A census of hidden and discoverable microbial diversity beyond genome-centric approaches

Prasoodanan P.K., V[1], Maistrenko OM[2,3], Fullam A[4], Mende DR[5], Kartal E[6], Coelho LP[7], Spang A[2], Bork P[4,8,9] & Schmidt TSB[1,4] *

[1] APC Microbiome & School of Medicine, University College Cork, Ireland
[2] Department of Marine Microbiology & Biogeochemistry, Royal Netherlands Institute for Sea Research(NIOZ), The Netherlands
[3] *present address*: Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands
[4] Molecular Systems Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[5] Human Biology-Microbiome-Quantum Research Center (WPI-Bio2Q), Keio University, Tokyo 108-8345, Japan
[6] Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany
[7] Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, Woolloongabba, Queensland 4102, Australia
[8] Max Delbrück Centre for Molecular Medicine, Berlin, Germany
[9] Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

equal contributions
*correspondence: sebastian.schmidt@ucc.ie

Abstract

Cataloguing Earth's biodiversity remains one of biology's most formidable challenges, and the greatest diversity is expected to reside among the smallest organisms: microbes. Yet the ongoing census of microbial life is hampered by disparate sampling of Earth's habitats, challenges in isolating uncultivated organisms, limited resolution in taxonomic marker gene amplicons, and incomplete recovery of metagenome-assembled genomes (MAGs). Here, we quantified discoverable bacterial and archaeal diversity in a comprehensive, curated cross-habitat dataset of 92,187 metagenomes. Clustering 502M sequences of 130 marker genes, we detected 705k bacterial and 27k archaeal species-level clades, the vast majority of which was hidden among 'unbinned' contigs, beyond current genome-centric approaches. At deeper taxonomic levels, we estimate that 10 archaeal and 145 bacterial novel phyla and around 80k novel genera are discoverable in current data. We identified soils and aquatic environments as novel lineage recovery hotspots, yet predict that discovery will remain in full swing across habitats as more data accrues.

Introduction

Microbial life on Earth has deep evolutionary roots, and is ubiquitous and abundant in extant ecosystems: prokaryotes (that is, Bacteria and Archaea) likely emerged more than 4 billion years ago [1], have been detected in virtually every environment on the planet, and are estimated to account for ~10^{30} cells and 10^{16}-10^{18} g of biomass [2,3]. This long history of evolution among large numbers of individuals across a comprehensive ecological range is reflected in a great accumulated biodiversity, vastly exceeding that of multicellular life [4]. Yet it remains unclear just how diverse extant Bacteria and Archaea really are: estimates based on extrapolations from abundance distributions of individual samples vary by orders of magnitude, predicting anywhere between 10^6 and 10^{12} prokaryotic species worldwide [5–8].

Only a fraction of this predicted diversity has been accounted for in existing sequence data. Using 16S rRNA amplicon sequences, previous studies estimated the number of species-level Operational Taxonomic Units (OTUs) in publicly available datasets at 35.5k in 2004 [9], 210k in 2014 [8], 109k in 2016 [10] and 740k in 2019 [7]. Based on *rarefaction curves*, tracking the number of newly discovered types (i.e., species) as more data (i.e., samples) are added to the survey, these analyses concluded that at least in some microbial habitats, the rate of species discovery was slowing down (reflected in a 'flattening off' in the rarefaction curves). However, besides amplification-related bias (targeting different gene subregions, differential taxonomic coverage of primers, introduction of 'chimeric' artefacts, [11]), unclear definitions of similarity cutoffs for 'species-level' resolution [12] where different lineages may indeed share identical amplified subsequences, 16S rRNA amplicon data provides limited information as it focusses on just one taxonomic marker gene, without links to functional or genomic context in particular for uncultivated taxa.

Surveys of prokaryotic diversity therefore increasingly rely on isolate and metagenome-assembled genomes (MAGs) that continue to accrue in public repositories, defining clade-level diversity (e.g., species) based on sequence similarity in either a subset of near-universal taxonomic marker genes [12–14] or of entire genomes [15]. For example, proGenomes3 [16] encompasses 41k species-level clusters of high-quality isolate genomes from NCBI's RefSeq and GenBank in its most recent release (v3, 01-2023), whereas the Genome Taxonomy Database (GTDB) r220 (04-2024) delineates 113k species, combining isolate genomes and high-quality MAGs into a systematic reference taxonomy [17]. In addition, large-scale MAG catalogues generated from publicly available data continue to add to the census, reporting e.g. 4.6k species-level genomes in the human gut microbiome [18], 8.3k species in the ocean [19], 21k species in soils [20], between 0.1-13k species for the various habitats in MGnify's Genome collection [21], or 18k species integrated across the habitats represented in the Joint Genome Institute's IMG/M database [22]. We recently developed SPIRE [23] which provides an integrated survey of microbial diversity across 99k manually annotated metagenomes from all sampled habitats on Earth, describing 107k species-level clusters of which 92k are entirely based on MAGs (i.e., lacking a cultivated representative).

Such genome-centric surveys have greatly expanded the scope of study for microbial diversity, ecology and evolution, in particular for uncultivated taxa: in the GTDB r220, arguably the broadest available

reference for prokaryotic taxonomy, just 73 out of 194 phyla and 30.6k out of 113.1k species contain cultivated representatives. In other words, 62% of currently recognized phyla and 73% of species have been described solely based on MAGs. Nevertheless, MAG recovery remains limited in terms of both precision (due to residual binning artefacts [24]) and recall (most bins remain incomplete with current automated binning tools [25]). Indeed, with common workflows, most metagenomically assembled contigs remain *unbinned*: in SPIRE, just 2.36Tbp (or 10%) out of a total assembly length of 23.3Tbp end up in medium or high-quality MAGs. Arguably, these unbinned contigs circumscribe a space of 'discoverable' microbial diversity, i.e. the reservoir of recoverable genomes that remains untapped by current workflows, potentially comprising representatives of the so-called 'rare biosphere' [26]. For example, recent studies estimated this genomically unrepresented microbial diversity at around 135k species in 18k metagenomes [27] and 83k species in 249k metagenomes [28]. Indeed, novel tools emerge to make these genomes accessible: the recently developed Bin Chicken workflow of targeted co-assembly and co-binning recovered genomes for 24k previously unrepresented species from just 800 sample groups [29].

Here, we conducted a survey of discoverable bacterial and archaeal diversity among a comprehensive set of 92,187 manually annotated metagenomes, sampled across Earth's microbial habitats. We track the diversity represented in 120 bacterial [13] and 53 archaeal [14] marker genes originating from reference isolate genomes, MAGs and the large fraction of unbinned metagenomic contigs to quantify the number of novel lineages to be incrementally discovered in each set. Using habitat-stratified rarefactions, we tally the amount of novel genomic diversity that remains to be discovered in available public sequence data (and eludes state-of-the-art binning algorithms) and estimate the degree of novelty expected to emerge in different environments as new metagenomes are added to the survey. Finally, we track discoverable lineages that lack genomic representation across marker gene trees to estimate their phylogenetic depth and quantify how many 'deeper' novel clades (i.e., unrecognized phyla, classes, orders, families and genera) remain undiscovered across different habitats. In other words, we ask how many and which novel microbial lineages are 'hiding in plain sight' among unbinned contigs because they are missed by current toolkits, and how many more lineages we are poised to discover as we continue to sample Earth's habitats.

Results

Just 20-50% of currently metagenomically discoverable prokaryotic species are captured by genome-centric approaches.

We tracked the incremental discovery of bacterial and archaeal diversity among 92,187 metagenomic samples in SPIRE [23] along rarefaction curves using species-level clusters of assembled taxonomic marker genes (Figure 1A; see Methods). We estimated the total discoverable diversity in the studied dataset at ~705k bacterial and ~27k archaeal species (see table S1 and Fig 1B). Of these, just 20k bacterial and 0.8k archaeal species (corresponding to 2.8%) encompassed cultivated representatives in proGenomes3, an additional 25k and 1.7k species (for a total representation of 6.4% and 8.8%) mapped to the GTDB r220, and a further 80.5k and 4.2k species were contributed by SPIRE MAGs, for a total species-level diversity representation within genome-based datasets of 17.8% among bacteria and 24.6% among Archaea. In other words, up to ~75-80% of species that were discoverable among metagenomically assembled contigs were not in fact captured by genomes of cultivated organisms or MAGs.

Species count estimates were remarkably consistent across the different independently queried taxonomic marker genes (figure S1), driven by a large fraction of singleton clusters (i.e., clusters containing just one sequence). While a long-tailed cluster size distribution is consistent with expectations and previous observations on the so-called 'rare biosphere' [26], a subset of singleton clusters may arise from artifacts or spurious sequences. When conservatively considering only non-singleton marker gene clusters (i.e., containing ≥2 sequences from independent sources), diversity estimates were accordingly reduced to a total of 249k bacterial and 12.8k archaeal species; this can be considered a lower bound, yet still corresponded to an increase of 98% relative to the discoverable species set represented by genomes.

At the same time, around half of the species defined by (isolate) genomes in proGenomes3 and GTDB r220 were *not* represented in metagenomic assemblies (total reference dataset sizes indicated by dotted lines in figure 1A). Arguably, this is because those with cultivated representatives and those pinpointed for isolation are themselves often part of the rare biosphere, occurring at low prevalence and/or low abundance in the wild. Thus, the discovery gap goes both ways: isolate genomes and MAGs captured only 20-50% of discoverable species in metagenomic contigs, but metagenomic assemblies likewise failed to capture around half of the diversity represented by isolates.

Novel species discovery remains in full swing for most of Earth's microbial habitats.

The total number of discoverable species, as well as their relative representation among cultivated organisms and MAGs, varied greatly between individual habitats (Fig 1B), with moderate correlation to sampling effort ($\rho_{Spearman}$ = 0.51 for Bacteria and 0.09 for Archaea). In line with expectations, estimated archaeal species counts were generally 2-3 orders of magnitude below those for Bacteria in host-associated and anthropogenic habitats (including some environments with almost no detectable Archaea), roughly within 1 log in aquatic habitats and soils, and almost on par in extreme environments like hot springs and hydrothermal vents. The degree of genomic representation of discoverable species likewise varied across habitats, but was generally within one order of magnitude of the total.

We next tested whether novel species discovery showed any signs of slowing down by calculating species discovery coefficients (termed 'α') based on individual, habitat-stratified rarefaction curves (Fig 1C; see Methods). We observed that genome-based discovery of archaeal species was fully saturated ($\alpha \leq 0$) or approaching saturation (low positive α) in most host-associated and anthropogenic habitats. Specifically, all archaeal species represented in proGenomes3 and GTDB r220 that *can* be metagenomically discovered using current approaches, have already been accounted for in habitats like human infant gut, oral or skin samples; adding more samples to the survey is not expected to increase metagenomic representation for these reference species. Similarly, anthropogenic environments (built environment and wastewater), wetlands and extreme environments (hot springs and hydrothermal vents) appear saturated in terms of metagenomic discovery of 'known' isolate-derived archaeal species, whereas even MAG-based archaeal species discovery showed clear signs of slowing down (low positive α). In contrast, archaeal species discovery coefficients among unbinned contigs were significantly higher than for genome-centric data sources across most habitats, indicating that the existing large discovery gap between binned and unbinned contigs will continue to widen, as the accumulation of unbinned novel diversity shows only moderate signs of slowing down.

Bacterial species discovery coefficients were generally higher, indicating that with current approaches, novel bacterial diversity will continue to be discovered at higher rates than for Archaea. Interestingly, the human mouth, urogenital tract and adult gut, as well as porcine and bovine gut environments are the only tested habitats in which the metagenomic discovery of isolate-represented diversity is approaching saturation. Yet even in these habitats, as well as across all others, MAG-based species discovery is expected to continue at high rates, outpaced only by species discovery among unbinned contigs ($\alpha \leq 0.5$ across most habitats). Soils, wetlands, the rhizosphere, freshwater habitats and the GI tracts of non-mammalian, non-avian animals stood out as particular hotspots of untapped diversity, with virtually no signs of a slowdown in species discovery ($\alpha \geq 0.8$). Overall, the significant difference in species discovery coefficients between reference genome sets, MAGs and unbinned contigs suggests a widening gap between *discovered* (i.e., genomically represented) and *discoverable* (in unbinned contigs) diversity, as the latter will continue to outpace the former.

Soils and aquatic habitats remain major reservoirs of unexplored microbial diversity.

Beyond species discovery rates within individual habitats, we next explored how much discoverable diversity each habitat contributed to the total survey. We tracked habitat-stratified species accumulation curves – quantifying how much diversity each sequentially added habitat contributed to the survey *incrementally*, beyond what was already contained in other habitats – which revealed striking discrepancies between sampling effort and observed unique diversity (Fig 2). Although our dataset contained 53,949 gastrointestinal metagenomes (corresponding to 59% of the total), these accounted for just ~253k (or 36%) of discoverable species for Bacteria and ~2k (or 7%) for Archaea. This corresponds to an average of 4.7 newly added bacterial and 0.04 archaeal species per sample. Indeed, the human gut alone (across age groups, geography and disease states) accounted for 49% of total samples, but just 16% of discoverable bacterial and 0.8% of archaeal diversity. A further 12,098 metagenomes from non-intestinal human body sites and the built environment increased the tally to ~72% of total samples, but contributed only an additional 13k (~2% of the total, or 1.05 per sample) bacterial and 0.2k (<1%, or 0.02 per sample) archaeal species. In other words, while human-associated, gastrointestinal and built environment metagenomes represented almost three quarters of sampling effort, they only accounted for 38% of discoverable bacterial and 8% of archaeal diversity.

In contrast, plant-associated habitats represented just 1.5k samples (1.7% of the total), but accounted for 47k (6.7%) bacterial species discoverable beyond those contained in animal-associated and anthropogenic habitats, with the rhizosphere standing out as particular hotspot of bacterial diversity at 74.4 new species added per sample. Yet the largest steps in the collector's curve were observed for soils (~6% samples contributing 16% of unique bacterial and 9% of archaeal diversity, at 21.8 and 0.45 added species per sample) and aquatic habitats (~16% of samples contributing 32% of bacterial and 78% of archaeal diversity, at 15.1 and 1.4 novel species per sample). Wetlands, including e.g. salt marshes and peatlands, were further discovery hotspots for both Bacteria and Archaea (at 53.7 and 5.0 uniquely added species per sample, respectively), as were hydrothermal vents (21.5 and 6.2 species per sample). Although these and other comparatively undersampled habitats generally contributed more novelty per sample than well-represented environments (like the human gut), the overall correlation between sampling effort and newly discovered species was only moderately negative ($\rho_{Spearman}$ = -0.21 for Bacteria and -0.26 for Archaea).

Uncaptured deeply branching lineages substantially enrich the Tree of Life

We next explored how these trends in species discovery translated to deeper taxonomic levels. We inferred gene phylogenies for each of the considered 53 archaeal and 122 bacterial markers, estimated the number of represented genus-, family-, order-, class- and phylum-level clades by cutting each tree at level-specific relative evolutionary divergence (RED, [30]) cutoffs, and summarised the resulting clade counts across a filtered subset of phylogenies (see Methods). As shown in Figure 3 for archaeal marker DNA-directed RNA polymerase subunit E (*rpoE* [TIGR00448]) as a representative example, unbinned sequences added considerable tip-level (i.e., approximately species-level) diversity for most recognized clades, while often also providing relevant ecological context via the origin of underlying samples. For example, for the family *Nitrosopumilaceae* (within the Thermoproteota phylum or TACK superphylum [31]) that encompasses several known ammonia-oxidizing lineages, we observed a significant expansion of covered diversity among marine genera (e.g. *Nitrosopelagicus* or *Nitrosopumilus*), while linking other subclades to wetland, groundwater, soil and plant-associated habitats, supported across several phylogenies (beyond the rpoE tree in Figure 3).

Marker gene phylogenies recovered recognized deep-branching clades as mostly homogenous groups and recapitulated common hypotheses regarding the order of branches near their respective putative roots [32]. Our estimates of reference phylum-level clades (i.e., containing sequences from proGenomes3 or GTDB) exceeded the number of phyla recognized in the GTDB r220 based on full species trees for both Archaea (estimate: 36±12; GTDB r220: 19) and Bacteria (238±82; GTDB: 175), in part because our RED partitioning algorithm tended to split large clades like the TACK superphylum (represented as single phylum Thermoproteota in the GTDB) or groups predicted to have skewed evolutionary rates such as DPANN archaea into multiple 'phyla'. At the same time, all explored marker gene phylogenies contained varying numbers of deep (phylum-level) branches that were novel relative to both proGenomes3 and GTDB r220, i.e. encompassing only sequences in unbinned contigs or SPIRE MAGs. For example, the archaeal rpoE tree in Figure 3 contains, among others, a predicted novel phylum-level branch comprising 758 sequences from different habitats, organized into 109 tips (i.e., species-level gene clusters) and five predicted family-level clades, as a sibling group to phylum Thermoplasmatota. Overall, we estimated a median of 10 novel phylum-level clades in Archaea (4±3 from SPIRE MAGs, 6±8 among unbinned contigs), corresponding to a ~28% increase over estimated reference clades, and 145 novel 'phyla' in Bacteria (28±29 among MAGs, 117±137 in unbinned contigs), a ~61% increase over reference estimates. Towards shallower taxonomic levels, our estimates of reference clade counts increasingly approximated those in the GTDB r220, whereas estimates on discoverable clades among SPIRE MAGs and unbinned contigs strongly increased with increasing taxonomic resolution (see Figure 4A and Table 1).

The recovery of predicted clades was heterogeneous across habitats (Figure 4B). Gut samples, non-intestinal human sites, and the phyllosphere each covered only a small fraction of both reference and discoverable predicted archaeal and bacterial clades, at both deep (i.e., phylum-level) and shallower (genus-level) resolution. Thus, despite comprising a majority of the sequencing effort (see above), these habitats encompassed only a small fraction of known and discoverable diversity. In contrast, the majority of predicted phylum-level clades was observable in soils, rhizosphere, wastewater and aquatic

environments. Hydrothermal vents stood out as particular hotspots of archaeal diversity: although accounting for just 296 metagenomes in our dataset (0.3% of the total), they contained representatives of four fifths of reference and two thirds of MAG-based and unbinned predicted archaeal phylum-level clades, the widest coverage of deep lineages observed for any tested habitat. Most discoverable genera were predicted in aquatic habitats (marine, freshwater and groundwater) for Archaea and soils for Bacteria. In other words, soils and aquatic habitats remain the major reservoirs of untapped prokaryotic diversity, irrespective of prior sampling efforts.

Discussion

There are multiple layers of 'discovery' when it comes to the description of novel prokaryotic lineages: the isolation of organisms for experimental access and formal taxonomic approval remains standard [33–36], yet the delineation of genomes directly from metagenomic sequencing data has enabled the functional and evolutionary characterisation for a greatly extended range of clades without cultivated representatives, and even variants in (amplified) marker gene sequences can be ecologically informative and indicative of further diversity. In this context, our study addresses three basic questions: (i) how many prokaryotic lineages can be discovered in currently available data; (ii) how much of this diversity is missed by genome-centric approaches; and (iii) should we expect the discovery of novelty to slow down as more data comes in?

Our results extend previous observations [18,20,23,37] that isolates significantly underestimate tractable genomic diversity in metagenomes. Yet strikingly, this discovery gap goes both ways: only around half of the species clusters encompassed by genomes of cultivated representatives (and by extension, reference MAGs) in proGenomes3 and GTDB r220 were detected in our metagenomic dataset. These levels of isolate-derived species detection among metagenomes are higher than previous reports [27], likely because the present dataset is more comprehensive, although saturating discovery curves for several environments indicate that further detections are not expected when adding more samples (at least when assuming consistent bias in sampling, see below). This tallies with the idea that cultivation efforts and metagenomic sampling often target different (micro-)environments and that cultivability does not necessarily reflect abundance (or ecological importance) in natural communities [38], as has been observed for well-studied organisms such as *E. coli* [39] or *C. difficile* [40].

The vast majority of discoverable diversity in our dataset extended beyond *all* genome-centric approaches: for every species detectable among genomes of cultivated organisms (proGenomes3), reference MAGs (GTDB r220) or dataset-specific MAGs (SPIRE), between one and four additional species-level marker gene clusters were found among unbinned contigs. As a lower bound, defined via the independent detection of at least two closely related marker gene sequences in distinct samples, one species remains 'unbinned' for every species discovered using genome-centric approaches. Notably, even this conservative estimate exceeds recent reports of 83k [28] and 135k [27] novel species based on raw reads mapping to conserved marker gene sequence windows and metagenomic assembly, respectively. When considering all detected sequence variants, estimates go up to four discoverable species per genomically defined species. At deeper taxonomic levels, our data likewise suggests a wealth of discoverable clades in available data: we estimate that less than one in four bacterial genera, and just three in five bacterial phyla that can be inferred from our assemblies are currently recognized in reference databases – although these numbers carry increasing uncertainties at deeper taxonomic levels. This puts the total number of (currently discoverable) bacterial phyla in the range of 300-400, compared to 30-40 among Archaea, although we note that RED-based clade definitions at phylum and class levels were broader in Archaea than in Bacteria. Indeed, this ratio of roughly one order of magnitude difference between bacterial and archaeal clade counts held across all taxonomic levels,

although towards shallower levels (genus and species), bacterial diversity increased disproportionately. Therefore, our data provides the broadest support yet for the observation that Archaea are significantly less diverse than Bacteria [7,41,42]. This aligns with the hypothesis that the last archaeal common ancestor, recently estimated to be slightly younger than the bacterial common ancestor [43], was a hyperthermophile with the adaptation of archaea to moderate environments suggested to having been supported by gene import from Bacteria [44].

We note that although our methodology includes filtering and correction for various error sources, and although assembled sequences greatly underestimate the total sequence variation represented among raw metagenomic reads [45], the presence of assembly and clustering artefacts and spurious gene variants (including paralogs, or eukaryotic or viral orthologs) are possible sources of noise that may lead to an overestimation of diversity. Either way, the gap between genomically captured and unbinned discoverable diversity was growing (rather than closing) in virtually all tested habitats. And given that most metagenomically derived genes [45] and inferred gene families [46,47] are likewise unaccounted for by genomes, our results define a range of how many additional lineages may be genomically disentangled from the vast space of unbinned and unclassified genome reads in metagenomic data with improved (or targeted) algorithms. Indeed, constantly refined sampling and sample processing protocols, the increasing use of longread technologies [48,49], faster mapping tools that render multi-sample co-binning more computationally tractable [50,51], a novel generation of binning tools [52–54] and innovations like the iterative targeted co-assembly workflow in Bin Chicken [29] promise to make major inroads into this bulk of untapped diversity.

Our dataset reflects existing sampling bias both *between* habitats (human fecal samples dominate the survey) and *within* environments (uneven coverage geographically, along depth-elevation gradients and regarding host species) as individual datasets are usually generated for specific purposes, rather than as part of global surveys. Indeed, the habitat definitions used in the present work are – by design – pragmatic and guided by data availability, and not necessarily reflective of the underlying diversity of environmental conditions: for example, soils are structured into many heterogeneous sub-environments with physicochemical and biotic parameters that vary at micrometer scale [55,56], whereas well-mixed ocean layers can present homogeneous environments across several square kilometers [57]. Moreover, while we use metagenomic samples as units of reference, we note that established protocols capture very different scopes of underlying communities: while typical ocean water samples (filtered from 10-100L [57]) and human fecal samples (~1g of feces) carry total microbial loads of 10^10-10^11 cells [58], typical soil samples of 1-2g [59] contain 2-3 orders of magnitude fewer cells [2]. Given that Earth is expected to harbour a total of ~10^29 microbial cells each in ocean water and soils, but just 10^23-10^24 cells in the global human gut microbiome [2], this puts both the abscissa (i.e., the number of samples considered) and the ordinate (i.e., the number of clades discovered) in the presented rarefaction curves into further perspective. Finally, both physiological and technical factors may reduce (or bias) recall: among others, different membrane and cell wall types yield differentially to common extraction protocols [60], genomic composition (including GC content) and structure impact sequencing protocols [61], and gene calling algorithms often struggle with non-standard genetic codes [47].

Therefore, we caution against extrapolating from our rarefaction curves, as future data may break out from previous sampling bias by adding samples from uncharacterized sub-environments. This also means that our inferred rarefaction curves and novel species discovery coefficients should be considered conservative lower bound estimates, as they assume continued sampling from the same underlying distribution, while in fact the vast majority of the planet remains metagenomic *terra incognita* and overlooked endemic clades in unsampled sites are expected to disproportionately contribute to global biodiversity [62,63]. Despite these caveats, soils, aquatic habitats and non-mammalian guts clearly stood out among studied habitats as both key reservoirs of discoverable diversity in existing data and as hotspots primed for future discovery, and our data suggest that for almost all considered habitats, the discovery of novel diversity will remain in full swing for the foreseeable future.

The present work takes stock of characterized and discoverable microbial diversity based on a snapshot of available data. It is encouraging – and, from a data science perspective, slightly intimidating – that the underlying data continues to accrue near-exponentially and that the numbers reported above will be outdated within months of publication. But as we have shown, even existing data may hold countless surprises yet, as a vast amount of discoverable prokaryotic diversity is 'hiding in plain sight', flying under the radar of current genome-centric approaches. Most undiscovered diversity – both that within existing data and that not yet sampled – will likely be part of the 'rare biosphere', and thus cover a disproportionately large functional diversity (of potential medical and biotechnological relevance), but also be more vulnerable and at risk of extinction on a rapidly changing planet. Cataloguing and characterizing these organisms, with a focus on understudied environments, is therefore not only relevant to the study of microbial and ecology and evolution, but can be considered an integral aspect of global conservation efforts.

Methods

Data sources and preprocessing

We used three principal data sources in our analysis: (i) proGenomes v3 [16], a quality-filtered subset of 907,388 isolate genomes sourced from NCBI's RefSeq and GenBank databases [64] organised into 41,171 species-level clusters based on a set of 40 'specl' marker genes [12]; (ii) the Genome Taxonomy Database (GTDB) release 220 (April 2024, [17]), a curated set of 596,859 isolate and metagenome-assembled genomes (MAGs), clustered into 107,235 bacterial and 5,869 archaeal species-level clusters sharing ≥95% whole-genome Average Nucleotide Identity (ANI); and (iii) SPIRE v1.1 [23], encompassing a curated set of 92,187 shotgun metagenomic samples assembled into ~23.3Tbp of contigs, a subset of which were binned into 1,158,468 medium or high quality MAGs, organised into 107,078 species-level clusters that partly overlap with those in proGenomes3.

Genomes from all three datasets were downloaded and taxonomically (re-)classified against the GTDB r220 using GTDB-tk v2.4.0 [65], and consensus taxonomies for species-level clusters were inferred based on adjusted majority votes, as described previously [23]. Compared to SPIRE v1.0, 6,959 metagenomic samples were excluded for the present analysis based on data type and provenance (e.g., excluding samples explicitly enriched for viruses) or insufficient assembly size. For the remaining 92,187 samples, habitat annotations and contextual data were updated by manually curating against an extended *microntology* v0.3.0 encompassing 103 terms and categories (reference set and updated annotations available via https://spire.embl.de/downloads and table S2) and further organised into 32 higher-level categories (see table S1).

Estimates of bacterial and archaeal diversity based on taxonomic marker genes

The majority of analyses presented in the main text are based on 168 near-universal taxonomic marker genes as established by the GTDB: 120 bacterial ('bac120', [13]) and 53 archaeal ('arc53', [14]) markers, with an overlap of five genes. We downloaded profile Hidden Markov Models (HMMs) for these marker sets as part of the GTDB-tk v2.4.0 database and used the HMMER v3.4 [66] *hmmsearch* routine to identify and extract marker gene sequences among predicted Open Reading Frames (ORFs). For the final filtered set of marker genes (see below), this yielded on average 904,654±98,879 sequences per bacterial marker in proGenomes3, 93,198±4,733 among GTDB r220 species representatives, 840,023±133,904 in medium and high-quality SPIRE MAGs, and 3,051,465±1,290,024 among 'unbinned' SPIRE contigs, i.e. those not assigned to a MAG passing genome quality filters in SPIRE v1. For Archaea, we obtained 2,451±150 sequences from proGenomes3, 4,898±410 from GTDB r220 species representatives, 15,504±2,593 from SPIRE MAGs and 107,419±75,823 from unbinned SPIRE contigs (see table S3 for a per-gene overview).

For each marker gene, we clustered extracted sequences using the MMseqs2 cascading clustering routine [67] with a cutoff of 96.5% sequence similarity (see below) and otherwise default parameters. To convert the number of marker gene sequence clusters into a corresponding number of species-level clusters, we estimated conversion factors as follows. First, we generated marker gene cluster discovery curves via iterative logarithmic rarefaction, i.e. we downsampled the number of considered gene sequences along a logarithmic scale (10, 20, …, 100 sequences; 200, 300, …, 1000 sequences; etc) with 10 iterations at each step. At each rarefaction point, we recorded the number of 'discovered' marker gene sequence clusters and the number of represented species or species-level genome clusters in proGenomes3, GTDB r220 and among SPIRE MAGs, considering each data source individually. We then used linear regression models of the type *number_of_species ~ number_of_gene_clusters* with a forced intercept at 0 along these rarefactions to estimate gene cluster to species conversion factors (i.e., the number of newly discovered species per newly discovered marker gene cluster). Based on benchmarks of marker gene sequence similarity cutoffs ranging from 95% to 99.5%, we based further analyses on 96.5% clusters as these showed very robust linear fits (with standard errors in the range of 10^-3) with conversion factors closest to identity (i.e., roughly one species discovered per marker gene cluster discovered), and high consistency across the different species-level reference clusterings in the underlying datasets (based on 40 specl marker genes in proGenomes3, 95% whole-genome ANI in GTDB and a combination of both approaches among SPIRE MAGs). Finally, we filtered the set of considered marker genes by excluding redundant markers (part of both the arc53 and bac120 sets and therefore prone to cross-mappings) and by removing markers for which predicted species numbers deviated by more than 20% from the domain-level median (in particular for some archaeal markers, detected sequence counts were inflated, possibly due to the detection of bacterial paralogs). Sequence counts per data source for the remaining set of 28 archaeal and 102 bacterial marker genes are available in table S3.

Inference of global and habitat-resolved species discovery curves

For each of 32 broadly defined habitat categories (ranging from 267 to 19,659 samples per group, see table S1), as well as for the combined set of all 92,187 metagenomes in SPIRE v1.1, we generated bacterial and archaeal species discovery curves. We iteratively subsampled the number of considered metagenomes along a logarithmic rarefaction scale, with 5 random permutations per step. At each rarefaction point and for each considered marker gene, we inferred the number of discovered species based on the number of discovered marker gene clusters, using gene-specific conversion factors as described above. Rarefaction permutations were averaged per marker gene and then summarised per marker gene set (bac120 for Bacteria and arc53 for Archaea) as median predicted species counts at each rarefaction step across marker genes.

In an independent approach, we performed an 'incremental' rarefaction, stratified by habitats: we sequentially added samples from each of the 32 broadly defined habitat categories, starting with 'adult, healthy human gut' samples, randomized discovery order within each habitat block and then tracked the *globally* discovered species (see Figure 2). In other words, we tracked the contribution of each habitat to overall discovered species diversity, beyond the diversity discovered in previously considered habitats.

Counts of discovered marker gene clusters (and, by proxy, inferred discovered species) were hierarchically stratified by data source to account for overlap between datasets as follows: all clusters containing at least one sequence originating from a genome in proGenomes3 were labelled as 'proGenomes3', irrespective of the origin of other sequences within the same cluster (data series marked in dark blue in main figures 1 & 2); clusters containing sequences from the GTDB r220 (and any other data source *except* proGenomes3) were labelled as 'GTDB' (light blue in figures 1 & 2); clusters containing sequences from SPIRE MAGs (but neither from proGenomes3 nor GTDB) were labelled as 'SPIRE MAG' (orange data series); clusters containing only sequences from unbinned SPIRE contigs were labelled as 'unbinned, non-singleton' if they contained at least two sequences from different contigs (dark yellow data series), and 'unbinned, all' otherwise (light yellow data series). Thus, clusters labelled as 'GTDB' represent sequence diversity contained in GTDB r220 *beyond* that already represented in proGenomes3, while a significant subset of 'proGenomes3'-labelled clusters also contained sequences originating from GTDB, SPIRE MAGs or unbinned contigs. Similarly, 'SPIRE MAG' clusters corresponded to diversity *not* already covered by proGenomes3 and GTDB r220; and 'unbinned' clusters encompassed diversity not represented in any genome or MAG in the dataset. Moreover, by design we only tracked marker gene clusters in metagenomic samples, meaning that a non-negligible subset of genome-based clusters from proGenomes3 and GTDB r220 were not represented in rarefaction runs and calculations, as corresponding sequences were not detected in any of the 92,187 considered metagenomic assemblies (see main text).

Calculation of species discovery coefficients

We quantified the rate at which novel species-level clusters are discovered in each habitat (and globally, across all habitats) using the equation

$$= \quad * \quad ^{-} \quad (1)$$

where S is the number of newly 'discovered' species per N samples added to the survey; k is a proportionality constant; and γ is a saturation coefficient. An analogous formula is commonly used to describe a (microbial) pangenome's *openness* [68,69], i.e. the degree to which novel genes are discovered as more genomes from the same species are considered. Moreover, the approach is conceptually and mathematically related to the Heridan-Heaps law in linguistics which describes the number of distinct words in a document as a function of that document's length [70].

We fit equation (1) to each habitat-specific and global rarefaction curve described above, stratified by data source, resulting in estimates for γ for each considered marker gene. For more intuitive interpretability, we calculated a *species discovery coefficient* α as

$$\alpha = 1 - \quad (2)$$

and summarised values across marker genes within each domain-specific set as median values. Thus defined, α scales on [-∞, 1], although only mildly negative values are expected to be observed in practice. If α ≤ 0, species discovery in a habitat is *saturated*, meaning that adding more samples of the same type is not expected to add novel species to the survey – analogous to a 'closed' microbial pangenome where additional strains do not add novel genes. Values for α in [0, 1] correspond to *unsaturated* species discovery curves where additional samples continue to add novel species to the survey (analogous to 'open' pangenomes); lower α values indicate a more pronounced 'flattening off' in the rarefaction curve, indicating a more pronounced slowdown in novel species discovery; higher α values indicate a less pronounced decrease in the rate of species discovery. For α ≈ 1, species discovery is *fully unsaturated*, meaning that each newly added sample adds novel species to the survey, with no discernible 'flattening off' in the species discovery curve.

Phylogenetic inference and analyses

Phylogenetic trees for each marker gene were constructed as follows. We first re-aligned cluster representative amino acid sequences against the respective marker gene HMMs using HMMER v3.4's [66] *hmmalign* routine, trimmed the resulting pseudo-multiple sequence alignments to informative columns using clipkit [71] in 'kpi-gappy' mode and removed sequences with >70% gaps among the remaining columns. Based on the resulting alignments, we inferred phylogenetic trees using FastTree2 [72] under the Whelan-Goldman model [73].

We calculated *relative evolutionary divergence* (RED) values [30] for each node in the resulting trees using the castor package v1.8.3 in R [74]. To infer marker-gene specific relative evolutionary divergence (RED) cutoffs at phylum, class, order, family, and genus levels, we subset phylogenies to sequences originating from fully taxonomically classified genomes in proGenomes3, the GTDB r220 or among SPIRE MAGs, cut the trees at incremental RED values (with a RED tolerance of ±0.1) and compared the resulting partitions (i.e., tip sets) versus assigned taxonomic labels, quantifying partition consistency as *adjusted mutual information* (AMI) [75]. These steps were performed iteratively on trees re-rooted using each recognized phylum in turn as an outgroup and then summarised across alternative root positions (ignoring the current outgroup) to obtain average RED values per internal node, in a workflow adapted from [30].

After confirming that the resulting RED cutoffs at peak AMI (i.e., best capturing the distribution of established taxonomic labels) well approximated those reported in [30] across individual archaeal marker genes, we used GTDB's reference cutoffs (with a ±0.1 tolerance interval) for bacterial marker gene trees and cut the full trees at the respective levels. The number of resulting clusters (i.e., nodes where the tree was cut) and their composition (i.e., each node's descendant tips) were then used as estimates of clade-level groups at each taxonomic level, for each marker gene. We manually selected genes that showed the most consistent and robust profiles across iterations (and rootings) and whose phylogenies were not ostensibly impacted by paralogs, bringing the final sets of considered genes to 29 archaeal and 53 bacterial markers. Marker gene phylogenies were visualized using the ggtree and ggtreeExtra packages in R [76].

Acknowledgements

The authors thank Askarbek N Orakov (Harvard T.H. Chan School of Public Health) and Jens Walter (University College Cork) for helpful discussions and feedback on the manuscript. This research was conducted with the financial support of Research Ireland under Grant Number 12/RC/2273-P2 (V.P.P.K and T.S.B.S) and from the Australian Research Council (grant #FT230100724, to L.P.C). A.S. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947317, ASymbEL), which also supported the position of O.M. This work made use of EMBL IT Services HPC resources [77].

Data Availability

Metagenomic assemblies, Metagenome-Assembled Genomes (MAGs), gene calls and corresponding annotations are available via spire.embl.de. Inferred marker gene phylogenies with annotations, as well as pre-generated tree visualizations for archaeal markers are available via the EBI BioStudies repository under accessions S-BSST2111 and S-BSST2112.

Figure Legends

Figure 1. The discovery of novel prokaryotic diversity remains in full swing. (A) Rarefaction curves of archaeal (top) and bacterial (bottom) species discovery across ~90k metagenomes. Based on marker gene cluster composition, we distinguished five hierarchically overlapping categories of data provenance for every inferred species (indicated as Venn diagrams): (i) those including an isolate genome represented in proGenomes3 (dark blue data series); (ii) additional species (beyond proGenomes3) represented in GTDB r220 isolate genomes and MAGs (light blue); (iii) species defined from SPIRE MAGs (orange); and species-level marker gene clusters exclusively based on unbinned contigs (i.e., not assigned to medium or high-quality MAGs), further distinguishing (iv) non-singleton clusters (containing ≥2 sequences from independent sources; dark yellow) among (v) all marker gene clusters (light yellow). Reference levels are indicated as dashed lines, i.e. the total number of species in proGenomes3, GTDB r220 and SPIRE MAGs including those *not* discoverable in metagenomes. **(B)** Total number of species-equivalent marker gene clusters per habitat; see figures S2 & S3 for corresponding habitat-stratified rarefaction curves. **(C)** Species discovery coefficients 'α', calculated from habitat-stratified rarefaction curves (see Methods). Values for α in [0, 1] correspond to *unsaturated* species discovery curves where additional samples continue to add novel species to the survey (analogous to 'open' pangenomes); lower α values indicate a more pronounced 'flattening off' in the rarefaction curve, indicating a more pronounced slowdown in novel species discovery; higher α values indicate a less pronounced decrease in the rate of species discovery. For α ≈ 1, species discovery is *fully unsaturated*, meaning that each newly added sample adds novel species to the survey, with no discernible 'flattening off' in the species discovery curve.

Figure 2. Soils and aquatic habitats are major reservoirs of unexplored microbial diversity. 'Habitat-cumulative' rarefaction curves for Archaea (top) and Bacteria (bottom). Samples were added to the survey sequentially by habitat and newly discovered species were tracked *incrementally*, i.e. in addition to those contained in previously added habitats (see Methods). Horizontal dotted lines indicate reference levels, i.e. species counts including those not detectable in our metagenomic dataset.

Figure 3. Unbinned contigs flesh out the Tree of Life among both known and deep novel clades. Example phylogeny of the archaeal RNA polymerase E gene (rpoE, [TIGR00448]), inferred from pre-clustered reference and metagenomic sequences (see Methods). Sequence clusters at tips are annotated according to their source (proGenomes3, GTDB r220 or SPIRE) and the habitat categories from which they were recovered among the full set of metagenomes. Nodes marked with dots indicate clade-level groups at phylum (red) and family (blue) level, inferred based on pre-calibrated cutoffs in relative evolutionary divergence (RED; see Methods). Phylum groups were taxonomically classified based on encompassed sequences from proGenomes3, GTDB r220 or SPIRE MAGs. Nitrosopumilaceae are indicated as an example clade where unbinned contigs substantially enrich known taxa. Deep novel (unclassified) phylum-level clades are highlighted at various points.

573 **Figure 4. Unbinned contigs and MAGs suggest thousands of discoverable deeper clades. (A)** Numbers
574 of predicted clade-level groups at phylum, class, order, family and genus level, based on relative
575 evolutionary divergence (RED) cutoffs (see Methods). Each point indicates an estimate for one marker
576 gene. Colours indicate data source: blue for clades containing at least one reference sequence (from
577 proGenomes3 or GTDB r220); orange for those containing no reference sequences, but at least one
578 SPIRE MAG; yellow for clades containing only sequences from unbinned SPIRE contigs. Dotted blue lines
579 indicate the reference number of clades per level in GTDB r220. **(B)** Distribution of discoverable phylum-
580 and genus-level clades across habitats. Bars indicate how many clades for each data source are
581 discoverable across all habitats (leftmost column) or when only considering data from individual
582 habitats.

583 # Table Legends

584 **Table 1. Estimated clade counts at deeper taxonomic levels.** Clade counts were estimated based on
585 relative evolutionary divergence (RED) cutoffs from individual marker gene phylogenies (see Methods).
586 Values shown are the median ± standard deviation across domain-specific marker gene sets.

587

Supplementary Figure Legends

Figure S1. Species count estimates are remarkably consistent across different marker genes. Each dot represents the estimated number of species based on observed marker gene clusters (see Methods) for one marker gene, normalized by the median across marker genes within category. Archaeal outliers to the top were marker genes with putative cross-mapping to bacterial orthologs (leading to an overestimation of diversity). For analyses in the main text, only a subset of consistent marker genes, further filtered based on additional criteria (see Methods) was used.

Figure S2. Habitat-stratified rarefaction curves for Archaea. Species discovery curves as described in the main text and shown in Figure 1, but subset by individual habitats.

Figure S3. Habitat-stratified rarefaction curves for Bacteria. Species discovery curves as described in the main text and shown in Figure 1, but subset by individual habitats.

Supplementary Table Legends

Table S1. Definitions and data overview for microbial habitats considered in this study. The table contains a full list of the various habitat categories, defined based on *microntology* annotations of samples (see Methods). It further provides a data overview including the number of samples per habitat and estimated species counts (across marker genes) in different data sources.

Table S2. Sample metadata. *microntology* annotations and assigned habitat category for each sample considered in this study.

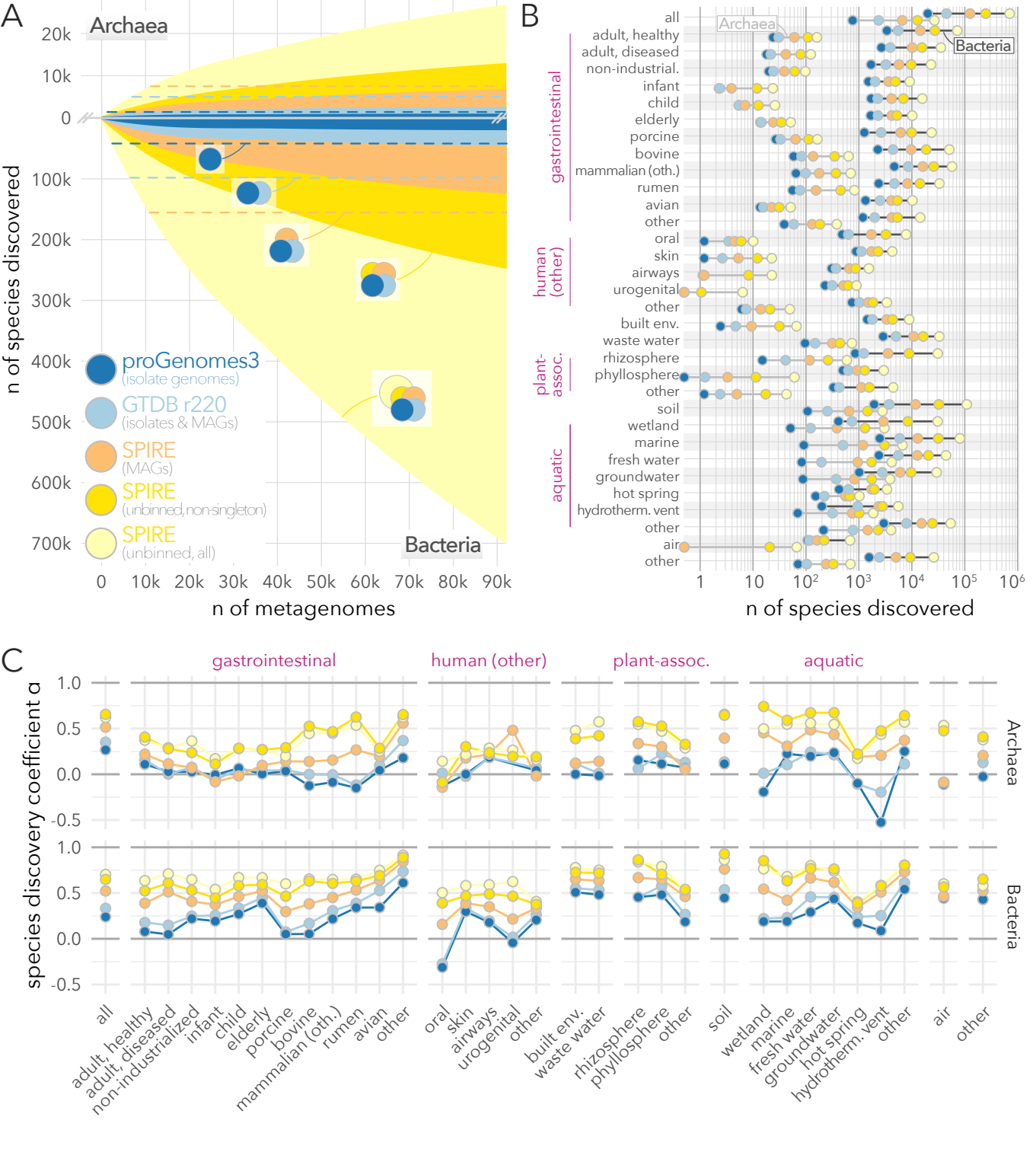
Table S3. Marker gene overview. Number of extracted sequences per data source for each marker gene considered in this study.

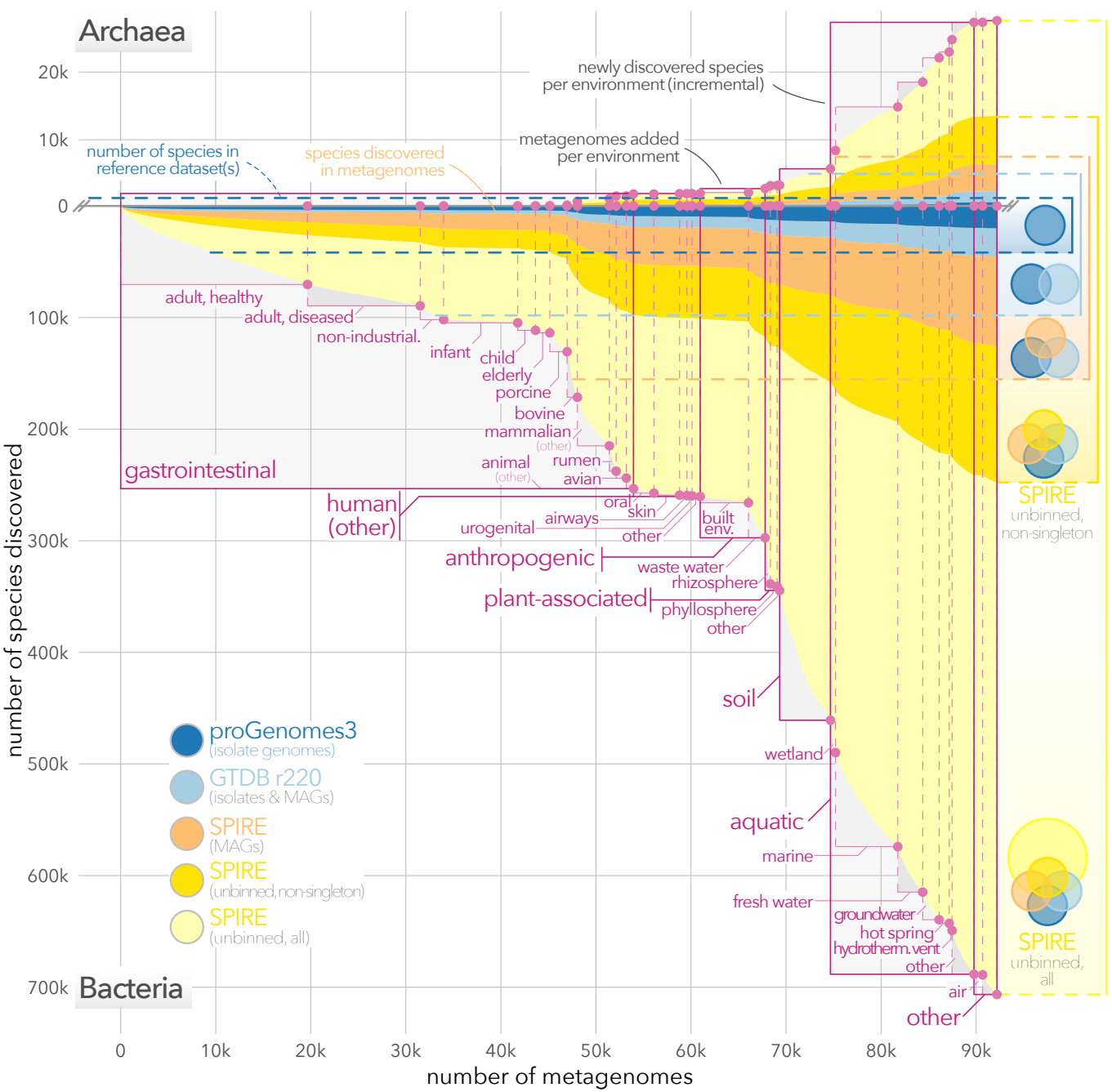
References

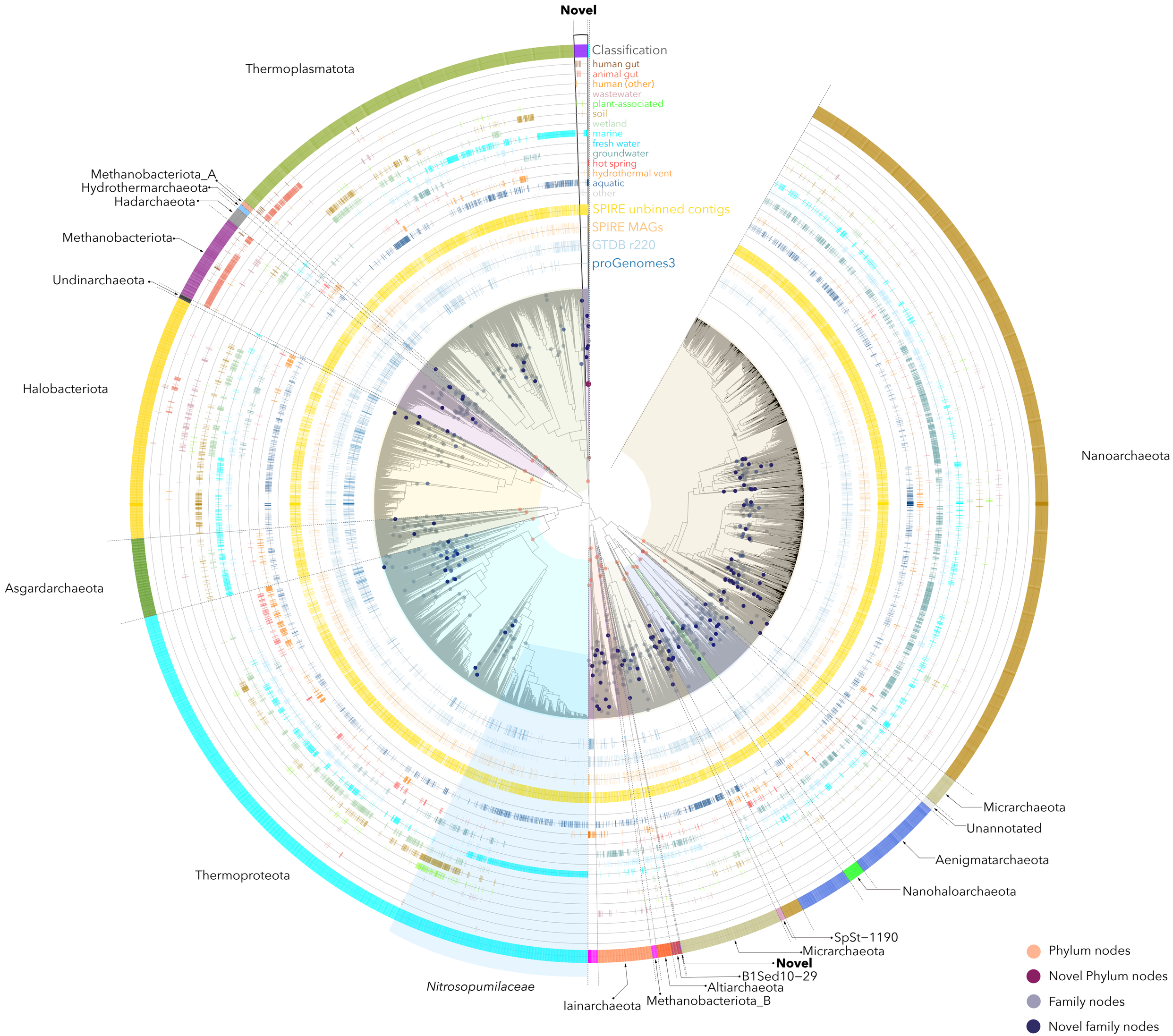
1. Moody, E. R. R. *et al.* The nature of the last universal common ancestor and its impact on the early Earth system. *Nat. Ecol. Evol.* **8**, 1654–1666 (2024).
2. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6578–6583 (1998).
3. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6506–6511 (2018).
4. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* 16048 (2016).
5. Curtis, T. P., Sloan, W. T. & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences* **99**, 10494–10499 (2002).
6. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 201521291–201525975 (2016).
7. Louca, S., Mazel, F., Doebeli, M. & Parfrey, L. W. A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biol.* **17**, e3000106 (2019).
8. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).
9. Schloss, P. D. & Handelsman, J. Status of the microbial census. *Microbiol. Mol. Biol. Rev.* **68**, 686–691 (2004).
10. Schloss, P. D., Girard, R. A., Martin, T., Edwards, J. & Thrash, J. C. Status of the Archaeal and Bacterial Census: an Update. *MBio* **7**, e00201–16 (2016).
11. Abellan-Schneyder, I. *et al.* Primer, pipelines, parameters: Issues in 16S rRNA gene sequencing. *mSphere* **6**, (2021).
12. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nature Publishing Group* **10**, 881–884 (2013).
13. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **31**, 1 (2017).
14. Dombrowski, N. *et al.* Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).
15. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
16. Fullam, A. *et al.* proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac1078.
17. Parks, D. H. *et al.* 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).
18. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0603-3.
19. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
20. Ma, B. *et al.* A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.* **14**, 7318 (2023).
21. Gurbich, T. A. *et al.* MGnify Genomes: A resource for biome-specific microbial genome catalogues. *J. Mol. Biol.* **435**, 168016 (2023).
22. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
23. Schmidt, T. S. B. *et al.* SPIRE: A Searchable, Planetary-scale mIcrobiome REsource. *Nucleic Acids Res.* **52**, D777–D783 (2024).
24. Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
25. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
26. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* **13**, 217–229 (2015).
27. Wu, D., Seshadri, R., Kyrpides, N. C. & Ivanova, N. N. A metagenomic perspective on the microbial prokaryotic genome census. *Sci. Adv.* **11**, eadq2166 (2025).
28. Woodcroft, B. J. *et al.* SingleM and Sandpiper: Robust microbial taxonomic profiles from metagenomic data. *Microbiology*

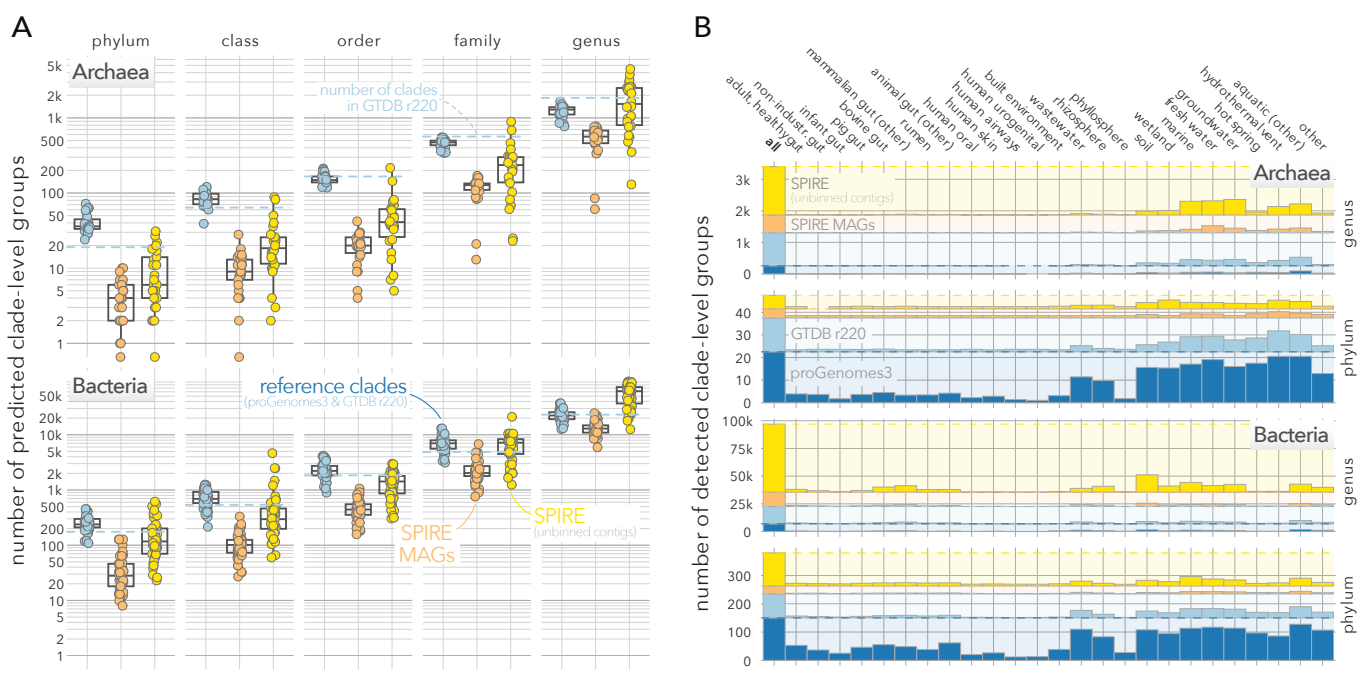
(2024).
29. Aroney, S. T. N., Newell, R. J. P., Tyson, G. W. & Woodcroft, B. J. Bin Chicken: targeted metagenomic coassembly for the efficient recovery of novel genomes. *bioRxiv* (2024) doi:10.1101/2024.11.24.625082.
30. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
31. Guy, L. & Ettema, T. J. G. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).
32. Moody, E. R. R. *et al.* An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *Elife* **11**, e66695 (2022).
33. *Bergey's Manual of Systematics of Archaea and Bacteria*. (Wiley, 2015). doi:10.1002/9781118960608.
34. Meier-Kolthoff, J. P., Carbasse, J. S., Peinado-Olarte, R. L. & Göker, M. TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Res.* **50**, D801–D807 (2022).
35. Parte, A. C., Sardà Carbasse, J., Meier-Kolthoff, J. P., Reimer, L. C. & Göker, M. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *Int. J. Syst. Evol. Microbiol.* **70**, 5607–5612 (2020).
36. Hedlund, B. P. *et al.* SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nat Microbiol* **7**, 1702–1708 (2022).
37. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
38. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *mSystems* **3**, (2018).
39. Ishii, S. & Sadowsky, M. J. Escherichia coli in the environment: Implications for water quality and human health. *Microbes Environ.* **23**, 101–108 (2008).
40. Ferretti, P. *et al.* C. difficile may be overdiagnosed in adults and is a prevalent commensal in infants. *eLife* (2023) doi:10.7554/elife.90111.
41. Aller, J. Y. & Kemp, P. F. Are Archaea inherently less diverse than Bacteria in the same environments?: Archaeal vs. bacterial diversity. *FEMS Microbiol. Ecol.* **65**, 74–87 (2008).
42. Spang, A. & Offre, P. Towards a systematic understanding of differences between archaeal and bacterial diversity: Systematic understanding of differences. *Environ. Microbiol. Rep.* **11**, 9–12 (2019).
43. Mahendrarajah, T. A. *et al.* ATP synthase evolution on a cross-braced dated tree of life. *Nat. Commun.* **14**, 7456 (2023).
44. López-García, P., Zivanovic, Y., Deschamps, P. & Moreira, D. Bacterial gene import and mesophilic adaptation in archaea. *Nat. Rev. Microbiol.* **13**, 447–456 (2015).
45. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
46. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**, (2022).
47. Rodríguez Del Río, Á. *et al.* Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* **626**, 377–384 (2024).
48. Sereika, M. *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).
49. Sereika, M. *et al.* Recovery of highly contiguous genomes from complex terrestrial habitats reveals over 15,000 novel prokaryotic species and expands characterization of soil and sediment microbial communities. *Bioinformatics* (2024).
50. Shaw, J. & Yu, Y. W. Fairy: fast approximate coverage for multi-sample metagenomic binning. *Microbiome* **12**, 151 (2024).
51. Sahlin, K. Strobealign: flexible seed size enables ultra-fast and accurate read alignment. *Genome Biol.* **23**, 260 (2022).
52. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
53. Pan, S., Zhao, X.-M. & Coelho, L. P. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* **39**, i21–i29 (2023).
54. Wang, Z. *et al.* Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nat. Commun.* **15**, 585 (2024).
55. Ruamps, L. S., Nunan, N. & Chenu, C. Microbial biogeography at the soil pore scale. *Soil Biol. Biochem.* **43**, 280–286 (2011).
56. Bach, E. M., Williams, R. J., Hargreaves, S. K., Yang, F. & Hofmockel, K. S. Greatest soil microbial diversity found in micro-

714 habitats. *Soil Biol. Biochem.* **118**, 217–226 (2018).
715 57. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359
716 (2015).
717 58. Nishijima, S. *et al.* Fecal microbial load is a major determinant of gut microbiome variation and a confounder for disease
718 associations. *Cell* **188**, 222–236.e15 (2025).
719 59. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
720 60. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**,
721 1069–1076 (2017).
722 61. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic*
723 *Acids Res.* **40**, e72 (2012).
724 62. Hedlund, B. P. & Staley, J. T. Microbial Endemism and Biogeography. in *Microbial Diversity and Bioprospecting* 225–231
725 (ASM Press, Washington, DC, USA, 2014). doi:10.1128/9781555817770.ch22.
726 63. Tamames, J., Abellán, J. J., Pignatelli, M., Camacho, A. & Moya, A. Environmental distribution of prokaryotic taxa. *BMC*
727 *Microbiol.* **10**, 85 (2010).
728 64. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **52**, D33–
729 D43 (2024).
730 65. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome
731 taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
732 66. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
733 67. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, (2018).
734 68. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the
735 microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950–13955 (2005).
736 69. Maistrenko, O. M. *et al.* Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-
737 species diversity. *ISME J.* **14**, 1247–1259 (2020).
738 70. Egghe, L. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. *J. Am. Soc. Inf. Sci. Technol.*
739 **58**, 702–709 (2007).
740 71. Steenwyk, J. L., Buida, T. J., 3rd, Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: A multiple sequence alignment trimming software
741 for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).
742 72. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS*
743 *One* **5**, e9490 (2010).
744 73. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a
745 maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
746 74. Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2018).
747 75. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance
748 necessary? in *Proceedings of the 26th Annual International Conference on Machine Learning* (ACM, New York, NY, USA,
749 2009). doi:10.1145/1553374.1553511.
750 76. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree: An r package for visualization and annotation of phylogenetic
751 trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
752 77. European Molecular Biology Laboratory, Pečar, J., Lueck, R. & Wahlers, M. EMBL Heidelberg HPC Cluster. Preprint at
753 https://doi.org/10.5281/ZENODO.12785829 (2020).
754


A
Archaea
Bacteria
n of species discovered
n of metagenomes
proGenomes3
(isolate genomes)
GTDB r220
(isolates & MAGs)
SPIRE
(MAGs)
SPIRE
(unbinned, non-singleton)
SPIRE
(unbinned, all)
B
gastrointestinal
human (other)
plant-assoc.
aquatic
all
adult, healthy
adult, diseased
non-industrial.
infant
child
elderly
porcine
bovine
mammalian (oth.)
rumen
avian
other
oral
skin
airways
urogenital
other
built env.
waste water
rhizosphere
phyllosphere
other
soil
wetland
marine
fresh water
groundwater
hot spring
hydrotherm. vent
other
air
other
n of species discovered
C
gastrointestinal
human (other)
plant-assoc.
aquatic
species discovery coefficient α
Archaea
Bacteria
all
adult, healthy
adult, diseased
non-industrialized
infant
child
elderly
porcine
bovine
mammalian (oth.)
rumen
avian
other
oral
skin
airways
urogenital
other
built env.
waste water
rhizosphere
phyllosphere
other
soil
wetland
marine
fresh water
groundwater
hot spring
hydrotherm. vent
other
air
other


Archaea
Bacteria
number of species discovered
number of metagenomes
0
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
200k
300k
400k
500k
600k
700k
newly discovered species per environment (incremental)
metagenomes added per environment
number of species in reference dataset(s)
species discovered in metagenomes
adult, healthy
adult, diseased
non-industrial.
infant
child
elderly
porcine
bovine
mammalian
(other)
animal
(other)
rumen
avian
gastrointestinal
human (other)
oral
skin
airways
urogenital
other
built env.
anthropogenic
waste water
rhizosphere
plant-associated
phyllosphere
other
soil
wetland
aquatic
marine
fresh water
groundwater
hot spring
hydrotherm. vent
other
air
other
SPIRE
unbinned, non-singleton
SPIRE
unbinned, all
proGenomes3
(isolate genomes)
GTDB r220
(isolates & MAGs)
SPIRE
(MAGs)
SPIRE
(unbinned, non-singleton)
SPIRE
(unbinned, all)


Novel
Classification
human gut
animal gut
human (other)
wastewater
plant-associated
soil
wetland
marine
fresh water
groundwater
hot spring
hydrothermal vent
aquatic
other
SPIRE unbinned contigs
SPIRE MAGs
GTDB r220
proGenomes3
Thermoplasmatota
Methanobacteriota_A
Hydrothermarchaeota
Hadarchaeota
Methanobacteriota
Undinarchaeota
Halobacteriota
Asgardarchaeota
Thermoproteota
Nitrosopumilaceae
Iainarchaeota
Methanobacteriota_B
Altiarchaeota
B1Sed10−29
Novel
Micrarchaeota
SpSt−1190
Nanohaloarchaeota
Aenigmatarchaeota
Unannotated
Micrarchaeota
Nanoarchaeota
Phylum nodes
Novel Phylum nodes
Family nodes
Novel family nodes


A
phylum
class
order
family
genus
Archaea
Bacteria
number of clades in GTDB r220
reference clades
(proGenomes3 & GTDB r220)
SPIRE MAGs
SPIRE
(unbinned contigs)
number of predicted clade-level groups
B
adult, healthy gut
non-industr. gut
infant gut
mammalian gut (other)
pig gut
bovine gut
animal gut (other)
rumen
human oral
human skin
human airways
human urogenital
built environment
wastewater
rhizosphere
phyllosphere
wetland
soil
marine
fresh water
groundwater
hot spring
hydrothermal vent
aquatic (other)
other
all
SPIRE
(unbinned contigs)
SPIRE MAGs
GTDB r220
proGenomes3
Archaea
Bacteria
genus
phylum
number of detected clade-level groups