

## Databases and Ontologies

# GlobDB: a comprehensive species-dereplicated microbial genome resource

Daan R. Speth<sup>1,\*</sup>, Nick Pullen<sup>1</sup>, Samuel T.N. Aroney<sup>2</sup>, Benjamin L. Coltman<sup>1</sup>, Jay Osvatic<sup>3,4</sup>, Ben J. Woodcroft<sup>2</sup>, Thomas Rattei<sup>1</sup>, Michael Wagner<sup>1,5</sup>

<sup>1</sup>Centre for Microbiology and Environmental Systems Science, University of Vienna, 1030 Vienna, Austria

<sup>2</sup>Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, QLD 4102, Australia

<sup>3</sup>Joint Microbiome Facility of the Medical University of Vienna and the University of Vienna, 1030 Vienna, Austria

<sup>4</sup>Department of Laboratory Medicine, Medical University of Vienna, 1090 Vienna, Austria

<sup>5</sup>Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, 9220 Aalborg, Denmark

\*Corresponding author. Centre for Microbiology and Environmental Systems Science, University of Vienna, Djerassiplatz 1, 1030 Vienna, Austria.

E-mail: daan.speth@univie.ac.at.

Associate Editor: Nicola Mulder

## Abstract

**Motivation:** Over the past years, substantial numbers of microbial species' genomes have been deposited outside of conventional INSDC databases.

**Results:** The GlobDB aggregates 14 independent genomic catalogues to provide a comprehensive database of species-dereplicated microbial genomes, with consistent taxonomy, annotations, and additional analysis resources. The GlobDB more than doubles the number of microbial species represented by genomes relative to the field standard genome taxonomy database.

**Availability and implementation:** The GlobDB is available at <https://globdb.org/>.

## 1 Introduction

Over the last twenty years, advances in DNA sequencing and data processing have led to an explosion in the available genome information of cultivated and uncultivated microorganisms. These genomes offer an unprecedented window into microbial diversity and its metabolic potential. Assembling genomes from metagenomes is now a routine practice, yielding thousands of genomes in studies focused on specific environments (Gurbich *et al.* 2023, Ma *et al.* 2023, Carlino *et al.* 2024, Chen *et al.* 2024, Cheng *et al.* 2024, Kim *et al.* 2024, Ma *et al.* 2024, Wang *et al.* 2024, Zhang *et al.* 2024, Michoud *et al.* 2025). The research community has greatly benefited from the field standard of depositing both raw and processed sequence data in International Nucleotide Sequence Database Collaboration (INSDC) databases (i.e. NCBI, ENA, and DDBJ) (Karsch-Mizrachi *et al.* 2025). Several ongoing meta-analysis efforts are processing data from many studies across environments, and have aggregated millions of microbial genomes into widely used genomic catalogues (Nayfach *et al.* 2021, Parks *et al.* 2022, Schmidt *et al.* 2023, Aroney *et al.* 2024, Dmitrijeva *et al.* 2025). While submitting unprocessed sequencing data to INSDC databases is standard practice, depositing assembled and binned metagenome recovered genomes (MAGs) of large-scale studies is less common. We have observed a recent trend that MAGs from

large scale studies are deposited in generic data repositories such as Zenodo or figshare, dedicated project websites, or national databases that are not (yet) INSDC members. Lack of submission to INSDC databases leads to omission of genomes from commonly used tools such as the NCBI BLAST webserver, or INSDC-derived databases such as the genome taxonomy database (GTDB), hindering further analysis and discovery based on these data. Furthermore, the fragmented deposition of microbial genomes means that there is no longer a central resource for species-dereplicated genomes that includes the full extent of microbial diversity represented by genome sequences.

## 2 Results

### 2.1 Database description

We have compiled the GlobDB, a comprehensive database of species-dereplicated microbial genomes. The GlobDB aims at maximal representation of phylogenetic diversity to facilitate genome-based analysis and discovery. The GlobDB follows the annual GTDB release schedule, and as of the current release integrates 14 resources (Table 1). These were sequentially dereplicated at 96% average nucleotide identity (ANI) and 50% aligned fractions (Jain *et al.* 2018) in the order listed in Table 1. Therefore, the GTDB species representatives are the base of the GlobDB, and the other 13 datasets extend

**Table 1.** Resources included in the GlobDB.

| Rank | Name   | Number of species reps | Included in the GlobDB | Environment   | References                                                                         |
|------|--------|------------------------|------------------------|---------------|------------------------------------------------------------------------------------|
| 1    | GTDB   | 143 614                | 143 607                | Global        | <a href="#">Parks <i>et al.</i> (2022)</a>                                         |
| 2    | mOTU   | 124 295                | 63 092                 | Global        | <a href="#">Dmitrijeva <i>et al.</i> (2025)</a>                                    |
| 3    | SPIRE  | 107 078                | 49 369                 | Global        | <a href="#">Schmidt <i>et al.</i> (2023)</a>                                       |
| 4    | RBG    | 38 494                 | 23 406                 | Global        | <a href="#">Aroney <i>et al.</i> (2024)</a>                                        |
| 5    | GEM    | 45 599                 | 3762                   | Global        | <a href="#">Nayfach <i>et al.</i> (2021)</a>                                       |
| 6    | MGnify | 29 556                 | 1732                   | Env. specific | <a href="#">Gurbich <i>et al.</i> (2023)</a>                                       |
| 7    | GOMC   | 24 195                 | 2858                   | Env. specific | <a href="#">Chen <i>et al.</i> (2024)</a>                                          |
| 8    | SMAG   | 21 078                 | 6216                   | Env. specific | <a href="#">Ma <i>et al.</i> (2023)</a>                                            |
| 9    | TPMC   | 10 723                 | 8644                   | Env. specific | <a href="#">Cheng <i>et al.</i> (2024)</a>                                         |
| 10   | KRGM   | 6348                   | 334                    | Env. specific | <a href="#">Kim <i>et al.</i> (2024)</a> ; <a href="#">Ma <i>et al.</i> (2024)</a> |
| 11   | cFMD   | 10 112 <sup>a</sup>    | 336                    | Env. specific | <a href="#">Carlino <i>et al.</i> (2024)</a>                                       |
| 12   | SHGO   | 5810 <sup>a</sup>      | 1021                   | Env. specific | <a href="#">Zhang <i>et al.</i> (2024)</a>                                         |
| 13   | AMXMAG | 1768 <sup>a</sup>      | 506                    | Env. specific | <a href="#">Wang <i>et al.</i> (2024)</a>                                          |
| 14   | GFS    | 2862                   | 1377                   | Env. specific | <a href="#">Michoud <i>et al.</i> (2025)</a>                                       |

<sup>a</sup> No species-dereplicated set was available for download.

upon this. Several datasets did not provide a species-dereplicated set for download. In these cases, genomes redundant with the GlobDB were first removed, followed by dereplication of the remainder using dRep ([Olm \*et al.\* 2017](#)). Finally, 420 genomes were removed from the dataset due to low genome quality or putatively representing eukaryotes. The full dereplication process is described in detail on <https://globdb.org/methods>. The current GlobDB release 226 contains more than double the number of species-representative genomes (306 260) compared to the GTDB release 226 species representative set (143 614), indicating that >50% of known genomic species diversity is not included in INSDC databases ([Fig. 1](#)). This also leads to underrepresentation of specific environments, such as the Tibetan Plateau (TPMC) or sheep and goat guts (SHGO), in these databases ([Table 1](#)). To our knowledge, the GlobDB is the only resource that integrates large MAG collections across all environments, with a focus on bringing in MAGs that have not been included in INSDC databases and maximizing representation of microbial diversity.

The GlobDB includes resources beyond aggregation of species representative genomes ([Fig. 1](#)). We provide standardized identifiers of each resource, and a complete 7-level taxonomy built upon and extending the GTDB taxonomy. The taxonomy of the species representatives sourced from the other 13 resources is determined using a custom approach based on the GTDB workflow ([Chaumeil \*et al.\* 2022](#), [Aroney \*et al.\* 2024](#)), automatically assigning new taxon labels according to their Relative Evolutionary Divergence (RED) value. In practice, this results in propagating GlobDB genome identifiers upwards through the taxonomic ranks, until a node in the tree is reached that already has GTDB taxonomy assigned. This approach ensures that the GlobDB taxonomy is an extension of the GTDB taxonomy and fully compatible with it.

Sylph ([Shaw and Yu 2025](#)) databases and a SingleM metapackage ([Woodcroft \*et al.\* 2025](#)) of GlobDB species representative genomes are available for taxonomic profiling of metagenomes. Additional genome metadata includes: completeness, contamination, and basic genome statistics calculated using CheckM2 ([Chklovski \*et al.\* 2023](#)); and whether the genome accession is linked to an isolate deposited in a major culture collection assessed via BacDive ([Schober \*et al.\* 2025](#)).

Furthermore, anvi'o ([Eren \*et al.\* 2021](#)) “contigs databases” of all GlobDB genomes are made available. These SQL

databases contain the genome sequences and contextual data, including: gene calls for rRNA, tRNA, and protein coding genes as well as functional annotations of protein coding genes generated using Pfam ([Mistry \*et al.\* 2021](#)), COG ([Galperin \*et al.\* 2021](#)), KEGG ([Kanehisa \*et al.\* 2023](#), [Kananen \*et al.\* 2025](#)), and dbCAN2 CAZymes ([Zhang \*et al.\* 2018](#)). GFF files with gene calls and COG IDs, fastA files with amino acid sequences of protein coding genes, and tab delimited files with functional annotations of each GlobDB genome are also available for download separately.

Finally, to be able to generate protein language model (pLM) embeddings of the GlobDB proteins, we clustered the 838 615 274 GlobDB amino acid sequences using a 40% identity threshold over 80% of the length of both sequences ([Steinegger and Söding 2018](#)), yielding 82 973 016 representatives of clusters larger than one sequence. We generated ProtT5-XL-U50 ([Elnaggar \*et al.\* 2022](#)) pLM embeddings for 82 972 511 of these cluster representatives. Sequences, annotations, cluster membership, and pLM embeddings of this clustered sequence set are available for download. Full methods for the GlobDB can be found on <https://globdb.org/methods>.

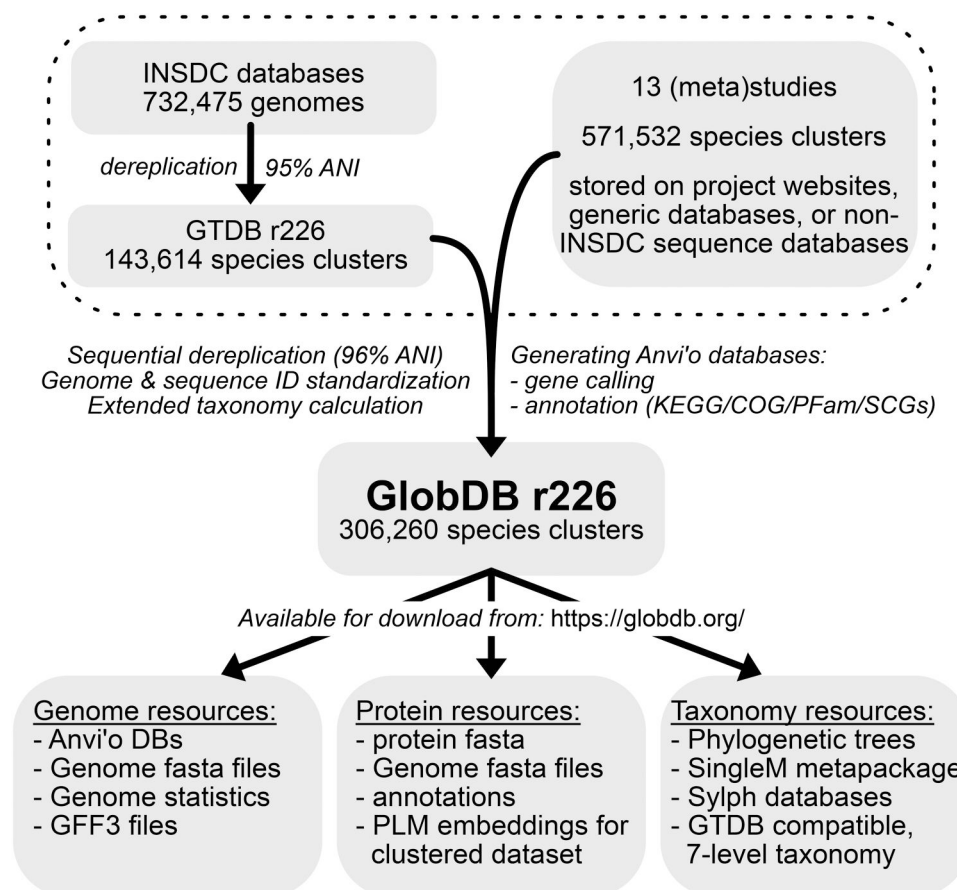
In summary, we introduce the GlobDB, a comprehensive database of species-dereplicated microbial genomes and several analysis products. We expect the GlobDB to be useful in future work that benefits from a broad representation of microbial genomic diversity at any taxonomic rank. This includes studies involving phylogenomics, comparative genomics, taxonomic profiling of metagenomes, or gene and protein family analyses.

## Acknowledgements

We thank Petra Pjevac for helpful discussion and critical review of the manuscript text. We thank Phil Hugenholz, Donovan Parks, and Pierre Chaumeil for early access to the GTDB species representatives, and Taylor Priest for sharing the mOTU species representatives. Computational results of this work have been achieved using the Life Science Compute Cluster (LiSC) of the University of Vienna.

## Conflict of interest

None declared.



**Figure 1.** Overview of data included in, and resources available from, the GlobDB. The work indicated within the dashed box at the top of the figure indicates the source data for the GlobDB, text in italics indicates data processing, and the text within grey boxes describes resources.

## Funding

This work was supported by the Austrian Science Fund FWF [10.55776/COE7 to D.R.S., B.L.C., T.R., and M.W.]; the Austrian Federal Ministry for Women, Science and Research [DataLife—Data Infrastruktur for Life Sciences to N.P., and T.R.]; the EMERGE National Science Foundation (NSF) Biology Integration Institute [#2022070 to S.T.N.A.]; and the Australian Research Council Future Fellowship [#FT210100521 to B.W.]. This research was funded in whole or in part by the Austrian Science Fund (FWF) [10.55776/COE7]. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

## Data availability

The GlobDB is hosted on the Life Science Compute Cluster (LiSC) of the University of Vienna. The previous (220) and current (226) releases are available at <https://globdb.org/>. Currently available releases will be hosted for 10 years after publication. In addition, we intend to keep following the annual update schedule of the GTDB, and will keep previous releases available for reproducibility of work that used the GlobDB.

## References

Aroney STN, Newell RJP, Tyson GW *et al.* Bin Chicken: targeted meta-genomic coassembly for the efficient recovery of novel genomes.

- bioRxiv, <https://doi.org/10.1101/2024.11.24.625082>, 2024, pre-print: not peer reviewed.
- Carlino N, Blanco-Míguez A, Punčochář M *et al.*; MASTER EU Consortium. Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. *Cell* 2024;187: 5775–95.e15.
- 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.
- Cheng M, Luo S, Zhang P *et al.* A genome and gene catalog of the aquatic microbiomes of the Tibetan Plateau. *Nat Commun* 2024; 15:1438.
- Chen J, Jia Y, Sun Y *et al.* Global marine microbial diversity and its potential in bioprospecting. *Nature* 2024;633:371–9.
- 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.
- 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* 2025; 53:D797–805.
- Elnaggar A, Heinzinger M, Dallago C *et al.* ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;44:7112–27.
- Eren AM, Kiehl E, Shaiber A *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 2021;6:3–6.
- Galperin MY, Wolf YI, Makarova KS *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 2021;49:D274–81.
- Gurbich TA, Almeida A, Beracochea M *et al.* MGnify genomes: a resource for biome-specific microbial genome catalogues. *J Mol Biol* 2023;435:168016.

- Jain C, Rodriguez-R LM, Phillippy AM *et al.* High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.
- Kananen K, Veseli I, Quiles Pérez CJ *et al.* Adaptive adjustment of profile HMM significance thresholds improves functional and metabolic insights into microbial genomes. *Bioinform Adv* 2025;5:vbaf039.
- Kanehisa M, Furumichi M, Sato Y *et al.* KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* 2023; 51:D587–92.
- Karsch-Mizrachi I, Arita M, Burdett T *et al.* The international nucleotide sequence database collaboration (INSDC): enhancing global participation. *Nucleic Acids Res* 2025;53:D62–6.
- Kim N, Kim CY, Ma J *et al.* MRGM: an enhanced catalog of mouse gut microbial genomes substantially broadening taxonomic and functional landscapes. *Gut Microbes* 2024;16:2393791.
- Ma B, Lu C, Wang Y *et al.* A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat Commun* 2023;14:7318.
- Ma J, Kim N, Cha JH *et al.* Exclusive cataloging of near-complete genomes unveils novel functional insights into the human gut microbiome. *bioRxiv*, <https://doi.org/10.1101/2024.12.11.627901>, 2024, preprint: not peer reviewed.
- Michoud G, Peter H, Busi SB *et al.*; Vanishing Glaciers Field Team. Mapping the metagenomic diversity of the multi-kingdom glacier-fed stream microbiome. *Nat. Microbiol* 2025;10:217–30.
- Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;49:D412–9.
- Nayfach S, Roux S, Seshadri R *et al.*; IMG/M Data Consortium. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021; 39:499–509.
- Olm MR, Brown CT, Brooks B *et al.* dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 2017; 11:2864–8.
- 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.
- Schmidt TSB, Fullam A, Ferretti P *et al.* SPIRE: a searchable, planetary-scale mIcrobiome REsource. *Nucleic Acids Res* 2023;52:D777–83.
- Schober I, Koblit J, Sardà Carbasse J *et al.* BacDive in 2025: the core database for prokaryotic strain data. *Nucleic Acids Res* 2025; 53:D748–56.
- Shaw J, Yu YW. Rapid species-level metagenome profiling and containment estimation with sylph. *Nat Biotechnol* 2025;43:1348–59.
- Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun* 2018;9:2542.
- Wang D, Meng Y, Huang L-N *et al.* A comprehensive catalog encompassing 1376 species-level genomes reveals the core community and functional diversity of anammox microbiota. *Water Res* 2024; 266:122356.
- Woodcroft BJ, Aroney STN, Zhao R *et al.* Comprehensive taxonomic identification of microbial species in metagenomic data using SingleM and Sandpiper. *Nat Biotechnol* 2025;1–6.
- Zhang H, Yohe T, Huang L *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 2018;46:W95–101.
- Zhang K, He C, Wang L *et al.* Compendium of 5810 genomes of sheep and goat gut microbiomes provides new insights into the glycan and mucin utilization. *Microbiome* 2024;12:104.