Nucleic Acids Research, 2026, **54**, D835–D841
https://doi.org/10.1093/nar/gkaf1241
Advance access publication date: 26 November 2025
Database issue

metaTraits: a large-scale integration of microbial phenotypic trait information

Daniel Podlesny[1,*,†], **Chan Yeong Kim**[1,†], **Shahriyar Mahdi Robbani**[1], **Christian Schudoma**[1], **Anthony Fullam**[1], **Lorenz C. Reimer**[2], **Julia Koblitz**[2], **Isabel Schober**[2], **Anandhi Iyappan**[1], **Thea Van Rossum**[1,3], **Jonas Schiller**[1], **Anastasia Grekova**[1], **Michael Kuhn**[1], **Peer Bork**[1,4,*]

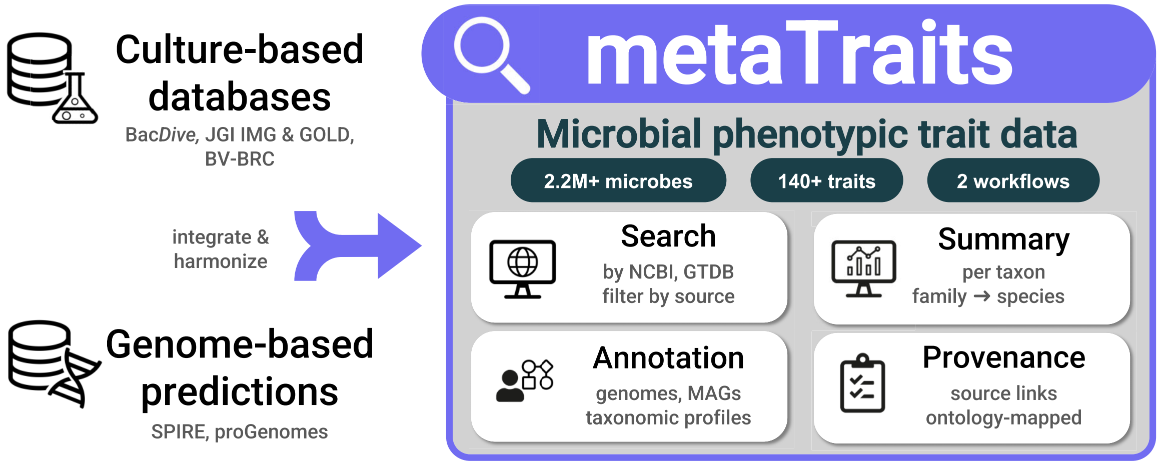
[1] European Molecular Biology Laboratory, Molecular Systems Biology Unit, 69117 Heidelberg, Germany
[2] Leibniz Institute DSMZ, 38124 Braunschweig, Germany
[3] Present address: Koonkie, Vancouver, Canada
[4] Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

[*] To whom correspondence should be addressed. Email: peer.bork@embl.org
Correspondence may also be addressed to Daniel Podlesny. Email: daniel.podlesny@embl.de
[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

Microbes differ greatly in their organismal structure, physiology, and environmental adaptation, yet information about these phenotypic traits is dispersed across multiple databases and is largely unavailable for taxa that remain uncultured. Here, we present metaTraits, a unified and accessible trait resource that integrates culture-derived trait information from Bac*Dive*, BV-BRC, JGI IMG, and GOLD with genome-based predictions for medium and high-quality isolate and metagenome-assembled genomes (MAGs) from proGenomes and SPIRE. metaTraits covers over 2.2 million genomes and >140 harmonized traits mapped to standardized ontologies, spanning cell morphology (e.g. shape, size, and Gram staining), physiology (e.g. motility and sporulation), metabolic and enzymatic activities, environmental preferences (e.g. temperature, salinity, and oxygen tolerance), and lifestyle categories. All records are linked to the original evidence, and species are cross-linked to NCBI and GTDB taxonomies. The interactive metaTraits website provides search and visualization tools, taxonomy-level summaries, and two workflows for annotating user-submitted genomes or community profiles. metaTraits substantially advances accessibility and interoperability of microbial trait data, enabling comprehensive trait-based analyses of microbiomes across diverse environments. metaTraits is accessible via https://metatraits.embl.de.

Graphical abstract



Introduction

Knowledge of microbial phenotypic traits provides essential insights into microbial functions, ecology, and interactions within both environmental and human-associated ecosystems. Trait information has been used to inform ecological modeling [1, 2], to investigate the codiversification of humans and their gut microbes [3], to characterize diverse global microbial habitats [4], to study functional microbiome shifts in disease contexts [5, 6], and to reveal links between traits and biogeographical and social patterns in microbial strain sharing networks [7, 8], among many other applications. Despite their broad utility and recognized importance, microbial phenotypic trait data remain fragmented across several culture-based

Received: August 14, 2025. **Revised:** October 20, 2025. **Accepted:** October 21, 2025

repositories, limiting comprehensive and large-scale biological analysis.

Among the most widely used trait databases, Bac*Dive* [9] provides detailed trait data for ∼100 000 strains (i.e. records describing individual cultivated microbial entities, with or without genome sequences) from 21 000 species, while resources such as BV-BRC (formerly PATRIC [10]), JGI's IMG [11], and GOLD [12] also capture trait metadata as part of their genome and sample submissions. However, these databases vary in their data models and curation standards, and trait records are not harmonized across sources, making integration into comparative genomics and microbiome research challenging. As a result, most published analyses rely on a single source and are often limited to well-studied, cultivated isolates.

Several efforts have aimed to unify microbial trait data. Madin *et al.* [13] standardized 26 data sources into a single trait resource for roughly 170 000 microbial records, while BactoTraits [14] mined Bac*Dive* and other datasets to assemble 19 core traits for nearly 20 000 microbes. Earlier projects, such as the Microbe Directory [15], pioneered a community curation approach to trait annotation. Unfortunately, these resources have either not been updated for years, focus exclusively on isolates, cover a limited set of traits, or lack interactive platforms for data exploration and annotation.

Crucially, the vast majority of prokaryotic diversity remains uncultured. Recent estimates suggest that 62% of recognized microbial phyla and 73% of species are represented only by metagenome-assembled genomes (MAGs) [16]. Existing isolate-focused databases thus overlook much of the microbial world, leading to large gaps in trait coverage. Computational methods such as Bac*Dive*-AI [17], GenomeSPOT [18], MICROPHERRET [19], and Traitar [20] have emerged to enable the prediction of microbial traits from genome sequences, potentially extending trait annotation to hundreds of traits (albeit with different accuracy) across millions of publicly available genomes. Yet, no resource systematically integrates both curated culture-derived traits and genome-based predictions across large-scale genome catalogs.

Here, we present metaTraits, a unified and accessible microbial trait resource that harmonizes trait data from major culture-based databases and systematically extends trait annotation to both isolate genomes and MAGs using genome-based prediction tools. metaTraits integrates phenotypic trait data for 2.2 million genomes from over 100 000 SPIRE species-level clusters (≥95% average nucleotide identity), encompassing >140 harmonized traits relevant to microbial morphology, metabolism, lifestyle, and ecology, among others. All data are mapped to standardized ontologies, cross-linked to external resources, and made available in alignment with FAIR principles [21]. By bridging the gap between culture-based and genome-based data, metaTraits enables trait-based microbiome analyses at an unprecedented scale and scope. The resource is accessible via an interactive website that supports user-friendly trait exploration and annotation workflows. metaTraits is openly available at https://metatraits.embl.de.

Database construction

Integrating phenotypic trait data from public databases

metaTraits integrates microbial phenotypic trait information from major public databases, including Bac*Dive*, JGI IMG and GOLD, and BV-BRC. The integrated data encompass microbial physiology (e.g. motility and sporulation), environmental preferences (e.g. oxygen requirements, pH, salinity, and temperature), morphology, metabolic and enzymatic activities, and more. Source data were collected via APIs or downloaded as tabular metadata files. To ensure consistency, trait data were manually recoded and converted to common data types and units, with harmonization of trait and category naming, correction of spelling and number formatting errors, and removal of significant outliers. For the isolate-centric databases, metaTraits captures 1 182 280 trait observations across 17 159 species (GTDB; Table 1).

Genome-based prediction of phenotypic traits

To address the substantial gaps resulting from uncultured taxa, metaTraits systematically incorporates genome-based phenotypic trait predictions using state-of-the-art computational tools, including Bac*Dive*-AI [17], GenomeSPOT [18], MICROPHERRET [19], and Traitar [20]. These predictions enable extrapolation of phenotypic traits to novel taxa based solely on genomic data, vastly expanding trait coverage beyond what is possible from cultured species alone. As these tools have been trained on different sets of genomes and phenotypes, their predictions inevitably vary in scope and accuracy. To reflect this, metaTraits emphasizes transparency by retaining provenance information for every trait and by summarizing predictions through an aggregation strategy (see below) that conveys both consistency and uncertainty across taxa.

Genome-based trait predictions were generated for 906 855 high-quality (CheckM2 [22] completeness > 90%, contamination < 5%, and GUNC [23] passed) isolate genomes from proGenomes3 [24], and 1 102 679 medium- and high-quality (CheckM2 [22] completeness > 50%, contamination < 5%, and GUNC [23] passed) MAGs from SPIRE [25]. In total, this approach yielded 207 million trait predictions. By incorporating MAGs derived from environmental samples, the dataset encompassed a broader phylogenetic diversity, enabling trait coverage across a wider range of taxa (Fig. 1A and B). Notably, the increase in trait annotation coverage relative to experiment-based datasets was particularly pronounced in clades that are rarely isolated or cultivated, such as Patescibacteria and Nanoarchaeota (Fig. 1A and Supplementary Figure S1).

Overall, metaTraits captures more than 144 traits (2855 if individual chemical compounds are counted separately) that are organized into 46 groups and observed across 64 126 named species in GTDB r220 (104 722 by SPIRE clusters, 54 654 by NCBI taxonomy), representing the largest publicly available collection of microbial phenotypic trait data to date.

Data standardization

To enhance interpretability and machine-readability, trait names were mapped wherever possible to standardized vocabulary terms. The Ontology of Microbial Phenotypes (OMP, [27]) served as the primary framework for capturing microbial phenotypic traits and experimental observations, supplemented as needed by terms (including composites thereof) such as from MICRO [28], SNOMED [29], and Gene Ontology (GO, [30, 31]). For each mapped term, metaTraits provides a direct link-out to the corresponding Ontology Lookup Service (OLS, [32]) page, enabling users to easily explore definitions and relationships for each trait.

For taxonomy harmonization, we used taxonkit [33] via pytaxonkit [34] to assign standardized taxonomy identifiers

Table 1. Overview of microbial phenotypic traits and taxonomy covered by metaTraits

Dataset	Trait groups	Observations/ Predictions	Records/Strains	NCBI species	GTDB species
All	144	207 340 808	2 238 622	54 654	65 349
Culture-based records	58	1 182 280	229 088	25 204	17 159
Bac*Dive*	55	859 472	58 192	14 565	9 468
BV-BRC	11	84 411	18 823	5 153	4 770
JGI IMG/GOLD	20	238 397	152 073	19 247	13 748
Genome-based predictions	130	206 158 528	2 009 534	48 029	64 126
proGenomes3	130	68 697 060	906 855	43 305	44 884
SPIREv1	130	137 461 468	1 102 679	14 023	35 416
Genome statistics	6				
Bac*Dive*-AI	9				
GenomeSPOT	10				
MICROPHERRET	52				
Traitar	62				

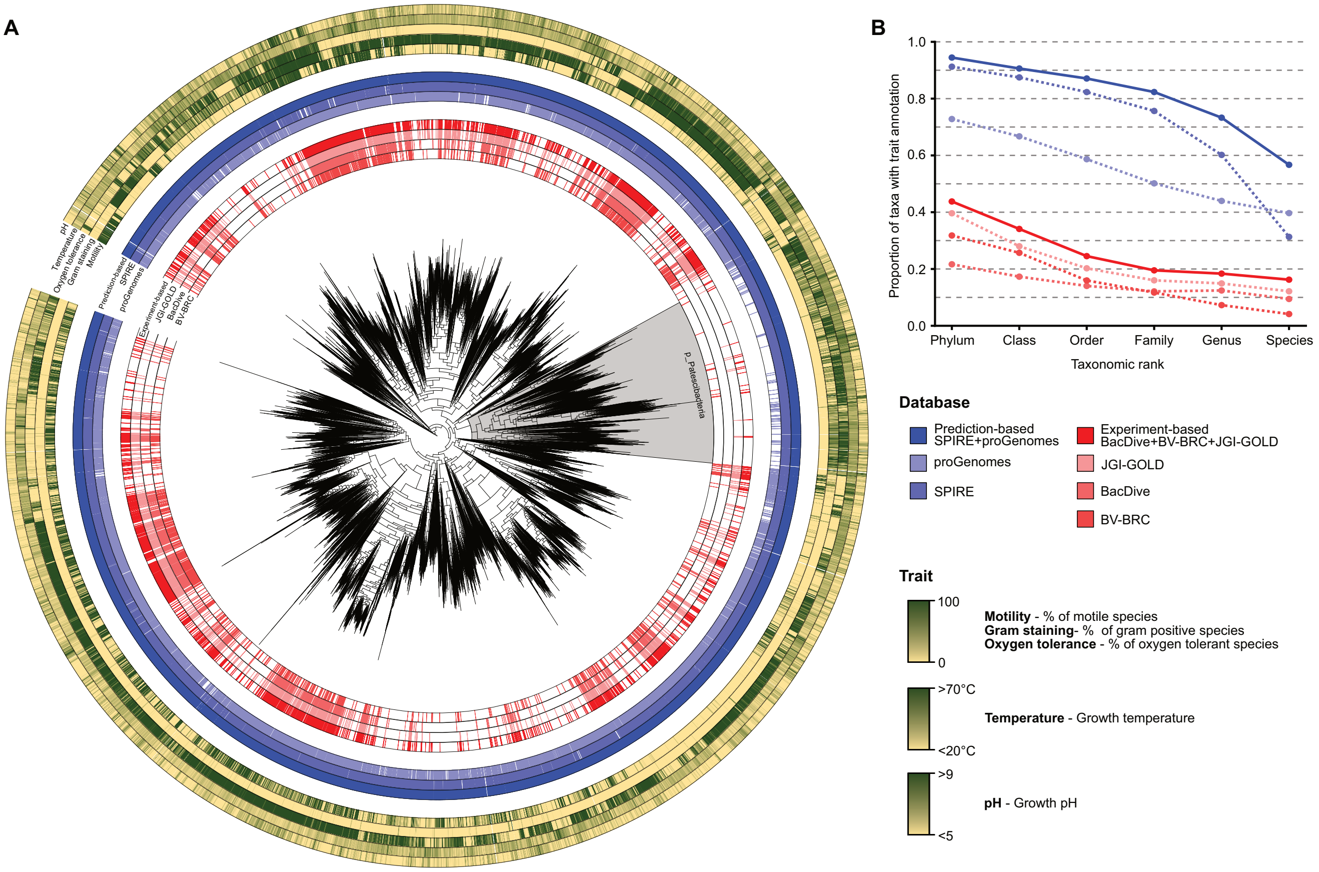


Figure 1. Experimental and predicted trait coverage across the bacterial tree of life. (**A**) Phylogenetic tree of 23 112 bacterial genera based on GTDB taxonomy release r220. From the innermost ring outward, red strips indicate genera for which experimentally derived trait information is available from databases such as JGI GOLD, Bac*Dive*, and BV-BRC, while blue strips indicate genera with traits predicted from genome sequences using proGenomes and SPIRE. Subsequent concentric gradient rings represent the proportion of species in each genus exhibiting the following traits: motility, Gram staining, oxygen tolerance, growth temperature, and growth pH. The gray shade highlights the phylum Patescibacteria, for which experimentally derived trait data are available for only 17 of 4581 species (0.37%), whereas prediction-based approaches provide trait annotations for 1677 species (36.7%). For genus-level trees, the lowest common ancestor (LCA) of all species within each genus from the GTDB-species tree was designated as the genus node, which was treated as a leaf in the reconstructed tree. As the GTDB taxonomy was used, all the LCAs were monophyletic with respect to their corresponding genera by definition. The resulting trees were visualized using iTOL [26]. (**B**) Proportion of GTDB taxa with trait annotations across taxonomic ranks. Blue lines represent coverage from prediction-based sources (proGenomes, SPIRE, and their combination), and red lines represent coverage from experimental sources (Bac*Dive*, BV-BRC, JGI GOLD, and their combination). Prediction-based methods consistently achieve broader coverage than experiment-based sources across all ranks.

for both NCBI ([35], 2025-07-28) and GTDB ([36], release r220). As there is no official taxonomic identifier system for GTDB, we employed the GTDB taxdump (gtdb-taxdump v0.5.0 r80-r220) provided by taxonkit for consistent mapping across the dataset. Not all trait records were annotated with both NCBI and GTDB taxonomy, or had an associated genome for taxonomic classification. However, since we had a large set of fully classified genomes, we created a mapping between the two systems: for each taxon in one taxonomy, a corresponding taxon in the other was assigned if at least 85% of genomes shared the same ID in both. This approach, with an average agreement rate of 99.47% (GTDB to NCBI) and 99.87% (NCBI to GTDB), respectively, enables robust cross-referencing between NCBI and GTDB taxonomies, which can be useful for many applications beyond trait analysis.

Database content

Website

The metaTraits website (https://metatraits.embl.de) centralizes trait data in a unified and accessible resource. Users can search and explore the database using either NCBI or GTDB taxonomy, with the option to flexibly include or exclude specific data sources depending on the specific research requirements (e.g. to focus on culture-derived or genome-based predictions). Trait data are aggregated by taxonomy, enabling trait summaries from species to phylum level and displaying distributions within relevant context. To reflect underlying genomic variability, numerical traits are summarized by their median, while binary and categorical traits report the fraction of observations assigned to each class, along with the total number of contributing observations and databases. When trait data for a clade are consistent (≥85% of observations in one category), a summary trait label is provided. This aggregation strategy also makes prediction uncertainties visible: inconsistent predictions within a clade result in broader distributions, and when fewer than 85% of genomes agree on a state, the trait is flagged as “no robust majority.” Each trait estimate is linked to its original evidence in the source databases, ensuring transparency and verifiability, and additional link-outs to relevant external resources are available. Downloadable taxonomy-level summaries support broad accessibility and seamless integration into downstream analyses.

Annotation of user-submitted data

Two annotation workflows for user-submitted data extend the practical utility of metaTraits:

(i) Genome annotation: The Nextflow-based porTraits workflow predicts microbial phenotypic traits for user-submitted isolate genomes and MAGs by integrating multiple genome-based prediction tools. Users provide genome or MAG FASTA files as input, porTraits calls genes with Prodigal [37], generates KO and PFAM matrices via eggNOG-mapper [38], and computes trait predictions using the models from Bac*Dive*-AI, Traitar, and MICROPHERRET; GenomeSPOT is run directly on the genome input. Taxonomic assignments are obtained using reCOGnise (for NCBI taxonomy) and GTDB-Tk [39] (for GTDB r220), which also enable retrieval of similar trait records from the metaTraits database for contextualization. For NCBI taxonomy assignment, reCOGnise extracts mOTUs [40] markers with fetchMGs [41], aligns them to the COG database using MAPseq [42], and assigns taxonomic IDs. The porTraits workflow can be executed directly on the metaTraits website (no registration required), via the interface of the Cloud-based Workflow Manager (https://clowm.bi.denbi.de, [43]), or their API. All workflow code is openly available at https://github.com/grp-bork/porTraits.

(ii) Microbial community annotation: This workflow enables users to annotate entire taxonomic profiles of microbial communities with trait information. It parses outputs from commonly used taxonomic profiling tools, maps features to taxonomic IDs, and annotates taxa with trait data from selected sources. Supported profilers currently include mOTUs [44], MetaPhlAn [45], Kraken [46], Krakenuniq [47], Bracken [48], Kaiju [49], as well as generic OTU tables with matching taxonomies.

These workflows make metaTraits not only a comprehensive reference resource but also an integrative tool compatible with widely adopted microbiome analysis software.

Use cases and outlook

metaTraits provides a foundation for diverse research applications and integration into existing microbiome analysis workflows. Key use cases include:

- **Quick taxonomic trait characterization:** Researchers can rapidly look up and summarize microbial traits at multiple taxonomic levels. This is especially valuable for studies based on 16S rRNA gene amplicon data, which often have resolution limited to the genus level. The ability to query both NCBI and GTDB taxonomies within metaTraits further extends utility to a broad user base.
- **Distribution and evolution of traits:** Microbiologists can assess whether traits observed or predicted for their strains are typical or unusual within their phylogenetic context, and readily identify exceptions or outliers within lineages. Overlaying traits with phylogenetic trees can provide insights into their evolution (e.g. motility as shown in Fig. 1A).
- **Community trait annotation:** Microbiome researchers can annotate entire microbial communities with trait prevalence and abundance profiles, enabling analyses of functional and ecological patterns at the community level. As shown in Fig. 2, such comparisons reveal distinct trait distributions across ocean, hot spring, soil, and host-associated microbiomes (Fig. 2B) and highlight functional shifts in health contexts, such as the increased prevalence of oxygen-tolerant microbes during infection (Fig. 2A). These profiles therefore facilitate the study of associations between microbial functions, taxa, and environmental variables.
- **Cross-database integration:** Integration with resources like SPIRE [25], GMGC [50], and Metalog (https://metalog.embl.de/) supports comprehensive multi-omic analyses, enabling researchers to relate traits and trait profiles to community composition, gene prevalence, and host or environmental data across databases.
- **Traits in context:** By combining traits of genomes or microbial communities with rich contextual metadata from resources like Metalog, researchers can investigate how

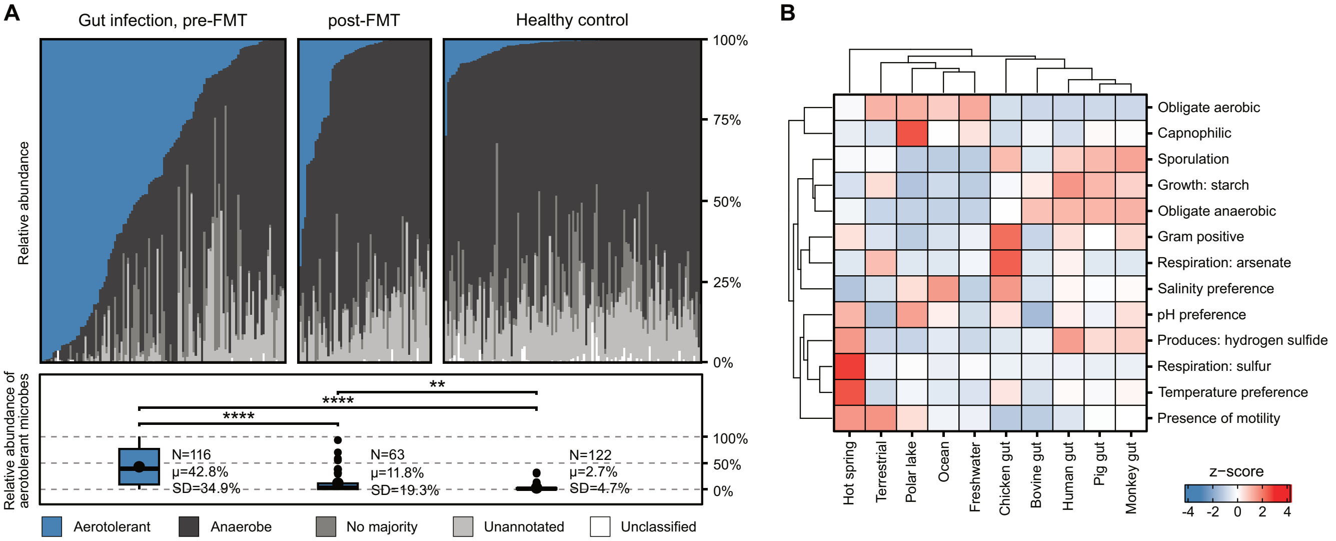


Figure 2. Trait-based comparisons of microbial communities using metaTraits. (**A**) Compositional shifts in microbes with distinct oxygen preference in the human gut (data from Metalog [51]) among individuals with *Clostridioides difficile* infection before and after fecal microbiota transplantation (FMT), compared to healthy controls. Bars show community-level relative abundances of aerotolerant (blue) and anaerobic (dark gray) taxa, as well as those that are taxonomically unclassified, lack trait annotations, or have no annotation majority. The data reveal a marked enrichment of aerotolerant and depletion of obligate anaerobic microbes during infection, which reverses following FMT. (**B**) Trait distributions across microbiomes from aquatic and terrestrial environments, and various animal hosts (data from Kim *et al.* [4]). Shown are *z*-scored trait abundances across habitats for selected traits, including Gram staining, sporulation, motility, and preferences for oxygen, temperature, and pH. Data are summarized across samples within each habitat, reflecting distinct ecological and physiological characteristics of the microbes and their environments.

functional traits are distributed across ecosystems, biogeography, disease, or health status. For example, one could explore whether certain traits are more common in gut microbiomes of individuals with inflammatory bowel disease, or how metabolic strategies differ between marine and freshwater samples. Ecologists can also utilize trait information to build and test hypotheses about the ecology and biogeography of prokaryotes.

Future updates to metaTraits will expand trait coverage, incorporate new prediction tools and updated genome catalogs, and enhance interoperability with external databases, supporting the continued growth of trait-based microbiome research.

Acknowledgements

The authors would like to thank members of the Bork Group for helpful discussions. We also acknowledge the EMBL IT Services for providing support and access to the EMBL high-performance computing infrastructure. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions: Daniel Podlesny (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Chan Yeong Kim (Data curation [equal], Formal analysis [equal], Methodology [equal], Validation [equal], Visualization [equal], Writing—review & editing [equal]), Shahriyar Mahdi Robbani (Software [equal], Visualization [equal]), Christian Schudoma (Software [equal]), Anthony Fullam (Data curation [equal], Resources [equal]), Lorenz Christian Reimer (Conceptualization [equal], Funding acquisition [equal]), Julia Koblitz (Resources [equal], Writing—review & editing [equal]), Isabel Schober (Resources [equal], Writing—review & editing [equal]), Anandhi Iyappan (Data curation [equal]), Thea Van Rossum (Conceptualization [equal], Funding acquisition [equal], Supervision [equal]), Jonas Schiller (Data curation [equal], Formal analysis [equal], Writing—review & editing [equal]), Anastasia Grekova (Data curation [equal]), Michael Kuhn (Conceptualization [equal], Writing—review & editing [equal]), and Peer Bork (Conceptualization [equal], Funding acquisition [equal], Supervision [equal], Writing—review & editing [equal]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 460129525 (NFDI4Microbiota), by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00240807), by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI, by the German Federal Ministry of Research, Technology and Space in the frame of de.NBI & ELIXIR-DE (W-de.NBI-014) and the European Molecular Biology Laboratory (EMBL). Funding to pay the Open Access publication charges for this article was provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 460129525 (NFDI4Microbiota).

Data availability

All raw data underlying metaTraits v1 is publicly available via proGenomes (https://progenomes.embl.de/), SPIRE (https://spire.embl.de/), and the culture databases listed in Table 1. No new sequencing data were generated for this study. The derived and curated data described above are freely accessible and downloadable via https://metatraits.embl.de. No registration is required. metaTraits is released under a Creative Commons Attribution-ShareAlike 4.0 International License. Source code for the trait annotation workflow porTraits is available at https://github.com/grp-bork/porTraits and on Zenodo (10.5281/zenodo.16809307).

References

1. Zwart JA, Solomon CT, Jones SE. Phytoplankton traits predict ecosystem function in a global set of lakes. *Ecology* 2015;96:2257–64. https://doi.org/10.1890/14-2102.1
2. Martiny JBH, Jones SE, Lennon JT *et al.* Microbiomes in light of traits: a phylogenetic perspective. *Science* 2015;350:aac9323. https://doi.org/10.1126/science.aac9323
3. Suzuki TA, Fitzstevens JL, Schmidt VT *et al.* Codiversification of gut microbiota with humans. *Science* 2022;377:1328–32. https://doi.org/10.1126/science.abm7759
4. Kim CY, Podlesny D, Schiller J *et al.* Planetary microbiome structure and generalist-driven gene flow across disparate habitats. bioRxiv, https://doi.org/10.1101/2025.07.18.664989, 18 July 2025, preprint: not peer reviewed.
5. Podlesny D, Fricke WF. Strain inheritance and neonatal gut microbiota development: a meta-analysis. *Int J Med Microbiol* 2021;311:151483. https://doi.org/10.1016/j.ijmm.2021.151483
6. Podlesny D, Durdevic M, Paramsothy S *et al.* Identification of clinical and ecological determinants of strain engraftment after fecal microbiota transplantation using metagenomics. *Cell Rep Med* 2022;3:100711. https://doi.org/10.1016/j.xcrm.2022.100711
7. Andreu-Sánchez S, Blanco-Míguez A, Wang D *et al.* Global genetic diversity of human gut microbiome species is related to geographic location and host health. *Cell* 2025;188:3942–59. https://doi.org/10.1016/j.cell.2025.04.014
8. Valles-Colomer M, Blanco-Míguez A, Manghi P *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* 2023;614:125–35. https://doi.org/10.1038/s41586-022-05620-1
9. Schober I, Koblitz J, Sardà Carbasse J *et al.* Bac*Dive* in 2025: the core database for prokaryotic strain data. *Nucleic Acids Res* 2025;53:D748–56. https://doi.org/10.1093/nar/gkae959
10. Olson RD, Assaf R, Brettin T *et al.* Introducing the bacterial and viral bioinformatics resource center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res* 2022;51:D678–89. https://doi.org/10.1093/nar/gkac1003
11. Chen I-MA, Chu K, Palaniappan K *et al.* The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res* 2023;51:D723–32. https://doi.org/10.1093/nar/gkac976
12. Mukherjee S, Stamatis D, Li CT *et al.* Genomes OnLine database (GOLD) v.10: new features and updates. *Nucleic Acids Res* 2025;53:D989–97. https://doi.org/10.1093/nar/gkae1000
13. Madin JS, Nielsen DA, Brbic M *et al.* A synthesis of bacterial and archaeal phenotypic trait data. *Sci Data* 2020;7:170. https://doi.org/10.1038/s41597-020-0497-4
14. Cébron A, Zeghal E, Usseglio-Polatera P *et al.* BactoTraits—a functional trait database to evaluate how natural and man-induced changes influence the assembly of bacterial communities. *Ecol Indic* 2021;130:108047.
15. Shaaban H, Westfall DA, Mohammad R *et al.* The microbe directory: an annotated, searchable inventory of microbes' characteristics. *Gates Open Res* 2018;2:3. https://doi.org/10.12688/gatesopenres.12772.1
16. Prasoodanan PKV, Maistrenko OM, Fullam A *et al.* A census of hidden and discoverable microbial diversity beyond genome-centric approaches. bioRxiv, https://doi.org/10.1101/2025.06.26.661807 , 26 June 2025, preprint: not peer reviewed.
17. Koblitz J, Reimer LC, Pukall R *et al.* Predicting bacterial phenotypic traits through improved machine learning using high-quality, curated datasets. *Commun Biol* 2025;8:897. https://doi.org/10.1038/s42003-025-08313-3
18. Barnum TP, Crits-Christoph A, Molla M *et al.* Predicting microbial growth conditions from amino acid composition. bioRxiv, https://doi.org/10.1101/2024.03.22.586313, 22 March 2024, preprint: not peer reviewed.
19. Bizzotto E, Fraulini S, Zampieri G *et al.* MICROPHERRET: MICRObial PHEnotypic tRait ClassifieR using machine lEarning techniques. *Environ Microbiome* 2024;19:58. https://doi.org/10.1186/s40793-024-00600-6
20. Weimann A, Mooren K, Frank J *et al.* From genomes to phenotypes: traitar, the microbial trait analyzer. *Msystems* 2016;1:e00101–16. https://doi.org/10.1128/mSystems.00101-16
21. Wilkinson MD, Dumontier M, Aalbersberg IJJ *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. https://doi.org/10.1038/sdata.2016.18
22. Chklovski A, Parks DH, Woodcroft BJ *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;20:1203–12. https://doi.org/10.1038/s41592-023-01940-w
23. Orakov A, Fullam A, Coelho LP *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol* 2021;22:178. https://doi.org/10.1186/s13059-021-02393-0
24. Fullam A, Letunic I, Schmidt TSB *et al.* proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Res* 2022;51:D760–6. https://doi.org/10.1093/nar/gkac1078
25. Schmidt TSB, Fullam A, Ferretti P *et al.* SPIRE: a searchable, planetary-scale mIcrobiome REsource. *Nucleic Acids Res* 2024; 52:D777–83. https://doi.org/10.1093/nar/gkad943
26. Letunic I, Bork P. Interactive tree of life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* 2024; 52:W78–82. https://doi.org/10.1093/nar/gkae268
27. Chibucos MC, Zweifel AE, Herrera JC *et al.* An ontology for microbial phenotypes. *BMC Microbiol* 2014;14:294. https://doi.org/10.1186/s12866-014-0294-3
28. Blank CE, Cui H, Moore LR *et al.* MicrO: an ontology of phenotypic and metabolic characters, assays, and culture media found in prokaryotic taxonomic descriptions. *J Biomed Semant* 2016;7:18. https://doi.org/10.1186/s13326-016-0060-6
29. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. *Stud Health Technol Inform* 2006;121:279–90.
30. Gene Ontology Consortium, Aleksander SA, Balhoff J *et al.* The gene ontology knowledgebase in 2023. *Genetics* 2023;224:iyad031.
31. Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;25:25–9. https://doi.org/10.1038/75556
32. McLaughlin J, Lagrimas J, Iqbal H *et al.* OLS4: a new ontology lookup service for a growing interdisciplinary knowledge ecosystem. *Bioinformatics* 2025;41:btaf279. https://doi.org/10.1093/bioinformatics/btaf279
33. Shen W, Ren H. TaxonKit: a practical and efficient NCBI taxonomy toolkit. *J Genet Genom* 2021;48:844–50. https://doi.org/10.1016/j.jgg.2021.03.006

34. Standage D. PyTaxonKit. 2024. https://doi.org/10.17605/OSF.IO/7UTG9
35. Schoch CL, Ciufo S, Domrachev M *et al.* NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020;2020:baaa062. https://doi.org/10.1093/database/baaa062
36. Parks DH, Chuvochina M, Rinke C *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022; 50:D785–94. https://doi.org/10.1093/nar/gkab776
37. Hyatt D, Chen G-L, Locascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf* 2010;11:119. https://doi.org/10.1186/1471-2105-11-119
38. Cantalapiedra CP, Hernández-Plaza A, Letunic I *et al.* eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 2021;38:5825–9. https://doi.org/10.1093/molbev/msab293
39. Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;38:5315–6. https://doi.org/10.1093/bioinformatics/btac672
40. Ruscheweyh H-J, Milanese A, Paoli L *et al.* Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* 2022;10:212. https://doi.org/10.1186/s40168-022-01410-z
41. Kultima JR, Sunagawa S, Li J *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* 2012;7:e47656. https://doi.org/10.1371/journal.pone.0047656
42. Matias Rodrigues JF, Schmidt TSB, Tackmann J *et al.* MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* 2017;33:3808–10. https://doi.org/10.1093/bioinformatics/btx517
43. Göbel D, Stoye J, Sczyrba A *et al.* The cloud-based workflow manager (CloWM)—an integrated platform for highly scalable workflow execution. https://doi.org/10.5281/zenodo.14039069
44. Dmitrijeva M, Ruscheweyh H-J, Feer L *et al.* The mOTUs online database provides web-accessible genomic context to taxonomic profiling of microbial communities. *Nucleic Acids Res* 2024;53:D797–805. https://doi.org/10.1093/nar/gkae1004
45. Blanco-Míguez A, Beghini F, Cumbo F *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 2023;41:1633–44. https://doi.org/10.1038/s41587-023-01688-w
46. Lu J, Rincon N, Wood DE *et al.* Metagenome analysis using the Kraken software suite. *Nat Protoc* 2022;17:2815–39. https://doi.org/10.1038/s41596-022-00738-y
47. Pockrandt C, Zimin AV, Salzberg SL. Metagenomic classification with KrakenUniq on low-memory computers. *JOSS* 2022;7:4908. https://doi.org/10.21105/joss.04908.
48. Lu J, Breitwieser FP, Thielen P *et al.* Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci* 2017;3:e104. https://doi.org/10.7717/peerj-cs.104
49. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* 2016;7:11257. https://doi.org/10.1038/ncomms11257
50. Coelho LP, Alves R, Del Río ÁR *et al.* Towards the biogeography of prokaryotic genes. *Nature* 2022;601:252–6. https://doi.org/10.1038/s41586-021-04233-4
51. Kuhn M, Schmidt TSB, Ferretti P *et al.* Metalog: curated and harmonised contextual data for global metagenomics samples. *Nucleic Acids Res* 2025. https://doi.org/10.1093/nar/gkaf1118.

Received: August 14, 2025. **Revised:** October 20, 2025. **Accepted:** October 21, 2025