Ultra-deep long-read metagenomics captures diverse taxonomic and biosynthetic potential of soil microbes

Caner Bagci[1,2, 5], Timo Negri[1], Elena Buena Atienza[3,4], Caspar Gross[2,3,4], Stephan Ossowski[2,3,4], Nadine Ziemert[1,2,5*]

[1]Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), University of Tübingen, Sand 14, Tübingen, 72076, Baden-Württemberg, Germany.
[2]Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Auf der Morgenstelle 24, Tübingen, 72076, Baden-Württemberg, Germany.
[3]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Calwerstraße 7, Tübingen, 72076, Baden-Württemberg, Germany.
[4]NGS Competence Center Tübingen University of Tübingen, Calwerstraße 7, Tübingen, 72076, Baden-Württemberg, Germany.
[5]Partner Site Tübingen, German Center for Infection Research (DZIF), Auf der Morgenstelle 28, Tübingen, 72076, Baden-Württemberg, Germany.

*Corresponding author(s). E-mail(s): nadine.ziemert@uni-tuebingen.de; Contributing authors: caner.bagci@uni-tuebingen.de; timo.negri@uni-tuebingen.de; Elena.Buena-Atienza@med.uni-tuebingen.de; Caspar.Gross@med.uni-tuebingen.de; stephan.ossowski@uni-tuebingen.de;

Abstract

Background Soil ecosystems have long been recognized as hotspots of microbial diversity, but most estimates of their complexity remain speculative, relying on limited data and extrapolation from shallow sequencing. Here, we revisit this question using one of the deepest metagenomic sequencing efforts to date, applying 148 Gbp of Nanopore long-read and 122 Gbp of Illumina short-read data to a single forest soil sample.

Results Our hybrid assembly reconstructed 837 metagenome-assembled genomes (MAGs), including 466 high- and medium-quality genomes, nearly all lacking close relatives among cultivated taxa. Rarefaction and k-mer analyses reveal that, even at this depth, we capture only a fraction of the extant diversity: nonparametric models project that over 10 Tbp would be required to approach saturation. These findings offer a quantitative, technology-enabled update to long-standing diversity estimates and demonstrate that conventional metagenomic sequencing efforts likely miss the majority of microbial and biosynthetic potential in soil. We further identify over 11,000 biosynthetic gene clusters (BGCs), >99% of which have no match in current databases, underscoring the breadth of unexplored metabolic capacity.

Conclusions Taken together, our results emphasize both the power and the present limitations of metagenomics in resolving natural microbial complexity, and they provide a new baseline for evaluating future advances in microbial genome recovery, taxonomic classification, and natural product discovery.

Keywords: microbiome, soil, metagenomics, nanopore, ultra-deep, diversity

1 Background

Soil is one of the most biologically diverse ecosystems on Earth, hosting a massive and largely uncharacterized diversity of microbial life. These microbial communities play critical roles in biogeochemical cycling, soil formation, nutrient turnover, plant health, and climate regulation [1]. Despite their ecological and biotechnological importance, the full extent and functional potential of soil microbiomes remain poorly understood.

A single gram of soil can contain up to 10^9 microbial cells and hundreds of thousands of species, spanning all domains of life [2, 3]. These communities are not only taxonomically complex but also exhibit highly uneven abundance distributions, where rare taxa - collectively termed the “rare biosphere” - may be functionally consequential despite their low abundance [4].

Although the vast diversity of soil microbes has been widely acknowledged for decades, most estimates of their richness are based on indirect approaches, including 16S rRNA gene surveys, limited shotgun datasets, or predictive modeling. As a result, foundational claims about the extent of soil microbial diversity are frequently cited but rarely re-examined with contemporary data. Much of the soil microbiome remains part of the so-called microbial dark matter: lineages with no cultured representatives and no reference genomes [5–7].

Metagenomics has enabled the cultivation-independent study of microbial communities, allowing for genome-resolved insights into uncultured organisms [8–10]. However, most metagenomic studies rely on short-read sequencing, which presents challenges in assembling highly complex, strain-rich communities like soil. These limitations hinder the recovery of low-abundance taxa, reduce the completeness of biosynthetic gene clusters, and often lead to fragmented or ambiguous assemblies [9]. Recent advances in long-read sequencing, particularly with Oxford Nanopore Technologies (ONT), have opened new opportunities for soil microbiome research. Long reads can span repeat-rich regions and operons, improving genome recovery and assembly contiguity [11–13]. Hybrid approaches combining long and short reads can further mitigate the high error rates of long reads while leveraging their structural advantages [14, 15]. Despite these improvements, it remains unclear how far even the most advanced sequencing technologies can go in capturing the full taxonomic and functional diversity of soil.

This study addresses this knowledge gap using one of the deepest metagenomic datasets generated from a single soil sample to date, combining 148 Gbp of ONT long-read and 122 Gbp of Illumina short-read data. We use this dataset to empirically assess the power and current limitations of metagenomics in resolving complex microbial communities. Specifically, we ask: How much taxonomic and biosynthetic diversity can

be recovered from a single soil sample using state-of-the-art sequencing? What proportion of this diversity is represented in existing databases? Can ultra-deep sequencing approach saturation, or are we still only scratching the surface? By addressing these questions, we aim to provide a data-driven reassessment of microbial diversity in soil and to establish a benchmark for future metagenomic studies of complex ecosystems.

2 Results

2.1 Sequencing Summary and Data Quality

To assess how deeply current sequencing technologies can probe soil microbial communities, we generated an ultra-deep metagenomic dataset from a single sample collected in the Schönbuch forest, a temperate mixed forest in southwestern Germany. The sample was taken from the A horizon of a Cambisol soil, which is known for its rich microbial diversity and favorable characteristics for DNA extraction. This particular site and soil type were previously shown to harbor high biosynthetic potential and taxonomic richness [16], and were therefore selected as a representative model system for deep metagenomic exploration.

We extracted total environmental DNA from this soil and sequenced it using a hybrid strategy that combined 148 Gbp of Oxford Nanopore long-read and 122 Gbp of Illumina short-read sequencing. Nanopore reads had a read length N50 of 12.2 kb, while Illumina reads achieved a mean Q-score of 35. Together, this 270 Gbp dataset represents one of the deepest single-sample soil metagenomes to date and provides a unique opportunity to empirically assess taxonomic and functional complexity.

2.2 Assembly and Recovery of Metagenome-Assembled Genomes

We employed a hybrid assembly strategy to combine the complementary strengths of long and short reads. Long-read data were assembled and polished using Illumina

data, resulting in 10.5 Gbp of assembled sequence with a contig N50 of 77.8 kb. For comparison, an Illumina-only assembly yielded a more fragmented assembly with an N50 of 865 bp and a total length of 9.2 Gbp, confirming the advantages of long-read inclusion.

Additionally, we recovered 837 metagenome-assembled genomes (MAGs) from the hybrid assembly, of which 466 met the MIMAG high- or medium-quality criteria [17]. All MAGs recovered represented a unique species (> 5% Mash distance). These MAGs spanned a wide range of bacterial phyla typical of soils, including Acidobacteriota, Pseudomonadota, Verrucomicrobiota, Actinomycetota, and Myxococcota. Remarkably, only 4 of the 837 MAGs had species-level matches (>= 95% ANI) in the GTDB database, and none had cultured representatives. Additionally, 13 MAGs belong to phyla with no cultured members, including Desulfobacterota, and CPR phyla FCPU426 and JAJYCY01, harboring between 2 and 19 BGCs. At other levels, 9, 45, 155, and 611 MAGs belong to novel lineages that correspond to class, order, family, and genus levels, respectively. The genome sizes of HQ and MQ MAGs range from 1 Mbp (Patescibacteria) to 15 Mbp (Planctomycetota) (Figure 1 and Supplementary Material S2).

Figure 1 illustrates a phylogenetic tree constructed from the recovered MAGs, also showing their phylogenetic novelty levels. Many of the MAGs that we recovered constitute novel families or even classes, with no currently known species, highlighting the vast unexplored microbial diversity in soil ecosystems. Additionally, several MAGs harbor a large number of biosynthetic gene clusters (BGCs) - up to 30 - suggesting a high capacity for the production of diverse and novel secondary metabolites.

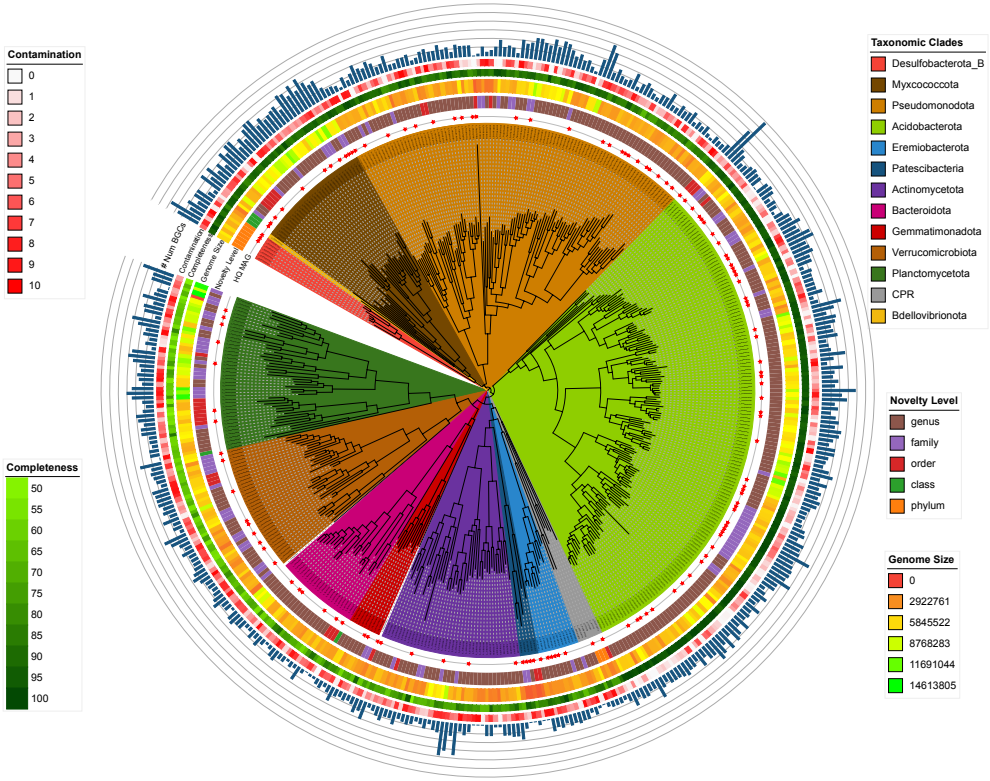

Fig. 1: Genome quality statistics and taxonomic classification of newly reconstructed high-quality and medium-quality metagenome-assembled genomes (MAGs). Each leaf represents a MAG, colored by its assigned phylum. Concentric rings around the leaves indicate, from innermost to outermost: (i) HQ MAGs, (ii) novelty level (genus to phylum), (iii) genome size (bp), (iv) completeness (%), (v) contamination (%), and (vi) the number of BGCs encoded by the MAG (range: 0-30).

2.3 Taxonomic analysis reveals unprecedented levels of novelty

Taxonomic classification of the raw reads and assembled contigs revealed a highly diverse microbial community, comprising members from 69 different phyla. As expected for soil environments, the community was dominated by taxa such as Pseudomonodata and Acidobacteriota (Figure 2A), along with representatives from many other phyla. Approximately 1% of the reads were of eukaryotic origin, and 8% of the

reads could not be classified at all, indicating the presence of deeply novel sequences not represented in current databases.

Figure 2B illustrates the taxonomic origin of assembled contigs and highlights the high fraction of sequences from uncultivated organisms across all taxonomic ranks. While soil has long been recognized as a reservoir of uncultivated diversity, with historical estimates suggesting 90–99% of taxa remain uncultured, our data now provide quantitative, genome-resolved confirmation of this. At the species level, over 99% of the assembled contigs could not be assigned to a taxon with any cultured representative. Only 0.7% of contigs could be linked to species that include at least one cultured strain, and just 16.7% could be linked to a cultured genus.

Even at broader taxonomic levels, substantial novelty persists: 9% of contigs could not be assigned to a known class, and 6.2% to a known phylum. Moreover, for those contigs that were classified, the majority belong to taxa composed exclusively of uncultivated representatives, often known only from metagenomic or single-cell sequencing. These patterns are in contrast to metagenomic datasets from more tractable environments, such as the human gut or ocean microbiomes, where a larger fraction of genomes can be linked to cultivated taxa [18, 19]. Despite the expansion of public genome resources like GTDB [20] and the GEM catalog [9], fewer than five of the 837 MAGs recovered here share >=95% ANI with known species. This confirms that soil remains one of the most genomically undercharacterized microbial ecosystems.

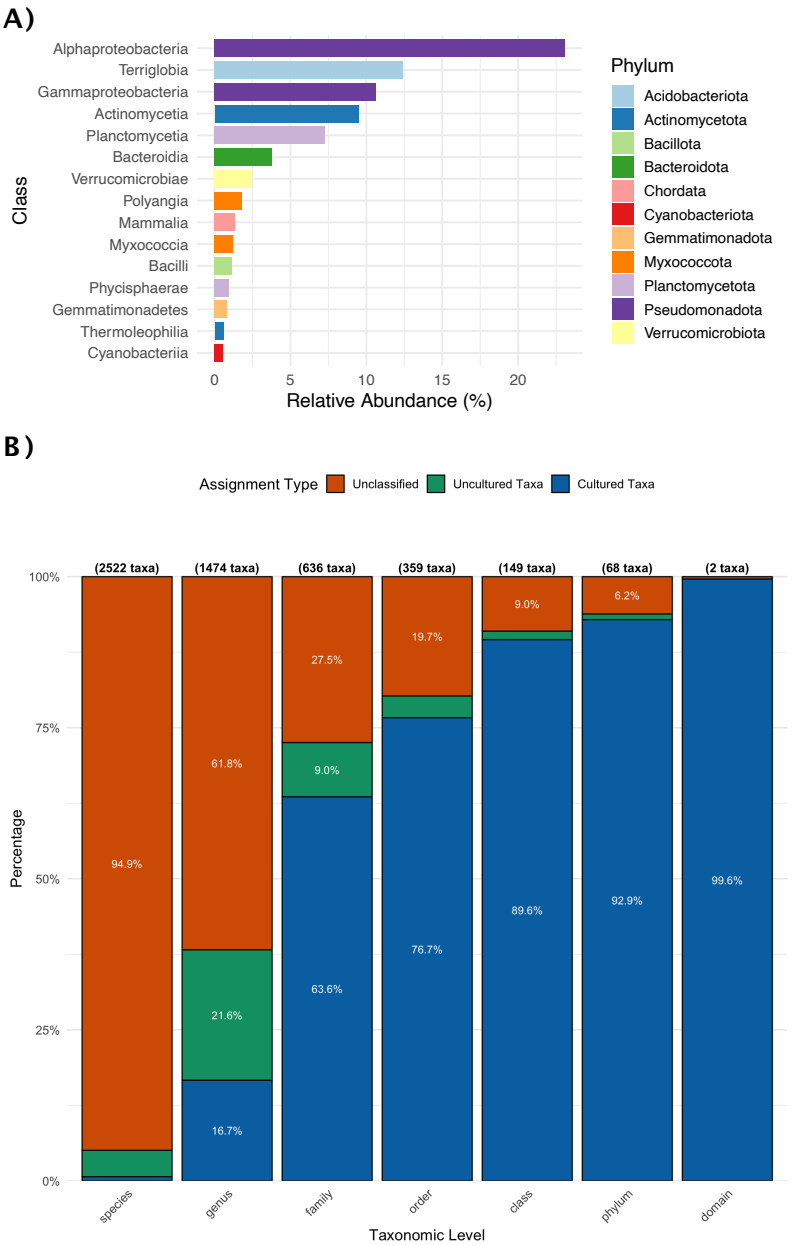

Fig. 2: Taxonomic summary of the assembled contigs. **A)** Class-level relative abundances of the assembled contigs, coloured by the phyla they belong to. **B)** Taxonomic novelty levels of the assembled contigs across different ranks. Bars represent the proportion of contigs assigned to cultured taxa (blue), uncultured taxa (only observed in other metagenome studies) (green), or remaining unclassified (completely novel) (orange) at each taxonomic level, from species to domain. The number of contigs evaluated at each level is shown above each bar. Most contigs could only be classified at higher taxonomic ranks, while species- and genus-level assignments were predominantly unclassified.

2.4 Sequencing depth and coverage limitations

To evaluate whether ultra-deep sequencing could fully capture the microbial diversity of the sample, we conducted k-mer analysis, rarefaction of marker genes, and 16S rRNA gene based profiling. All approaches consistently indicated that the sequencing depth, while unprecedented, was still insufficient to reach diversity saturation.

K-mer analysis, which examines the frequency distribution of short subsequences within the reads, revealed a pronounced left-skewed distribution. Most k-mers occurred only once or a few times, indicating a high prevalence of unique or rare sequences. This pattern is characteristic of highly diverse communities and reflects the presence of many low-abundance taxa whose genomes were either incompletely covered or missed entirely in assembly. The right-hand tail of the distribution corresponded to high-frequency k-mers, likely originating from abundant taxa or repetitive genomic regions. This strong skewness, observed in both Nanopore and Illumina datasets, underscores the immense sequence complexity of the sample (Figure 3A, and Supplementary Figures 1 and 2).

Rarefaction analysis using Nonpareil further supported this conclusion. Based on the redundancy of reads, the model predicted that approximately 10 Tbp of sequencing data would be required to achieve near-complete coverage of the microbial diversity in this single soil sample — a nearly 50-fold increase over our already deep dataset (Figure 3B). These results quantitatively highlight the limitations of even the most comprehensive sequencing efforts to date.

In parallel, we clustered full-length 16S rRNA gene sequences extracted from Nanopore reads into operational taxonomic units (OTUs) at various sequence identity thresholds. At 100% and 97% identity, the number of observed OTUs continued to increase linearly with sequencing depth, showing no indication of saturation (Figure 4A). While higher-order groupings (e.g., family or phylum) began to plateau,

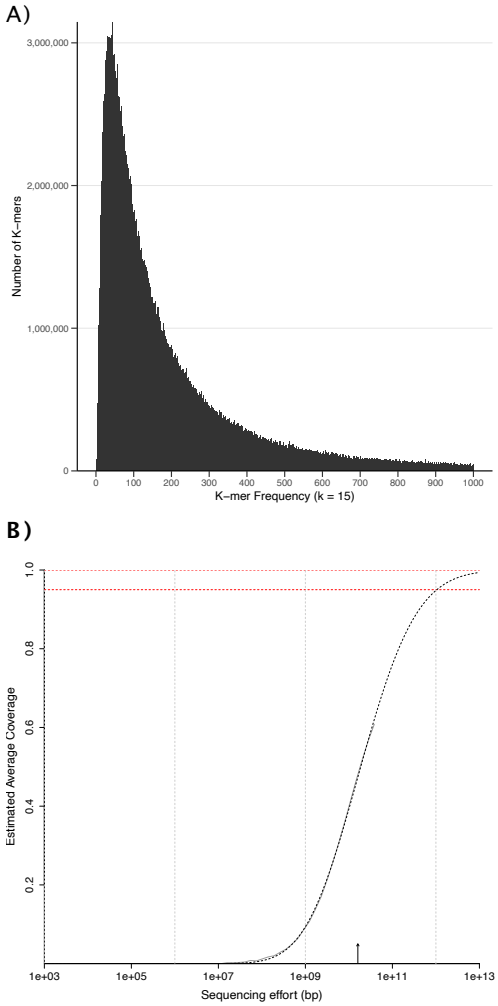

Fig. 3: Sequencing depth diagnostics show that ultra-deep sequencing still undersamples the total soil diversity. **A)** 15-mer spectrum of Nanopore reads from the soil metagenome. Bars show the number of distinct 15-mers observed at each copy number. The left-skewed peak at low frequencies reflects a very large pool of unique or low-coverage k-mers, showing extreme sequence diversity. The long right-hand tail originates from higher coverage k-mers, possibly from abundant taxa and repetitive genomic regions. **B)** Nonpareil curve estimating the sequencing coverage of the combined Illumina dataset. The curve shows the relationship between sequencing effort and estimated coverage, with the solid line representing the observed dataset and the dashed extrapolation indicating expected gains with additional sequencing. The curve suggests that substantially more sequencing effort would be required to cover the whole estimated diversity at least once.

even genus-level rarefaction curves remained unsaturated. A complementary analysis based on 120 conserved single-copy bacterial marker genes showed similar trends. For each marker gene, we detected between 9,000 and 59,000 unique copies across the dataset (Supplementary Figure 3).

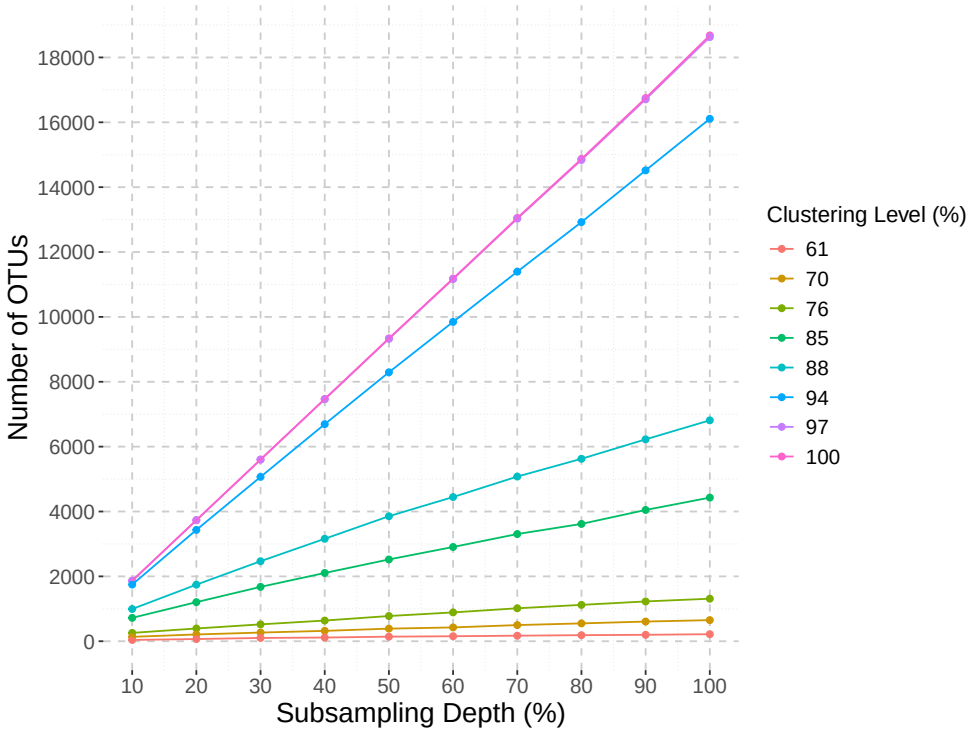

Fig. 4: Rarefaction curves of 16S rRNA gene sequences at different clustering thresholds. The number of observed operational taxonomic units (OTUs) is plotted against subsampling depth for clustering thresholds ranging from 61% to 100% sequence identity. Higher clustering thresholds (e.g., 97% and 100%) result in a greater number of OTUs detected, gradually increasing by the subsampling depth, while the increase is less steep at lower clustering thresholds. This highlights the number of observed species is not saturated despite the deep sequencing efforts.

Taken together, these results provide robust, multi-layered evidence that the soil microbial community remains vastly undersampled, even with sequencing efforts far exceeding typical practice. The failure to reach saturation was especially evident at

the species level, where OTU-based and marker gene rarefaction curves continued to rise linearly across all subsampling depths. This indicates an enormous degree of fine-scale taxonomic resolution, with many closely related but genomically distinct lineages persisting at low abundance. Such strain- and species-level diversity is particularly important given that many ecologically and functionally relevant traits - such as niche specialization, symbiosis, and secondary metabolite biosynthesis - can differ even between closely related genomes. In contrast to more constrained environments such as the human gut, where species-level diversity tends to plateau with a few hundred genomes [21] or marine planktonic communities, which show relatively stable core microbiomes [22], the soil microbiome appears to exhibit near-limitless diversity at high resolution. These findings not only underscore the limitations of current sequencing depths but also call into question the accuracy of historical species richness estimates derived from much shallower data. They suggest that achieving anything approaching species-level saturation in soil will likely require terabase- to petabase-scale efforts, combined with refined analytical tools to resolve low-abundance and highly diverse taxa, even with sequencing efforts far exceeding typical practice. They demonstrate that many of the diversity estimates historically used in soil microbiome research - often derived from much shallower sequencing - likely underestimate the true scale of taxonomic complexity.

2.5 Biosynthetic potential of soil microbes

We identified a total of 11,381 biosynthetic gene clusters (BGCs) in the hybrid metagenomic assembly. Of these, 5,652 were classified as complete, reflecting the improved contiguity made possible through long-read sequencing. This proportion of complete clusters ($\sim$50%) represents a substantial improvement over previous metagenomic efforts, where complete BGCs typically account for less than 10% of predictions [23].

Clustering the identified BGCs using BiG-SCAPE resulted in 10,215 gene cluster families (GCFs), indicating that most clusters in our dataset are non-redundant and functionally distinct. Figure 5A illustrates the size distribution of these GCFs, with the majority being singletons or doubletons, and the largest being a RiPP cluster 115 members - further underscoring the functional uniqueness and fine-grained metabolic specialization within the community. The cumulative ordering of the GCFs by size shows that the largest eight families account for 3.1% of all BGCs, whereas each of the remaining tens of thousands of families contributes only very little to the total number (Figure 5B).

Taxonomic annotation of the BGC-containing contigs revealed that biosynthetic potential is broadly distributed across the phylogenetic spectrum. While well-known producers such as Actinomycetota and Pseudomonadata were represented, we also detected numerous BGCs in underexplored phyla such as Verrucomicrobiota and Acidobacteriota (Figure 5C). These groups have historically been overlooked due to cultivation barriers but appear to harbor substantial secondary metabolic capacity.

Importantly, only 108 BGCs ($\sim$ 1%) had significant matches to known GCFs in the BGC Atlas database, which compiles nearly 2 million clusters from metagenomic sources. This confirms earlier observations that metagenomes are rich in biosynthetic novelty [12, 24], but also demonstrates that even with massive reference expansions, a typical soil still contains an overwhelming majority of uncharacterized BGCs. Notably, our dataset contains more novel BGCs than reported in most large-scale studies of marine or human microbiomes, where the rate of novel BGC discovery is often constrained by lower microbial diversity and shorter contigs [23, 24].

Of the total BGCs identified, 6,279 were linked to metagenome-assembled genomes (MAGs), enabling genome-resolved exploration of natural product potential. Several MAGs contained up to 30 BGCs each - including members of poorly characterized lineages - highlighting taxa with unusually high biosynthetic richness. This represents

one of the largest genome-resolved inventories of BGCs from a single environmental sample to date.

Together, these findings underscore not only the scale of biosynthetic capacity in soil but also the extent to which it remains hidden in standard analyses. The ability to recover thousands of complete and genome-linked BGCs highlights the value of ultra-deep, long-read-enabled metagenomics for natural product discovery. In addition, the high novelty rate and wide taxonomic distribution of these clusters point to a vast and largely untapped chemical repertoire - one that is unlikely to be captured through shallow sequencing or cultivation-based surveys alone.

3 Discussion and Conclusion

Our study presents one of the most comprehensive metagenomic investigations of a single soil microbiome to date, combining ultra-deep Oxford Nanopore and Illumina sequencing to yield nearly 300 Gbp of data. This effort enabled high-quality hybrid assemblies and genome-resolved analysis of taxonomic and functional potential in a temperate forest soil sample, providing an empirical benchmark for the capabilities and current limitations of metagenomic sequencing in one of the most complex microbial habitats on Earth.

Despite the ultra-deep sequencing depth, our analyses demonstrate that the microbial diversity of the sample remains far from saturated. Rarefaction curves, k-mer profiles, and 16S rRNA gene and marker gene analyses consistently indicate that an estimated 10 Tbp of sequencing would be required to approach near-complete community recovery. By comparison, MAG recovery in human-gut communities reach saturation with 5-10 Gbp per sample, and 25 Gbp is estimated to be enough even to classify taxa of very low abundances ($<$ 1e-06) [25]. Similarly, Ni et al. [26] estimate this number to be around 7 Gbp to detect gene content of organisms with relative abundances over 1% in human faecal samples.

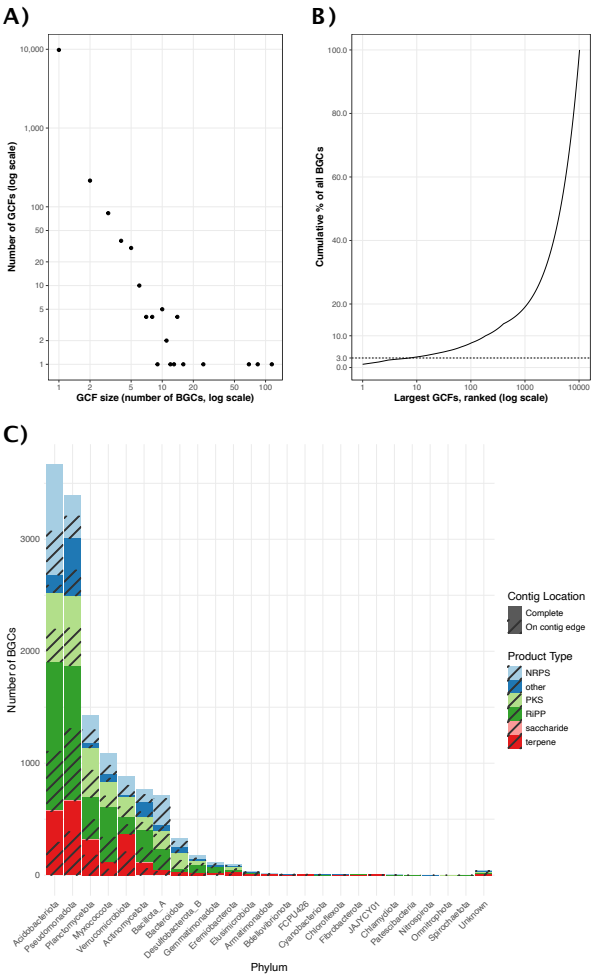

Fig. 5: The diversity and taxonomic composition of the biosynthetic gene clusters. **A)** The size distribution of gene cluster families (GCFs). Each dot represents one distinct GCF size in the dataset (11,381 biosynthetic gene clusters grouped into 10,215 GCFs). The x-axis shows the size of the family (number of BGCs), the y-axis shows the number of families that have that exact size; both axes are log10 scaled. The plot reveals a heavy tail: 9,813 families (96.1 %) are singletons, only 20 families contain >= 10 BGCs, and the largest family comprises 115 BGCs. **B)** Cumulative contribution of the largest GCFs to the total BGC dataset. Families are ordered from largest to smallest along the log10-scaled x-axis; the y-axis tracks the running percentage of all 11,381 BGCs accounted for as each additional family is added. The first eight GCFs explain 3.1% of all BGCs, whereas 7,939 families are needed to reach 80%. The rapid rise followed by a very shallow tail show that a small handful of families contribute to about 3% of BGCs in the dataset, while thousands of minor families are unique. **C)** Distribution of biosynthetic gene cluster (BGC) product types across bacterial phyla. Each bar represents the total number of BGCs assigned to a given phylum, colored by predicted product type. The striped lines indicate the proportion of BGCs that were classified as "on contig edge" by antiSMASH. Phyla are ordered by total BGC count, with "Unknown" indicating contigs for which no confident phylum-level classification could be assigned.

Ultra-deep campaigns such as Tara Oceans have shown that planktonic marine communities also benefit from multi-terabase sequencing - yielding thousands of novel genes and genomes and greatly refining oceanic reference catalogs. Yet, the incremental discovery of truly new sequence space in the ocean has begun to flatten at these depths [22, 27, 28]. In soils, by contrast, every additional sequencing effort still brings a steeper growth of previously unseen sequence diversity, and MAGs, with no hint of saturation.

All in all, these comparisons highlight not only the extraordinary complexity of soil microbiomes but also how typical sequencing depths remain inadequate for capturing their fine-scale taxonomic and functional diversity. Especially at the species and strain level - where many ecologically critical traits are encoded - our results imply that the vast majority of soil microbial diversity is still waiting to be uncovered.

Taxonomic profiling of assembled contigs revealed the presence of a diverse range of phyla in the soil, including the expected dominance of well-known soil phyla, such as Acidobacteriota, Actinomycetota, and Pseudomonadota. Yet, over 95% of contigs could not be assigned to a species present currently in databases, and fewer than 0.7% could be linked to a species with at least one cultured representative. Notably, we recovered MAGs that represent candidate novel phyla, underscoring the ability of genome-resolved metagenomics to reveal deeply divergent lineages missed by conventional approaches. Only four MAGs recovered in this study share >= 95% ANI with entries in GTDB, and none correspond to cultured isolates, reaffirming how incompletely current databases represent soil ecosystems. These lineages not only expand the bacterial tree of life but, in some cases, also harbor unique biosynthetic gene clusters, suggesting functional capacities that remain entirely uncharted.

The scale of biosynthetic diversity uncovered is similarly striking. We identified over 11,000 BGCs, more than 5,600 of which are complete and many of which could be linked to high-quality MAGs. Fewer than 1% of these BGCs matched known families in public databases, reinforcing the notion that soils remain a vast and underexplored

reservoir for natural product discovery. Importantly, our dataset shows that long-read sequencing can overcome common limitations of metagenomics - such as incomplete BGC recovery and fragmentation - thereby enabling genome-resolved assessments of secondary metabolism. In contrast to earlier studies relying solely on short-read data, our approach yields a high proportion of complete BGCs and facilitates the association of biosynthetic capacity with uncultured taxa.

Together, these results redefine the scale of the unknown in soil microbiomes. They demonstrate that even the deepest sequencing efforts to date barely scratch the surface of microbial and metabolic diversity, particularly in complex environments. Moving forward, terabase- to petabase-scale sequencing will likely be necessary to achieve species-level saturation in soil. However, deeper sequencing alone will not be sufficient. Advances in bioinformatics for assembly, binning, and functional annotation - alongside complementary data from metatranscriptomics, metabolomics, and single-cell approaches - will be essential for translating sequence data into ecological and biochemical insight.

In conclusion, this study provides a new benchmark for what is currently achievable in soil metagenomics. It quantitatively updates long-held assumptions about diversity, reveals the power of long-read sequencing for functional genome recovery, and underscores the vast untapped potential that soil microbial communities continue to hold. These findings have broad implications for microbial ecology, taxonomy, and natural product research, and call for a renewed investment in both sequencing scale and analytical innovation to fully explore the hidden majority of microbial life beneath our feet.

4 Materials and Methods

4.1 DNA Extraction and Sequencing

The A horizon of the soil type Cambisol was sampled from the Schönbuch Forest, near Tübingen, Germany (the same site previously reported by [16]), on 31.05.2022. High molecular weight (HMW) metagenomic DNA was isolated using a protocol described in detail in our previous studies [16, 29]. Genomic integrity was assessed using pulse-field capillary electrophoresis with the Genomic DNA 165 kb Analysis Kit on a FemtoPulse (Agilent) instrument. Quantitation of DNA was assessed using the dsDNA High Sensitivity assay on a Qubit 3 fluorometer (Thermo Fischer), and purity was assessed by Nanodrop. A total of 2.4 μg of genomic DNA was used as input for the library preparation with the 1D Ligation Kit SQK-LSK109-XL Sequencing Kit (Oxford Nanopore Technologies). Four PromethION R9 flow cells were utilized for long-read Nanopore sequencing, each loaded with 600 ng (50 fmol) of genomic DNA. The raw signal data from the PromethION runs were basecalled with guppy (v 5.0.7) in high-accuracy mode. For complementary short-read sequencing, three NovaSeq 6000 flowcells were loaded with the same isolated DNA; two of them ran with 200 cycles, and one with 300 cycles.

A total of 148 Gbp of Nanopore sequencing data was generated, with a read length N50 of 12.2 kb. The three Illumina sequencing runs yielded a total of 122 Gbp raw data, with a mean Q-score of 35. The raw Illumina reads were adapter and quality trimmed using fastp [30] (v 0.23.4) with default settings.

4.2 Metagenomic Assembly

Nanopore reads from all four runs were pooled together and assembled using metaFlye (v2.9.5-b1801) [31] with the '–meta' option optimized for metagenomic data, and the –nano-raw option for the error-prone Nanopore reads. The resulting draft assembly was polished using one round of medaka [32] (v 2.0.1) using r941_prom_hac_g507 model

with the –bacteria flag. The medaka polished assembly was further corrected using the error-corrected Illumina reads with one round of NextPolish [33] (v 1.4.1), with parameters -max_depth 100 for short-read mapping with bwa [34] (v 0.7.18), and -min_read_len 1k -max_depth 100 -x map-ont for long-read mapping minimap2 [35] (v 2.28-r1209).

In parallel, an Illumina-only assembly was performed using MEGAHIT [36] (v1.2.9) with default metagenomic assembly parameters. All subsequent analyses were based on the hybrid Nanopore-Illumina assembly, which exhibited superior contiguity (contig N50 = 77.8 kb) and total assembled length (10.5 Gbp).

4.3 Taxonomic Analysis

The raw reads were assigned taxonomy using metabuli [37] (v 1.0.8) against the pre-computed database provided by the authors, which contains GTDB release 214.1 [20] and the human T2T genome [38], with parameters –seq-mode 2 for Illumina reads, and –seq-mode 3 for Nanopore reads. The contigs from the final assembly were assigned taxonomy using MMSeqs2 [39] (v 16.747c6) against the GTDB 214.1 database [20] in easy-taxonomy mode.

4.4 Binning and Genome Reconstruction

Metagenome-assembled genomes (MAGs) were recovered using a combination of binning strategies to maximize bin quality. First, binning was performed independently on the final NextPolish-corrected assembly, using COMEBin [40], MetaDecoder [41], SemiBin2 [42], and VAMB [43], on a CUDA-enabled environment when possible. COMEBIN [40] (v 1.0.4, CheckM v1.1.3 [44]), on mapping files produced by minimap2 [35], using parameters NUM_VIEWS 6, TEMPERATURE 0.07, EMBEDDING_SIZE 2048, COVERAGE_EMBEDDING_SIZE 2048, BATCH_SIZE 1024. MetaDecoder [41]

(v 1.1.0) was run with default parameters according to the developers' recommendations. SemiBin2 [42] (v 2.1.0) was run in two different modes, –self-supervised and –semi-supervised with soil chosen as the environment type, and long_read chosen as the sequencing type, using prodigal [45] (v 2.6.3) as the gene-caller and GTDB database v95 [20] with the previously described minimap2 mappings of the long reads. Six different binning results were computed using VAMB [43] (v 4.1.4) - the combinations of the default and taxonomy-enabled binning modes, with the input types of long- and short-reads individually and combined. The abundance profiles were computed using strobealign [46] (v 0.15.0). For TaxVAMB [47], the taxonomic assignment of contigs was computed using metabuli [37] against the GTDB database [20].

In order to produce a final, unified metagenome-assembled-genome binning, the binning results from the above-mentioned ten runs were collected, and refined using MAGScoT [48] (v 1.1). MAGScot was run on gene-calling results from prodigal [45] (v 2.6.3) and hmmsearch results from the HMMER package (v 2.6.3) [49] against PFAM [50] and TIGRFAMs [51] domains.

The resulting MAGs were assessed for their completeness and contamination with CheckM2 [52] (v 1.0.2). Their pairwise distances were calculated using Mash [53] (v 2.3) with default parameters. Their taxonomic assignments were performed with GTDB-Tk [54] (v 2.4.0) using the GTDB database version r220 [20] in classify mode. A phylogenetic tree of HQ and MQ MAGs was built with GTDB-Tk in *de novo* mode and drawn using iTOL [55].

4.5 Rarefaction Analysis

Microbial diversity and sequencing coverage were assessed using multiple complementary approaches. K-mer frequency analysis was conducted with ntCard [56] using 15-mers and 17-mers for both Nanopore and Illumina data. Nonpareil [57] (v3.3.3) was used to estimate sequence coverage and project additional sequencing efforts

needed for complete community recovery. Full-length 16S rRNA gene sequences were extracted from Nanopore reads using Barrnap [58] (v0.9) and clustered at various identity thresholds (100%, 97%, 95%, etc.) using VSEARCH [59] (v2.15.2) to generate operational taxonomic units (OTUs). Diversity rarefaction curves were constructed based on cumulative OTU counts across subsampled sequencing depths. Additionally, conserved single-copy marker genes (bac120 set) [10] were identified from the raw Nanopore reads using LAST aligner [60] (v 1615) against the GTDB bac120 reference set in frameshift mode (-F 15) and quantified to independently assess taxonomic diversity.

4.6 Annotation, Identification and Clustering of Biosynthetic Gene Clusters

The resulting final assembly, and the refined MAGs were annotated using Bakta [61] (v 1.11.0, database version 5.1). Biosynthetic gene clusters (BGCs) were identified using antiSMASH [62] (v7) with the options "–fullhmmer –clusterhmmer –tigrfam –asf –cc-mibig –cb-general –cb-subclusters –cb-knownclusters –pfam2go –rre –smcog-trees –tfb". BGCs were subsequently clustered into gene cluster families (GCFs) using BiG-SCAPE [24] (v2 beta5) based on domain architecture and sequence similarity thresholds. The novelty of BGCs was assessed by comparisons against the MIBiG 4.0 [63] and BGC Atlas [23] databases.

Acknowledgments. The authors acknowledge the support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the Federal Ministry of Education and Research (BMBF) through grant no. 031 A535A. We thank the Interfaculty Institute for Biomedical Informatics (IBMI) at the University of Tübingen for providing the computational resources essential for this study. NGS sequencing methods were performed with the support of the DFG-funded NGS Competence Center Tübingen (INST 37/1049-1). Additionally,

we thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2124—390838134 for structural support.

Authors' contributions. C.B. and N.Z. wrote the main manuscript. C.B. performed the data analysis and prepared the figures. T.N. conducted the soil sampling and DNA isolation, while E.A.B. carried out the Nanopore sequencing. C.G. performed basecalling and preliminary quality control of the sequencing data. C.B., S.O., and N.Z. conceptualized the study. All authors reviewed the manuscript.

Data availability. The raw sequencing data, full metagenome assembly, and extracted metagenome-assembled genomes (MAGs) that support the findings of this study have been deposited in the European Nucleotide Archive under the primary accession code PRJEB89893. The extracted MAGs and identified biosynthetic gene clusters (BGCs) are available at Zenodo (DOI: 10.5281/zenodo.15533781 for the total metagenome assembly and DOI: 10.5281/zenodo.15529477 for MAGs) and in the BGC Atlas repository at https://bgc-atlas.cs.uni-tuebingen.de/ultra-deep-soil/.

Supplementary information. Supplementary Material S1. Supplementary Figure S1. 17-mer spectrum of Illumina reads. Supplementary Figure S2. 17-mer spectrum of Nanopore reads. Supplementary Figure S3. The number of occurrences of each marker gene from the bac120 set in raw Nanopore reads. Supplementary Material 2. Supplementary Table 1. The details of the extracted MAGs, including their completeness and contamination values, taxonomic classification, assembly statistics, genome annotation statistics, and the number of BGCs identified.

Declarations

References

[1] Fierer, N.: Embracing the unknown: disentangling the complexities of the soil microbiome. Nature Reviews Microbiology **15**(10), 579–590 (2017)

[2] Roesch, L.F., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K., Kent, A.D., Daroub, S.H., Camargo, F.A., Farmerie, W.G., Triplett, E.W.: Pyrosequencing enumerates and contrasts soil microbial diversity. The ISME journal **1**(4), 283–290 (2007)

[3] Torsvik, V., Øvreås, L.: Microbial diversity and function in soil: from genes to ecosystems. Current opinion in microbiology **5**(3), 240–245 (2002)

[4] Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., Küsel, K., Rillig, M.C., Rivett, D.W., Salles, J.F., *et al.*: Where less may be more: how the rare biosphere pulls ecosystems strings. The ISME journal **11**(4), 853–862 (2017)

[5] Lok, C.: Mining the microbial dark matter. Nature **522**(7556), 270 (2015)

[6] Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., *et al.*: A new view of the tree of life. Nature microbiology **1**(5), 1–6 (2016)

[7] Lloyd, K.G., Steen, A.D., Ladau, J., Yin, J., Crosby, L.: Phylogenetically novel uncultured microbial cells dominate earth microbiomes. MSystems **3**(5), 10–1128 (2018)

[8] Nascimento Lemos, L., Manoharan, L., William Mendes, L., Monteiro Venturini, A., Satler Pylro, V., Tsai, S.M.: Metagenome assembled-genomes reveal similar functional profiles of CPR/Patescibacteria phyla in soils. Environmental microbiology reports **12**(6), 651–655 (2020)

[9] Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., *et al.*: A genomic catalog of Earth's microbiomes. Nature biotechnology **39**(4), 499–509 (2021)

[10] Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W.: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature microbiology **2**(11), 1533–1542 (2017)

[11] Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A.H.Q., Kumar, M.S., Li, C., Dvornicic, M., Soldo, J.P., Koh, J.Y., Tong, C., *et al.*: Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nature biotechnology **37**(8), 937–944 (2019)

[12] Van Goethem, M.W., Osborn, A.R., Bowen, B.P., Andeer, P.F., Swenson, T.L., Clum, A., Riley, R., He, G., Koriabine, M., Sandor, L., *et al.*: Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics. Communications biology **4**(1), 1302 (2021)

[13] Agustinho, D.P., Fu, Y., Menon, V.K., Metcalf, G.A., Treangen, T.J., Sedlazeck, F.J.: Unveiling microbial diversity: harnessing long-read sequencing technology. Nature methods **21**(6), 954–966 (2024)

[14] Wick, R.R., Judd, L.M., Holt, K.E.: Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. PLOS Computational Biology **19**(3), 1010905 (2023)

[15] Zhang, T., Li, H., Jiang, M., Hou, H., Gao, Y., Li, Y., Wang, F., Wang, J., Peng, K., Liu, Y.-X.: Nanopore sequencing: Flourishing in its teenage years. Journal of Genetics and Genomics (2024)

[16] Mantri, S.S., Negri, T., Sales-Ortells, H., Angelov, A., Peter, S., Neidhardt, H., Oelmann, Y., Ziemert, N.: Metagenomic sequencing of multiple soil horizons and sites in close vicinity revealed novel secondary metabolite diversity. Msystems

6(5), 10–1128 (2021)

[17] Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., *et al.*: Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature biotechnology **35**(8), 725–731 (2017)

[18] Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., *et al.*: Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nature Biotechnology **41**(11), 1633–1644 (2023)

[19] Chen, J., Jia, Y., Sun, Y., Liu, K., Zhou, C., Liu, C., Li, D., Liu, G., Zhang, C., Yang, T., *et al.*: Global marine microbial diversity and its potential in bioprospecting. Nature **633**(8029), 371–379 (2024)

[20] Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P.: GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic acids research **50**(D1), 785–794 (2022)

[21] Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., *et al.*: Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell **176**(3), 649–662 (2019)

[22] Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., *et al.*: Structure and function of the global ocean microbiome. Science **348**(6237), 1261359 (2015)

[23] Bağcı, C., Nuhamunada, M., Goyat, H., Ladanyi, C., Sehnal, L., Blin, K., Kautsar, S.A., Tagirdzhanov, A., Gurevich, A., Mantri, S., *et al.*: BGC Atlas: a web resource for exploring the global chemical diversity encoded in bacterial genomes. Nucleic acids research **53**(D1), 618–624 (2025)

[24] Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L., Yeong, M., Cruz-Morales, P., Abubucker, S., *et al.*: A computational framework to explore large-scale biosynthetic diversity. Nature chemical biology **16**(1), 60–68 (2020)

[25] Kim, C.Y., Lee, M., Yang, S., Kim, K., Yong, D., Kim, H.R., Lee, I.: Human reference gut microbiome catalog including newly assembled genomes from under-represented Asian metagenomes. Genome Medicine **13**, 1–20 (2021)

[26] Ni, J., Yan, Q., Yu, Y.: How much metagenomic sequencing is enough to achieve a given goal? Scientific reports **3**(1), 1968 (2013)

[27] Duarte, C.M., Ngugi, D.K., Alam, I., Pearman, J., Kamau, A., Eguiluz, V.M., Gojobori, T., Acinas, S.G., Gasol, J.M., Bajic, V., *et al.*: Sequencing effort dictates gene discovery in marine microbial metagenomes. Environmental Microbiology **22**(11), 4589–4603 (2020)

[28] Nishimura, Y., Yoshizawa, S.: The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. Scientific Data **9**(1), 305 (2022)

[29] Negri, T., Mantri, S., Angelov, A., Peter, S., Muth, G., Eustáquio, A.S., Ziemert, N.: A rapid and efficient strategy to identify and recover biosynthetic gene clusters from soil metagenomes. Applied Microbiology and Biotechnology **106**(8), 3293–3306 (2022)

[30] Chen, S.: Ultrafast one-pass fastq data preprocessing, quality control, and deduplication using fastp. Imeta **2**(2), 107 (2023)

[31] Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P., *et al.*: metaFlye: scalable long-read metagenome assembly using repeat graphs. Nature methods **17**(11), 1103–1110 (2020)

[32] Oxford Nanopore Technologies: Medaka: sequence correction provided by ONT Research. https://github.com/nanoporetech/medaka. Accessed: 2025-05-26 (2023)

[33] Hu, J., Fan, J., Sun, Z., Liu, S.: NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics **36**(7), 2253–2255 (2020)

[34] Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics **25**(14), 1754–1760 (2009)

[35] Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**(18), 3094–3100 (2018)

[36] Li, D., Liu, C.-M., Luo, R., Sadakane, K., Lam, T.-W.: MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinformatics **31**(10), 1674–1676 (2015)

[37] Kim, J., Steinegger, M.: Metabuli: sensitive and specific metagenomic classification via joint analysis of amino acid and dna. Nature Methods **21**(6), 971–973 (2024)

[38] Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., *et al.*: The complete sequence

of a human Y chromosome. Nature **621**(7978), 344–354 (2023)

[39] Steinegger, M., Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology **35**(11), 1026–1028 (2017)

[40] Wang, Z., You, R., Han, H., Liu, W., Sun, F., Zhu, S.: Effective binning of metagenomic contigs using contrastive multi-view representation learning. Nature Communications **15**(1), 585 (2024)

[41] Liu, C.-C., Dong, S.-S., Chen, J.-B., Wang, C., Ning, P., Guo, Y., Yang, T.-L.: MetaDecoder: a novel method for clustering metagenomic contigs. Microbiome **10**(1), 46 (2022)

[42] 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**(Supplement_1), 21–29 (2023)

[43] Nissen, J.N., Johansen, J., Allesøe, R.L., Sønderby, C.K., Armenteros, J.J.A., Grønbech, C.H., Jensen, L.J., Nielsen, H.B., Petersen, T.N., Winther, O., *et al.*: Improved metagenome binning and assembly using deep variational autoencoders. Nature biotechnology **39**(5), 555–560 (2021)

[44] 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 research **25**(7), 1043–1055 (2015)

[45] Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J.: Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics **11**, 1–11 (2010)

[46] Sahlin, K.: Strobealign: flexible seed size enables ultra-fast and accurate read alignment. Genome Biology **23**(1), 260 (2022)

[47] Kutuzova, S., Piera, P., Nor Nielsen, K., Olsen, N.S., Riber, L., Gobbi, A., Forero-Junco, L.M., Dougherty, P.E., Westergaard, J.C., Christensen, S., et al.: Binning meets taxonomy: TaxVAMB improves metagenome binning using bi-modal variational autoencoder. bioRxiv, 2024–10 (2024)

[48] Rühlemann, M.C., Wacker, E.M., Ellinghaus, D., Franke, A.: MAGScoT: a fast, lightweight and accurate bin-refinement tool. Bioinformatics **38**(24), 5430–5433 (2022)

[49] Eddy, S.R.: Accelerated profile HMM searches. PLoS computational biology **7**(10), 1002195 (2011)

[50] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L., Tosatto, S.C., Paladin, L., Raj, S., Richardson, L.J., *et al.*: Pfam: The protein families database in 2021. Nucleic acids research **49**(D1), 412–419 (2021)

[51] Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., Beck, E.: TIGRFAMs and genome properties in 2013. Nucleic acids research **41**(D1), 387–395 (2012)

[52] 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. Nature methods **20**(8), 1203–1212 (2023)

[53] Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M.: Mash: fast genome and metagenome distance estimation using MinHash. Genome biology **17**, 1–14 (2016)

[54] Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H.: GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics **38**(23), 5315–5316 (2022)

[55] Letunic, I., Bork, P.: Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic acids research **52**(W1), 78–82 (2024)

[56] Mohamadi, H., Khan, H., Birol, I.: ntCard: a streaming algorithm for cardinality estimation in genomics data. Bioinformatics **33**(9), 1324–1330 (2017)

[57] Rodriguez-R, L.M., Gunturu, S., Tiedje, J.M., Cole, J.R., Konstantinidis, K.T.: Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. MSystems **3**(3), 10–1128 (2018)

[58] Seeman, T.: Barrnap 0.9: rapid ribosomal RNA prediction. https://github.com/tseemann/barrnap (2013)

[59] Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F.: VSEARCH: a versatile open source tool for metagenomics. PeerJ **4**, 2584 (2016)

[60] Sheetlin, S.L., Park, Y., Frith, M.C., Spouge, J.L.: Frameshift alignment: statistics and post-genomic applications. Bioinformatics **30**(24), 3575–3582 (2014)

[61] Schwengers, O., Jelonek, L., Dieckmann, M.A., Beyvers, S., Blom, J., Goesmann, A.: Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial genomics **7**(11), 000685 (2021)

[62] Blin, K., Shaw, S., Augustijn, H.E., Reitz, Z.L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B.R., Metcalf, W.W., Helfrich, E.J., *et al.*: antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and

visualisation. Nucleic acids research **51**(W1), 46–50 (2023)

[63] Zdouc, M.M., Blin, K., Louwen, N.L., Navarro, J., Loureiro, C., Bader, C.D., Bailey, C.B., Barra, L., Booth, T.J., Bozhüyük, K.A., *et al.*: MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic acids research **53**(D1), 678–690 (2025)