Nucleic Acids Research, 2026, **54**, D852–D857
https://doi.org/10.1093/nar/gkaf1208
Advance access publication date: 20 November 2025
Database issue

proGenomes4: providing 2 million accurately and consistently annotated high-quality prokaryotic genomes

Anthony Fullam [1], Ivica Letunic [2], Oleksandr M Maistrenko [3], Alexandre Areias Castro [4], Luis Pedro Coelho [4], Anastasiia Grekova [1], Christian Schudoma [1], Supriya Khedkar [5], Mahdi Robbani [1], Michael Kuhn [1], Thomas SB Schmidt [6], Peer Bork [1,7,8,9,*], Daniel R Mende [10,*]

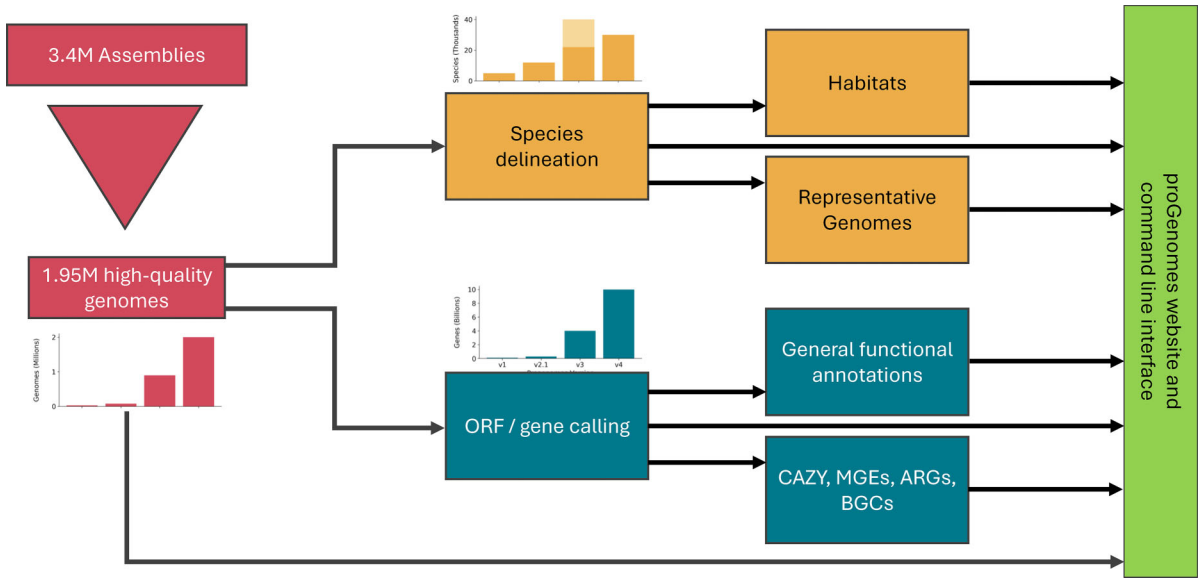
[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg 69117, Germany
[2]Biobyte solutions GmbH, Bothestraße 142, Heidelberg 69117, Germany
[3]Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam 1000 BE, The Netherlands
[4]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, 37 Kent St., Brisbane, Queensland, Woollongabba 4102, Australia
[5]BioQuant, University of Heidelberg, 69120 Heidelberg, Germany
[6]APC Microbiome and School of Medicine, University College Cork, T12 YT20 Cork, Ireland
[7]Max Delbrück Centre for Molecular Medicine, Berlin 13125, Germany
[8]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg 97074, Germany
[9]Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, Heidelberg 69120, Germany
[10]Human Biology-Microbiome-Quantum Research Center (WPI-Bio2Q), Keio University, Tokyo 108-8345, Japan

*To whom correspondence should be addressed. Email: mende@keio.jp
Correspondence may also be addressed to Peer Bork. Email: bork@embl.de

Abstract

The pervasive availability of publicly available microbial genomes has opened many new avenues for microbiology research, yet it also demands robust quality control and consistent annotation pipelines to ensure meaningful biological insights. proGenomes4 (prokaryotic Genomes v4) addresses this challenge by providing a resource of nearly 2 million high-quality microbial genomes, a doubling in scale from previous versions, encompassing over 7 billion genes. Each genome underwent rigorous quality assessment and comprehensive functional annotation by applying multiple standardized annotation workflows, including the systematic identification of mobile genetic elements and biosynthetic gene clusters. proGenomes4 contains 32 887 species with ecological habitat metadata as well as precomputed pan-genomes. This substantially expanded resource provides the microbiology community with a foundation for large-scale comparative studies and is freely accessible via a newly developed command line interface and at https://progenomes.embl.de/.

Graphical abstract



Received: September 17, 2025. **Revised:** October 12, 2025. **Accepted:** October 13, 2025

Introduction

The first microbial genomes were sequenced >30 years ago [1], yet the availability of large-scale genomics data remains a driver for discovery and innovation [2]. Nowadays, the availability of high-quality, low-cost genome sequencing using both short and long reads ensures that high-quality genomes can be assembled for any cultured organism [3]. Nevertheless, to gain biological insight from these ever-growing data sets, scientists need access to high-quality genomes with consistent annotations [4].

Microbial genomes are available from a variety of databases. The NCBI RefSeq database [5] is a prominent example providing access to a wide range of genome sequences, yet it only provides gene names, gene symbols, and EC numbers as annotations for coding sequences. Similar information, often enriched by specific annotations, is available from the PATRIC (Pathosystems Resource Integration Center) database [6], Ensembl Bacteria [7], and the Joint Genome Institute Integrated Microbial Genomes & Microbiomes database [8]. More recently, the AllTheBacteria database has combined *de novo* assemblies of genomes from public repositories with functional annotations [9]. In addition, databases such as SPIRE, MGnify, and motus-db provide millions of metagenome-assembled genomes (MAGs) [10, 11]. The Genome Taxonomy Database (GTDB) [12] is a dedicated, genomics- and phylogeny-based consistent taxonomy covering bacteria and archaea, comprising both reference genomes and MAGs. GTDB resolves many inconsistencies in previous taxonomies and provides an automated way to find and correct submitter errors.

In addition to such taxonomic annotation consistency, functional annotations have the same requirement, as has habitat information. Habitat information in particular is often inconsistent or missing as it depends on submitters. To address this issue, multiple resources linking microbes to environments have been established, including Microbe Atlas Project (MAP) [13] and Omnicrobe [14]. proGenomes4 (prokaryotic Genomes v4) integrates and links to MAP, which uses a comprehensive, annotated 16S ribosomal RNA (rRNA) catalog to link taxa to habitats, which are organized in an ontology.

Another important yet often neglected issue in microbial genomic databases is genome quality. Tools such as CheckM [15, 16] and GUNC [17] provide the means to consistently assess the quality of genomes at scale.

Here, we present the proGenomes4 database, which provides nearly 2 million high-quality bacterial and archaeal reference genomes of isolates (twice as many as in the previous version). proGenomes enables researchers' direct access to deeply and consistently annotated, high-quality genomes, providing information relevant for many different disciplines including but not limited to microbial evolution, ecology, and clinical and applied microbiology [18–20]. Genome quality is assured by using checkM2 and GUNC [16, 17]. Multiple annotation layers provide both general functional annotations via eggNOG [21] as well as specialized information about e.g. mobile genetic elements (MGEs) [22] and biosynthetic gene clusters (BGCs) [23] for over 8 billion genes. Further, the genomes are linked to other databases and resources, providing comprehensive access to information. proGenomes4 is designed to provide direct and easy access to the data and information needed for comparative analyses of prokaryotic genomes at any scale. The database can be accessed via a newly developed command line interface for bulk downloads and is available at https://progenomes.embl.de/.

Database construction and characteristics

proGenomes4 is accessible via its website (https://progenomes.embl.de/), which gives users access to all data and enables them to browse the available microbial genomes. By specifying NCBI assembly ID or the taxonomic name of the organism, species or clade in the search bar, users can interactively find and explore information about their desired organisms. In addition, the downloads section allows users to download precomputed bulk files providing e.g. the genomes and annotations of all species representatives. Major upgrades of the underlying computational pipeline and dataset are planned every two years.

Genome collection

All 3.1 million bacterial and archaeal genomes available in the NCBI Nucleotide database were downloaded on 3 April 2025 using NCBI Datasets CLI (version 18.4.0) [24]. First, we removed duplicate genomes (i.e. genomes present in multiple NCBI databases under different but linked accessions) and those marked as "suppressed" or "derived from metagenomes." Open reading frames were predicted for these genomes using prodigal (v2.6.3) (default parameters) [25]. Genomes with a circular, closed assembly were treated as high quality by default. Other incomplete genomes were filtered to only retain high-quality genomes using CheckM2 (v1.1.0) [16] and GUNC (v1.0.6) [17] (CheckM: completeness >90% and contamination <5%; GUNC: contamination <5% and clade separation score <0.45). This quality control step removed 1.2M genomes, with 1.9M high-quality genomes remaining (Fig. 1).

Delineating species

We used an ANI (average nucleotide identity)-based approach to consistently delineate species [26]. In short, we first preclustered genomes using Mash v2.3 (parameter -d 0.1) [27] under a *single linkage* algorithm, ensuring that all pairs of genomes sharing ≥90% mash similarity are part of the same precluster. Preclusters were then resolved using *average linkage* clustering at ≥95% ANI [calculated using fastANI (1.32) [26]] to obtain species-level clusters.

Selection of representative genomes

With the steady increase in genomes, there are many species now for which genomes from multiple reference strains are now available, leading to a certain degree of redundancy. While proGenomes4 continues to provide annotations for all high-quality genomes, many applications require nonredundant genome set (e.g. metagenomic read mapping or metagenomics-based strain tracking [28, 29])

As for previous versions, proGenomes4 provides a nonredundant set of representative genomes. Precomputed FASTA files of all representative genomes can be downloaded directly on the proGenomes website.

For the majority of species, we chose representatives by genome quality and citation statistics. For 820 species, however, we utilized a manually curated list of strains that are *de facto* representatives of their respective species, e.g. *Mycobacterium tuberculosis* H37Rv. Otherwise, we filtered each

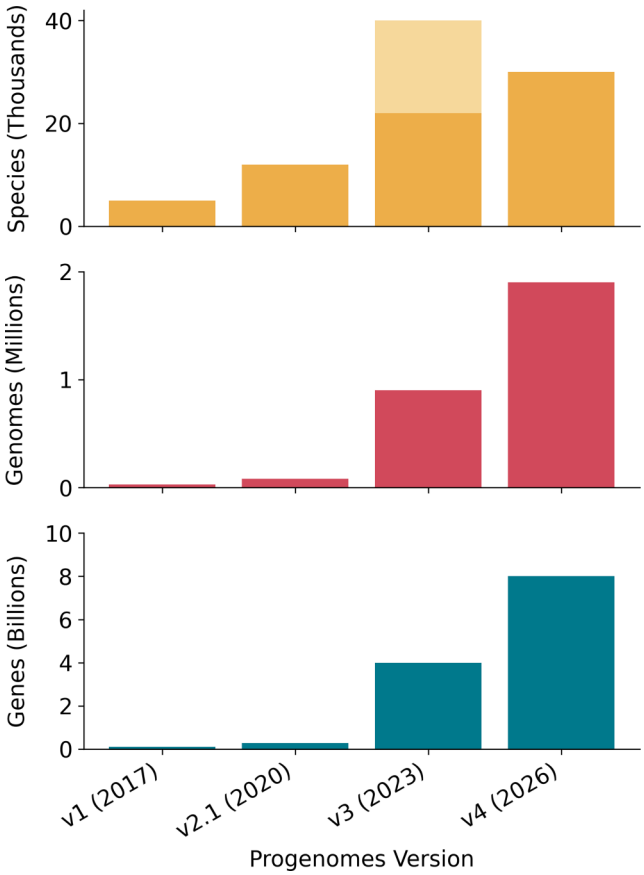


Figure 1. Growth of proGenomes across versions. Overall number of species, high-quality genomes, and genes in proGenomes (2017), proGenomes2 (2020), proGenomes3 (2023), and proGenomes4. The number of species in proGenomes3 is shown with and without MAGs, for proGenomes4 all possible MAGs were excluded.

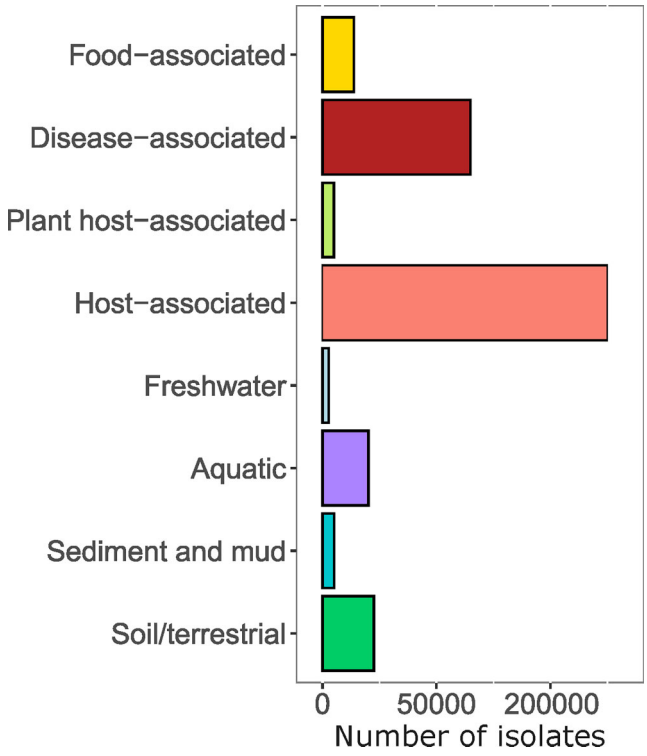


Figure 2. Curated habitat annotations. The number of isolates annotated to the different habitats is shown.

species genome set to only contain complete genomes and chose the most highly cited strain out of those [30]. If a species did not have a single complete genome, the genome assembly with the highest N50 statistic was selected.

Pan-genomes

Many species in proGenomes4 are represented by many different genomes. To provide users with a simple way to study all genes encoded by a species, we precompiled pan-genomes for every species in the form of a nonredundant gene set. The nonredundant gene sets were generated using mmseqs2 [31] (version: 18) using following parameters: –min-seq-id 0.95 -c 0.90 –cov-mode 0

Functional annotation

Functional annotations are one main aspect of proGenomes, with the goal of providing accurate, consistent, and broad annotations. General functional annotations were generated using eggNOG-mapper [32] (version:2.1.12). This results in assignments to functionally annotated orthologous groups from eggNOG 5.0 [33]. This broad annotation effort resulted in ~6 billion predicted proteins to be annotated to existing orthologous groups.

Carbohydrate utilization is one of the major sources for microbial energy generation. Hence, there are dedicated tools and databases providing carbohydrate-active enzyme annotations. For proGenomes4, we are utilizing Cayman, which has been shown to be highly accurate [34].

Antimicrobial resistance genes were annotated using Abricate 1.0.1 with three different databases (i.e. vfdb, megares, and deeparg 1.0.4)

We used proMGE to identify MGEs across representative genomes, using recombinases as annotation anchors and pangenome information for determining MGE boundaries [22].

Similar to MGEs, enzymes often function in conjunction with the enzymes encoded by neighboring genes. Such BGCs often encode for the cellular machinery producing ecologically or clinically relevant metabolites. We predicted the presence of BGCs using GECCO (v0.9.8) [23].

Habitat information

There has been a growing interest in consistent habitat annotations for microbial genomes, both for large scale analyses and classical microbiology applications. proGenomes4 uses both source information (via BV-BRC v3.53.3, accessed on 11 September 2025 [35]) and species detection in environmental sequencing data (via MAP [13]) to provide such annotations. BV-BRC habitat annotation enabled us to annotate 379 068 out of 1.9 M genomes (Fig. 2), which represents a two-fold increase of annotated genomes compared to the previous version of the database (proGenomes3).

To annotate habitats using the MAP, we extracted 16S rRNA genes from genomes and matched them to the sequences in the MAP v3 database using MAPseq [36]. Next, we linked proGenomes species clusters and 98% MAP Operational Taxonomic Units (OTUs) using the mapped 16S sequences. By applying a majority rule, the best matching MAP OTU was identified for each proGenomes species cluster if at

least 80% of the 16S sequences of a given proGenomes species cluster were mapped to the same 98% MAP OTU.

Links to outside databases

Even though proGenomes4 provides many different annotation tracks, dedicated databases often provide details that cannot be mirrored. Instead, we chose to add additional links to outside databases such as NCBI Genome [24], BacDive [37], GTDB [12], and MAP [36] enabling direct access.

Database design

At its core, proGenomes4 is a PostgreSQL-powered relational database system, which stores all obtained information on the included genomes and their features. The website can directly interact with the database and make the information available to the users. To efficiently access the sequence information (genomes, gene, and protein sequences), indexed FASTA flatfiles are utilized.

Programmatic access

In this version, we make available a command line tool and Python package to facilitate high-throughput use of the database. Users can easily download genome sets by habitat and genomic datasets. The tool maps the various artifacts in the database to facilitate download by users. It contains two main subcommands: view and download. The first allows users to preview datasets from proGenomes4 in the command line. The download command allows the selection of either datasets or genomic sets as the download target. For genomic sets, it also enables users to select which components to download for each habitat: contigs, genes, and proteins. The functional annotations for all representative genomes are also available via the command line tool.

The command line tool is available at https://github.com/BigDataBiology/progenomes-cli and can be readily installed via pypi.

Website

proGenomes4 can be accessed via its dedicated website (https://progenomes.embl.de). The genomes of taxonomic groups as well as specI clusters can be accessed easily via a search function. For each genome, we provide the information stored within proGenomes3 as well as direct links to external database entries.

Future outlook

proGenomes will continue to be developed actively, ocaften in response to user feedback. We aim to add additional annotation tracks in future releases as well as to improve the options for programmatic access introduced in the current release. Further, we are striving to improve integration with other resources and will determine best practices for this purpose.

Discussion

proGenomes4 offers easy and direct access to over 2 million high-quality genomes, including multiple functional annotation tracks, as well as taxonomic and habitat assignments via a dedicated website. The website further provides links to relevant entries in related databases and direct download of bulk data. proGenomes serves a broad user base ranging from microbiologists interested in well-annotated genomes for strains used for wet-lab experiments to artificial intelligence researchers interested in microbial genome dynamics. Across its different versions, proGenomes has been used as a foundation for multiple widely used resources such as eggNOG, mOTUs, and Spire.

proGenomes4 will continue to be a valuable resource, which we expect to be widely used by a broad range of researchers.

Acknowledgements

The authors would like to thank all proGenomes users as well the members of the different research groups involved, in particular Yan-Ping Yuan (EMBL) for technical support. We acknowledge the use of EMBL Heidelberg HPC Cluster [38].

Author contributions: Anthony Fullam (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Resources [equal], Software [equal], Visualization [equal], Writing – review & editing [equal]), Oleksandr Maistrenko (Formal analysis [equal], Methodology [equal], Visualization [equal], Writing – review & editing [equal]), Alexandre A. Castro (Resources [equal], Software [equal], Writing – review & editing [equal]), Luis Pedro Coelho (Software [equal], Supervision [equal], Validation [equal], Writing – review & editing [equal]), Anastasiia Grekova (Methodology [equal]), Christian Schudoma (Data curation [equal], Formal analysis [equal]), Supriya Khedkar (Methodology [equal], Validation [equal]), Shahriyar Mahdi Robbani (Data curation [equal], Formal analysis [equal]), Michael Kuhn (Conceptualization [equal], Resources [equal], Supervision [equal], Writing – review & editing [equal]), Thomas S. B. Schmidt (Conceptualization [equal], Software [equal], Supervision [equal], Writing – review & editing [equal]), Peer Bork (Conceptualization [equal], Project administration [equal], Resources [equal], Supervision [equal], Writing – review & editing [equal]), and Daniel R Mende (Conceptualization [equal], Methodology [equal], Project administration [equal], Software [equal], Supervision [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]

Conflict of interest

The authors declare no conflict of interest.

Funding

This research was conducted with the financial support of Research Ireland under Grant Number 12/RC/2273-P2 (T.S.B.S.), the Australian Research Council Future Fellowship FT230100724 (L.P.C.), Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 460129525 (NFDI4Microbiota), the Ministry of Science, Research and Art Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI, and the German Federal Ministry of Research, Technology and Space in the frame of de.NBI & ELIXIR-DE (W-de.NBI-014). Funding to pay the Open Access publication charges for this article was provided by MEXT/World Premier International Research Center Initiative (WPI), Keio University. The Human Biology-Microbiome-Quantum Research Center (WPI-Bio2Q) is

supported by World Premier International Research Center Initiative (WPI), MEXT, Japan (D.R.M).

Data availability

Progenomes4 can be accessed via its dedicated website (https://progenomes.embl.de). The source code for the command line tool is available at https://github.com/BigDataBiology/progenomes-cli.

References

1. Fleischmann RD, Adams MD, White O *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496–512. https://doi.org/10.1126/science.7542800
2. Eren AM, Banfield JF Modern microbiology: embracing complexity through integration across scales. *Cell* 2024;187:5151–70. https://doi.org/10.1016/j.cell.2024.08.028
3. Wick RR, Judd LM, Holt KE Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *PLoS Comput Biol* 2023;19:e1010905. https://doi.org/10.1371/journal.pcbi.1010905
4. Overbeek R, Bartels D, Vonstein V *et al.* Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem Rev* 2007;107:3431–47. https://doi.org/10.1021/cr068308h
5. Goldfarb T, Kodali VK, Pujar S *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res* 2025;53:D243–57. https://doi.org/10.1093/nar/gkae1038
6. Davis JJ, Wattam AR, Aziz RK *et al.* The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res* 2020;48:D606–12.
7. Yates AD, Allen J, Amode RM *et al.* Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res* 2022;50:D996–1003. https://doi.org/10.1093/nar/gkab1007
8. 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
9. Hunt M, Lima L, Anderson D *et al.* AllTheBacteria—all bacterial genomes assembled, available, and searchable. bioRxiv, https://doi.org/10.1101/2024.03.08.584059, 11 March 2024, preprint: not peer reviewed.
10. 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
11. 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. https://doi.org/10.1093/nar/gkae1004
12. 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
13. Rodrigues JFM, Tackmann J, Malfertheiner L *et al.* The MicrobeAtlas database: global trends and insights into Earth's microbial ecosystems. bioRxiv, https://doi.org/10.1101/2025.07.18.665519, 18 July 2025, preprint: not peer reviewed.
14. Dérozier S, Bossy R, Deléger L *et al.* Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. *PLoS One* 2023;18:e0272473. https://doi.org/10.1371/journal.pone.0272473
15. Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–55. https://doi.org/10.1101/gr.186072.114
16. Chklovski A, Parks DH, Woodcroft BJ *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning *Nature Methods* 2023;20:1203–12. https://doi.org/10.1038/s41592-023-01940-w
17. 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
18. Mende DR, Letunic I, Huerta-Cepas J *et al.* proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res* 2017;45:D529–34. https://doi.org/10.1093/nar/gkw989
19. Mende DR, Letunic I, Maistrenko OM *et al.* proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res* 2020;48:D621–5.
20. Fullam A, Letunic I, Schmidt TSB *et al.* proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Res* 2023;51:D760–6. https://doi.org/10.1093/nar/gkac1078
21. Hernández-Plaza A, Szklarczyk D, Botas J *et al.* eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res* 2023;51:D389–94. https://doi.org/10.1093/nar/gkac1022
22. Khedkar S, Smyshlyaev G, Letunic I *et al.* Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res* 2022;50:3155–68. https://doi.org/10.1093/nar/gkac163
23. Carroll LM, Larralde M, Fleck JS *et al.* Accurate *de novo* identification of biosynthetic gene clusters with GECCO. bioRxiv, https://doi.org/10.1101/2021.05.03.442509, 4 May 2021, preprint: not peer reviewed.
24. O'Leary NA, Cox E, Holmes JB *et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci Data* 2024;11:732.
25. Hyatt D, Chen G-L, Locascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119. https://doi.org/10.1186/1471-2105-11-119
26. 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. https://doi.org/10.1038/s41467-018-07641-9
27. Ondov BD, Treangen TJ, Melsted P *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132. https://doi.org/10.1186/s13059-016-0997-x
28. Shaw J, Yu YW Rapid species-level metagenome profiling and containment estimation with sylph. *Nat Biotechnol* 2025;43:1348–59. https://doi.org/10.1038/s41587-024-02412-y
29. Olm MR, Crits-Christoph A, Bouma-Gregson K *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* 2021;39:727–36. https://doi.org/10.1038/s41587-020-00797-0
30. Pafilis E, Frankild SP, Fanini L *et al.* The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS One* 2013;8:e65390. https://doi.org/10.1371/journal.pone.0065390
31. Steinegger M, Söding J MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–8. https://doi.org/10.1038/nbt.3988
32. 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
33. Huerta-Cepas J, Szklarczyk D, Heller D *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.

Nucleic Acids Res 2019; 47:D309–14. https://doi.org/10.1093/nar/gky1085
34. Ducarmon QR, Karcher N, Tytgat HLP *et al.* Large-scale computational analyses of gut microbial CAZyme repertoires enabled by Cayman. bioRxiv, https://doi.org/10.1101/2024.01.08.574624, 8 January 2024, preprint: not peer reviewed.
35. 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* 2023;51:D678–89. https://doi.org/10.1093/nar/gkac1003
36. 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
37. Reimer LC, Sardà Carbasse J, Koblitz J *et al.* BacDive in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res* 2022; 50:D741–6. https://doi.org/10.1093/nar/gkab961
38. European Molecular Biology Laboratory, Pečar J, Lueck R *et al.* . EMBL Heidelberg HPC cluster. version v1. Zenodo. https://doi.org/10.5281/ZENODO.12785830. 1 January 2020.

Received: September 17, 2025. **Revised:** October 12, 2025. **Accepted:** October 13, 2025