI bacterial and archaeal isolates and MAGs, the following links were accessed: https://ftp.ncbi.nlm.nih.gov/genomes/genbank/archaea/assembly_summary.txt for archaea and https://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/assembly_summary.txt for bacteria.

NCBI genomes of isolates marked with comments such as "fragmented assembly," "refseq annotation failure," "contaminated," or "many frameshifted proteins" were excluded from the study. The NCBI MAGs were designated as "derived from metagenome" in the

summary files. To screen out mitochondria and plastid genes, the mitochondrial and plastid genomes were downloaded as reference sequences from the respective ftp links provided by NCBI: ftp://ftp.ncbi.nih.gov/genomes/refseq/mitochondrion and ftp://ftp.ncbi.nih.gov/genomes/refseq/plastid.

For NCBI isolates, mitochondrial and plastid genomes, and MAGs, gene sequences were predicted by prodigal with default settings (*69*). Hmm profiles of the selected phylogenetic markers (Table 1) were searched against protein sequences of genes from these datasets using hmmsearch (option: -cut_ga -A) from the hmmer 3.2.2 package (*70*). For IMG datasets, precomputed annotations were used. During hmm searches, alignments were constructed, and alignments from the same genome were concatenated in the case of COG0013 and COG0086 because they are each composed of multiple pfam hmm profiles (Table 1) (*71*). Only genes that had at least 60% non-gapped positions across the entire alignment length were included in the study, while fragmented genes corresponding to the same phylogenetic markers were excluded.

Tree building and PD analysis

Marker protein sequences from IMG metagenomes include those from HQ MAGs (checkM completeness ≥ 90% and contamination < 5%), MQ MAGs (checkM completeness ≥ 50% and contamination <10%), and metagenomic scaffolds not assigned to any MAGs. To eliminate redundancy, sequences of each phylogenetic marker were deduplicated using cdhit (option: -c 1.00 -aS 1.00) (*72*). From each cdhit cluster, a representative sequence was selected on the basis of the following order of preference: NCBI mitochondria, NCBI plastids, IMG isolates, NCBI isolates, IMG HQ MAGs, IMG MQ MAGs, NCBI MAGs, and IMG metagenome scaffolds.

Sequences of each phylogenetic marker were aligned using hmmsearch (-A option), and maximum likelihood trees were inferred using FastTree2 with the WAG substitution module (*73*). Eukaryotic clades within each phylogenetic marker tree were identified and trimmed out. To achieve this, a subtree consisting of 50,000 representatives was extracted, and the eukaryotic clades (including nucleus, mitochondria, and plastidic sequences) were manually determined. These determinations were based on domain assignments obtained from IMG annotation of isolates and metagenomic scaffolds (*74*). Representative sequences from each eukaryotic clade within the subtree were used to identify their common ancestral nodes in the marker trees. All the leaves originating from these common ancestral nodes were considered as eukaryotic sequences and removed. Potential horizontal gene transfer clades across archaea and bacteria domains were also removed during this step. Last, the trees were rooted in the middle of the archaeal and bacterial clades, and each tree was split into two trees: one for bacteria and one for archaea. The same protocol for identifying eukaryotic clades was applied to identify and extract subtrees for the *Pseudomonadota*, *Bacillota*, *Bacteroidota*, and *Actinomycetota* phyla separately.

The PD contribution of different groups, such as IMG isolates, NCBI isolates, MAGs from IMG/M (MQ and HQ), NCBI MAGs, and IMG metagenomes, was calculated and plotted individually for the *Bacteria*, *Archaea*, *Pseudomonadota*, *Bacillota*, and *Actinomycetota* trees. These calculations follow the methodology described by Wu *et al.* (*2*). The original code used for these analyses can be accessed at https://doi.org/10.5281/zenodo.7058177. NCBI MAGs are maintained as a distinct category because we were unable to ascertain quality and sample source for these genomes unlike IMG-derived MAGs.

RED analysis and OTUing

GTDB taxonomy (*75*) was assigned to both the isolates and MAGs included in the bacterial and archaeal trees. For IMG isolates and MAGs, the GTDB taxonomy was obtained from IMG/M (*68*). For NCBI isolates and MAGs, the GTDB taxonomy was either downloaded from https://data.gtdb.ecogenomic.org/releases/latest/ (*75*) or assigned using GTDB-Tk (*76*).

Tree-based OTUs were constructed for the rooted bacterial and archaeal trees. RED was computed for all the nodes in the tree (*17*). RED cutoffs ranging from 0.001 to 0.999 at intervals of 0.001 were used to group the leaves into OTUs based on the RED values of their common ancestor nodes. The leaves that received GDTB assignments from the tree were clustered into reference groups at the following taxonomic levels: phylum, class, order, family, and genus. Each reference cluster at a specific level was then compared with the OTUs generated using different RED cutoffs. To assess the similarity between the reference clusters and the OTUs, the AMI was calculated using the metrics.adjusted_mutual_info_score function from the scikit-learn library (version 1.2.0) in Python. The OTUs associated with the RED cutoff that produced the highest AMI value were selected as the OTUs representing the taxonomic level for the phylogenetic tree.

To compare species-level OTUs derived from two distinct phylogenetic markers, only single-copy genes sourced from isolates and MAGs were considered. Isolates and MAGs shared between both OTU datasets are grouped on the basis of the distribution of single-copy genes among the OTUs. Subsequently, the AMI is calculated to assess the dissimilarity between these two sets of species-level OTUs. A similar approach is used to compute AMI values for all phylogenetic marker pairs across various taxonomic levels, including species, genus, family, order, class, and phylum, for both bacteria and archaea (table S1). The distribution of AMI values is visualized in fig. S4.

GTDB taxonomy (*75*) was assigned to each OTU on the basis of the GTDB taxonomy assignments of IMG/NCBI isolates and MAGs within the same OTU, using a majority rule method. For OTUs that consist solely of metagenome genes (mOTUs), the GTDB assignments of related higher taxonomic level OTUs were adapted.

Supplementary Materials

The PDF file includes:

Figs. S1 to S4
Tables S1 to S4
Legends for tables S2 and S3

Other Supplementary Material for this manuscript includes the following:

Tables S2 and S3

REFERENCES AND NOTES

1. N. C. Kyrpides, Fifteen years of microbial genomics: Meeting the challenges and fulfilling the dream. *Nat. Biotechnol.* **27**, 627–632 (2009).
2. D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.-F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H.-P. Klenk, J. A. Eisen, A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).
3. S. Nayfach, S. Roux, R. Seshadri, D. Udwary, N. Varghese, F. Schulz, D. Wu, D. Paez-Espino, I.-M. Chen, M. Huntemann, K. Palaniappan, J. Ladau, S. Mukherjee, T. B. K. Reddy, T. Nielsen, E. Kirton, J. P. Faria, J. N. Edirisinghe, C. S. Henry, S. P. Jungbluth, D. Chivian, P. Dehal, E. M. Wood-Charlson, A. P. Arkin, S. G. Tringe, A. Visel, IMG/M Data Consortium, T. Woyke, N. J. Mouncey, N. N. Ivanova, N. C. Kyrpides, E. A. Eloe-Fadrosh, A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).

4. A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, R. D. Finn, A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
5. Y.-H. Chen, P.-W. Chiang, D. Y. Rogozin, A. G. Degermendzhy, H.-H. Chiu, S.-L. Tang, Salvaging high-quality genomes of microbial species from a meromictic lake using a hybrid sequencing approach. *Commun. Biol.* **4**, 996 (2021).
6. R. D. Stewart, M. D. Auffret, A. Warr, A. W. Walker, R. Roehe, M. Watson, Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
7. Y. Nishimura, S. Yoshizawa, The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Sci. Data* **9**, 305 (2022).
8. G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, J. F. Banfield, Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
9. C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, T. Woyke, Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
10. J. Schultz, F. Modolon, R. S. Peixoto, A. S. Rosado, Shedding light on the composition of extreme microbial dark matter: Alternative approaches for culturing extremophiles. *Front. Microbiol.* **14**, 1167718 (2023).
11. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
12. C. J. Castelle, J. F. Banfield, Major new microbial groups expand diversity and alter our understanding of the tree of Life. *Cell* **172**, 1181–1197 (2018).
13. R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. Murat Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, T. Woyke, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
14. D. P. Faith, Biodiversity and evolutionary history: Useful extensions of the PD phylogenetic diversity assessment framework. *Ann. N. Y. Acad. Sci.* **1289**, 69–89 (2013).
15. O. Bartoš, M. Chmel, I. Swierczková, The overlooked evolutionary dynamics of 16S rRNA revises its role as the "gold standard" for bacterial species identification. *Sci. Rep.* **14**, 9067 (2024).
16. J. M. Janda, S. L. Abbott, 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45**, 2761–2764 (2007).
17. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
18. J. T. Staley, A. Konopka, Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* **39**, 321–346 (1985).
19. A. Orakov, A. Fullam, L. P. Coelho, S. Khedkar, D. Szklarczyk, D. R. Mende, T. S. B. Schmidt, P. Bork, GUNC: Detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
20. A. Meziti, L. M. Rodriguez-R, J. K. Hatt, A. Peña-Gonzalez, K. Levy, K. T. Konstantinidis, The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: Insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Appl. Environ. Microbiol.* **87**, e02593-20 (2021).
21. J. C. Setubal, Metagenome-assembled genomes: Concepts, analogies, and challenges. *Biophys. Rev.* **13**, 905–909 (2021).
22. R. Seshadri, S. Roux, K. J. Huber, D. Wu, S. Yu, D. Udwary, L. Call, S. Nayfach, R. L. Hahnke, R. Pukall, J. R. White, N. J. Varghese, C. Webb, K. Palaniappan, L. C. Reimer, J. Sardà, J. Bertsch, S. Mukherjee, T. B. K. Reddy, P. P. Hajek, M. Huntemann, I.-M. A. Chen, A. Spunde, A. Clum, N. Shapiro, Z.-Y. Wu, Z. Zhao, Y. Zhou, L. Evtushenko, S. Thijs, V. Stevens, E. A. Eloe-Fadrosh, N. J. Mouncey, Y. Yoshikuni, W. B. Whitman, H.-P. Klenk, T. Woyke, M. Göker, N. C. Kyrpides, N. N. Ivanova, Expanding the genomic encyclopedia of *Actinobacteria* with 824 isolate reference genomes. *Cell Genom.* **2**, 100213 (2022).
23. S. Nayfach, K. S. Pollard, Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* **16**, 51 (2015).
24. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, J. I. Gordon, The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
25. P. D. Browne, T. K. Nielsen, W. Kot, A. Aggerholm, M. T. P. Gilbert, L. Puetz, M. Rasmussen, A. Zervas, L. H. Hansen, GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *Gigascience* **9**, giaa008 (2020).
26. R. Seshadri, S. C. Leahy, G. T. Attwood, K. H. Teh, S. C. Lambie, A. L. Cookson, E. A. Eloe-Fadrosh, G. A. Pavlopoulos, M. Hadjithomas, N. J. Varghese, D. Paez-Espino, Hungate 1000 project collaborators, R. Perry, G. Henderson, C. J. Creevey, N. Terrapon, P. Lapebie, E. Drula, V. Lombard, E. Rubin, N. C. Kyrpides, B. Henrissat, T. Woyke, N. N. Ivanova, W. J. Kelly, Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367 (2018).
27. B. Feehan, Q. Ran, V. Dorman, K. Rumback, S. Pogranichniy, K. Ward, R. Goodband, M. C. Niederwerder, S. T. M. Lee, Novel complete methanogenic pathways in longitudinal genomic study of monogastric age-associated archaea. *Anim. Microbiome* **5**, 35 (2023).
28. G. Borrel, P. S. Adam, L. J. McKay, L.-X. Chen, I. N. Sierra-García, C. M. K. Sieber, Q. Letourneur, A. Ghozlane, G. L. Andersen, W.-J. Li, S. J. Hallam, G. Muyzer, V. M. de Oliveira, W. P. Inskeep, J. F. Banfield, S. Gribaldo, Wide diversity of methane and short-chain alkane metabolisms in uncultured archaea. *Nat. Microbiol.* **4**, 603–613 (2019).
29. Z.-S. Hua, Y.-L. Wang, P. N. Evans, Y.-N. Qu, K. M. Goh, Y.-Z. Rao, Y.-L. Qi, Y.-X. Li, M.-J. Huang, J.-Y. Jiao, Y.-T. Chen, Y.-P. Mao, W.-S. Shu, W. Hozzein, B. P. Hedlund, G. W. Tyson, T. Zhang, W.-J. Li, Insights into the ecological roles and evolution of methyl-coenzyme M reductase-containing hot spring Archaea. *Nat. Commun.* **10**, 4574 (2019).
30. S. Wei, H. Cui, Y. Zhu, Z. Lu, S. Pang, S. Zhang, H. Dong, X. Su, Shifts of methanogenic communities in response to permafrost thaw results in rising methane emissions and soil property changes. *Extremophiles* **22**, 447–459 (2018).
31. L. M. Rodriguez-R, S. Gunturu, J. M. Tiedje, J. R. Cole, K. T. Konstantinidis, Nonpareil 3: Fast estimation of metagenomic coverage and sequence diversity. *Am. Soc. Microbiol. J.* **22**, mSystems 3:10.1128/msystems.00039-18.
32. The Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
33. K. A. Beauchemin, E. M. Ungerfeld, R. J. Eckard, M. Wang, Review: Fifty years of research on rumen methanogenesis: Lessons learned and future challenges for mitigation. *Animal* **14**, s2–s16 (2020).
34. E. A. Davidson, J. D. Semrau, N. K. Nguyen, On behalf of steering committee and participants of the ASM/AGU colloquium: The roles of microbes in mediating methane emissions, improved scientific knowledge of methanogenesis and methanotrophy needed to slow climate change during the next 30 years. *mBio* **14**, e02059-23 (2023).
35. *The Role of Microbes in Mediating Methane Emissions* (American Academy of Microbiology Colloquia Reports, American Society for Microbiology, Washington, DC, 2023); www.ncbi.nlm.nih.gov/books/NBK598985/.
36. C. J. Creevey, W. J. Kelly, G. Henderson, S. C. Leahy, Determining the culturability of the rumen bacterial microbiome. *Microb. Biotechnol.* **7**, 467–479 (2014).
37. F. Baquero, T. M. Coque, J. C. Galán, J. L. Martinez, The origin of niches and species in the bacterial world. *Front. Microbiol.* **12**, 657986 (2021).
38. S. J. Salter, M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. J. Loman, A. W. Walker, Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
39. L. C. Reimer, A. Lissin, I. Schober, J. F. Witte, A. Podstawka, B. Bunk, H. Lüken, J. Overmann, StrainInfo SI-ID 408098.1, version 1.0, Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures (2023); https://doi.org/10.60712/SI-ID408098.1.
40. L. C. Reimer, A. Lissin, I. Schober, J. F. Witte, A. Podstawka, B. Bunk, H. Lüken, J. Overmann, StrainInfo SI-ID 47142.1, version 1.0, Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures (2023); https://doi.org/10.60712/SI-ID47142.1.
41. R. Biedendieck, T. Knuuti, S. J. Moore, D. Jahn, The "beauty in the beast"—the multiple uses of Priestia megaterium in biotechnology. *Appl. Microbiol. Biotechnol.* **105**, 5719–5737 (2021).
42. L. J. Bird, Z. Wang, A. P. Malanoski, E. L. Onderko, B. J. Johnson, M. H. Moore, D. A. Phillips, B. J. Chu, J. F. Doyle, B. J. Eddie, S. M. Glaven, Development of a genetic system for *Marinobacter atlanticus* CP1 (*sp. nov.*), a wax ester producing strain isolated from an autotrophic biocathode. *Front. Microbiol.* **9**, 3176 (2018).
43. S. Louca, F. Mazel, M. Doebeli, L. W. Parfrey, A census-based estimate of Earth's bacterial and archaeal diversity. *PLOS Biol.* **17**, e3000106 (2019).
44. A. Jousset, C. Bienhold, A. Chatzinotas, L. Gallien, A. Gobet, V. Kurm, K. Küsel, M. C. Rillig, D. W. Rivett, J. F. Salles, M. G. A. van der Heijden, N. H. Youssef, X. Zhang, Z. Wei, W. H. G. Hol, Where less may be more: How the rare biosphere pulls ecosystems strings. *ISME J.* **11**, 853–862 (2017).
45. C. Y. Kim, J. Ma, I. Lee, HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nat. Commun.* **13**, 6367 (2022).
46. N. V. Patin, K. D. Goodwin, Long-read sequencing improves recovery of picoeukaryotic genomes and zooplankton marker genes from marine metagenomes. *mSystems* **7**, e00595-22 (2022).
47. L. H. Orellana, K. Krüger, C. Sidhu, R. Amann, Comparing genomes recovered from time-series metagenomes using long- and short-read sequencing technologies. *Microbiome* **11**, 105 (2023).

48. L. F. Delgado, A. F. Andersson, Evaluating metagenomic assembly approaches for biome-specific gene catalogues. *Microbiome* **10**, 72 (2022).
49. S. Hofmeyr, R. Egan, E. Georganas, A. C. Copeland, R. Riley, A. Clum, E. Eloe-Fadrosh, S. Roux, E. Goltsman, A. Buluç, D. Rokhsar, L. Oliker, K. Yelick, Terabase-scale metagenome coassembly with *MetaHipMer*. *Sci. Rep.* **10**, 10689 (2020).
50. R. Riley, R. M. Bowers, A. P. Camargo, A. Campbell, R. Egan, E. A. Eloe-Fadrosh, B. Foster, S. Hofmeyr, M. Huntemann, M. Kellom, J. A. Kimbrel, L. Oliker, K. Yelick, J. Pett-Ridge, A. Salamov, N. J. Varghese, A. Clum, Terabase-scale coassembly of a tropical soil microbiome. *Microbiol. Spectr.* **11**, e00200-23 (2023).
51. R. Lettich, R. Egan, R. Riley, Z. Wang, A. Tritt, L. Oliker, K. Yelick, A. Buluç, GenomeFace: A deep learning-based metagenome binner trained on 43,000 microbial genomes. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.02.07.579326.
52. Z. Wang, R. You, H. Han, W. Liu, F. Sun, S. Zhu, Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nat. Commun.* **15**, 585 (2024).
53. D. D. Kang, E. M. Rubin, Z. Wang, Reconstructing single genomes from complex microbial communities. *It. Inf. Technol.* **58**, 133–139 (2016).
54. S. Pan, C. Zhu, X.-M. Zhao, L. P. Coelho, A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* **13**, 2326 (2022).
55. W. C. Nelson, B. J. Tully, J. M. Mobberley, Biases in genome reconstruction from metagenomic data. *PeerJ* **8**, e10119 (2020).
56. C. Liu, N. Zhou, M.-X. Du, Y.-T. Sun, K. Wang, Y.-J. Wang, D.-H. Li, H.-Y. Yu, Y. Song, B.-B. Bai, Y. Xin, L. Wu, C.-Y. Jiang, J. Feng, H. Xiang, Y. Zhou, J. Ma, J. Wang, H.-W. Liu, S.-J. Liu, The mouse gut microbial biobank expands the coverage of cultured bacteria. *Nat. Commun.* **11**, 79 (2020).
57. S. R. Vartoukian, Cultivation strategies for growth of uncultivated bacteria. *J. Oral Biosci.* **58**, 142–149 (2016).
58. K. L. Cross, J. H. Campbell, M. Balachandran, A. G. Campbell, C. J. Cooper, A. Griffen, M. Heaton, S. Joshi, D. Klingeman, E. Leys, Z. Yang, J. M. Parks, M. Podar, Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat. Biotechnol.* **37**, 1314–1321 (2019).
59. H. J. Tripp, J. B. Kitner, M. S. Schwalbach, J. W. H. Dacey, L. J. Wilhelm, S. J. Giovannoni, SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452**, 741–744 (2008).
60. O. V. Karnachuk, A. P. Lukina, V. V. Kadnikov, V. A. Sherbakova, A. V. Beletsky, A. V. Mardanov, N. V. Ravin, Targeted isolation based on metagenome-assembled genomes reveals a phylogenetically distinct group of thermophilic spirochetes from deep biosphere. *Environ. Microbiol.* **23**, 3585–3598 (2021).
61. W.-D. Xian, N. Salam, M.-M. Li, E.-M. Zhou, Y.-R. Yin, Z.-T. Liu, Y.-Z. Ming, X.-T. Zhang, G. Wu, L. Liu, M. Xiao, H.-C. Jiang, W.-J. Li, Network-directed efficient isolation of previously uncultivated *Chloroflexi* and related bacteria in hot spring microbial mats. *NPJ Biofilms Microbiomes* **6**, 20 (2020).
62. A. Hanke, E. Hamann, R. Sharma, J. S. Geelhoed, T. Hargesheimer, B. Kraft, V. Meyer, S. Lenk, H. Osmers, R. Wu, K. Makinwa, R. L. Hettich, J. F. Banfield, H. E. Tegetmeyer, M. Strous, Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat. *Front. Microbiol.* **5**, 231 (2014).
63. S. Liu, C. D. Moon, N. Zheng, S. Huws, S. Zhao, J. Wang, Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome* **10**, 76 (2022).
64. H. Leng, Y. Wang, W. Zhao, S. M. Sievert, X. Xiao, Identification of a deep-branching thermophilic clade sheds light on early bacterial evolution. *Nat. Commun.* **14**, 4354 (2023).
65. N. C. Kyrpides, P. Hugenholtz, J. A. Eisen, T. Woyke, M. Göker, C. T. Parker, R. Amann, B. J. Beck, P. S. G. Chain, J. Chun, R. R. Colwell, A. Danchin, P. Dawyndt, T. Dedeurwaerdere, E. F. DeLong, J. C. Detter, P. D. Vos, T. J. Donohue, X.-Z. Dong, D. S. Ehrlich, C. Fraser, R. Gibbs, J. Gilbert, P. Gilna, F. O. Glöckner, J. K. Jansson, J. D. Keasling, R. Knight, D. Labeda, A. Lapidus, J.-S. Lee, W.-J. Li, J. Ma, V. Markowitz, E. R. B. Moore, M. Morrison, F. Meyer, K. E. Nelson, M. Ohkuma, C. A. Ouzounis, N. Pace, J. Parkhill, N. Qin, R. Rossello-Mora, J. Sikorski, D. Smith, M. Sogin, R. Stevens, U. Stingl, K. Suzuki, D. Taylor, J. M. Tiedje, B. Tindall, M. Wagner, G. Weinstock, J. Weissenbach, O. White, J. Wang, L. Zhang, Y.-G. Zhou, D. Field, W. B. Whitman, G. M. Garrity, H.-P. Klenk, Genomic encyclopedia of bacteria and archaea: Sequencing a myriad of type strains. *PLOS Biol.* **12**, e1001920 (2014).
66. E. J. Stewart, Growing Unculturable Bacteria. *J. Bacteriol.* **194**, 4151–4160 (2012).
67. G. Kapinusova, M. A. Lopez Marin, O. Uhlik, Reaching unreachables: Obstacles and successes of microbial cultivation and their reasons. *Front. Microbiol.* **14**, 1089630 (2023).
68. I.-M. A. Chen, K. Chu, K. Palaniappan, A. Ratner, J. Huang, M. Huntemann, P. Hajek, S. J. Ritter, C. Webb, D. Wu, N. J. Varghese, T. B. K. Reddy, S. Mukherjee, G. Ovchinnikova, M. Nolan, R. Seshadri, S. Roux, A. Visel, T. Woyke, E. A. Eloe-Fadrosh, N. C. Kyrpides, N. N. Ivanova, The IMG/M data management and analysis system v.7: Content updates and new features. *Nucleic Acids Res.* **51**, D723–D732 (2023).
69. D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
70. S. R. Eddy, Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
71. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
72. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
73. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010).
74. A. Clum, M. Huntemann, B. Bushnell, B. Foster, B. Foster, S. Roux, P. P. Hajek, N. Varghese, S. Mukherjee, T. B. K. Reddy, C. Daum, Y. Yoshinaga, R. O'Malley, R. Seshadri, N. C. Kyrpides, E. A. Eloe-Fadrosh, I.-M. A. Chen, A. Copeland, N. N. Ivanova, DOE JGI metagenome workflow. *mSystems* **6**, e00804 (2021).
75. D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, P. Hugenholtz, 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 (2022).
76. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk v2: Memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

Acknowledgments: We thank I.-M. Chen, K. Palaniappan, and the IMG and GOLD teams for providing access or staging data for easy retrieval. **Funding:** The work conducted by the US Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under contract no. DE-AC02-05CH11231. **Author contributions:** Conceptualization: N.C.K., N.N.I., D.W., and R.S. Methodology: D.W., N.N.I., and R.S. Software: D.W. Validation: D.W., R.S., and N.N.I. Formal analysis: D.W. and R.S. Investigation: R.S. and N.N.I. Resources: R.S. Data curation: R.S. and N.N.I. Project administration: N.C.K., N.N.I., and R.S. Writing—original draft: R.S. and D.W. Writing—review and editing: R.S., D.W., N.N.I., and N.C.K. Visualization: D.W. and R.S. Supervision: N.C.K. and N.N.I. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials, as well as from the IMG/M database (https://img.jgi.doe.gov). The list of the JGI Proposal Award DOIs is available in table S4.

Submitted 2 May 2024
Accepted 17 December 2024
Published 17 January 2025
10.1126/sciadv.adq2166