

Genome-resolved long-read sequencing expands known microbial diversity across terrestrial habitats

Received: 23 December 2024

Accepted: 16 June 2025

Published online: 24 July 2025

 Check for updates

Mantas Sereika¹, Aaron James Mussig², Chenjing Jiang¹, Kalinka Sand Knudsen¹, Thomas Bygh Nymann Jensen¹, Francesca Petriglieri¹, Yu Yang¹, Vibeke Rudkjøbing Jørgensen¹, Francesco Delogu¹, Emil Aarre Sørensen¹, Per Halkjær Nielsen¹, Caitlin Margaret Singleton¹, Philip Hugenholtz² & Mads Albertsen¹ ✉

The emergence of high-throughput, long-read DNA sequencing has enabled recovery of microbial genomes from environmental samples at scale. However, expanding the terrestrial microbial genome catalogue has been challenging due to the enormous complexity of these environments. Here we performed deep, long-read Nanopore sequencing of 154 soil and sediment samples collected during the Microflora Danica project, yielding genomes of 15,314 previously undescribed microbial species, recovered using our custom mmlong2 workflow. The recovered microbial genomes span 1,086 previously uncharacterized genera and expand the phylogenetic diversity of the prokaryotic tree of life by 8%. The long-read assemblies also enabled the recovery of thousands of complete ribosomal RNA operons, biosynthetic gene clusters and CRISPR-Cas systems. Furthermore, the incorporation of the recovered genomes into public genomic databases substantially improved species-level classification rates for soil and sediment metagenomic datasets. These findings demonstrate that long-read sequencing allows cost-effective recovery of high-quality microbial genomes from highly complex ecosystems, which remain an untapped source of biodiversity.

The vast majority of microorganisms are predicted to be undiscovered¹. Traditionally, achieving genomes of previously uncharacterized microbial species involves isolating and cultivating the microorganisms, followed by sequencing². While this method has successfully yielded thousands of previously undescribed genomes³, culturing can be labour intensive and time consuming⁴, and most microbes are estimated to be unsuitable for isolation⁵. In the past decade, genome-centric metagenomics has emerged as an alternative and expedient means of characterizing microbial diversity through recovery of metagenome-assembled genomes (MAGs)^{6,7}. Despite potential issues of contamination and

incompleteness (for example, different microbial strains⁸ or species⁹), metagenomics allows large-scale recovery of previously undescribed genomes from uncultured microorganisms^{10,11}. So far, the Genome Taxonomy Database (GTDB, release 220) comprises 113,104 prokaryotic species, of which 72.5% are represented exclusively by MAGs¹², highlighting the current limitations in culture-based genomics. Therefore, MAGs will be indispensable to obtain genomic coverage of the estimated 2–4 million prokaryotic species inhabiting the biosphere¹³.

Soil has the potential to greatly increase the number of microbial species in the databases given its enormous microbial diversity¹⁴.

¹Center for Microbial Communities, Aalborg University, Aalborg, Denmark. ²Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Queensland, Australia. ✉e-mail: ma@bio.aau.dk

However, this complexity also makes soil exceptionally challenging for MAG recovery¹⁴. Several attempts have been made to improve MAG recovery from soil, such as reducing the complexity of the sample through species enrichment¹⁵, cell sorting¹⁶ or deep short-read sequencing (for example, over 100 Gbp to several Tbp of sequencing data)^{17,18}. However, none of these approaches have resulted in the cost-effective recovery of high-quality microbial genomes. Hence, developing a solution for the efficient recovery of high-quality MAGs from soil and other microbially complex habitats is considered the ‘grand challenge’ of metagenomics¹⁹.

In recent years, long-read sequencing has substantially enhanced our ability to recover high-quality microbial genomes from medium-complexity samples^{20,21}. This has been complemented by the development of bioinformatic methods that improve MAG recovery from challenging samples through the use of deep-learning algorithms^{22,23} or additional binning features^{24–26}. Therefore, multiple sequencing and bioinformatic approaches have now become available for tackling the ‘terrestrial metagenome challenge’.

Here we performed deep long-read Nanopore sequencing (~100 Gbp per sample) of 154 complex environmental samples collected as part of the Microflora Danica project, which aims to genomically catalogue microbial diversity in Denmark²⁷. By developing a bioinformatics workflow that uses state-of-the-art metagenomic binning tools, combined with multicoverage and iterative binning, we obtained over 15,000 species-level MAGs. The great majority (97.9%) of these MAGs represent previously undescribed microbial genera or species, substantially expanding the microbial tree of life.

Results

High-throughput MAG recovery from soils and sediments

Of the 10,683 environmental samples collected during the Microflora Danica sampling campaign²⁷, 154 samples (125 soil, 28 sediment, 1 water) from 15 distinct habitats (Supplementary Table 1, Fig. 1 and Dataset 1) were selected (see Methods for selection criteria) for deep long-read Nanopore sequencing (Fig. 1a) to explore assembly performance across a wide breadth of sample types. A total of 14.4 Tbp long-read data was generated, with a median of 94.9 Gbp and an interquartile range (IQR) of 56.3–133.1 Gbp (Fig. 1b). The sequence reads had a median N50 (length cutoff where reads of that size or longer cover at least 50% of the total number of bases in the read dataset) of 6.1 kbp (IQR: 4.6–7.3 kbp) (Fig. 1c) and assembled into a total of 295.7 Gbp of metagenomic contigs, with a median contig N50 (length cutoff where contigs of that size or longer cover at least 50% of the total assembly) of 79.8 kbp (IQR: 45.8–110.1 kbp) per sample. The majority of reads were assembled into contigs, as a median 62.2% (IQR: 53.1–69.8%) of the sequence data was mapped back to the assemblies (Fig. 1d).

To improve MAG recovery from high-complexity environmental samples, we developed *mmlong2*, a metagenomics workflow that features multiple optimizations for recovering prokaryotic MAGs from extremely complex metagenomic datasets. Briefly, *mmlong2* performs metagenome assembly, polishing, removal of eukaryotic contigs and extraction of circular MAGs (cMAGs) as separate genome bins (Fig. 1g). It then performs differential coverage binning (incorporating read mapping information from multisample datasets, Supplementary Dataset 2), ensemble binning (using multiple bidders on the same metagenome, Supplementary Fig. 2a) and iterative binning (metagenome gets binned multiple times iteratively, Supplementary Fig. 2b), which all contribute to increased MAG recovery (see Methods for details). Compared with other contemporary metagenomic binning workflows, *mmlong2* enables recovery of more MAGs from terrestrial metagenomes, with the trade-off of moderately increased compute times (Supplementary Table 2).

In total, 6,076 high-quality (HQ) and 17,767 medium-quality (MQ) MAGs (23,843 total, Supplementary Dataset 3) were recovered by the *mmlong2* workflow from the 154 sequenced samples, including 3,349

(14.0%) MAGs recovered by iterative binning (Supplementary Table 3), with a median of 154 (IQR: 89–204) high- or medium-quality MAGs recovered per sample (Fig. 1e). The obtained MAGs were estimated to account for a median of 24.0% (IQR: 16.7–32.9%) of the sequence data within individual samples (Fig. 1f).

Lower per-sample sequencing yields were observed for samples originating from the two habitat categories of agricultural fields as well as the bogs, mires and fens habitat (Supplementary Fig. 3a), which might be attributed to suboptimal DNA extraction leaving contaminants that compromise the DNA sequencing. In addition, the agricultural field samples had low amounts of sequence data assembled into contigs (median 45.0%, IQR: 39.3–50.1%, Supplementary Fig. 3b) and also the lowest per-sample count of high- or medium-quality MAGs (median 56 MAGs, IQR: 34–89, Supplementary Fig. 3c), whereas coastal habitat samples yielded the highest MAG recovery metrics (Supplementary Fig. 3). To investigate whether the relatively poor MAG yield from agricultural field samples (Supplementary Fig. 3d) was only due to sequencing yield, three agricultural and coastal samples were selected and subsampled to specific sequencing depths (from 20 to 100 Gbp, see Methods for more details). Despite normalization for sequencing effort, the coastal habitat samples still exhibited greater MAG yield (Supplementary Fig. 4a).

Between the two habitat types, there were no substantial differences in non-prokaryotic DNA (Supplementary Fig. 4b) and overall, a comparable number of prokaryotic species was observed in the reads (Supplementary Fig. 4c) or contigs (Supplementary Fig. 4d), without signs of full microbial diversity capture at 100 Gbp sequencing depth. However, *k*-mer redundancy analysis indicated that for the coastal samples, more species were abundant, compared with the agricultural samples (Supplementary Fig. 4e–g). Furthermore, MAGs from the coastal samples also had lower rates of MAG polymorphism (proxy for microdiversity, Supplementary Fig. 4h). Hence, the relatively poor MAG recovery from the agricultural field samples was influenced by reduced sequencing yields, higher microdiversity and the absence of dominant species (Supplementary Fig. 4i).

The multiple reasons for variation in MAG recovery were assumed to be facilitated by extensive ecological differences between the two soil habitats, as the highly saline and low-nutrient coastal ecosystems select for salt-tolerant organisms²⁸, while the microbial communities from high-nutrient agricultural fields are shaped by agricultural practices^{29,30}. Across the Microflora Danica 10,000 shallow metagenomes²⁷, the agricultural and coastal samples (Supplementary Fig. 5a) featured distinct microbial community compositions (Supplementary Fig. 5b, analysis of similarities (ANOSIM) $R = 0.755$, $p = 0.001$, $n = 1,046$), while notable differences were also observed between the coastal habitats of salt marshes or meadows and the habitats of sea cliffs, shingle or stony beaches (ANOSIM $R = 0.329$, $p = 0.001$, $n = 235$). Furthermore, phylum-level differences were pronounced (Supplementary Fig. 5c), as the agricultural habitats exhibited greater relative abundances of Firmicutes (Supplementary Fig. 5d) and fewer Proteobacteria (Supplementary Fig. 5e) or Bacteroidota (Supplementary Fig. 5f). Hence, variation in microbial community composition and taxonomic diversity was also expected to affect MAG yield from terrestrial habitats.

Contribution to terrestrial microbiomes

The recovered 23,843 MAGs (Fig. 2a) were dereplicated into 15,640 different species-level MAGs (Fig. 2b), comprising 4,894 HQ and 10,746 MQ MAGs (Fig. 2c–f and Supplementary Fig. 6a,b). Since the MAGs were recovered with long-read Nanopore sequencing, we refer to the dereplicated genome set as the Microflora Danica long-read (MFD-LR) MAG catalogue. The genomic catalogue was inspected for potential Nanopore-associated sequencing errors, and several instances ($n = 73$, 0.5% of genomes) of coding density values <75% were detected for MAGs with lower coverage (Supplementary Fig. 6c) and reduced

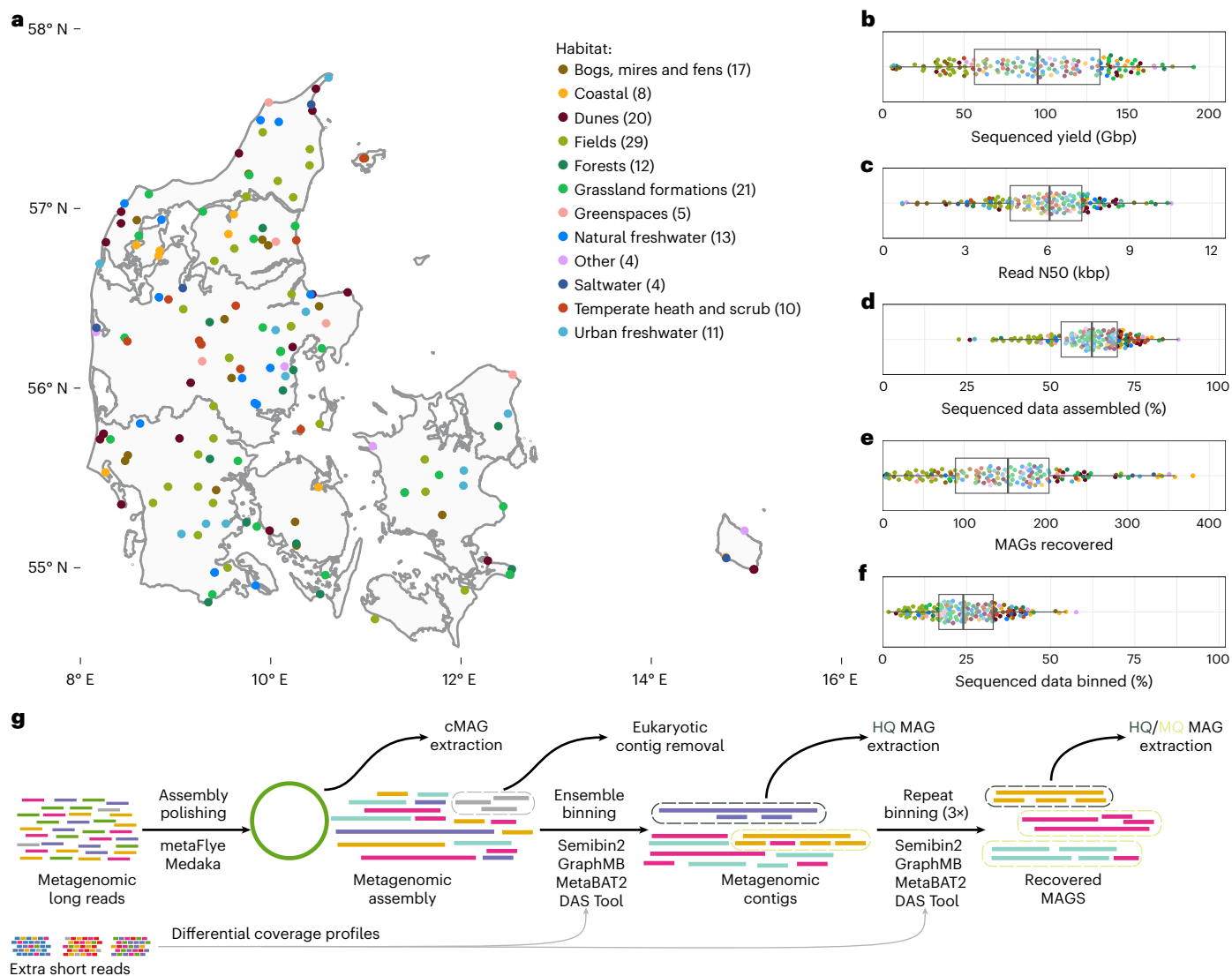


Fig. 1 | Overview of the sequenced environmental samples. **a**, Geographic distribution of the samples, coloured by sample habitat (Microflora Danica habitat descriptor level 1). Samples from habitats that occur once in the dataset are grouped into the ‘Other’ habitat category. **b**, Per-sample sequencing yield in Gbp. **c**, Per-sample sequenced read N50 values in kbp. **d**, Per-sample percentage of sequencing data that was estimated as assembled into contigs. **e**, Per-sample count of the recovered high- or medium-quality MAGs (see Methods for quality criteria). **f**, Estimates for per-sample percentage of sequenced data represented

by the recovered high- or medium-quality MAGs. For the boxplots, results are shown for the sequenced samples ($n = 154$); the central line represents the median, the hinges correspond to the 25th and 75th percentiles, and the whiskers extend up to 1.5 times the IQR. **g**, Schematic overview of the mmlong2 metagenomics workflow. ‘Extra short reads’ refers to the shallow metagenome datasets from the Microflora Danica project²⁷ used for differential coverage binning (see Methods).

guanine-cytosine (GC) content (Supplementary Fig. 6d). However, the reduced coding density was also found to be mostly prevalent in archaeal MAGs with increased rates of long homopolymers (>6 repeating nucleotides, Supplementary Fig. 6e), which in turn were more frequent in MAGs with low GC content (Supplementary Fig. 6f).

After species-level MAG dereplication, 51.4% ($n = 12,255$) of the recovered MAGs were singletons. Plotting a rarefaction curve for the MAGs showcased a near-linear (unsaturated) relationship between the number of recovered MAGs and the number of species-level clusters (Fig. 2b and Supplementary Fig. 7). The largest species-level cluster of 39 MAGs was recovered for the *Pseudolabrys* genus in the order Rhizobiales, and in total, 126 species-level clusters with >10 MAGs per cluster were obtained (Supplementary Fig. 8).

An advantage of long-read generated MAGs is that they mostly include ribosomal RNA (rRNA) operons, enabling direct comparison to the thousands of available 16S rRNA datasets and large-scale databases.

Of the recovered dereplicated MAGs, 12,823 (82.0%) included at least one 16S rRNA gene and were taxonomically classified against the Microflora Global 16S rRNA database, which features 16S rRNA gene sequences from the original Microflora Danica project as well as major publicly available 16S rRNA databases²⁷. Overall, 12,460 (97.2%) of these MAGs were classified to the genus level (>94.5% 16S rRNA gene identity), while 10,438 (81.4%) of the MAGs were assigned a species-level match (>98.7% 16S rRNA gene identity). Coverage of Microflora Danica core genera (across ~10,000 metagenomes²⁷) by the MAG dataset in this study varied from 72.3% to 93.0%, depending on the metadata description category (Fig. 2g), and exceeded 90% for all soil habitats (Supplementary Fig. 9).

Overall, 183.4% more dereplicated MAGs were recovered from the 154 deeply sequenced terrestrial samples than the Microflora Danica short-read (MFD-SR) shallow metagenome study²⁷ that sequenced close to 70-fold more samples (10,683 samples at ~5 Gbp each) (15,640

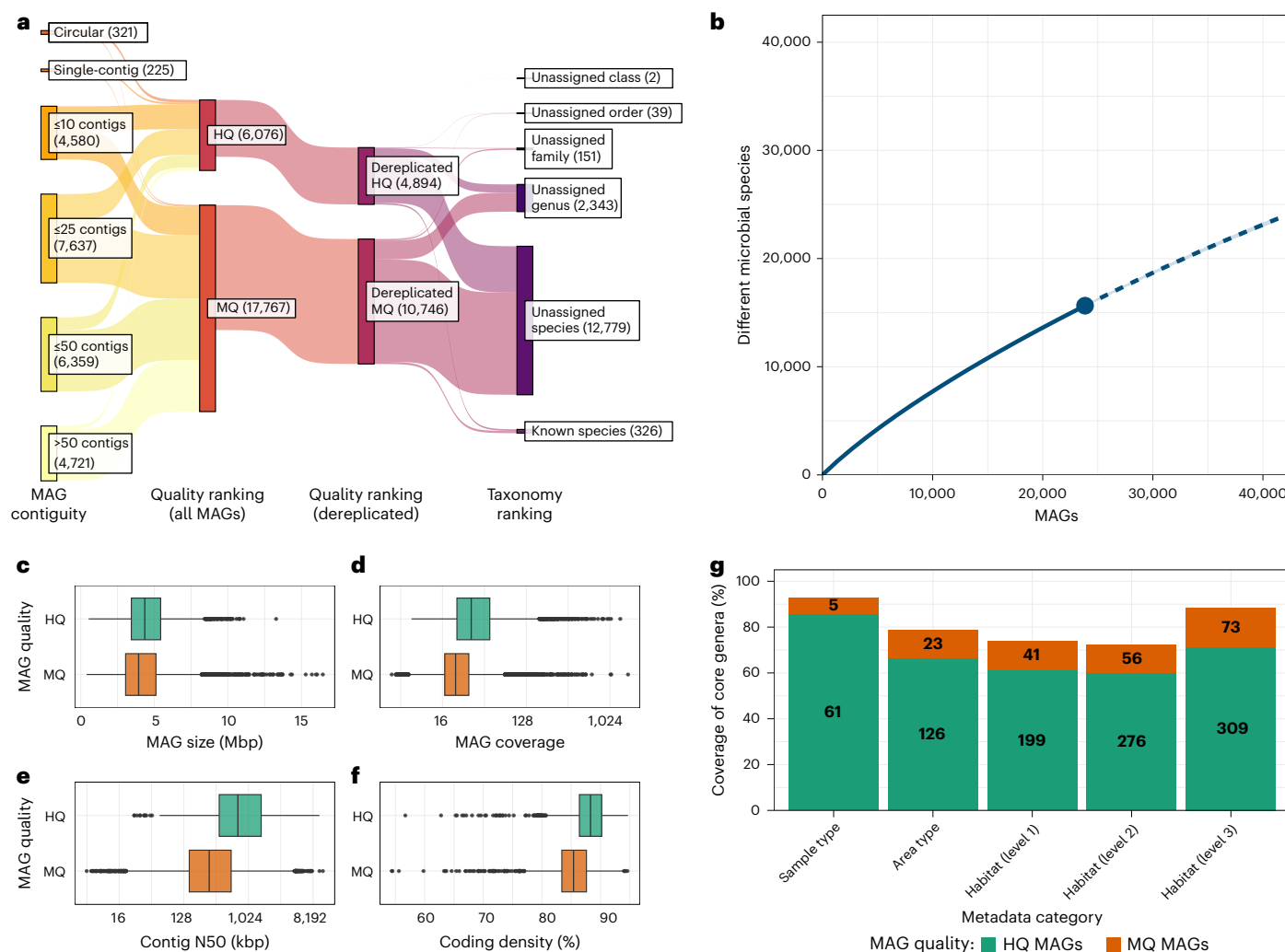


Fig. 2 | Overview of the MAGs recovered from the sequenced samples.

a, Aggregated MAG counts by MAG contiguity, MAG quality rankings before and after MAG dereplication, together with MAG taxonomy rankings according to GTDB R220 classification. ‘Single-contig’ refers to MAGs that are composed of one contig, which was not reported as circular by the assembler. **b**, Species-level (>95% ANI) rarefaction curve for HQ and MQ MAGs recovered in this study. The solid line depicts the rarefaction (interpolation), while the dotted line shows the extrapolation of the curve. **c**, Distribution of dereplicated MAG size values

in Mbp, grouped by MAG quality. **d**, Distribution of dereplicated MAG coverage values. **e**, Contig N50 for dereplicated MAGs in kbp. **f**, Coding density values for dereplicated MAGs. For the boxplots, results are shown for the dereplicated HQ ($n = 4,894$) and MQ ($n = 10,746$) MAGs; the central line represents the median, the hinges correspond to the 25th and 75th percentiles, the whiskers extend up to 1.5 times the IQR, and individual outliers are shown. **g**, Percentage coverage of Microflora Danica core genera²⁷ at different sample metadata description levels by the recovered MAGs.

vs 5,518 HQ and MQ MAGs), and 11-fold more (4,894 vs 422) dereplicated HQ MAGs were recovered from this study. Also, more MAGs were recovered in this project than the recent genome catalogues of the Tibetan Plateau Microbial Catalogue (TPMC³¹) and the Old Woman Creek wetland microbial genome catalogue (OWC³²; Fig. 3a and Table 1). Compared with global genomic catalogues that aggregate vast numbers of previously published sequencing data (including low-complexity samples), such as the Searchable, Planetary-scale Microbiome Resource (SPIRE³³), the Genomes from Earth’s Microbiome (GEM³⁴), Rare Biosphere Genomes (RBG³⁴) and Soil Microbial Dark Matter Metagenome Assembled Genome (SMAG³⁵), our 154 long-read sequenced samples still produced similar or higher numbers of HQ genomes, despite, for example, SPIRE utilizing almost 100,000 individual samples (Fig. 3a and Table 1). The inferred sequencing costs per HQ MAG recovered were also estimated to be the lowest for the MFD-LR catalogue (Supplementary Table 4).

Dereplicating MAGs between catalogues resulted in 138,407 species-level clusters, with most species-level overlaps occurring between the genome catalogues of SPIRE, SMAG and GEM, due to large

overlaps in primary data sources (Fig. 3b). For MAGs from this study, most species-level overlaps occurred with the short-read Microflora Danica MAG catalogue ($n = 1,423$), although 12,750 dereplicated MAGs (and 3,653 HQ MAGs) from this project represent distinct species.

MAGs from this study also featured greater assembly contiguity with a median contig count of 20 (IQR: 10–36), compared with >100 for short-read MAG catalogues (Table 1). Improved genome contiguity suggests enhanced assembly of complex genomic regions and indeed, we observed greatly improved recovery of rRNA genes as part of complete operons (Fig. 3c) and more complete defence gene islands, especially CRISPR-Cas clusters (Fig. 3d). More complete biosynthetic gene clusters (BGCs) were also a feature of the long-read assemblies, and a median of 6.1-fold (IQR: 3.8–14.8) more complete BGCs were observed in the MAGs from this study than other short-read MAG catalogues (Fig. 3e).

The aforementioned genome catalogues were used as reference databases for classifying the ~10,000 shallow metagenome datasets from the Microflora Danica project²⁷. Using the GTDB R220 database alone for read classification resulted in a median species-level

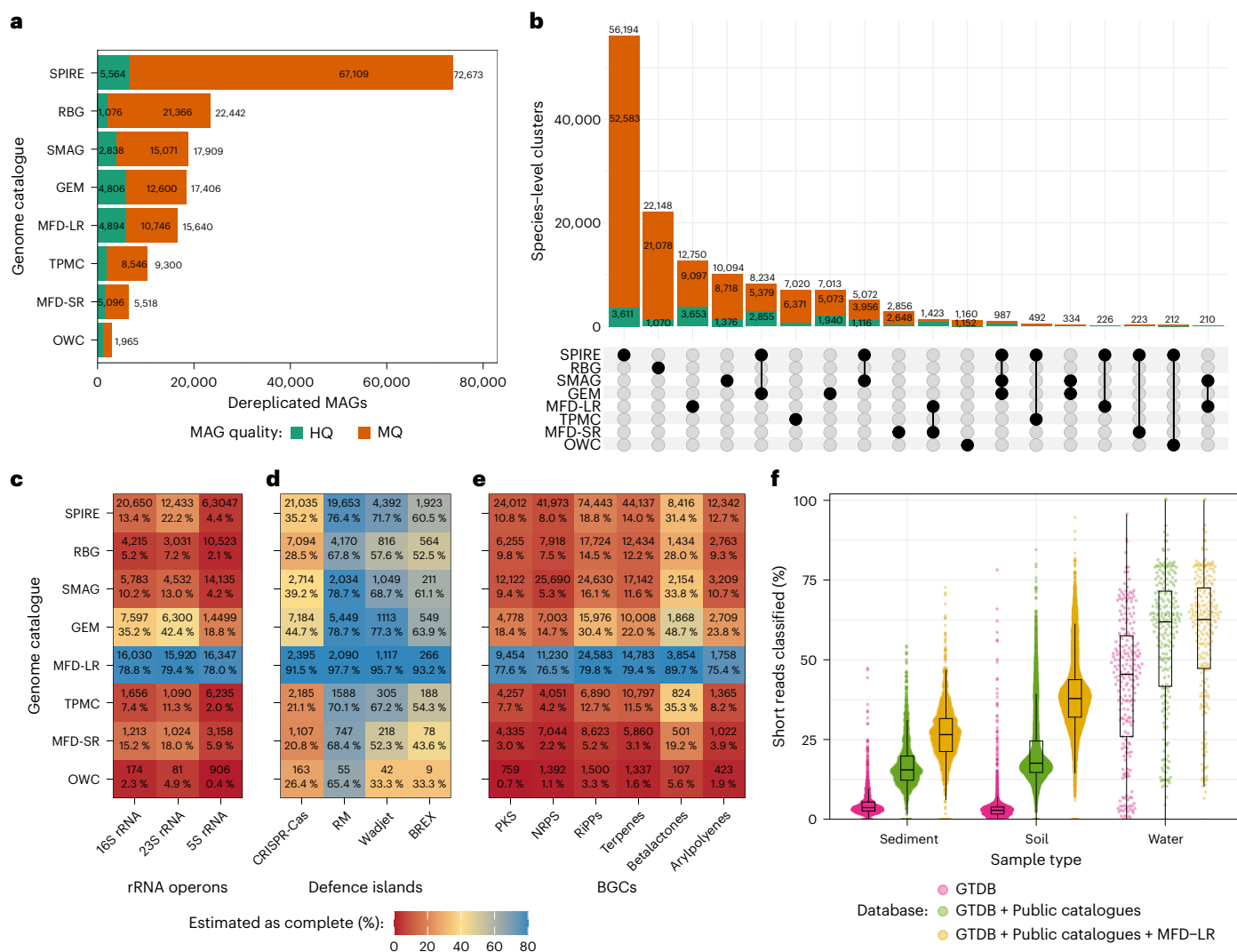


Fig. 3 | Comparison of MAG catalogues from large-scale terrestrial environment studies. a, Dereplicated MAGs (HQ and MQ) per genome catalogue. Individual counts for groups with >1,000 MAGs are presented when possible. The dereplicated MAG counts were achieved by applying the same genome quality control for all catalogues (see Methods). **b**, Upset plot for species-level MAG overlap between the catalogues. A genome cluster is marked as HQ if at least one of the genomes in the cluster is an HQ MAG. Groups with <200 MAGs were omitted from the plot. Total counts of gene clusters for rRNA operons (**c**), defence islands (**d**) and BGCs (**e**) predicted in the MAGs, grouped by gene

cluster type and coloured by the fraction of clusters estimated as complete (see Methods). **f**, Classification rates for Microflora Danica shotgun metagenome datasets (9,916 samples above 1 Gbp yield) using different genome databases, grouped by sample type. 'Public catalogues' refers to previously described terrestrial genome catalogues (see Methods). For the boxplots, results are shown for the soil ($n = 8,179$), sediment ($n = 1,518$) and water ($n = 219$) sample types; the central line represents the median, the hinges correspond to the 25th and 75th percentiles, and the whiskers extend up to 1.5 times the IQR.

classification rate of 3.0% (IQR: 1.8–4.1%), whereas including the short-read MAG catalogues increased the median classification rate to 17.4% (IQR: 14.3–24.4%). Addition of the long-read MAGs from this study resulted in a database of 229,714 non-redundant genomes and increased species classification to a median of 36.6% (IQR: 29.6–42.9%, Fig. 3f), with the greatest improvements occurring for soil samples (Supplementary Fig. 10).

Previously undescribed and expanded microbial lineages

Taxonomic classification using GTDB R220 resulted in average nucleotide identity (ANI)-based species-level assignments for 326 MAGs, which comprise 2.1% of the dereplicated MAGs. To determine the phylogenomic gain and diversity for the remaining 15,198 (97.9%) dereplicated MAGs that could not be assigned a species-level taxonomic label, de novo phylogenetic trees were constructed using MAGs from this study and GTDB R220 species representatives (Fig. 4). MAGs recovered in this study were found to increase the total branch length of the GTDB

prokaryotic genome tree by 8.1% (Supplementary Fig. 11), with most of the branch expansion occurring at genus or species level in both the bacterial and archaeal domains (Supplementary Fig. 12). Based on relative evolutionary divergence (RED), this added diversity comprises 1 phylum, 21 orders, 91 families and 1,086 genera (Table 2).

The microbial lineages represented by MFD-LR MAGs were widely distributed across terrestrial habitats, with 98.8% ($n = 9,930$) of MFD-SR samples containing reads classified to MAGs from at least one previously undescribed genus (Supplementary Fig. 13a–c). Reads for genomes of previously uncharacterized families and orders were found in 68.2% ($n = 6,849$) and 24.7% ($n = 2,480$) of samples, respectively. Urban soils had the highest frequency of previously undescribed genera (median 35 per sample, IQR: 25–45, Supplementary Fig. 13d), contributing a median of 2.9% (IQR: 2.0–3.8%) of sequenced reads (Supplementary Fig. 13e), with the roadside habitat featuring the most previously undescribed genera (median 41, IQR: 26–55) and families (median 5, IQR: 4–6, Supplementary Fig. 14a). In contrast, uncharacterized

Table 1 | Summary of terrestrial prokaryotic genome catalogues

Genome catalogue	Dereplicated MAGs ^a	HQ MAGs ^b	Median contigs ^c	Total samples ^d	Original samples ^e	Yield (Tbp)
SPIRE	72,673	5,564	163	99,146	0	~500
RBG	22,442	1,076	193	1,286	0	17.8
SMAG	17,909	2,838	306	3,304	363	46.7 ^f
GEM	17,406	4,806	124	10,450	0	NA
MFD-LR	15,640	4,894	20	154	154	14.4
TPMC	9,300	754	303	498	498	25.7 ^f
MFD-SR	5,518	422	666	10,683	10,683	48.2
OWC	1,965	12	456	42	42	3.1

^aIncludes species-representative HQ and MQ MAGs. ^bHQ MAG count after standardized genome quality assessment (see Methods). ^cRefers to the median contig count per dereplicated MAG. ^dIncludes the number of samples used for MAG recovery. ^eRefers to the number of samples sequenced as part of the study. ^fSequencing yield statistics were extrapolated from the reported read counts.

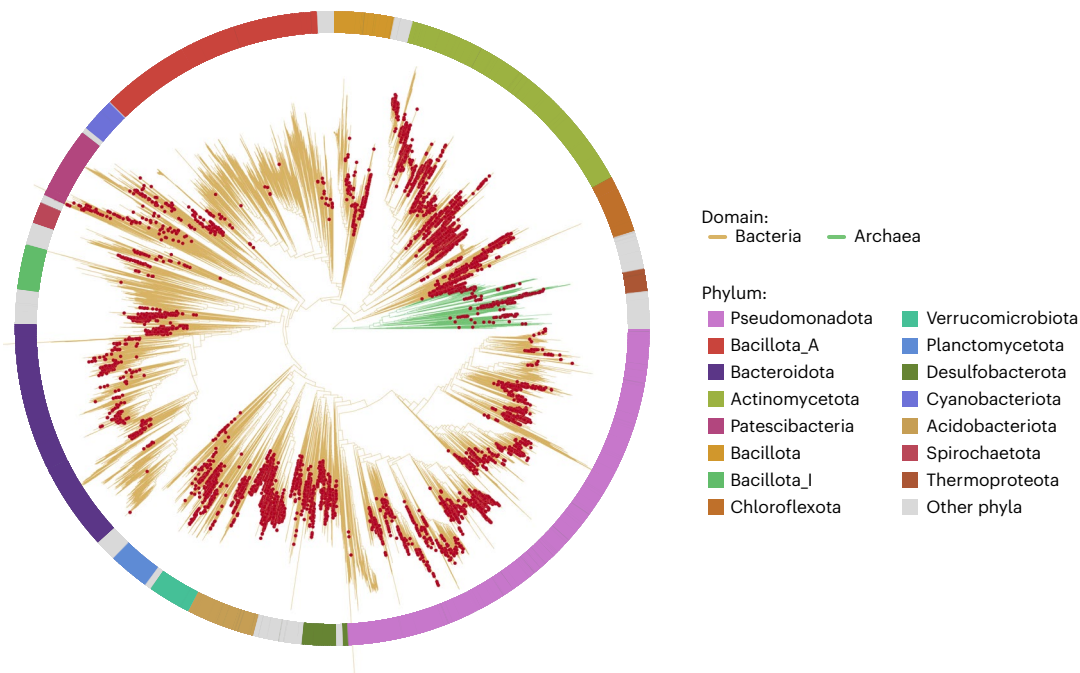


Fig. 4 | Distribution of the recovered MAGs across the microbial genome tree of life. Microbial genome trees with GTDB R220 representative species for Bacteria (120 marker genes) and Archaea (53 marker genes) were built separately with 100 bootstraps and merged into a single tree, spanning both domains.

The tree branches are coloured by domain (Bacteria or Archaea). Tree tips of dereplicated MFD-LR MAGs are marked with red dots within the tree. The outer circle highlights tree tips of the 15 phyla with the highest number of genomes. Reproduced from [GTDB](#), [CC BY-SA 4.0](#).

genera and families, identified only by the 16S rRNA gene in terrestrial habitats²⁷, were most common in sediment samples (Supplementary Figs. 13f,g and 14b), with 22,437 genera and 1,095 families currently lacking genomic representation.

The MAG representing a previously undescribed phylum was successfully re-assembled into a circular 2.9 Mbp genome with a GC content of 51.3% and a single rRNA operon. The coding density was 91.8%, although 57.7% of the predicted genes ($n = 2,473$) were hypothetical. We detected species-level matches of the MAG in 7 of the ~10,000 MFD-SR environmental metagenomes, representing four geographic locations (Supplementary Fig. 13a). Six of these metagenomes were from dystrophic lakes (characterized by high organic acid content, low nutrients and low pH), suggesting relatively low environmental prevalence of the lineage and habitat specificity. This was reflected in the genomic potential of the MAG, as metabolic reconstruction indicated that the bacterium is probably motile, Gram-negative and adapted to an anaerobic environment with available dissolved organic carbon. The MAG encoded the potential to ferment glucose to acetate, use ethanolamine as a source of nitrogen and energy, and fix or detoxify

formaldehyde using the ribulose monophosphate pathway³⁶ (Supplementary Dataset 4). Due to the considerable phylogenetic distinctiveness of the cMAG, we propose the name *Oederibacterium danicum* sp. nov. in honour of Georg Christian Oeder, a scientist who led the original Flora Danica project³⁷.

A total of 207 previously undescribed genera and 1,170 species were represented by at least one HQ MAG comprising ≤ 10 contigs, for which we proposed names (Supplementary Dataset 5) under the SeqCode³⁸. Since the MFD-LR MAGs were recovered from Danish habitats, genus names were derived from Danish towns that were nearby the sampling locations, and species names were derived from environmental features of the samples from which the MAGs were obtained (see Methods). For genomes that could be assigned to GTDB lineages with placeholder names, we also proposed higher rank names on the basis of the genus stems under the SeqCode to provide taxonomic congruence (Supplementary Dataset 5).

MAGs recovered in this study spanned 75 of the 217 currently recognized phyla, with 50% or higher increases in species-level MAGs for 10 phyla (Supplementary Figs. 15, 16 and 17a, and Table 5). Notably, MAGs

Table 2 | Summary for the contribution of MFD-LR MAGs to GTDB R220

Rank	Lineages in GTDB ^a	Lineages added ^b	Lineages without HQ genomes ^a	HQ MAGs added ^b	Lineages without 16S rRNA ^a	MAGs with 16S rRNA added ^b
Phylum	217	1	72	2	26	1
Class	607	1	277	8	86	3
Order	2,030	21	1,142	46	324	12
Family	5,460	91	3,581	158	1,095	54
Genus	24,959	1,086	18,228	612	7,398	436
Species	113,104	15,314	85,812	4,757	61,767	12,541

^aRefers to the prokaryotic lineages in GTDB R220. ^bRefers to the addition of MFD-LR MAGs to GTDB R220.

were recovered for underrepresented phyla with placeholder names, such as JAUUVQV01, CAKKQC01 and UBP4, all of which featured only 2 species in GTDB R220. Furthermore, the inclusion of species-level MAGs from this study has substantially expanded several highly populated phyla. Actinomycetota increased by 42.1% (from 11,737 to 16,683 species-level genomes), Chloroflexota by 38.5% (from 2,749 to 3,808 genomes) and Acidobacteriota by 134.2% (from 1,891 to 4,429 genomes). Similar increases in microbial lineage genome counts were observed when examining class, order and family ranks (Supplementary Fig. 17b–d and Table 5).

A total of 12,779 dereplicated and previously undescribed species MAGs were classified as 2,052 different known genera. Of these genera, 682 (32.2%) were represented by a single genome in GTDB and inclusion of MAGs recovered in this study expanded the species-level representatives by more than 100% for 1,065 genera (51.9% of existing genera with MFD-LR MAGs). The highest number of recovered genomes for a known genus was for Palsa-744 (Actinomycetota), with an increase of 1,230.8% (from 26 to 346 genomes), whereas the highest increase in a microbial lineage of 5,000% (from 1 to 51 genomes) was observed for the genus RYN-230 (Actinomycetota).

This study provides HQ genomes for 158 known microbial families and 612 known genera that were previously represented only by MQ genomes (Table 2). Notable examples of such lineages include the orders of Pacearchaeales (Nanoarchaeota) and Micrarchaeales (Micrarchaeota), which in GTDB R220 are represented by 235 and 189 MQ MAGs, respectively. Similarly, this study provides genomes with complete 16S rRNA genes for 436 known genera (Table 2) that were previously lacking such representation, including the Actinomycetota genera *Gaiellasilicea*, *Gaiella* and *Desertimonas*, which were all expanded more than 10-fold.

Discussion

Here we developed mmlong2, a bioinformatic workflow that capitalizes on high-throughput deep long-read sequencing to recover MAGs from highly complex terrestrial samples. To evaluate performance, we sequenced 154 soil and sediment samples across 15 environmental habitats. Overall, hundreds of MAGs could be recovered from each sample, thereby enabling cost-efficient MAG recovery at scale from soils and sediments. However, MAG recovery varied between habitats, especially with agricultural soils consistently yielding fewer MAGs. We show that the variance in MAG recovery between habitats was influenced by sequencing yield, microdiversity and community composition. Furthermore, non-biological factors can also impact MAG recovery, as terrestrial habitats can feature vastly different chemical compositions and abiotic factors³⁹, which shape microbial communities^{30,40}. Hence, we recommend that researchers take into consideration the unique features of each terrestrial habitat when conducting experimental design for

future metagenomics projects (for example, habitat-optimized DNA extraction⁴¹, low-biomass-compatible sequencing protocols⁴²). In general, we recommend sequencing at least 60 Gbp per sample, as this ensures access to the genomes of both dominant terrestrial species and low-abundance species as evidenced by no indication of saturation observed in the sequencing depth investigated (up to 100 Gbp). We also note that high-throughput recovery of multipartite or plasmid-containing genomes from terrestrial environments remains challenging, although recent advances in methylation-based binning offer promising improvements²⁶.

Compared with other extensive genome-centric studies of terrestrial habitats^{14,31,33,35}, this study used long reads to recover MAGs from terrestrial samples at scale. The improved long-read MAG contiguity permits higher resolution of complex genomic regions⁴³, such as repeated operons and gene clusters. As the majority of the MAGs were recovered with 16S rRNA genes, most could be linked to the Microflora Global²⁷ and other 16S rRNA gene databases. Since rRNA gene databases are generally more diverse than genome databases⁴⁴, recovering more MAGs with complete 16S rRNA genes facilitates improved taxonomic classification and improved linkage between genome and 16S rRNA gene databases⁴⁵. Furthermore, unlike previous terrestrial MAG catalogues, the majority of BGCs and CRISPR-Cas defence islands recovered in this study were estimated to be complete due to improved assembly of the long reads⁴⁶ and represent the largest collection of complete BGCs from a MAG catalogue so far, which could facilitate the discovery of medically and industrially valuable biochemical compounds⁴⁷.

Previously undescribed HQ and MQ MAGs were recovered for the great majority of genera reported as constituting the core microbiome of different terrestrial habitats²⁷, thereby enabling further in-depth analysis of functional potential²⁰. The recovered MAGs can also be used to design targeted cultivation strategies to establish pure cultures of select microbial species⁴⁸. Furthermore, including MAGs from this study in taxonomically classifying the ~10,000 short-read Microflora Danica datasets increased median species-level classification from 17.3% to 36.8%, representing a substantial improvement in the ability to explore complex microbial communities at species level using short-read shotgun metagenomics. The considerable improvement in terrestrial metagenome classification also underscores the need for more localized metagenomics projects to acquire genomes of microbes unique to a particular environment or habitat type⁴⁹.

Most of the recovered MAGs from this study constitute previously undescribed microbial species or genera, which is a common finding of recent large-scale terrestrial microbiome studies^{33,35}, highlighting that each genome catalogue contributes substantially to characterizing the global microbiome. However, thousands of genera in terrestrial habitats still lack genomic representation, necessitating further genome recovery from complex environments. Although the addition of previously undescribed microbial lineages from this study occurred mainly at species or genus level, hundreds of recognized order or family level lineages were substantially expanded. As many microbial lineages are currently represented by a single placeholder MAG in GTDB, the expansion of these lineages is imperative to fill the gaps in the tree of life. This study also provides HQ MAGs for hundreds of GTDB lineages currently only represented by comparatively fragmented lower-quality MAGs. By proposing Latin names for microbial lineages under the SeqCode³⁸ using contiguous HQ MAGs as nomenclatural types, we help to address the contemporary issue of a rapidly growing number of unnamed microbial taxa in public databases⁵⁰. As microbial genome databases continuously improve³, the quality and not just the quantity of database additions should be emphasized. Hence, we anticipate this genome catalogue will serve as a valuable resource and template for gaining insights into the microbial ecology of the world's most complex environments.

Methods

Sample selection

Samples used in this project include terrestrial samples collected as part of the Microflora Danica sampling campaign²⁷. Briefly, bulk soil samples were collected using a weed extractor, which was cleaned with 70% ethanol before sampling, while also taking special care to avoid objects, such as sticks, leaves, grass and insects. The bulk sediment samples were collected using a gravity corer, followed by removal of any collected water or larger debris. A detailed description of the sample collection and processing is provided in the Microflora Danica study²⁷. All environmental samples used in this study were collected and handled in a responsible manner and in accordance with local laws.

Samples for deep, long-read sequencing were selected using the Microflora Danica shallow metagenome 16S rRNA gene observational tables, aggregated to the genus level (on the basis of classification to the Microflora Global 16S rRNA gene reference database) of 10,683 environmental samples²⁷. Initially, samples with >2 Gbp sequencing yield and with sample type of 'soil', 'sediment' or 'water' were selected to ensure that the picked samples are from environmental habitats and that the metagenomic-derived 16S rRNA gene profiles are adequately representative of the sample. Next, genera with at least 0.1% relative abundance and a minimum raw abundance (supporting read count) of 5 were counted, and samples that featured at least 75 of the selected genera with a combined relative abundance of 70% were further selected to omit samples that are mostly dominated by rare species or belong to a low-complexity metagenome. For the remaining samples, genera assigned with de novo taxonomy after classification to the Microflora Global 16S rRNA gene reference database²⁷ and featuring a minimum relative abundance of 0.2% as well as minimum raw abundance of 10 were counted, and 300 samples with the highest number of uncharacterized genera were selected to optimize the likelihood of recovering previously undescribed MAGs. The remaining samples were then manually curated to optimize for microbial diversity between the samples by omitting samples that overlap based on sampling location, or feature high overlapping genus counts with the rest of the selected samples.

DNA extraction and Nanopore sequencing

DNA from the selected environmental samples was extracted using the DNeasy PowerSoil Pro kit (QIAGEN, 47016), and the quality of the extracted DNA was evaluated using the NanoDrop One spectrophotometer (Thermo Fisher) and the Qubit dsDNA HS kit (Thermo Fisher, Q33231) with a Qubit 3.0 fluorometer (Thermo Fisher) to measure DNA concentration. The DNA was then prepared for sequencing using the SQK-LSK114 Ligation Sequencing kit (Oxford Nanopore), loaded into FLO-PRO114M Nanopore flow cells (Oxford Nanopore) and sequenced in 400 bps sequencing speed mode using either the P2 or the P24 (Supplementary Dataset 1) sequencers (Oxford Nanopore).

Read data processing

The raw Nanopore sequencing data were collected using the MinKnow software (v.22.07.4-23.04.5, Supplementary Dataset 1, <https://community.nanoporetech.com/downloads>) and basecalled with Guppy (v.6.2.1-6.5.7, Supplementary Dataset 1, <https://community.nanoporetech.com/downloads>) in super-accurate mode. Due to irreversible updates to the MinKnow software, some samples were sequenced with the 4 kHz sampling rate, while others were acquired using the 5 kHz rate (indicated in Supplementary Dataset 1). The sequenced reads were then split with duplex-tools (v.0.2.14, <https://github.com/nanoporetech/duplex-tools>) and trimmed using Porechop (v.0.2.3)⁵¹. Reads of Phred Quality score <7 or length <0.2 kbp were filtered out with NanoFilt (v.2.6.0)⁵². The split, trimmed and filtered Nanopore read summary statistics were acquired using NanoQ (v.0.10.0)⁵³.

MAG recovery with mmlong2

MAGs were recovered from the sequenced samples using a custom-developed mmlong2-lite metagenomics workflow v.1.0.2 (<https://github.com/Serka-M/mmlong2-lite>). Briefly, the mmlong2-lite metagenomics workflow v.1.0.2 is a Snakemake (v.7.26.0)⁵⁴ bioinformatics workflow that can take long reads (Nanopore or PacBio HiFi) and perform metagenome assembly, contig filtering, binning and initial MAG quality check. For Nanopore datasets, the reads are assembled into metagenomes using Flye (v.2.9.2)⁵⁵ with the '-meta' and '-nano-hq' options. Furthermore, the '-fmc' flag of mmlong2-lite controls the 'min_read_cov_cutoff' option of Flye, which can be increased to filter out more low-coverage contigs and thus speed up the metagenome assembly turnaround time. For this study, the '-fmc 8' option of the workflow was used with read datasets consisting of >50 Gbp of data to speed up the assembly.

The assembled Nanopore-only metagenomes were then polished with 1 round of Medaka (v.1.8.0, <https://github.com/nanoporetech/medaka>) to reduce the amount of indel errors in the initial assembly. Contigs <3 kbp were filtered out using SeqKit (v.2.4.0)⁵⁶ and the remaining contigs were then classified with Tiara (v.1.0.3)⁵⁷ to remove eukaryotic contigs from the assembly.

Before metagenomic binning, the 'assembly_info.txt' file outputted by Flye was used to extract circular contigs above the default length threshold of 250 kbp to be kept as separate bins. The remaining contigs were then used for iterative ensemble binning (Supplementary Fig. 2a) with MetaBAT2 (v.2.15)⁵⁸, SemiBin2 (v.1.5)⁵⁹, GraphMB (v.0.1.5)²⁴, and DAS Tool (v.1.1.3)⁶⁰ with the '-search_engine diamond' setting. For this study, multiple shallow metagenome read datasets (2,819 different samples) from the Microflora Danica study²⁷ were selected on the basis of overlapping genus-aggregated community profiles (specific samples indicated in Supplementary Dataset 2) and used as input for the workflow (with the '-cov' option) to perform multicoverage metagenomic binning for improved MAG recovery. The coverage profiles were generated by mapping the read datasets to the metagenome using Minimap2 (v.2.26)⁶¹ and SAMtools (v.1.16.1)⁶², followed by coverage calculation using the 'jgi_summarize_bam_contig_depths' function of MetaBAT2. The concatenated coverage profiles were then provided as input to MetaBAT2 and GraphMB, while for SemiBin2, the mapping files were provided directly.

After recovering the ensemble bins with DAS Tool, CheckM2 (v.1.0.2)⁶³ was used to acquire bin completeness and contamination metrics, followed by selection of bins meeting the requirements for HQ MAGs (>90% completeness, <5% contamination). The unselected contigs were then binned again using the same bidders and all HQ or MQ MAGs (>50% completeness, <10% contamination) were selected (Supplementary Fig. 2b). Ensemble binning of the unselected contigs was repeated for the third time with the genome quality score filtering feature of DAS Tool turned off as well as the use of a pre-trained binning model for SemiBin2 ('global' model by default, 'soil' for samples from this study), followed by retention of all HQ and MQ MAGs. The remaining unselected contigs were then binned one last time with MetaBAT2, and only HQ or MQ MAGs, as estimated by CheckM2, were kept as the final output of the MAG production workflow.

MAG relative abundance and coverage values were computed using CoverM (v.0.6.1)⁶⁴, while quality metrics were obtained by running Quast (v.5.2.0)⁶⁵ on the MAGs. Outputs from different tools were then aggregated to acquire a single dataframe with per-genome statistics.

The mmlong2-lite workflow is publicly available at <https://github.com/Serka-M/mmlong2-lite> and <https://zenodo.org/record/8013498>. The workflow is also the metagenomic binning component of the full mmlong2 pipeline (<https://github.com/Serka-M/mmlong2>), which supports automated MAG analyses that were omitted in this project to conserve computing resources.

MAG quality control and inspection

CheckM1 (v.1.2.2)⁶⁶ was run on all the recovered MAGs in lineage-specific workflow and all MAGs with <50% completeness or >10% contamination were omitted. The MAGs were then assigned a quality score using CheckM1 metrics as follows: completeness – (5 × contamination). MAGs with a quality score <30 were omitted. The remaining MAGs were then dereplicated using dRep (v.2.6.2)⁶⁷ with the following settings: ‘-comp 50’, ‘-con 10’, ‘-sa 0.95’, ‘-nc 0.4’. Furthermore, the MAGs were screened for tRNA genes with tRNAscan-SE (v.2.0.9)⁶⁸ using bacterial and archaeal models, while rRNA genes were detected with Barrnap (v.0.9, <https://github.com/tseemann/barrnap>) and Bakta (v.1.9.4)⁶⁹, which was run in metagenome mode.

Following the minimum information about metagenome-assembled genome (MIMAG) guidelines⁷⁰, MAGs were classified into HQ MAGs if they exhibited >90% completeness and <5% contamination estimates by CheckM2 (v.1.0.2)⁶³ while also featuring the 16S, 23S and 5S rRNA genes at least once, together with a minimum of 18 unique tRNA genes. MAGs not meeting these criteria but featuring >50% completeness and <10% contamination were classified as MQ MAGs. Only MIMAG HQ and MQ MAGs, as estimated by both CheckM1 and CheckM2, were used in this study (Supplementary Dataset 3).

Unless otherwise specified, MAG completeness, contamination and coding density values, as reported by CheckM2, were used in the plots and text. For cMAGs, dnaapler (v.1.1.0)⁷¹ was applied to verify that replication initiator genes are present in all recovered cMAGs.

MAG taxonomic classification

The recovered MAGs were classified with GTDB-Tk (v.2.4.0)⁷² against the GTDB R220 database using the ‘classify_wf’ workflow. The 16S rRNA sequences, which were extracted from the recovered MAGs, were classified with Usearch (v.11.0.667)⁷³ against the Microflora Global (v.1.0)²⁷ and SILVA (v.138.2)⁷⁴ 16S rRNA gene databases using the ‘-usearch_global -strand both -top_hit_only’ settings. The 16S rRNA taxonomic classification was considered species level if the top hit identity to the reference database sequence was ≥98.7%, genus level if ≥94.5%, family level if ≥86.5% and order level if ≥82.0% top hit identity.

Comparison of terrestrial habitats

To compare different soil habitats for MAG recovery at normalized sequencing depths, three sequenced samples per habitat for agricultural (MFD00392, MFD05176, MFD08497) and coastal (MFD02416, MFD05684, MFD01721) groups were randomly selected and subsampled to custom depths (20, 40, 60, 80, 100 Gbp) using Rasusa (v.2.0.0)⁷⁵, followed by MAG recovery with mm2long2-lite (v.1.0.2). Detection of eukaryotic sequences was performed by classifying the reads and assembled contigs with Kaiju (v.1.10.1)⁷⁶ using the ‘kaiju_db_nr_euk_2023-05-10’ database. Read and contig taxonomic profiling and species detection was done using Melon (v.0.2.0)⁷⁷. Variant detection in MAGs for microdiversity assessment was performed with Longshot (v.1.0.0)⁷⁸. Read *k*-mer counts were acquired using Jellyfish (v.2.2.10)⁷⁹, while general read and contig statistics were achieved with Nanoq (v.0.10.0)⁵³ and Cramino (v.0.14.1)⁸⁰. The code used for performing yield-normalized metagenomic comparisons is available at <https://github.com/Serka-M/mmcomp>.

For comparing microbial compositions between different habitats, metagenomic datasets from the short-read Microflora Danica study²⁷ were selected to include samples from coastal and agricultural habitats with at least 1 Gbp sequencing yield. To ensure similar sample counts per group, the agricultural samples were randomly subsetted to achieve up to 150 samples per habitat descriptor level 2. Next, the metagenomic-derived 16S rRNA profiles of the selected 1,046 samples were used to build a Bray–Curtis dissimilarity matrix, and the statistical significance of community composition differences was evaluated with the ANOSIM test (two-sided) using 999 permutations via vegan (v.2.6-6.1)⁸¹ in R (v.4.4.1)⁸². Principal coordinate decomposition of the dissimilarity matrix was performed with ape (v.5.8)⁸³.

Comparison of metagenomic binning workflows

The mm2long2 workflow was compared against similar metagenomic binning pipelines that feature support for Nanopore long-read assembly and binning (Supplementary Table 2). A test run with 100 Gbp of sample MFD02416 was used to recover MAGs with mm2long2-lite (v.1.0.2) using the settings ‘-sem soil’ and ‘-med r1041_e82_400bps_sup_g615’. A test run was also performed with Aviairy (v.0.11.0, <https://github.com/rhysnewell/aviary>) with the following settings: ‘aviary complete’, ‘-z ont_hq’, ‘-s 3000’, ‘-b 250000’, ‘-w recover_mags’, ‘-medaka-model r1041_e82_400bps_sup_g615’, ‘-skip-qc’, ‘-binning-only’. The SqueezeMeta workflow (v.1.6.5)⁸⁴ was also included by using the settings ‘-a flye’, ‘-m sequential’, ‘-contiglen 3000’, ‘-map minimap2-ont’, ‘-mapping_options -I 120 G -K 5 G’, ‘-binners concoct,metabat2,maxbin’, ‘-nocog’, ‘-nokegg’, ‘-nopfam’, ‘-test 15’. All test runs were performed using 100 CPUs to ensure comparable run times. The recovered MAGs from different workflows were then classified according to MIMAG guidelines, which included processing the MAGs from SqueezeMeta with CheckM2 while reusing the CheckM2 quality scores outputted by mm2long2 and Aviairy. HQ and MQ MAGs from different workflows were dereplicated to compare species-level genome recovery.

Comparison of MAG catalogues

Publicly available MAG catalogues were downloaded from the following studies that featured terrestrial MAGs and at least 1,000 dereplicated MAGs: SPIRE³³, RBC³⁴, GEM¹⁴, SMAG³⁵, TPMC³¹, OWC³². MAG quality assessment and quality filtering was performed in the same manner as with the MAGs recovered in this study. For the GEM catalogue, the full MAG dataset was downloaded and dereplication was performed separately to obtain non-redundant MAGs that were recovered in the study. For the SPIRE catalogue, genome entries from the proGenomes database⁸⁵ were omitted and the MAG catalogue was also dereplicated separately, as multiple instances of species-level redundancy were observed for the catalogue. For RBC, MAGs that were reported as representative of previously undescribed species by the authors were used.

MAGs from all catalogues were then annotated with Bakta (v.1.9.4)⁶⁹ and screened for defence islands via DefenseFinder (v.1.3.0)⁸⁶. Screening for secondary metabolites was performed with antiSMASH (v.7.1.0)⁸⁷ using the following options: ‘-cb-general’, ‘-cb-subclusters’, ‘-cb-knownclusters’, ‘-genefinding-tool prodigal-m’, ‘-asf’, ‘-pfam2go’, ‘-smcog-trees’, ‘-rre’, ‘-tfbs’. The secondary metabolite data were then parsed and aggregated into a dataframe via the ‘tabulate_regions.py’ script from multiSMASH (v.0.3.0, <https://github.com/zreitz/multi-smash>), and gene clusters located at the edge of the contig were considered potentially incomplete. Genomes from the catalogues and GTDB R220 were used for building reference databases to classify the Microflora Danica short-read datasets²⁷ (9,916 samples that had >1 Gbp yield) via Sylph (v.0.6.1)⁸⁸.

Phylogenomic analysis

Automated MAG phylogenetic assessment was performed using a custom pipeline available from <https://github.com/aaronmussig/mag-phylogeny>. Briefly, marker genes were extracted from the MAGs and aligned with the marker genes of GTDB R220 representative genomes using the ‘infer’ module of GTDB-Tk (v.2.4.0)⁷². The marker gene alignment was then used to build bacterial and archaeal genome trees via FastTree (v.2.1.11)⁸⁹ with the ‘WAG’ model and 100 bootstraps. RED values⁹⁰ for the recovered MAGs were determined with PhyloRank (v.0.1.12, <https://github.com/dparks1134/PhyloRank>) and phylogenetic classification was performed with the ‘summary_novelty_of_genomes.py’ script of the workflow. GTDB representative species classifications, lineage taxonomies and genome quality rankings were acquired from GTDB R220 metadata files. Phylogenies of highly divergent MFD-LR MAGs were manually inspected and curated.

Recovery and analysis of *Oederibacterium danicum* genome

Fragmented MAG MFD01231.bin.1.34, representing the *Oederibacterium danicum* lineage, was re-assembled into a cMAG by initially classifying the contigs of sample MFD01231 assembly using mmseqs2 (v.14.7e284)⁹¹ to the National Center for Biotechnology Information (NCBI) nr database (release 9 November 2023). Reads mapping to contigs, which were given the NCBI taxonomy of 'd_Bacteria' without any subordinate taxa, were then extracted using the 'view' and 'bam2fq' modules of SAMtools, and assembled with Flye using the '-meta-extra-params min_read_cov_cutoff=18' settings. The resulting assembly produced a linear contig, which could be linked to MFD01231.bin.1.34 through 16S rRNA gene sequence matching (rRNA prediction with Barrnap, alignment with Usearch). The reads mapping to the selected linear contig were then extracted using SAMtools 'view' and 'bam2fq' modules with '-q 20 -m 1000' settings, and assembled again via Flye with the '-extra-params min_read_cov_cutoff=26' option. The resulting circular contig (as reported by Flye) was polished with Medaka and contig circularity was further manually confirmed by inspecting read mappings to both ends of the contig. The initial and re-assembled genomes for *Oederibacterium danicum* were compared with Quast to check for irregularities.

Metabolic potential of the cMAG for *Oederibacterium danicum* was inferred through annotation with DRAM (v.1.4.6)⁹² (Supplementary Dataset 4) and the MicroScope Microbial Genome Annotation and Analysis Platform (v.3.17.3)⁹³. MicroScope classified the MAG as belonging to a Gram-negative population. Motility was inferred on the basis of extensive flagellar motility operons (for example, from K02391 in Supplementary Dataset 4). Nearly all genes involved in glycolysis (8/9) and the pentose phosphate (5/6) pathways were identified. Pyruvate oxidation to acetate was indicated by the presence of a gene for pyruvate:ferredoxin oxidoreductase (PFOR; KEGG orthology number K03737), positioned next to a group 3b NiFe hydrogenase operon⁹⁴ (for example, K04656 and syntenic region) and a gene for acetyl-CoA synthetase (ACS, K01895). A bacterial microcompartment for ethanolamine use⁹⁵ was identified on the basis of DRAM and MicroScope analysis (for example, K04021 and syntenic region). Anaerobic metabolism was inferred due to missing key tricarboxylic acid cycle and electron transport chain genes (for example, genes for succinate dehydrogenase and cytochrome c oxidase). In addition, genes for high-affinity cytochrome bd oxidase (K00426/K00425) were located next to a gene for superoxide dismutase (K04564), indicating the potential to combat oxidative stress⁹⁶. The ribulose monophosphate cycle was complete, including a gene for the key enzyme 3-hexulose-6-phosphate synthase (HPS-PHI, K13831). Several genes of the reductive acetyl-CoA or Wood–Ljungdahl pathway were also identified, but a gene for the key enzyme acetyl-CoA synthase was not detected. Due to the phylogenetic distinctiveness of the MAG, missing genes may result from homologues not currently in databases, or a variation of known C1 processing pathways⁹⁷.

Naming microbial lineages

HQ MAGs with 10 or less contigs that were unclassified in both GTDB R220 and Silva 138.2 were selected for naming under the SeqCode registry³⁸ (<https://registry.seqco.de/>). MAGs with 16S rRNA gene matches to undefined or placeholder taxonomic ranks in Silva 138.2 (for example, incertae sedis) were also included in the naming. Latinized genus names were derived from the names of Danish cities or parishes located within 15 km of the sampling sites. Whenever possible, different city or parish names were used to reduce name redundancy. The Latinized names were then given suffixes indicating a microbial lineage. The suffix *-bacter* was assigned only to lineages whose genomes contained the shape-determining *mreB* gene⁹⁸, while *-coccus* was used for lineages lacking *mreB*. Furthermore, the suffix *-monas* was proposed only for lineages with genomes containing flagellar genes, whereas *-plasma* was assigned to lineages lacking both *mreB* and genes associated with a peptidoglycan-containing cell wall (for example, genes for

peptidoglycan glycosyltransferases, peptidoglycan-binding protein LysM). If the named lineages featured parent taxa with placeholder names in GTDB, the genus names were used to generate Latinized replacements for the placeholders. However, if a parent taxon with a placeholder name featured any subordinate taxa with Latinized names, the replacements were not proposed due to disagreements in taxonomic opinion. Genera, which were proposed as nomenclatural types for phyla, were named in honour of the leading scientists of the Flora Danica project³⁷.

For species names, the species epithet that reflects the sample's environmental conditions (for example, sample type or habitat) was used whenever possible. If metadata-based naming was not possible, or to reduce name redundancy, generic species names (for example, *danicum*, *nordicum*) were assigned. Explanations for each genus and species name are provided in Supplementary Dataset 5, and the names are scheduled for manual registration under SeqCode post publication.

Statistics and reproducibility

No statistical method was used to predetermine sample size. No data were excluded from the analyses. The sequencing experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

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

Data availability

The sequencing read data (raw Nanopore data and basecalled reads), metagenomic assemblies and the recovered MAGs are available at ENA with BioProject ID: PRJEB58634. Individual accession numbers for raw nanopore data, basecalled reads and metagenomic assemblies can be found in Supplementary Dataset 1. Accession numbers for uploaded MAGs are available in Supplementary Dataset 3. Description of supplementary dataset column names is provided at <https://github.com/Serka-M/MFD-LR/blob/main/analysis/datasets/description.md>. Microflora Danica short-read sequencing data used in the study are available at NCBI with BioProject ID: PRJNA1071982.

Datasets used for plotting the figures are available in Zenodo at <https://doi.org/10.5281/zenodo.157822> (ref. 99). Relevant files too large to be hosted on GitHub are available in Zenodo at <https://zenodo.org/records/15064411> (ref. 100). The source data underlying the figures are provided with this paper.

The GTDB R220 database used for MAG taxonomy can be accessed at <https://data.ace.uq.edu.au/public/gtdb/data/releases/release220>. The Kaiju database used for contig taxonomy is accessible at <https://bioinformatics-centre.github.io/kaiju/downloads.html>. The NCBI nr database can be accessed at <https://ftp.ncbi.nlm.nih.gov/blat/db/>. The Microflora Global 16S rRNA gene reference database is available at <https://zenodo.org/records/15535748>. The SILVA 16S rRNA gene database can be accessed at https://www.arb-silva.de/no_cache/download/archive/release_138_2/.

The MAG catalogues used in this study are available as follows: MFD-SR: <https://zenodo.org/records/15535748>, SPIRE: <https://spire.embl.de/downloads>, TPMC: <https://download.cncb.ac.cn/bigd/TPMC/>, OWC: <https://zenodo.org/records/8194033>, SMAG: <https://zenodo.org/records/8223844>, GEM: <https://portal.nersc.gov/GEM>, RBG: <https://connectquedu.sharepoint.com/:f/s/BinChickensupplementarydata/ErJbGzlvzglMoLSAljv3dtlBaEmqdzXoUHRYQIYYbtSY5Q?e=alvFdH>.

Code availability

Code and datasets used for plotting the figures are available in Zenodo at <https://doi.org/10.5281/zenodo.157822> (ref. 99). The MAG recovery workflow used in the study is available in Zenodo at <https://doi.org/10.5281/zenodo.15782530> (ref. 101) and

<https://doi.org/10.5281/zenodo.15782610> (ref. 102). The yield-normalized comparative metagenomics workflow is available in Zenodo at <https://doi.org/10.5281/zenodo.15782326> (ref. 103). The MAG phylogeny workflow is available in Zenodo at <https://doi.org/10.5281/zenodo.15782786> (ref. 104).

References

- Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl Acad. Sci. USA* **113**, 5970–5975 (2016).
- Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z. & Ettema, T. J. G. Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.* **19**, 225–240 (2021).
- 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 (2021).
- Imachi, H. et al. Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature* **577**, 519–525 (2020).
- Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *mSystems* **3**, e00055–18 (2018).
- Tyson, G. W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
- Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
- Quince, C. et al. STRONG: metagenomics strain resolution on assembly graphs. *Genome Biol.* **22**, 214 (2021).
- Orakov, A. et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
- Chen, Y.-H. et al. Salvaging high-quality genomes of microbial species from a meromictic lake using a hybrid sequencing approach. *Commun. Biol.* **4**, 996 (2021).
- Stewart, R. D. et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
- Dmitriyeva, M. et al. The mOTUs online database provides web-accessible genomic context to taxonomic profiling of microbial communities. *Nucleic Acids Res.* **53**, D797–D805 (2024).
- 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).
- Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
- Delmont, T. O. et al. Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Front. Microbiol.* **6**, 358 (2015).
- Alteio, L. V. et al. Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. *mSystems* **5**, e00768–19 (2020).
- Howe, A. C. et al. Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl Acad. Sci. USA* **111**, 4904–4909 (2014).
- Riley, R. et al. Terabase-scale coassembly of a tropical soil microbiome. *Microbiol. Spectr.* **11**, e0020023 (2023).
- White, R. A. et al. Molecule long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. *mSystems* **1**, e00045–16 (2016).
- Singleton, C. M. et al. Connecting structure to function with the recovery of over 1,000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.* **12**, 2009 (2021).
- Kim, C. Y., Ma, J. & Lee, I. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nat. Commun.* **13**, 6367 (2022).
- Pan, S., Zhu, C., Zhao, X.-M. & Coelho, L. P. A deep Siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* **13**, 2326 (2022).
- Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
- Lamurias, A., Sereika, M., Albertsen, M., Hose, K. & Nielsen, T. D. Metagenomic binning with assembly graph embeddings. *Bioinformatics* **38**, 4481–4487 (2022).
- Beaulaurier, J. et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* **36**, 61–69 (2018).
- Heidelberg, S. et al. Nanomotif: identification and exploitation of DNA methylation motifs in metagenomes using Oxford nanopore sequencing. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.04.29.591623> (2024).
- Singleton, C. M. et al. Microflora Danica: the atlas of Danish environmental microbiomes. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.06.27.600767> (2024).
- Rath, K. M., Fierer, N., Murphy, D. V. & Rousk, J. Linking bacterial community composition to soil salinity along environmental gradients. *ISME J.* **13**, 836–846 (2019).
- Mo, Y. et al. Agricultural practices influence soil microbiome assembly and interactions at different depths identified by machine learning. *Commun. Biol.* **7**, 1349 (2024).
- Peng, Z. et al. The neglected role of micronutrients in predicting soil microbial structure. *npj Biofilms Microbiomes* **8**, 103 (2022).
- Cheng, M. et al. A genome and gene catalog of the aquatic microbiomes of the Tibetan Plateau. *Nat. Commun.* **15**, 1438 (2024).
- Oliverio, A. M. et al. Mapping the soil microbiome functions shaping wetland methane emissions. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.02.06.579222> (2024).
- Schmidt, T. S. B. et al. SPIRE: a Searchable, Planetary-scale microbiome REsource. *Nucleic Acids Res.* **52**, D777–D783 (2024).
- 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. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.11.24.625082> (2024).
- Ma, B. et al. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.* **14**, 7318 (2023).
- Orita, I. et al. The archaeon *Pyrococcus horikoshii* possesses a bifunctional enzyme for formaldehyde fixation via the ribulose monophosphate pathway. *J. Bacteriol.* **187**, 3636–3642 (2005).
- Knudsen, H. *The Story Behind Flora Danica* (Lindhardt og Ringhof, 2016).
- Hedlund, B. P. et al. SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nat. Microbiol.* **7**, 1702–1708 (2022).
- Ahmed, S. et al. How biotic, abiotic, and functional variables drive belowground soil carbon stocks along stress gradient in the Sundarbans Mangrove Forest? *J. Environ. Manage.* **337**, 117772 (2023).
- Riddley, M. et al. Differential roles of deterministic and stochastic processes in structuring soil bacterial ecotypes across terrestrial ecosystems. *Nat. Commun.* **16**, 2337 (2025).
- Chauhan, G., Arya, M., Kumar, V., Verma, D. & Sharma, M. An improved protocol for metagenomic DNA isolation from low microbial biomass alkaline hot-spring sediments and soil samples. *3 Biotech* **14**, 34 (2024).

42. Simon, S. A. et al. Dancing the Nanopore limbo – Nanopore metagenomics from small DNA quantities for bacterial genome reconstruction. *BMC Genomics* **24**, 727 (2023).
43. Koren, S. & Phillippy, A. M. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **23**, 110–120 (2015).
44. Robeson, M. S. et al. RESCRIPT: reproducible sequence taxonomy reference database management. *PLoS Comput. Biol.* **17**, e1009581 (2021).
45. McDonald, D. et al. Greengenes2 unifies microbial data in a single reference tree. *Nat. Biotechnol.* **42**, 715–718 (2024).
46. Sánchez-Navarro, R. et al. Long-read metagenome-assembled genomes improve identification of novel complete biosynthetic gene clusters in a complex microbial activated sludge ecosystem. *mSystems* **7**, e00632-22 (2022).
47. Chen, J. et al. Global marine microbial diversity and its potential in bioprospecting. *Nature* **633**, 371–379 (2024).
48. Cross, K. L. et al. Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat. Biotechnol.* **37**, 1314–1321 (2019).
49. Anthony, W. E. et al. From soil to sequence: filling the critical gap in genome-resolved metagenomics is essential to the future of soil microbial ecology. *Environ. Microbiome* **19**, 56 (2024).
50. Pallen, M. J., Rodriguez-R, L. M. & Alikhan, N.-F. Naming the unnamed: over 65,000 Candidatus names for unnamed Archaea and Bacteria in the Genome Taxonomy Database. *Int. J. Syst. Evol. Microbiol.* **72**, 005482 (2022).
51. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Y. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* **3**, e000132 (2017).
52. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
53. Steinig, E. & Coin, L. Nanoq: ultra-fast quality control for nanopore reads. *J. Open Source Softw.* **7**, 2991 (2022).
54. Mölder, F. et al. Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
55. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
56. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
57. Karlicki, M., Antonowicz, S. & Karnkowska, A. Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics* **38**, 344–350 (2022).
58. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
59. 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).
60. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
61. Li, H. Minimap and minimap: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
62. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
63. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
64. Aroney, S. T. N. et al. CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**, btaf147 (2025).
65. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
66. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
67. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
68. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* https://doi.org/10.1007/978-1-4939-9173-0_1 (2019).
69. Schwengers, O. et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genom.* **7**, 000685 (2021).
70. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
71. Bouras, G., Grigson, S. R., Papudeshi, B., Mallawaarachchi, V. & Roach, M. J. Dnaapler: a tool to reorient circular microbial genomes. *J. Open Source Softw.* **9**, 5968 (2024).
72. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
73. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
74. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
75. Hall, M. B. Rasusa: randomly subsample sequencing reads to a specified coverage. *J. Open Source Softw.* **7**, 3941 (2022).
76. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).
77. Chen, X. et al. Melon: metagenomic long-read-based taxonomic identification and quantification using marker genes. *Genome Biol.* **25**, 226 (2024).
78. Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019).
79. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
80. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
81. Oksanen, J. et al. vegan: community ecology package. Ordination methods, diversity analysis and other functions for community and vegetation ecologists. <https://doi.org/10.32614/CRAN.package.vegan> (2016).
82. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
83. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
84. Tamames, J. & Puente-Sánchez, F. SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front. Microbiol.* **9**, 3349 (2018).
85. Fullam, A. et al. proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Res.* **51**, D760–D766 (2023).

86. Tesson, F. et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).
87. Blin, K. et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50 (2023).
88. Shaw, J. & Yu, Y. W. Rapid species-level metagenome profiling and containment estimation with sylph. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-024-02412-y> (2024).
89. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
90. 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).
91. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
92. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
93. Vallenet, D. et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Res.* **48**, D579–D589 (2020).
94. Greening, C. et al. Minimal and hybrid hydrogenases are active from archaea. *Cell* **187**, 3357–3372.e19 (2024).
95. Pokhrel, A., Kang, S.-Y. & Schmidt-Dannert, C. Ethanolamine bacterial microcompartments: from structure, function studies to bioengineering applications. *Curr. Opin. Microbiol.* **62**, 28–37 (2021).
96. Das, A., Silaghi-Dumitrescu, R., Ljungdahl, L. G. & Kurtz, D. M. Cytochrome bd oxidase, oxidative stress, and dioxygen tolerance of the strictly anaerobic bacterium *Moorella thermoacetica*. *J. Bacteriol.* **187**, 2020–2029 (2005).
97. Zhuang, W.-Q. et al. Incomplete Wood–Ljungdahl pathway facilitates one-carbon metabolism in organohalide-respiring *Dehalococcoides mccartyi*. *Proc. Natl Acad. Sci. USA* **111**, 6419–6424 (2014).
98. Figge, R. M., Divakaruni, A. V. & Gober, J. W. MreB, the cell shape-determining bacterial actin homologue, co-ordinates cell wall morphogenesis in *Caulobacter crescentus*. *Mol. Microbiol.* **51**, 1321–1332 (2004).
99. Sereika, M. Repository for Microflora Danica long-read (MFD-LR) MAGs (1.0.0). *Zenodo* <https://doi.org/10.5281/zenodo.15782215> (2025).
100. Sereika, M. Supplementary data for MFD-LR study (1.0.). Data set. *Zenodo* <https://doi.org/10.5281/zenodo.15064411> (2025).
101. Sereika, M. Code for mmlong2-lite: lightweight bioinformatics pipeline for microbial genome recovery (1.1.0). *Zenodo* <https://doi.org/10.5281/zenodo.15782531> (2025).
102. Sereika, M. Code for mmlong2: bioinformatics pipeline for recovery and analysis of metagenome-assembled genomes (1.1.0). *Zenodo* <https://doi.org/10.5281/zenodo.15782610> (2025).
103. Sereika, M. Code for mmcomp: snakemake workflow for yield-normalized comparative genome-centric metagenomics (0.0.1). *Zenodo* <https://doi.org/10.5281/zenodo.15782326> (2025).
104. Mussig, A. & Sereika, M. Code for mag-phylogeny: a pipeline to infer novelty of genomes using the GTDB framework (1.0.0). *Zenodo* <https://doi.org/10.5281/zenodo.15782786> (2025).

Acknowledgements

This study was funded by research grants from Poul Due Jensen Foundation (grant Microflora Danica to M.A. and P.H.N.), Villum Foundation (grants 130690 and 50093 to M.A.) and the European

Union (ERC grant 101078234 to M.A.). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We thank the Microflora Danica Consortium for their contributions to sample and metadata collection across Denmark; S. R. Bielidt, R. H. Kirkegaard and K. S. Andersen for maintaining the laboratory and computational infrastructure used during the study.

Author contributions

M.A., P.H.N. and M.S. designed the study. T.B.N.J., E.A.S. and Y.Y. contributed to the generation or processing of the short-read shotgun metagenome or 16S amplicon sequencing data used in the study. T.B.N.J., V.R.J. and F.D. performed curation and validation of the sample metadata. M.S., C.M.S., K.S.K. and F.P. performed sample selection for sequencing. M.S. and C.J. performed sample DNA extraction and Nanopore sequencing. M.S. carried out the long-read sequencing data processing, MAG recovery, MAG analysis and writing of the initial paper. A.J.M., P.H. and M.S. performed phylogenetic analysis of the MAGs. P.H.N., P.H., M.A., F.P. and M.S. proposed names for select microbial lineages. C.M.S. and K.S.K. conducted metabolic reconstruction of the *Oederibacterium danicum* genome. All authors reviewed the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41564-025-02062-z>.

Correspondence and requests for materials should be addressed to Mads Albertsen.

Peer review information *Nature Microbiology* thanks Joseph Nesme, Joshua Quick and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025, corrected publication 2025

Reporting Summary

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

Statistics

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

n/a Confirmed

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection MinKNOW software v22.07.4-23.04.5 (Oxford Nanopore, UK), Guppy v6.2.1-6.5.7 (Oxford Nanopore, UK)

Data analysis duplex-tools v0.2.14 (Oxford Nanopore, England), Porechop v0.2.3, NanoFilt v2.6.0, NanoQ v0.10.0, mmlong2-lite v1.0.2, Snakemake v7.26.0, Flye v2.9.2, Minimap2 v2.26, SAMtools v1.16.1, Medaka v1.8.0 (Oxford Nanopore, England), SeqKit v2.4.0, Tiara v1.0.3, Aviary v0.11.0, SqueezeMeta v1.6.5, MetaBAT2 v2.15, SemiBin2 v1.5, GraphMB v0.1.5, DAS Tool v1.1.3, CoverM v0.6.1, Quast v5.2.0, CheckM1 v1.2.2, CheckM2 v1.0.2, dRep v2.6.2, tRNAseScan-SE v2.0.9, Barrnap v0.9, Bakta v1.9.4, dnaapl v1.1.0, Rasusa v2.0.0, Kaiju v1.10.1, Melon v0.2.0, Longshot v1.0.0, Jellyfish v2.2.10, Cramino v0.14.1, Usearch v11.0.667, GTDB-Tk v2.4.0, FastTree v2.1.11, PhyloRank v0.1.12, Sylph v0.6.1, DefenseFinder v1.3.0, mmseqs2 v.14.7e284, DRAM v1.4.6, antiSMASH v7.0.1, multiSMASH v0.3.0, MicroScope v3.17.3, R v4.4.1, vegan v2.6-6.1, ape v5.8, SeqCode Registry: <https://registry.seqco.de/>

Code written as part of the study:
<https://github.com/Serka-M/MFD-LR>
<https://github.com/Serka-M/mmlong2>
<https://github.com/Serka-M/mmlong2-lite>
<https://github.com/Serka-M/mmcomp>
<https://github.com/aaronmussig/mag-phylogeny>

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

The sequencing read data and the recovered MAGs are available at ENA bio project ID: PRJEB58634.

Individual accession numbers for raw nanopore data, basecalled reads and metagenomic assemblies can be found in Supplementary dataset 1 of the paper.

Accession numbers for uploaded MAGs are available at Supplementary dataset 3.

Description of supplementary dataset column names is provided at <https://github.com/Serka-M/MFD-LR/blob/main/analysis/datasets/description.md>.

Microflora Danica short-read sequencing data used in the study is available at NCBI with BioProject ID: PRJNA1071982.

Datasets used for plotting the figures are available at <https://github.com/Serka-M/MFD-LR>.

Relevant files too large to be hosted on GitHub are available at <https://zenodo.org/records/15064411>.

GTDB-tk database used for MAG taxonomy can be accessed at <https://data.ace.uq.edu.au/public/gtdb/data/releases/release220>.

NCBI nr database can be accessed at <https://ftp.ncbi.nlm.nih.gov/blast/db>.

Kaiju database is available at <https://bioinformatics-centre.github.io/kaiju/downloads.html>.

The Microflora Global 16S rRNA database is available at <https://zenodo.org/records/15535748>.

Silva 16S rRNA database can be accessed at https://www.arb-silva.de/no_cache/download/archive/release_138_2/.

Different MAG catalogues used in this study is available as follows:

MFD-SR — <https://zenodo.org/records/15535748>,

SPIRE — <https://spire.embl.de/downloads>,

TPMC — <https://download.cncb.ac.cn/bigd/TPMC>,

RBG — <https://connectquedu.sharepoint.com/:f:/s/BinChickensupplementarydata/Er.bGzlvzglMoLSAljv3dtlBaEmqdZXoUHRYRQIYYbtSY5Q?e=alvFdH>,

OWC — <https://zenodo.org/records/8194033>,

SMAG — <https://zenodo.org/records/8223844>,

GEM — <https://portal.nersc.gov/GEM>.

Research involving human participants, their data, or biological material

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

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

Population characteristics

Recruitment

Ethics oversight

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size

Data exclusions

Replication	Sequencing data analyses were performed programmatically, using custom scripts and workflows, which are publically available, thus ensuring reproducibility of the current study
Randomization	The sequenced samples were selected to optimize for microbial diversity and reduce potential redundancy (e.g. overlapping geographic location, community profiles) between the samples. Control for covariates was not applicable, as the primary aim was to descriptively catalogue microbial genomes and expand reference database representation. All subset analyses were conducted using random sampling to ensure unbiased representation.
Blinding	Blinding was not relevant to this study, as sequencing data processing and genome recovery was performed irrespective of the sample characteristics

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Plants

Seed stocks	Not applicable
Novel plant genotypes	Not applicable
Authentication	Not applicable