EcoFoldDB: Protein structure-guided functional profiling of ecologically relevant microbial traits at the metagenome scale

Timothy M. Ghaly[1,2]*, Vaheesan Rajabal[1,2], Dylan Russel[1,2], Elena Colombi[1,2], Sasha G. Tetu[1,2]*

[1]School of Natural Sciences, Macquarie University, Australia

[2]ARC Centre of Excellence in Synthetic Biology, Australia

*Corresponding Authors:

Timothy M. Ghaly: timothy.ghaly@mq.edu.au

Sasha G. Tetu: sasha.tetu@mq.edu.au

Keywords: Protein folding; protein function; protein homology detection; structural bioinformatics; metagenomic annotations; biogeochemical cycling; plant growth-promotion

Abstract

Microbial communities are fundamental to planetary health and ecosystem processes. High-throughput metagenomic sequencing has provided unprecedented insights into the structure and function of these communities. However, functional profiling of metagenomes remains constrained due to the limited sensitivity of existing sequence homology-based methods to annotate evolutionarily divergent genes. Protein structure, which is more conserved than sequence and intrinsically tied to molecular function, offers a solution. Capitalising on recent breakthroughs in structural bioinformatics, we present EcoFoldDB, a database of protein structures curated for ecologically relevant microbial traits, and its companion pipeline, *EcoFoldDB-annotate*, which leverages Foldseek with the ProstT5 protein language model for rapid structural homology searching directly from sequence data. *EcoFoldDB-annotate* outperforms state-of-the-art sequence-based methods in annotating metagenomic protein sequences, in terms of both sensitivity and precision. To demonstrate its utility and scalability, we performed structure-guided functional profiling of 32 million proteins encoded by 8,000 high-quality metagenome-assembled genomes from the global soil microbiome. *EcoFoldDB-annotate* could resolve the phylogenetic partitioning of important nitrogen cycling pathways, from taxonomically restricted nitrifiers to more widespread denitrifiers, as well as identifying novel, uncultivated bacterial taxa enriched in plant growth-promoting traits. We anticipate that EcoFoldDB will enable researchers to extract ecological insights from environmental genomes and metagenomes, and accelerate discoveries in microbial ecology. EcoFoldDB and its companion annotation pipeline are freely available via https://github.com/timghaly/EcoFoldDB.

Introduction

Microbial community functioning shapes Earth system health, influencing global carbon dynamics, driving essential nutrient cycles and ecosystem services, and underpinning the biodiversity of entire biomes [1-4]. Deciphering the complexity of microbial communities, including their composition, interactions, and functions, is critical for understanding ecosystem processes and their responses to environmental change. High-throughput shotgun metagenomic sequencing has revolutionised the field of microbial ecology, allowing a culture-independent view of the structure and function of microbiomes [5-8], and facilitating unprecedented insights into their ecology and evolution at scales ranging from individual strains to whole communities [9, 10].

Functional annotations of microbial genes represent a cornerstone of metagenomic analysis. Nevertheless, this task represents a formidable challenge due to the extensive size of metagenomic datasets, and the massive functional diversity encoded within metagenomes [11]. Several bioinformatic tools exist to enable large-scale functional profiling, e.g., InterProScan [12], eggNOG-mapper [13], KofamScan [14]. However, these tools prioritise exhaustive functional identification over ecological interpretation. Researchers thus often need to manually screen annotation results to gain meaningful insights into the ecological functions of microbial communities. This has led to the development of ecology-focused tools for functional profiling, such as METABOLIC [15] and DRAM [16], which curate annotations into ecologically coherent functional categories, enabling user-friendly interpretation and analysis. However, existing functional annotation tools rely on sequence homology-based methods (e.g., MMSeqs2 [17], DIAMOND [18], or HMMER [19]). Consequently, annotation of metagenomes remains constrained since they are

overwhelmingly composed of genes from uncultivated microbial taxa, and are enriched with proteins lacking detectable sequence homology to reference databases [11, 20].

The structure of proteins, which are inherently linked to their molecular function, allows homology discovery beyond the limits of sequence similarity detection, as structure evolves much slower than sequence [21]. However, until recent breakthroughs in structural bioinformatics, the use of protein structure for large-scale functional profiling has not been feasible. Notably, AlphaFold2 revolutionised protein structure predictions, achieving near-experimental accuracy from sequence data alone [22]. Yet, identifying similar protein structures at scale remained a significant challenge. Foldseek addressed this limitation by enabling protein structural homology searching at speeds comparable to sequence searching [23]. Foldseek achieves this by mapping the 3D coordinates for each residue in a protein structure to one of the 20 states in their 3D interaction (3Di) alphabet. Thus, Foldseek effectively represents the structure of a protein as a 20-letter alphabet, enabling the use of fast sequence searching algorithms that were originally developed for the 20-letter amino acid alphabet [17]. However, despite these advances, the computational bottleneck of AlphaFold2 structure predictions still hindered metagenomic applications, even with performance optimisations like ColabFold [24]. To that end, the ProstT5 protein language model was developed, which translates between protein amino acid and Foldseek's 3Di sequences, without the need for structure predictions [25]. ProstT5 was built from an existing metagenomic protein language model (ProtT5) [26], and further trained on 17 million non-redundant and diverse protein sequences, paired with their corresponding 3Di sequences from the AlphaFold Protein Structure Database [27]. Consequently, ProstT5 enables ~4,000× faster structural homology searching, compared to AlphaFold2-ColabFold, without sensitivity loss [25].

To take advantage of these recent breakthroughs for applications in microbial ecology, we created EcoFoldDB, a curated Foldseek database of protein structures spanning microbial traits of ecological relevance. The included traits cover biogeochemical cycling of carbon, nitrogen, sulphur, phosphorus, iron, and trace atmospheric gases, as well as key plant-microbe interactions, and osmotic stress tolerance mechanisms. Complementing this resource, we introduce *EcoFoldDB-annotate*, an annotation pipeline that leverages the speed and accuracy of ProstT5's 3Di sequence translation, and Foldseek's structural homology search capabilities. This annotation framework enables structure-guided functional profiling, scalable to metagenomic datasets.

Methods

Database construction

We included eight microbial functional categories, and 35 sub-categories, within EcoFoldDB that span several key ecological processes (Fig. 1; Supplementary Table S1). Pathways and associated genes were sourced from published literature or existing functional databases. These consisted of *(1) Trace gas oxidation*: atmospheric hydrogen oxidation [28, 29], carbon monoxide oxidation [29], and methane oxidation [29]; *(2) Carbon cycling*: carbon fixation [30-32], one-carbon (C1) metabolism [32], complex carbohydrate degradation [33], fatty acid degradation [32], and polyphenol and aromatic hydrocarbon cycling [34, 35]; *(3) Nitrogen cycling*: nitrogen fixation, nitrification, denitrification, dissimilatory nitrate reduction to ammonium (DNRA), nitrate assimilation, anaerobic ammonium oxidation (anammox), and organic nitrogen mineralisation [32, 36]; *(4) Sulphur cycling*: sulphate reduction, sulphite reduction, sulphide oxidation, sulphur disproportionation, sulphite oxidation, thiosulphate

oxidation, thiosulphate disproportionation, sulphur reduction, and organic sulphur mineralisation [15, 32, 37-41]; *(5) Phosphorus cycling*: organic phosphorus mineralisation and inorganic phosphorus solubilisation [42, 43]; *(6) Iron cycling*: siderophore biosynthesis [32]; *(7) Plant-microbe interactions*: ethylene suppression via ACC deaminase activity [44], indole-3-acetic acid production [45], gibberellin production [46], abscisic acid production [47], trans-Zeatin production [48], jasmonic acid production [47, 49], brassinolide production [50], GABA production [51-53], putrescine and spermidine production [54], and legume root nodulation [55, 56]; and (8) *Osmotic stress tolerance*: osmoprotectant uptake and osmoprotectant biosynthesis [57-60]. Included genes and pathways, and their functional categorisations are available as Supplementary Table S1.

Proteins with these functional annotations were selected from the manually curated databases, Swiss-Prot [61] and MetaCyc [32]. The corresponding predicted structures for these protein sequences were then retrieved from the AlphaFold Protein Structure Database (AFDB) [22, 27]. Finally, redundancy was removed from this set of protein structures to create EcoFoldDB using Foldseek *cluster* [23, 62] with a minimum TM-score threshold of 0.9 and a bidirectional coverage of 0.9 [parameters: foldseek cluster --alignment-type 1 --tmscore-threshold 0.9 -c 0.9 --cov-mode 0 --cluster-reassign 1].

EcoFoldDB-annotate pipeline and performance benchmarking

To enable metagenome-scale annotations using EcoFoldDB, we created *EcoFoldDB-annotate*, which employs Foldseek [23] with the ProstT5 protein language model [25]. Together, Foldseek with ProstT5, allow protein structural alignments from amino acid sequences without the need for structure predictions [25]. *EcoFoldDB-annotate* begins by filtering protein sequences exceeding 4,000 amino acids in length, archiving excluded

sequences in a dedicated output file for user inspection. Proteins larger than this drastically increase the runtime of ProstT5 inference, and are likely too large to represent any protein in EcoFoldDB, with its largest entry being 2,378 amino acids. *EcoFoldDB-annotate* then uses ProstT5 to translate the amino acid alphabet to Foldseek's 3D interaction (3Di) structural alphabet, allowing structural alignment of input proteins against EcoFoldDB with Foldseek. Finally, the pipeline assigns functionally categorised annotations based on EcoFoldDB ontology, using the best-hit approach with default query and target coverage cut-offs of 80%, and an e-value threshold of 1e-19 (see Supplementary Fig. S1 for impacts of e-value thresholds on annotation specificity).

To test its functional annotation capacity, we benchmarked *EcoFoldDB-annotate* against two widely used state-of-the-art annotation tools, eggNOG-mapper v2.1.1 [13] and KofamScan v1.3.0 [14]. To create a benchmarking dataset, we used the ESM Metagenomic Atlas 30 [63], a non-redundant database of 37 million high-confidence (pTM > 0.7 and pLDDT > 0.7) predicted protein structures from metagenomic assemblies (derived from the MGnify database [64]). The ESM Atlas 30 was chosen as it includes diverse proteins found in metagenomes, with protein entries sharing <30% sequence similarity with each other, and includes protein 'dark matter' sequences from genomes of uncultivated taxa. We created a positive benchmarking dataset by generating a subset of ESM Atlas proteins exhibiting high structural similarity to those in EcoFoldDB, as determined by Foldseek, selecting only hits with both a query and target TM-score > 0.8. This TM-score threshold was chosen because protein pairs with a TM-score >0.8 have a near certain (~100%) posterior probability of sharing the same fold topology [65]. Thus, in the absence of a validated reference dataset, these high-confidence matches were used as a proxy for ground truth functional annotations. A negative benchmarking set was generated by selecting ESM Atlas proteins with TM-Scores that fell within the range of 0.3–0.4. This thresholding strategy was set because the posterior

probability of proteins sharing the same fold approaches zero for pairs with TM-score < 0.4 [65], thus generating a set of distinct, but still structurally related (TM-score > 0.3) proteins. This resulted in 188,918 and 518,500 proteins in the positive and negative sets, respectively.

The corresponding amino acid sequences of the positive and negative benchmarking sets were then used as input for *EcoFoldDB-annotate*, eggNOG-mapper, and KofamScan. *EcoFoldDB-annotate* was run with default settings; eggNOG-mapper sequence searches were performed using DIAMOND [18] against eggNOG v5.0.2 [66]; and KofamScan was run using its model-specific score thresholds [14].

EcoFoldDB annotations of microbial genomes from the global soil microbiome

To demonstrate its utility and scalability, we employed *EcoFoldDB-annotate* on >8,000 high-quality metagenome-assembled genomes (MAGs) from the global soil microbiome. Genome sequences were obtained from the Soil Microbial Dark Matter MAG [67] and Microflora Danica long-read MAG [68] catalogues, collectively covering diverse soil habitats sampled across the globe. Genome sequences were filtered to retain only high-quality (HQ) MAGs based on MIMAG standards [69]. That is, genomes that are >90% complete, with <5% contamination, have the 23S, 16S, and 5S rRNA genes, and at least 18 different tRNA genes. Genome completion and contamination was determined for all MAGs using CheckM2 [70] with default settings. Genomes were screened for rRNA genes using barrnap (https://github.com/tseemann/barrnap) and bakta v1.8.1 [71] (which uses Infernal [72] against the Rfam database [73]), and tRNA genes were screened using bakta (which employs tRNAscan-SE 2.0 [74]) with default settings. MAGs were dereplicated at approximately species-level (95% average nucleotide identity, ANI, and 50% aligned fraction, AF) with dRep v3.5.0 [75]. dRep was run using only skani v0.2.2 [76] for ANI and AF calculations,

and CheckM2 results were provided to dRep for genome quality information [command: dRep dereplicate --SkipMash --S_algorithm skani -sa 0.95 -nc 0.5 --genomeInfo checkm2_info.csv -p 48].

Proteins were predicted from the HQ soil MAGs using Prodigal v2.6.3 [77] [parameters: -p meta], which was parallelised across 48 CPU cores using GNU Parallel [78]. This resulted in the prediction of 31,890,584 proteins. These proteins were functionally annotated against EcoFoldDB using *EcoFoldDB-annotate* with default settings. Genome sequences and predicted protein sequences of these HQ MAGs have been made available via https://doi.org/10.5281/zenodo.15054612, and EcoFoldDB annotations are available as Additional file 1.

Taxonomy of the HQ soil MAGs was assigned with GTDB-Tk v2.4.0 [79, 80] using GTDB release 220 [81-84] [parameters: classify_wf --skip_ani_screen]. A whole-genome phylogeny was generated using PhyloPhlAn 3.0 [85], employing DIAMOND [18] for marker detection, MAFFT [86] for sequence alignment, trimAl [87] for alignment trimming, and FastTree2 [88], with the WAG substitution model [89], for phylogenetic inference [PhyloPhlAn parameters: --diversity high --fast]. The resulting phylogenetic tree was rooted with minimal ancestor deviation (MAD) rooting [90] using MADroot [91] (https://github.com/davidjamesbryant/MADroot), and visualised using the *ggtree* [92] and *ggtreeExtra* [93] R packages.

Results and Discussion

Here, we present EcoFoldDB, a database of protein structures that cover microbial functions of ecological relevance, including those involved in: (1) trace gas oxidation, (2) carbon cycling, (3) nitrogen cycling, (4) sulphur cycling, (5) phosphorus cycling, (6) iron cycling, (7)

plant-microbe interactions, and (8) osmotic stress tolerance (Fig. 1). These eight functional categories are organised into 35 sub-categories, and encompass 168 different functional activities/pathways (Fig. 1; Supplementary Table S1). Together, EcoFoldDB comprises 867 protein structures linked to 651 genes (Supplementary Table S1). These functions cover key biogeochemical cycles important for ecosystem health, biodiversity, and climate change (i.e., carbon sequestration and greenhouse gas emissions), as well as microbial traits relevant to adaptive responses to environmental stress, and plant growth promotion (Fig. 1). Future versions of EcoFoldDB can be readily expanded to incorporate new functions, particularly as novel microbial traits influencing ecosystem dynamics are uncovered.

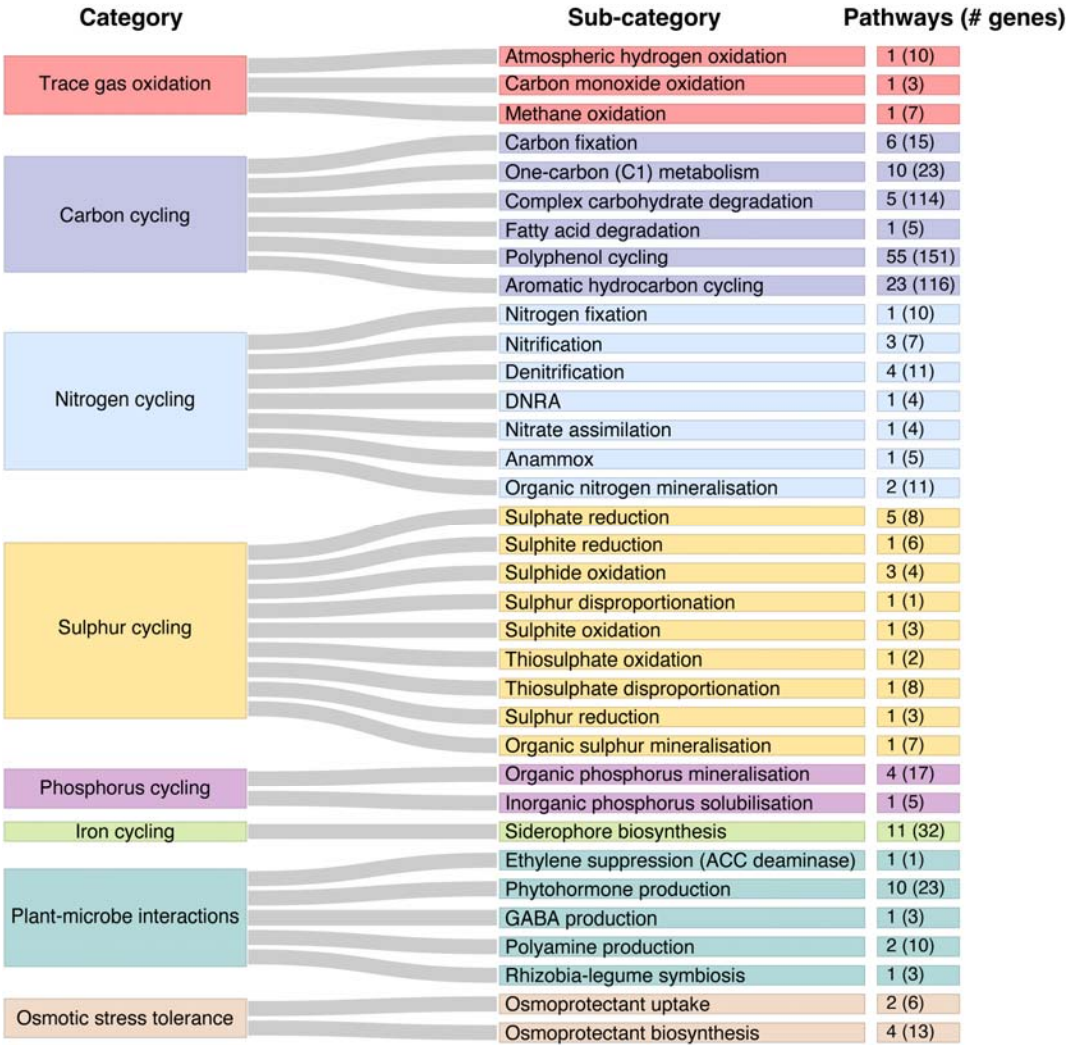

Fig. 1. Ecologically relevant functions covered by EcoFoldDB. The EcoFoldDB ontology is comprised of 8 functional categories (left), which are divided into 35 sub-categories (middle), covering 168 pathways/activities and 651 genes (right). For a full list of EcoFoldDB entries and their functional categorisations, see Supplementary Table S1.

'EcoFoldDB-annotate' performs highly sensitive and accurate structure-guided functional profiling

To enable scalable, structure-guided functional annotation using EcoFoldDB, we developed *EcoFoldDB-annotate* (Fig. 2a). This pipeline leverages Foldseek [23] and the ProstT5 protein language model [25] to directly translate amino acid sequences into 3Di structural representations, enabling homology searches against EcoFoldDB without the need for computationally intensive structure predictions. ProstT5 achieves this translation ~4,000-fold faster than AlphaFold2-ColabFold without losing sensitivity [24], allowing structural homology detection scalable to metagenomic datasets.

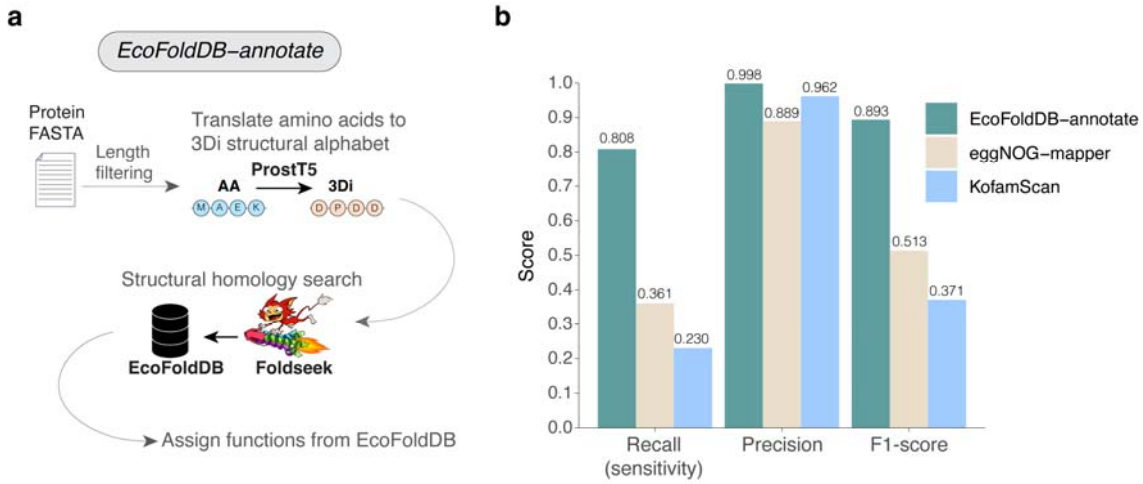

Fig. 2. *EcoFoldDB-annotate* pipeline and performance benchmarking. a, *EcoFoldDB-annotate* first filters protein sequences to remove those larger than 4,000 amino acids in length, then uses the ProstT5 protein language model to translate amino acid sequences to Foldseek's 3D interaction (3Di) structural alphabet for structural homology searching against EcoFoldDB, and finally assigns hierarchically categorised functional annotations based on EcoFoldDB ontology. **b**, The performance of *EcoFoldDB-annotate* was benchmarked against

eggNOG-mapper and KofamScan. The following benchmarking metrics were used: recall/sensitivity (proportion of all possible positives correctly identified), precision (proportion of predicted positives that were true), and F1-score (the harmonic mean of recall and precision).

EcoFoldDB-annotate outperforms sequence-based annotations for proteins derived from metagenomic assemblies (Fig. 2b). Since, *EcoFoldDB-annotate* profiling relies on structural homology, it achieves markedly higher sensitivity, with a recall of 0.808 compared to 0.361 (eggNOG-mapper) and 0.230 (KofamScan), while maintaining the highest precision (0.998) and F1-score (0.893) among the three tools (Fig. 2b). This combination of sensitivity and precision establishes *EcoFoldDB-annotate* as an effective solution for annotating metagenomes, which are frequently dominated by sequences lacking detectable amino acid homology to reference databases [11, 20]. Its performance, scalability, and hierarchical classification of ecologically relevant functions (Fig. 1) positions *EcoFoldDB-annotate* as an invaluable resource for environmental metagenomic studies.

EcoFoldDB-annotate is scalable to millions of proteins from soil metagenomes

To demonstrate the utility of *EcoFoldDB-annotate*, we used it to perform structure-based functional annotations of 32 million protein sequences encoded by 8,000 high-quality metagenome-assembled genomes (MAGs), derived from diverse soil habitats across the globe. Soil microbes harbor a diverse range of metabolic capabilities, and influence global carbon dynamics and biogeochemical cycles that shape ecosystem health and productivity [94]. Soil environments generally harbour greater phylogenetic diversity and genetic novelty compared to other environments [95-97], making sequenced-based functional annotations particularly difficult [95]. Leveraging the recent breakthroughs in protein structure predictions [22, 24] can improve functional annotations of proteins that lie beyond the limits

of sequence homology detection with known proteins [98]. To that end, we employed *EcoFoldDB-annotate* to functionally annotate MAGs from the global soil microbiome. We processed 32 million proteins in batches of 1 million sequences, taking ~5.3 hours per batch using a server with 4 Nvidia Tesla Volta V100 GPUs and 48 CPU cores (see Additional file 1 for all EcoFoldDB annotations). This demonstrates the scalability of *EcoFoldDB-annotate* to functionally profile large metagenomic datasets.

EcoFoldDB reveals the taxonomic distribution of nitrogen cycling functions in the global soil microbiome

As a use case example, we used the EcoFoldDB annotations of the high-quality MAGs to reveal the phylogenetic distribution of encoded nitrogen cycling pathways in soils. Our findings reveal distinct taxonomic partitioning, with some pathways, particularly nitrification, being taxonomically restricted, while others, like dissimilatory nitrate reduction (DNRA) and denitrification, were broadly distributed across taxa (Fig. 3). For instance, ammonia oxidation, the initial step in nitrification, was restricted to just three prokaryotic families. Archaea dominated this functional guild (80% of ammonia oxidisers), with all archaeal representatives belonging to the family Nitrososphaeraceae. This agrees with previous studies showing archaea to be the predominant drivers of ammonia oxidation on the planet [99, 100]. The remaining 20% belong to the bacterial families Nitrospiraceae (phylum Nitrospirota) and Nitrosomonadaceae (phylum Pseudomonadota). Similarly, nitrite oxidation (nitrification) was phylogenetically constrained, occurring predominantly within Nitrospirota (Fig. 3), a pattern consistent with prior reports identifying this phylum as the most widespread nitrite oxidisers [101]. In contrast, pathways like DNRA and denitrification (particularly nitrate, nitrite, and

nitrous oxide reduction) were distributed across diverse bacterial and archaeal taxa, underscoring their functional versatility across lineages and environmental niches.

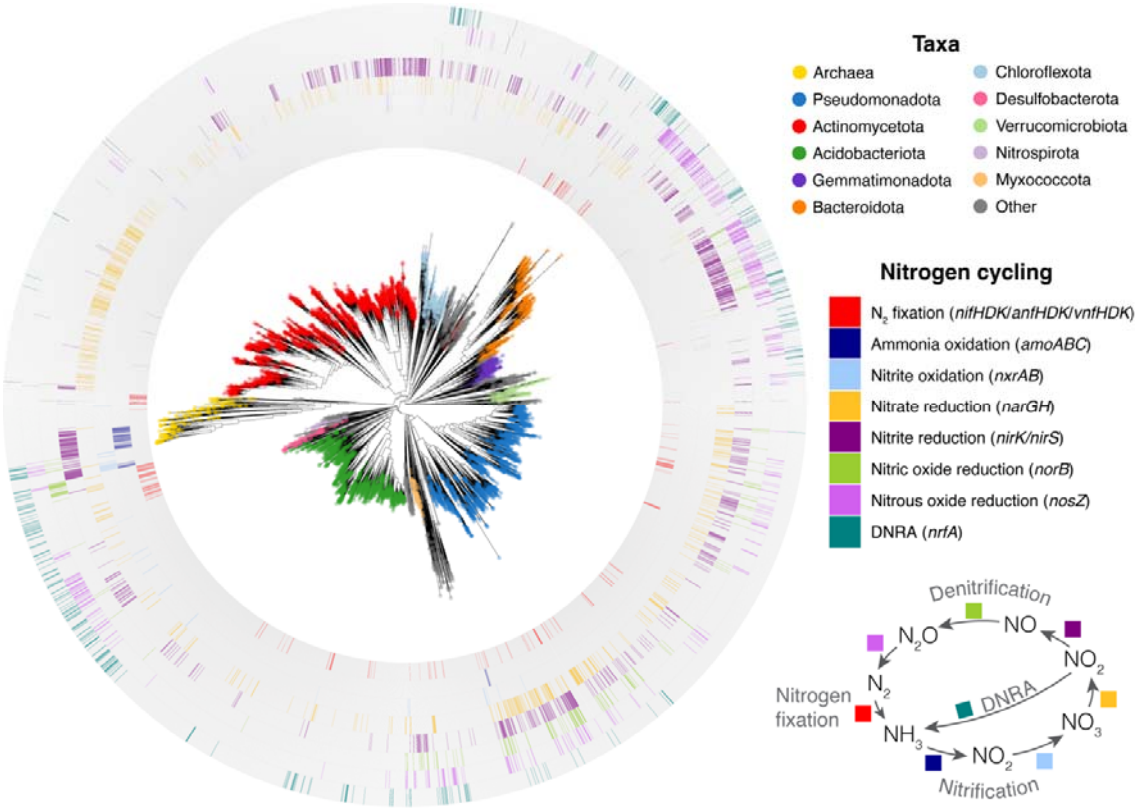

Fig. 3. Nitrogen cycling functions identified by *EcoFoldDB-annotate* among high-quality MAGs from the global soil microbiome. A whole genome phylogeny of high-quality bacterial and archaeal MAGs, derived from soil metagenomes across the globe. Coloured tip points on the tree indicate GTDB phyla, except for archaeal genomes, which have all been assigned a single domain-level colour (yellow tip points). Coloured rings indicate the genomes identified to have the genetic potential for nitrogen cycling functions. As a minimum, the following genes were required for N_2 fixation: (*nifH* and *nifD* and *nifK*) or (*anfH* and *anfD* and *anfK*) or (*vnfH* and *vnfD* and *vnfK*); ammonia oxidation: *amoA*, *amoB* and *amoC*; nitrite oxidation: *nxrA* and *nxrB*; nitrate reduction: *narG* and *narH*; nitrite reduction: *nirK* or *nirS*; nitric oxide reduction: *norB*; nitrous oxide reduction: *nosZ*; and DNRA (dissimilatory nitrite reduction to ammonia): *nrfA*.

The taxonomic and metabolic insights generated here, from phylogenetically limited nitrifying taxa to more widespread denitrifiers, demonstrate *EcoFoldDB-annotate*'s capacity to disentangle complex microbial roles in biogeochemical cycling. This scalable and sensitive

annotation framework thus allows efficient linking of microbial community structure to ecosystem-scale functions in large datasets.

EcoFoldDB identifies microbial genomes encoding plant-microbe interaction traits

As a second use case example, we sought to identify potential plant growth-promoting members from the global soil microbiome. We used the EcoFoldDB annotations to identify soil MAGs encoding plant-microbe interaction traits. We found that many genomes encode multiple plant-microbe interaction traits, with 19% encoding three or more (Fig. 4). A few of these traits were widespread, e.g., putrescine and spermidine production, which were present in most bacterial phyla. This is not surprising since biosynthesis of these polyamines is also prevalent in non-plant-associated bacteria given their diverse roles in bacterial physiology and growth, such as biofilm formation, motility, and siderophore production [102]. The genetic capacity to produce the phytohormone auxin (indole-3-acetic acid, IAA) was also widespread, with 57% of MAGs containing at least one of the known IAA biosynthesis pathways (Fig. 4). This agrees with experimental evidence, which suggest 48-65% of cultured isolates from soil and plant-associated environments can produce IAA [103, 104].

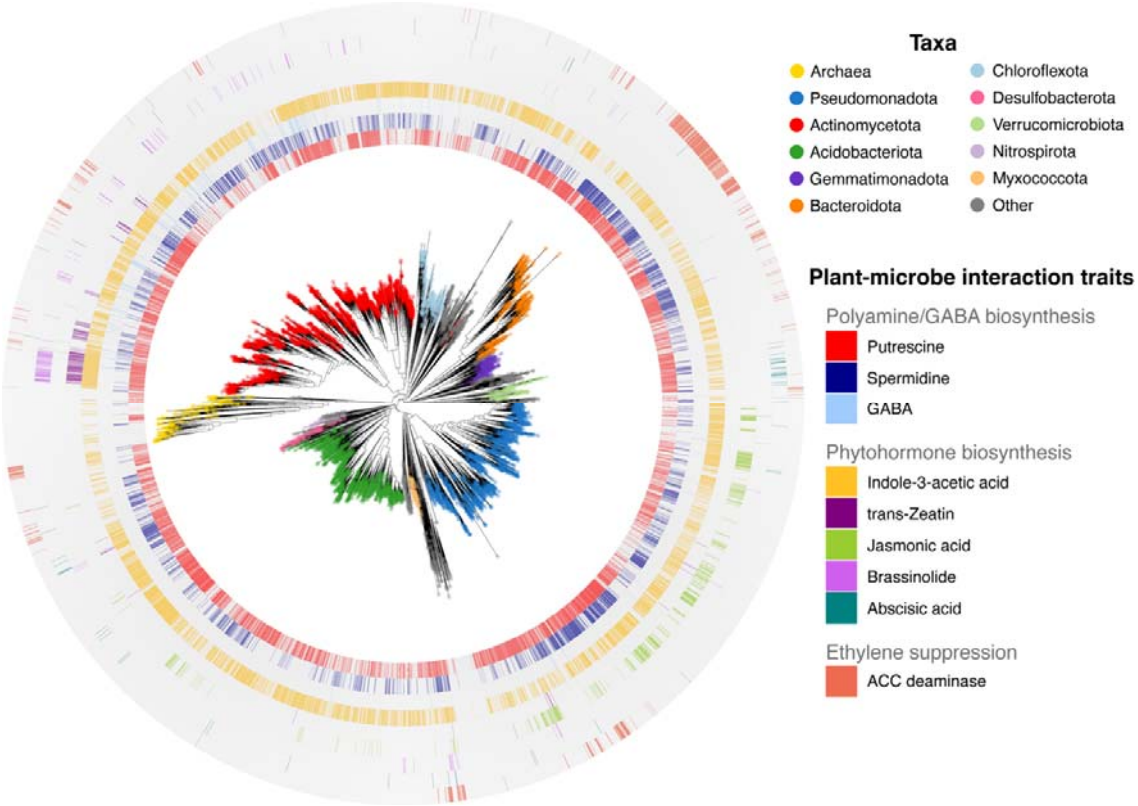

Fig. 4. Plant-microbe interaction functions identified by *EcoFoldDB-annotate* among high-quality MAGs from the global soil microbiome. A whole genome phylogeny of high-quality bacterial and archaeal MAGs, derived from soil metagenomes across the globe. Coloured tip points on the tree indicate GTDB phyla, except for archaeal genomes, which have all been assigned a single domain-level colour (yellow tip points). Coloured rings indicate the genomes identified to have the genetic potential for plant-microbe interaction traits. As a minimum, the following genes were required for putrescine production: *speC* or (*speA* and *speB*) or (*speA* and *aguA* and *aguB*); spermidine production: (*speD*/*speH* and *speE*) or (*cansdh* and *nspC*); GABA (gamma-aminobutyric acid) production: *gadA*/*gadB* and *gadC*; indole-3-acetic acid (IAA) production via the indole-3-acetamide (IAM) pathway: *iaaM* and *iaaH*; IAA production via the indole-3-pyruvic acid (IPA) pathway: *idpC* and *aldH*; IAA production via the tryptamine (TAM) pathway: *tdc* and *maoA* and *aldH*; IAA production via the indole-3-acetonitrile (IAN) pathway: *nit*; trans-Zeatin production: *log* and *cyp735A*; jasmonic acid production: *opr* and *ehhadh* and *kat*; brassinolide production: *br6ox*; abscisic acid production: *xan* and *aao*; gibberellin production: *ko* and *kao* (not shown as only three genomes were identified to carry both genes, Additional file 1); and ACC deaminase: *acdS*.

Other plant-microbe interaction traits, however, had strong taxonomic signals (Fig. 4). For example, 89% of trans-Zeatin producers and 82% of brassinolide producers belonged to the phylum Actinomycetota, largely from the order Mycobacteriales. Jasmonic acid production clustered within Pseudomonadota, notably the orders Rhizobiales, Burkholderiales, Steroidobacterales and Sphingomonadales. Abscisic acid producers were

primarily affiliated with the phyla Verrucomicrobiota and Acidobacteriota, and ACC deaminase activity showed a strong Bacteroidota signature. Indeed, over half (54%) of all Bacteroidota MAGs encoded ACC deaminase, known to induce plant tolerance to abiotic stress [44]. While ACC deaminase activity has been experimentally confirmed in Bacteroidota isolates [105-108], research on this trait has historically focused on Pseudomonadota [109, 110]. Thus, to our knowledge, the pronounced enrichment of ACC deaminase activity within Bacteroidota observed here has not been previously reported.

To identify possible novel plant growth promoting strains, we examined the subset of genomes encoding the greatest number of identified plant-microbe interaction traits, with 26 species-clustered MAGs being positive for five different traits (Fig. 5). Of these, 19 currently have no cultured species representative, with 11 lacking a cultured genus-level representative. The enrichment of encoded plant-microbe interaction traits, suggest that they likely have plant-growth promoting effects. Indeed, most of the species that do have a cultured representative have already been shown to comprise plant growth-promoting members (e.g., *Rhodococcus erythropolis* [111, 112], *Achromobacter kerstersii* [113], *Pseudomonas putida* [114-116], *Pseudomonas furukawaii* [117]). The remaining species, which are also likely to comprise plant growth-promoting members, represent ideal candidates for targeted isolation and culturing. Such isolation efforts can be guided by genomic information from the MAG sequences, e.g., via optimisation of culture medium and conditions, including substrate utilisation, oxygen requirements, resistance to antibiotics, etc. [118-120]. This demonstrates how EcoFoldDB can help bridge ecological discovery and real-world application, translating metagenomic insights into actionable strategies for sustainable management.

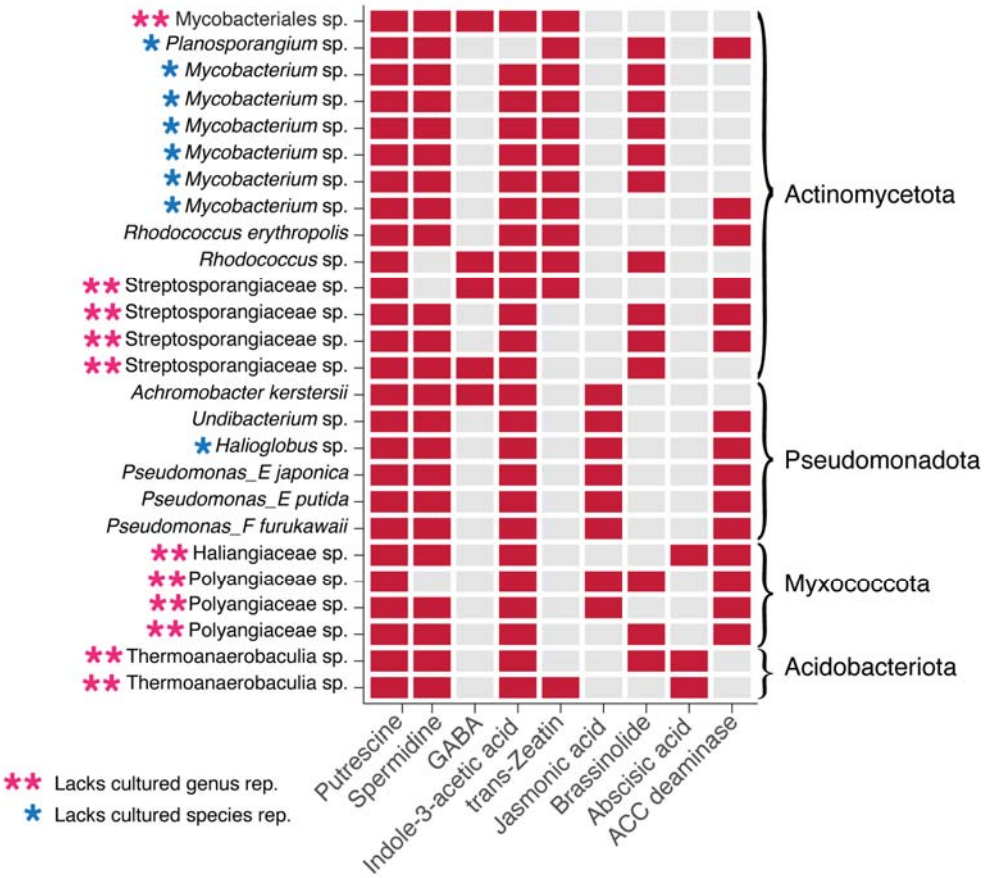

Fig. 5. Identification of potential plant growth-promoting bacteria as candidates for targeted culturing. Soil MAGs encoding the greatest number of plant-microbe interaction traits (x-axis), as determined by *EcoFoldDB-annotate* are shown, labelled with their lowest named GTDB taxonomic rank (y-axis). Each MAG represents a distinct species, sharing <95% average nucleotide identity. The minimum gene sets for a genome to be considered positive (red rectangles) for each trait are as listed in Fig. 4. Strains lacking a cultured species representative are denoted with a single blue asterisk, and those also lacking a cultured genus representative are shown with double pink asterisks. Figure data, including the genome identifiers for these 26 MAGs, are available as Supplementary Table S2.

Conclusion

Here, we introduce EcoFoldDB, a database of protein structures linked to ecologically relevant functions, and its companion pipeline, *EcoFoldDB-annotate*, for structure-guided functional profiling. *EcoFoldDB-annotate* achieves highly sensitive and accurate annotations, scalable to metagenomic datasets (millions of proteins). We showcase the use of *EcoFoldDB-*

annotate on a global soil microbiome MAG catalogue, revealing key insights into the taxonomic distribution of important nitrogen cycling pathways, and plant-microbe interaction traits, advancing our understanding of important ecosystem processes. We anticipate that EcoFoldDB will serve as a valuable resource for microbial ecology studies, enabling researchers to extract ecological insights from environmental genomes and metagenomes, and accelerating discoveries in microbiome science.

Data availability

EcoFoldDB and *EcoFoldDB-annotate* pipeline are available via https://github.com/timghaly/EcoFoldDB. Genomic and protein sequences for the high-quality soil MAGs are available via https://doi.org/10.5281/zenodo.15054612. EcoFoldDB annotations of these genomes are available as Additional file 1.

Acknowledgements

We would like to thank the developers of all the software used in this study, particularly Foldseek and ProstT5, without which, this work would not have been possible. TMG would also like to thank Mary, Saoirse, and Maebh Ghaly for their loving support.

References

1. Steffen W, Richardson K, Rockström J, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, de Vries W, de Wit CA, et al: **Planetary boundaries: Guiding human development on a changing planet.** *Science* 2015, **347:**1259855.
2. Wardle DA, Bardgett RD, Klironomos JN, Setälä H, van der Putten WH, Wall DH: **Ecological linkages between aboveground and belowground biota.** *Science* 2004, **304:**1629-1633.

3. Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, Behrenfeld MJ, Boetius A, Boyd PW, Classen AT, et al: **Scientists' warning to humanity: microorganisms and climate change.** *Nature Reviews Microbiology* 2019, **17:**569-586.
4. Gillings MR, Paulsen IT: **Microbiology of the Anthropocene.** *Anthropocene* 2014, **5:**1-8.
5. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H, et al: **Structure and function of the global topsoil microbiome.** *Nature* 2018, **560:**233-237.
6. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al: **Structure and function of the global ocean microbiome.** *Science* 2015, **348:**1261359.
7. Bahram M, Espenberg M, Pärn J, Lehtovirta-Morley L, Anslan S, Kasak K, Kõljalg U, Liira J, Maddison M, Moora M: **Structure and function of the soil microbiome underlying N_2O emissions from global wetlands.** *Nature Communications* 2022, **13:**1430.
8. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, et al: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, **486:**207-214.
9. Zhou Z, Tran PQ, Martin C, Rohwer RR, Baker BJ, McMahon KD, Anantharaman K: **Unravelling viral ecology and evolution over 20 years in a freshwater lake.** *Nature Microbiology* 2025, **10:**231-245.
10. Rohwer RR, Kirkpatrick M, Garcia SL, Kellom M, McMahon KD, Baker BJ: **Two decades of bacterial ecology and evolution in a freshwater lake.** *Nature Microbiology* 2025, **10:**246-257.
11. Pavlopoulos GA, Baltoumas FA, Liu S, Selvitopi O, Camargo AP, Nayfach S, Azad A, Roux S, Call L, Ivanova NN, et al: **Unraveling the functional dark matter through global metagenomics.** *Nature* 2023, **622:**594-602.
12. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30:**1236-1240.
13. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J: **eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale.** *Molecular Biology and Evolution* 2021, **38:**5825-5829.
14. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H: **KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold.** *Bioinformatics* 2019, **36:**2251-2252.
15. Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, Karaoz U, Anantharaman K: **METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks.** *Microbiome* 2022, **10:**33.
16. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa Sabina L, Solden LM, Liu P, Narrowe AB, Rodríguez-Ramos J, Bolduc B, et al: **DRAM for distilling microbial metabolism to automate the curation of microbiome function.** *Nucleic Acids Research* 2020, **48:**8883-8900.

17. Steinegger M, Söding J: **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.** *Nature biotechnology* 2017, **35:**1026-1028.
18. Buchfink B, Reuter K, Drost H-G: **Sensitive protein alignments at tree-of-life scale using DIAMOND.** *Nature Methods* 2021, **18:**366-368.
19. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic acids research* 2011, **39:**W29-W37.
20. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, Delmont TO, Duarte CM, Eren AM, Finn RD, et al: **Unifying the known and unknown microbial coding sequence space.** *eLife* 2022, **11:**e67667.
21. Illergård K, Ardell DH, Elofsson A: **Structure is three to ten times more conserved than sequence—a study of structural response in protein cores.** *Proteins: Structure, Function, and Bioinformatics* 2009, **77:**499-508.
22. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al: **Highly accurate protein structure prediction with AlphaFold.** *Nature* 2021, **596:**583-589.
23. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M: **Fast and accurate protein structure search with Foldseek.** *Nature Biotechnology* 2024, **42:**243-246.
24. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M: **ColabFold: making protein folding accessible to all.** *Nature Methods* 2022, **19:**679-682.
25. Heinzinger M, Weissenow K, Sanchez Joaquin G, Henkel A, Mirdita M, Steinegger M, Rost B: **Bilingual language model for protein sequence and structure.** *NAR Genomics and Bioinformatics* 2024, **6:**lqae150.
26. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M: **Prottrans: Toward understanding the language of life through self-supervised learning.** *IEEE transactions on pattern analysis and machine intelligence* 2021, **44:**7112-7127.
27. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al: **AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences.** *Nucleic Acids Research* 2023, **52:**D368-D375.
28. Søndergaard D, Pedersen CNS, Greening C: **HydDB: A web tool for hydrogenase classification and analysis.** *Scientific Reports* 2016, **6:**34212.
29. Bay SK, Dong X, Bradley JA, Leung PM, Grinter R, Jirapanjawat T, Arndt SK, Cook PLM, LaRowe DE, Nauer PA, et al: **Trace gas oxidizers are widespread and active members of soil microbial communities.** *Nature Microbiology* 2021, **6:**246-256.
30. Ruiz-Fernández P, Ramírez-Flandes S, Rodríguez-León E, Ulloa O: **Autotrophic carbon fixation pathways along the redox gradient in oxygen-depleted oceanic waters.** *Environmental Microbiology Reports* 2020, **12:**334-341.
31. Garritano AN, Song W, Thomas T: **Carbon fixation pathways across the bacterial and archaeal tree of life.** *PNAS Nexus* 2022, **1:**pgac226.
32. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, et al: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.** *Nucleic Acids Research* 2013, **42:**D459-D471.

33. Zheng J, Ge Q, Yan Y, Zhang X, Huang L, Yin Y: **dbCAN3: automated carbohydrate-active enzyme and substrate annotation.** *Nucleic Acids Research* 2023, **51:**W115-W121.
34. McGivern BB, Cronin DR, Ellenbogen JB, Borton MA, Knutson EL, Freire-Zapata V, Bouranis JA, Bernhardt L, Hernandez AI, Flynn RM, et al: **Microbial polyphenol metabolism is part of the thawing permafrost carbon cycle.** *Nature Microbiology* 2024, **9:**1454-1466.
35. McGivern BB, Woyda R, Flynn RM, Wrighton KC: **CAMPER: curated annotations for profiling microbial polyphenol metabolic potential.** *bioRxiv* 2024**:**2023.2009.2024.559193.
36. Tu Q, Lin L, Cheng L, Deng Y, He Z: **NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes.** *Bioinformatics* 2018, **35:**1040-1048.
37. Eichhorn E, van der Ploeg JR, Kertesz MA, Leisinger T: **Characterization of α-ketoglutarate-dependent taurine dioxygenase from *Escherichia coli*.** *Journal of Biological Chemistry* 1997, **272:**23031-23036.
38. Ellis HR: **Mechanism for sulfur acquisition by the alkanesulfonate monooxygenase system.** *Bioorganic Chemistry* 2011, **39:**178-184.
39. Jamshad M, Murrell JC, Fülöp V: **Purification and crystallization of the hydroxylase component of the methanesulfonate monooxygenase from *Methylosulfonomonas methylovora* strain M2.** *Protein Expression and Purification* 2007, **52:**472-477.
40. Soutourina J, Blanquet S, Plateau P: **Role of d-cysteine desulfhydrase in the adaptation of *Escherichia coli* to d-cysteine.** *Journal of Biological Chemistry* 2001, **276:**40864-40872.
41. Morra MJ, Dick WA: **Mechanisms of H_2S production from cysteine and cystine by microorganisms isolated from soil by selective enrichment.** *Applied and Environmental Microbiology* 1991, **57:**1413-1417.
42. Zeng J, Tu Q, Yu X, Qian L, Wang C, Shu L, Liu F, Liu S, Huang Z, He J, et al: **PCycDB: a comprehensive and accurate database for fast analysis of phosphorus cycling genes.** *Microbiome* 2022, **10:**101.
43. Siles JA, Starke R, Martinovic T, Parente Fernandes ML, Orgiazzi A, Bastida F: **Distribution of phosphorus cycling genes across land uses and microbial taxonomic groups based on metagenome and genome mining.** *Soil Biology and Biochemistry* 2022, **174:**108826.
44. Glick BR, Todorovic B, Czarny J, Cheng Z, Duan J, McConkey B: **Promotion of plant growth by bacterial ACC deaminase.** *Critical Reviews in Plant Sciences* 2007, **26:**227-242.
45. Tang J, Li Y, Zhang L, Mu J, Jiang Y, Fu H, Zhang Y, Cui H, Yu X, Ye Z: **Biosynthetic pathways and functions of indole-3-acetic acid in microorganisms.** *Microorganisms* 2023, **11:**2077.
46. Bottini R, Cassán F, Piccoli P: **Gibberellin production by bacteria and its involvement in plant growth promotion and yield increase.** *Applied Microbiology and Biotechnology* 2004, **65:**497-503.
47. Khalil A, Bramucci AR, Focardi A, Le Reun N, Willams NLR, Kuzhiumparambil U, Raina J-B, Seymour JR: **Widespread production of plant growth-promoting hormones among marine bacteria and their impacts on the growth of a marine diatom.** *Microbiome* 2024, **12:**205.
48. Frébortová J, Frébort I: **Biochemical and structural aspects of cytokinin biosynthesis and degradation in bacteria.** *Microorganisms* 2021, **9:**1314.

49. Li C, Schilmiller AL, Liu G, Lee GI, Jayanty S, Sageman C, Vrebalov J, Giovannoni JJ, Yagi K, Kobayashi Y, Howe GA: **Role of β-oxidation in jasmonate biosynthesis and systemic wound signaling in tomato.** *The Plant Cell* 2005, **17:**971-986.
50. Bajguz A, Chmur M, Gruszka D: **Comprehensive overview of the brassinosteroid biosynthesis pathways: Substrates, products, inhibitors, and connections.** *Frontiers in Plant Science* 2020, **11:**1034.
51. Ueno H: **Enzymatic and structural aspects on glutamate decarboxylase.** *Journal of Molecular Catalysis B: Enzymatic* 2000, **10:**67-79.
52. Ma D, Lu P, Yan C, Fan C, Yin P, Wang J, Shi Y: **Structure and mechanism of a glutamate–GABA antiporter.** *Nature* 2012, **483:**632-636.
53. Kinnersley AM, Turano FJ: **Gamma aminobutyric acid (GABA) and plant responses to stress.** *Critical Reviews in Plant Sciences* 2000, **19:**479-509.
54. Dunn MF, Becerra-Rivera VA: **The biosynthesis and functions of polyamines in the interaction of plant growth-promoting rhizobacteria with plants.** *Plants* 2023, **12:**2671.
55. Rossen L, Shearman C, Johnston A, Downie J: **The nodD gene of Rhizobium leguminosarum is autoregulatory and in the presence of plant exudate induces the nodA, B, C genes.** *The EMBO Journal* 1985, **4:**3369-3373.
56. Prell J, Poole P: **Metabolic changes of rhizobia in legume nodules.** *Trends in Microbiology* 2006, **14:**161-168.
57. Ruhal R, Kataria R, Choudhury B: **Trends in bacterial trehalose metabolism and significant nodes of metabolic pathway in the direction of trehalose accumulation.** *Microbial Biotechnology* 2013, **6:**493-502.
58. Chen C, Beattie GA: ***Pseudomonas syringae* BetT Is a low-affinity choline transporter that is responsible for superior osmoprotection by choline over glycine betaine.** *Journal of Bacteriology* 2008, **190:**2717-2725.
59. Jebbar M, Talibart R, Gloux K, Bernard T, Blanco C: **Osmoprotection of *Escherichia coli* by ectoine: uptake and accumulation characteristics.** *Journal of Bacteriology* 1992, **174:**5027-5035.
60. Yang T, Nian Y, Lin H, Li J, Lin X, Li T, Wang R, Wang L, Beattie GA, Zhang J, Fan M: **Structure and mechanism of the osmoregulated choline transporter BetT.** *Science Advances* 2024, **10:**eado6229.
61. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Research* 2003, **31:**365-370.
62. Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M: **Clustering predicted structures at the scale of the known protein universe.** *Nature* 2023, **622:**637-645.
63. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al: **Evolutionary-scale prediction of atomic-level protein structure with a language model.** *Science* 2023, **379:**1123-1130.
64. Richardson L, Allen B, Baldi G, Beracochea M, Bileschi Maxwell L, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell Lucy J, et al: **MGnify: the microbiome sequence data analysis resource in 2023.** *Nucleic Acids Research* 2022, **51:**D753-D759.
65. Xu J, Zhang Y: **How significant is a protein structure similarity with TM-score = 0.5?** *Bioinformatics* 2010, **26:**889-895.

66. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen Lars J, et al: **eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.** *Nucleic Acids Research* 2018, **47:**D309-D314.
67. Ma B, Lu C, Wang Y, Yu J, Zhao K, Xue R, Ren H, Lv X, Pan R, Zhang J, et al: **A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources.** *Nature Communications* 2023, **14:**7318.
68. Sereika M, Mussig AJ, Jiang C, Knudsen KS, Jensen TBN, Petriglieri F, Yang Y, Jørgensen VR, Delogu F, Sørensen EA, et al: **Recovery of highly contiguous genomes from complex terrestrial habitats reveals over 15,000 novel prokaryotic species and expands characterization of soil and sediment microbial communities.** *bioRxiv* 2024**:**2024.2012.2019.629313.
69. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, et al: **Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.** *Nature Biotechnology* 2017, **35:**725-731.
70. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW: **CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning.** *Nature Methods* 2023, **20:**1203-1212.
71. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A: **Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification.** *Microbial Genomics* 2021, **7**.
72. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics* 2013, **29:**2933-2935.
73. Ontiveros-Palacios N, Cooke E, Nawrocki EP, Triebel S, Marz M, Rivas E, Griffiths-Jones S, Petrov AI, Bateman A, Sweeney B: **Rfam 15: RNA families database in 2025.** *Nucleic acids research* 2025, **53:**D258-D267.
74. Chan Patricia P, Lin Brian Y, Mak Allysia J, Lowe Todd M: **tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes.** *Nucleic Acids Research* 2021, **49:**9077-9096.
75. Olm MR, Brown CT, Brooks B, Banfield JF: **dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication.** *The ISME Journal* 2017, **11:**2864-2868.
76. Shaw J, Yu YW: **Fast and robust metagenomic sequence comparison through sparse chaining with skani.** *Nature Methods* 2023, **20:**1661-1665.
77. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics* 2010, **11:**119.
78. Tange O: **GNU Parallel 2018, Mar 2018, ISBN 9781387509881, DOI https://doi.org/10.5281/zenodo.1146014.** 2018.
79. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH: **GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.** *Bioinformatics* 2019, **36:**1925-1927.

80. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH: **GTDB-Tk v2: memory friendly classification with the genome taxonomy database.** *Bioinformatics* 2022, **38:**5315-5316.
81. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P: **A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.** *Nature Biotechnology* 2018, **36:**996-1004.
82. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P: **A complete domain-to-species taxonomy for Bacteria and Archaea.** *Nature Biotechnology* 2020, **38:**1079-1086.
83. Rinke C, Chuvochina M, Mussig AJ, Chaumeil P-A, Davín AA, Waite DW, Whitman WB, Parks DH, Hugenholtz P: **A standardized archaeal taxonomy for the Genome Taxonomy Database.** *Nature Microbiology* 2021, **6:**946-959.
84. Parks DH, Chuvochina M, Rinke C, Mussig AJ, 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* 2021, **50:**D785-D794.
85. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U, et al: **Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0.** *Nature Communications* 2020, **11:**2500.
86. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Molecular biology and evolution* 2013, **30:**772-780.
87. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.** *Bioinformatics* 2009, **25:**1972-1973.
88. Price MN, Dehal PS, Arkin AP: **FastTree 2 – Approximately maximum-likelihood trees for large alignments.** *PLoS One* 2010, **5:**e9490.
89. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Molecular Biology and Evolution* 2001, **18:**691-699.
90. Tria FDK, Landan G, Dagan T: **Phylogenetic rooting using minimal ancestor deviation.** *Nature Ecology and Evolution* 2017, **1:**193.
91. Bryant D, Charleston M: **MAD roots for large trees.** *arXiv* 2018**:**1811.03174.
92. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY: **ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data.** *Methods in Ecology and Evolution* 2017, **8:**28-36.
93. Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, Tang W, Feng T, Chen M, Zhan L, et al: **ggtreeExtra: Compact visualization of richly annotated phylogenetic data.** *Molecular Biology and Evolution* 2021, **38:**4039-4042.
94. Fierer N: **Embracing the unknown: disentangling the complexities of the soil microbiome.** *Nature Reviews Microbiology* 2017, **15:**579-590.
95. Holland-Moritz H, Vanni C, Fernandez-Guerra A, Bissett A, Fierer N: **An ecological perspective on microbial genes of unknown function in soil.** *bioRxiv* 2021**:**2021.2012.2002.470747.

96. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, et al: **A communal catalogue reveals Earth's multiscale microbial diversity.** *Nature* 2017, **551:**457-463.
97. Brewer TE, Aronson EL, Arogyaswamy K, Billings SA, Botthoff JK, Campbell AN, Dove NC, Fairbanks D, Gallery RE, Hart SC, et al: **Ecological and genomic attributes of novel bacterial taxa that thrive in subsurface soil horizons.** *mBio* 2019, **10:**10.1128/mbio.01318-01319.
98. Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, Tauriello G, Akdel M, Andreeva A, Bateman A, et al: **Uncovering new families and folds in the natural protein universe.** *Nature* 2023, **622:**646-653.
99. Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C: **Archaea predominate among ammonia-oxidizing prokaryotes in soils.** *Nature* 2006, **442:**806-809.
100. Wuchter C, Abbas B, Coolen MJL, Herfort L, van Bleijswijk J, Timmers P, Strous M, Teira E, Herndl GJ, Middelburg JJ, et al: **Archaeal nitrification in the ocean.** *Proceedings of the National Academy of Sciences* 2006, **103:**12317-12322.
101. Leung PM, Daebeler A, Chiri E, Hanchapola I, Gillett DL, Schittenhelm RB, Daims H, Greening C: **A nitrite-oxidising bacterium constitutively consumes atmospheric hydrogen.** *The ISME Journal* 2022, **16:**2213-2219.
102. Michael AJ: **Polyamine function in archaea and bacteria.** *Journal of Biological Chemistry* 2018, **293:**18693-18701.
103. Abd-Alla MH, El-Sayed E-SA, Rasmey A-HM: **Indole-3-acetic acid (IAA) production by *Streptomyces atrovirens* isolated from rhizospheric soil in Egypt.** *Journal of Biology and Earth Sciences* 2013, **3:**182-193.
104. Russell D, Rajabal V, Alfonzetti M, van der Merwe MM, Gallagher RV, Tetu SG: **Seed banking impacts native *Acacia ulicifolia* seed microbiome composition and function.** *Environmental Microbiome* 2025, **20:**4.
105. Maimaiti J, Zhang Y, Yang J, Cen YP, Layzell DB, Peoples M, Dong Z: **Isolation and characterization of hydrogen-oxidizing bacteria induced following exposure of soil to hydrogen gas and their impact on plant growth.** *Environmental Microbiology* 2007, **9:**435-444.
106. Gandhi Pragash M, Narayanan KB, Naik PR, Sakthivel N: **Characterization of *Chryseobacterium aquaticum* strain PUPC1 producing a novel antifungal protease from rice rhizosphere soil.** *Journal of Microbiology and Biotechnology* 2009, **19:**99-107.
107. Nadeem SM, Zahir ZA, Naveed M, Arshad M: **Rhizobacteria containing ACC-deaminase confer salt tolerance in maize grown on salt-affected fields.** *Canadian Journal of Microbiology* 2009, **55:**1302-1309.
108. Marques AP, Pires C, Moreira H, Rangel AO, Castro PM: **Assessment of the plant growth promotion abilities of six bacterial isolates using *Zea mays* as indicator plant.** *Soil Biology and Biochemistry* 2010, **42:**1229-1235.
109. Blaha D, Prigent-Combaret C, Mirza MS, Moenne-Loccoz Y: **Phylogeny of the 1-aminocyclopropane-1-carboxylic acid deaminase-encoding gene *acdS* in phytobeneficial and pathogenic Proteobacteria and relation with strain biogeography.** *FEMS Microbiology Ecology* 2006, **56:**455-470.
110. Nascimento FX, Rossi MJ, Soares CRFS, McConkey BJ, Glick BR: **New insights into 1-aminocyclopropane-1-carboxylate (ACC) deaminase**

phylogeny, evolution and ecological significance. *PLoS One* 2014, **9:**e99168.

111. Behzad A, Diakite S, Astarkhanova TS, Pakina EN, Allen D, Mohammadi PM, Diarra O, Saquee FS: **Effectiveness of *Rhodococcus erythropolis* strain OPI-01 on the fungal development in winter wheat.** *Biodiversitas Journal of Biological Diversity* 2024, **25:**1063-1070.

112. Trivedi P, Pandey A, Sa T: **Chromate reducing and plant growth promoting activies of psychrotrophic *Rhodococcus erythropolis* MtCC 7905.** *Journal of Basic Microbiology* 2007, **47:**513-517.

113. Vijay K, Hassan S, Govarthanan M, Kavitha T: **Biosurfactant lipopeptides and polyketide biosynthetic gene in rhizobacterium *Achromobacter kerstersii* to induce systemic resistance in tomatoes.** *Rhizosphere* 2022, **23:**100558.

114. Ahemad M, Khan MS: **Evaluation of plant-growth-promoting activities of rhizobacterium *Pseudomonas putida* under herbicide stress.** *Annals of microbiology* 2012, **62:**1531-1540.

115. Hall JA, Peirson D, Ghosh S, Glick B: **Root elongation in various agronomic crops by the plant growth promoting rhizobacterium *Pseudomonas putida* GR12–2.** *Israel Journal of Plant Sciences* 1996, **44:**37-42.

116. Costa-Gutierrez SB, Lami MJ, Santo MCC-D, Zenoff AM, Vincent PA, Molina-Henares MA, Espinosa-Urgel M, de Cristóbal RE: **Plant growth promotion by *Pseudomonas putida* KT2440 under saline stress: role of *eptA*.** *Applied microbiology and biotechnology* 2020, **104:**4577-4592.

117. Rafique E, Mumtaz MZ, Ullah I, Rehman A, Qureshi KA, Kamran M, Rehman MU, Jaremko M, Alenezi MA: **Potential of mineral-solubilizing bacteria for physiology and growth promotion of *Chenopodium quinoa* Willd.** *Frontiers in Plant Science* 2022, **13:**1004833.

118. Liu S, Moon CD, Zheng N, Huws S, Zhao S, Wang J: **Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation.** *Microbiome* 2022, **10:**76.

119. Karnachuk OV, Lukina AP, Kadnikov VV, Sherbakova VA, Beletsky AV, Mardanov AV, Ravin NV: **Targeted isolation based on metagenome-assembled genomes reveals a phylogenetically distinct group of thermophilic spirochetes from deep biosphere.** *Environmental Microbiology* 2021, **23:**3585-3598.

120. Pope P, Smith W, Denman S, Tringe S, Barry K, Hugenholtz P, McSweeney C, McHardy A, Morrison M: **Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies.** *Science* 2011, **333:**646-648.