D760–D766 *Nucleic Acids Research, 2023, Vol. 51, Database issue* *Published online 21 November 2022*
https://doi.org/10.1093/nar/gkac1078

proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes

Anthony Fullam [1], Ivica Letunic [2], Thomas S.B. Schmidt [1], Quinten R. Ducarmon [1], Nicolai Karcher [1], Supriya Khedkar [1], Michael Kuhn [1], Martin Larralde [1], Oleksandr M. Maistrenko [3], Lukas Malfertheiner [4], Alessio Milanese [5], Joao Frederico Matias Rodrigues [4], Claudia Sanchis-López [6], Christian Schudoma [1], Damian Szklarczyk [4], Shinichi Sunagawa [5], Georg Zeller [1], Jaime Huerta-Cepas [6], Christian von Mering [4], Peer Bork [1,7,8,9,*] and Daniel R. Mende [10,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [2]Biobyte solutions GmbH, Bothestr. 142, 69117 Heidelberg, Germany, [3]Royal Netherlands Institute for Sea Research (NIOZ), Department of Marine Microbiology & Biogeochemistry, 1797 SZ, 't Horntje (Texel), Netherlands, [4]Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland, [5]Institute of Microbiology, Department of Biology and Swiss Institute of Bioinformatics, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland, [6]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Campus de Montegancedo-UPM, 28223, Pozuelo de Alarcón, Madrid, Spain, [7]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany, [8]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany, [9]Yonsei Frontier Lab (YFL), Yonsei University, 03722 Seoul, South Korea and [10]Department of Medical Microbiology, Amsterdam University Medical Centers, Amsterdam, The Netherlands

Received September 15, 2022; Revised October 15, 2022; Editorial Decision October 17, 2022; Accepted November 07, 2022

ABSTRACT

The interpretation of genomic, transcriptomic and other microbial 'omics data is highly dependent on the availability of well-annotated genomes. As the number of publicly available microbial genomes continues to increase exponentially, the need for quality control and consistent annotation is becoming critical. We present proGenomes3, a database of 907 388 high-quality genomes containing 4 billion genes that passed stringent criteria and have been consistently annotated using multiple functional and taxonomic databases including mobile genetic elements and biosynthetic gene clusters. proGenomes3 encompasses 41 171 species-level clusters, defined based on universal single copy marker genes, for which pan-genomes and contextual habitat annotations are provided. The database is available at http://progenomes.embl.de/

INTRODUCTION

Microbiology and microbiome research have made great advances over the recent decades, in large part thanks to the availability of large-scale genomics data (1). Nowadays, hundreds of thousands of genomes are available and microbiology has become a data-intensive as well as data-driven research field. Sequencing has become available at low costs, fueling the continued exponential increase of sequenced bacterial and archaeal genomes (2,3). This increase in data has led to many new discoveries and a better understanding of the biology of microbes facilitated by comparative genomics e.g. (4,5). To leverage comparative studies of these genomes for scientific discoveries (6), high-quality genomes with consistent annotations are required.

The proGenomes (prokaryotic genomes) database provides researchers with such high-quality genomes in a framework that can serve multiple biological disciplines ranging from evolution and ecology to medicine. By further providing easy access and many different annotation layers at once, proGenomes enables researchers of all levels

*To whom correspondence should be addressed. Tel: +31 6 5747341; Email: d.r.mende@amsterdamumc.nl
Correspondence may also be addressed to Peer Bork. Email: bork@embl.de

of expertise in genomics to perform comparative analyses and gain scientific insight. Other prominent examples of genomics databases are the NCBI RefSeq database (7), which enables public access to a comprehensive set of genomes but only provides minimal annotations, Ensembl Bacteria (8), the DOE's Joint Genome Institute Integrated Microbial Genomes & Microbiomes (JGI IMG/M) database (9), the PATRIC (Pathosystems Resource Integration Center) database (10) of the Genome Taxonomy Database (GTDB) (11,12), which focuses on a consistent taxonomy across the bacterial and archaeal tree of life. The latter is a highly important effort, as many other databases suffer from phylogenetic and taxonomic inconsistencies, often due to submitter errors (13–15). However, similar consistency is needed for other types of annotations, such as gene functions, phenotypic data and habitat information. Similarly, habitat annotation is often neglected in existing databases, and indeed for most isolates the habitat is insufficient for ecological analysis due to a lack of a unified ontology. Even if habitats are described, these are often incomparable. Different groups have set out to establish habitat databases and ontologies such as the Microbe Atlas Project (MAP), the Earth Microbiome Project (EMP), ENVO and JGI Gold (16–19). For example, the MAP uses 16S rRNA sequences from studies across the globe to map taxa to habitats (17). proGenomes3 integrates and links to MAP now, further improving the existing habitat annotations.

While general functional annotations are of utmost importance for comparative genomics (and are included in proGenomes via eggNOG annotations (20,21)), some genomic elements require focused and dedicated approaches. For example, mobile genetic elements (MGEs) cover on average 13% of prokaryotic genomes but their annotation still remains poor. Most available databases focus on annotation of a particular MGE type (22–24) and an overview of all MGEs within a genome for comparative analysis is missing. As a new feature within proGenomes3, we identified MGEs for all representative genomes using recombinase marker genes which were further annotated as transposable elements, phages, phage-like elements, conjugative elements, mobility islands, and integrons based on a previously described framework for mobile element annotation (25).

Ensuring a high quality of genomes requires an assessment of genomic completeness and contamination. Recent advances in this area have led to the development of the CheckM and GUNC tools (26–28). proGenomes3 applies these quality control tools to all included genomes and consistently annotates them taxonomically and functionally. These are combined and linked with habitat information, adding further value for comparative analyses and metagenomic studies. The updated version provides ten times as many genome sequences and annotations compared to proGenomes2 and has a higher phylogenetic coverage. Additionally, these genomes are now linked to a number of additional resources enabling direct access to a complete picture of genomes of interest. A number of workflows were improved for proGenomes3, enabling the processing of nearly one million genomes and four billion genes, while increasing the number of annotation tracks. In essence, proGenomes3 provides easy access to everything needed for comparative analyses of prokaryotic genomes. The database is available at http://progenomes.embl.de/

DATABASE CONSTRUCTION AND CHARACTERISTICS

The proGenomes3 website allows users to access and browse microbial genomes. A search function gives direct access via the NCBI assembly ID or the taxonomic name of the organism, species or clade which can be interactively explored. Subsets of genomes can be downloaded directly.

Future updates will be in regular intervals and major upgrades of the underlying computational pipeline are planned every 2 years. For the current release, proGenomes 3.0, genomes were downloaded on 30 September 2021.

Genome collection

We downloaded all bacterial and archaeal genomes that were available from the NCBI Nucleotide database on 30 Sep 2021. All genomes were annotated using Prokka (1.14.5). Closed genomes were accepted as high quality. Incomplete genomes were quality filtered using CheckM (1.0.13) and GUNC (1.0.1) (CheckM: completeness > 90% and contamination < 5%; GUNC: contamination < 5% and clade separation score < 0.45) (20,21). After removing 117 723 genomes (15 928 genomes were filtered out due to GUNC and 106 766 due to CheckM, overlap: 8319), this resulted in a total of 907 388 high-quality genomes. High-quality, yet incomplete genomes are suitable for most genomics analysis, but might still miss core genes, hence parameters should be adjusted accordingly when using these genomes for specific follow-up analyses.

Delineating species using specI clusters

The specI method delineates genomes into accurate and consistent species clusters (29). Generally, these agree with the existing species definitions based on morphological and phenotypic features. We employed a divide-and-conquer strategy to generate specI clusters: First, genomes were subdivided into broader clusters by using single linkage clustering at a 90% Average Nucleotide Identity (ANI) cutoff calculated using Mash (30). Afterwards, specI species clusters were generated for every one of these broader clusters as described for previous proGenomes versions (29). In short, a set of 40 universal, single-copy marker genes (31,32) was extracted from all genomes and pairwise genome-to-genome identities were calculated with vsearch (v1.8.0) (33) as a length-weighted average of the nucleotide identities. Pairwise identities were converted into distances and clustered using average linkage clustering with a distance cutoff of 3.5% (96.5% nucleotide identity). The 907 388 proGenomes3 genomes were delineated into 41 171 specI species clusters. This is >3-fold increase in specI clusters when compared to proGenomes2 (34). Genomes and specI clusters were taxonomically annotated using GTDB (version 202) (12) and the NCBI taxonomy (version from 1 Oct 2021).

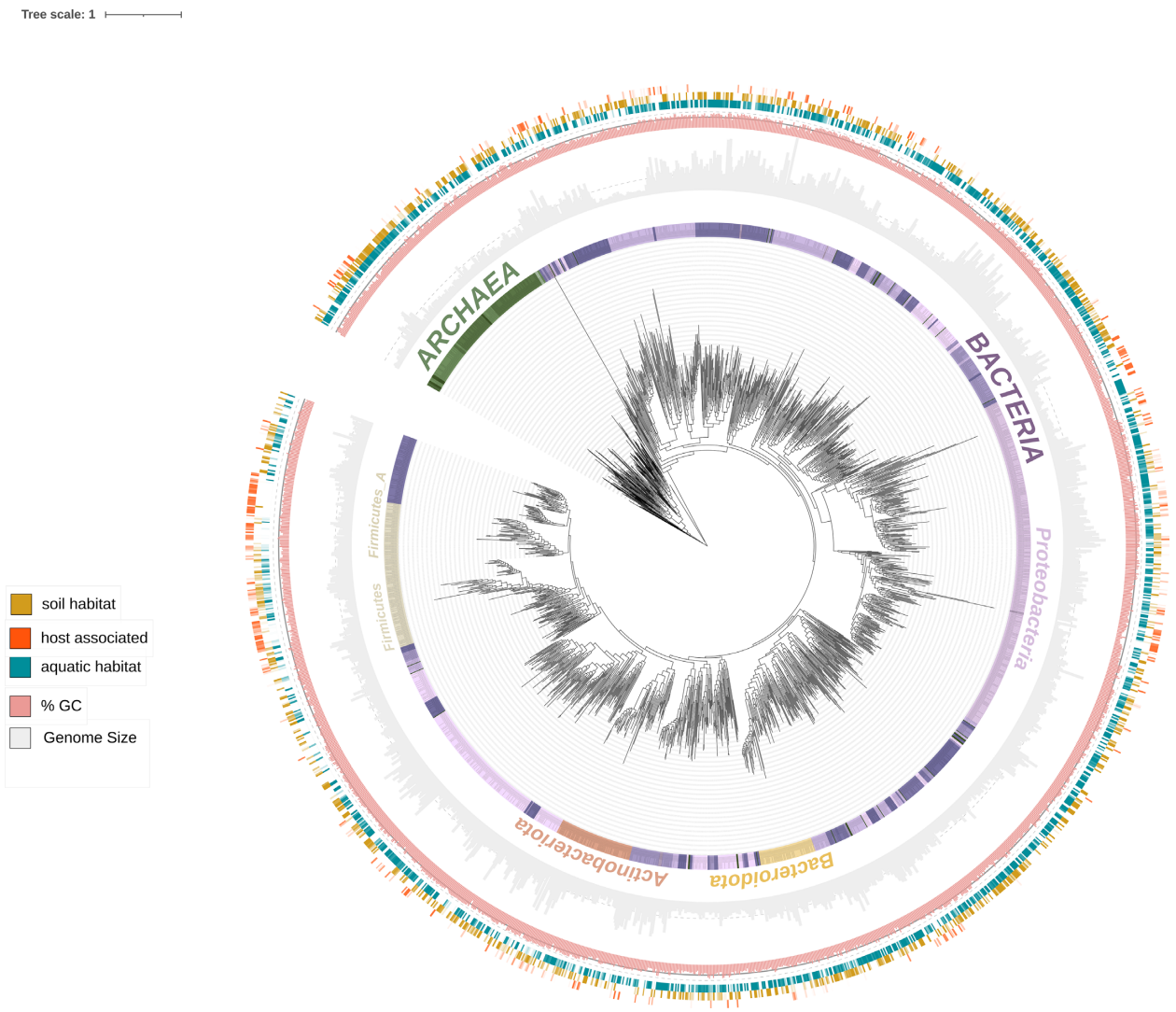

Figure 1. Phylogenetic tree of 41k representative genomes collapsed at the order level (GTDB taxonomy). Phylum, habitat as well as GC content and genome size were displayed as rings surrounding the tree.

Selection of representative genomes

Due to the availability of multiple genomes for many species and strains, genomic databases have to handle an increasing amount of redundancy. Many applications in genomics require non-redundant genomes (35,36), and accordingly proGenomes provides a non-redundant set of 41 171 representative genomes as well as habitat-specific subsets. These representative genome collections are readily available for direct download from the proGenomes website.

We selected one representative genome per specI cluster. Some strains are *de facto* representatives of a species within parts of the scientific community, for example *Mycobacterium tuberculosis* H37Rv. To make sure that these genomes are included in the set of representatives, a whitelist was compiled including genomes from highly important strains and is available on the proGenomes website. However, most clusters do not contain genomes in the whitelist. For these, we first sub-selected all complete genomes and then chose the genome of the most highly cited strain (37). If no complete genomes were part of the specI cluster, the genome with the highest N50 was selected. We provide a phylogenetic tree of all representative genomes to facilitate further analyses (Figure 1). The phylogenetic tree was built from a set of 40 universal, single-copy marker genes (29,32), which were separately aligned with FAMSA v2 (38). The concatenated alignment was used to generate a tree using FastTree/2.1.11-GCC-8.2.0–2.31.1 (39). The tree was annotated and visualized using ete4 (40).

Pan-genomes

Pan-genomes have been used to understand the genomic variability within species (4). Within proGenomes3, the pan-genome for every specI species cluster is provided as a non-redundant gene set.

These were generated by clustering using mmseqs2 (version 13.45111) (exact parameters used: –min-seq-id 0.95 -c 0.90 –cov-mode 0). Using this process, we reduced the total number of genes from ca. 4 billion to ca. 200 million while

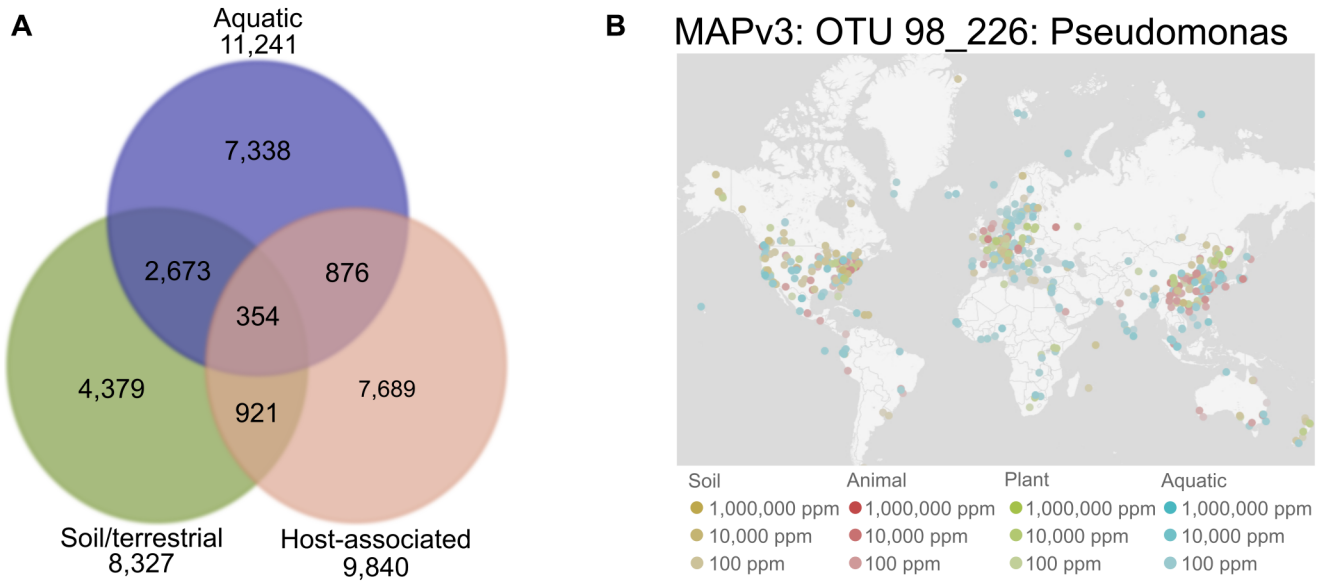

Figure 2. Habitat annotations. (**A**) Venn diagram of specI clusters annotated to different high-level habitat categories. (**B**) Global distribution of one MAP OTU linked to a specI cluster in proGenomes.

providing a more comprehensive coverage of each species' functional repertoire.

Functional annotation

Consistent functional annotation of microbial genomes is crucial for comparative analyses and to understand phenotype, lifestyle and ecological role. Providing these annotations is one of the main focal points of proGenomes. Overall annotations were assigned using eggNOG-mapper for eggNOG 5.0 (20) which assigned protein coding sequences to functionally annotated orthologous groups. A total of 3.7 billion protein-coding genes received eggNOG annotations.

To provide Carbohydrate-active Enzyme (CAZy) annotations, we utilized CAZy sequences obtained from dbCAN2 (41) to obtain optimal HMM *P*-values in a cross-validation scheme. Briefly, we divided module sequences of all (sub)families into training and testing sets and computed (sub)family-wise HMM *P*-value cutoffs that yield maximum classification performance using the testing set as positive instances (for a given (sub)family) and all other sequences as negative instances. Using these optimized *P*-values for each family, we then annotated pangenomes using the pyhmmer suite and transferred annotations to ORFs of all individual genomes.

proGenomes3 provides gene-level annotations of antimicrobial resistance based on two complementary tools with default parameters: (i) abricate v1.0.1 (https://github.com/tseemann/abricate using the Virulence Factor Database (42) (accessed 2020-04-19) and MEGARes v2.0 (43) as references; and (ii) DeepARG v1.0 (44).

Mobile genetic elements were identified by annotating recombinases using 68 high-accuracy profile HMM models and reconciling these results using pangenome information as described in (25). This yielded ∼33 million MGE recombinases across the entire database of which the ones belonging to representative genomes were subsequently used to annotate MGE types namely transposable elements, integrons, phages and conjugative elements including plasmids within the representative set.

Biosynthetic gene cluster (BGC) prediction was performed with GECCO v0.9.5 (45), using features from Pfam v35.0 (46).

Habitat information

Consistent habitat annotations are becoming more and more important for genomics analyses (4). Thus, proGenomes3 provides annotations of genomes and specI species clusters to habitats. For proGenomes3, we updated the habitat annotation process which now includes annotation based on both the PATRIC database (47) and the Microbe Atlas Project (MAP) (17).

For habitat annotations based on the PATRIC database, information regarding the isolation source was parsed from the PATRIC database version 3.6.12 (accessed on 29 August 2022). PATRIC habitat annotations are available for 25 314 out of the 41 171 specI clusters (187 808/907 388 genomes) with three main categories (soil-associated, aquatic, host-associated, Figure 2A) and several additional categories (mud/sediment, freshwater, disease-associated and food-associated). In more detail, we downloaded the PATRIC metadata including all metadata fields. The PATRIC habitat metadata was curated by finding keywords that allow to place an isolate into one of the habitat categories ('soil', 'aquatic', 'host-associated') in any of the columns 'habitat', 'isolation_source', 'disease' in the downloaded file from the PATRIC database.

For Microbe Atlas Project (MAP) annotations, we extracted 16S rRNA genes from the proGenomes3 genomes and matched them to the set of MAP OTUs clustered at 98% ID. When multiple 16S rRNA genes were found, the longest version was selected. 636 792 (84.5%) of the 753 909 16S rRNA (longer than 600 bp) sequences identified in

proGenomes3 confidently mapped to 16 366 MAP OTUs. The mapped 16S sequences were furthermore analyzed to create links between specI clusters and 98% MAP OTUs. A majority rule was employed to identify the best match for each specI cluster. A link was only generated if at least 80% of the 16S sequences within one specI cluster were mapped to the same 98% MAP OTU. This led to a reliable assignment of 19 902 specI clusters to 9511 MAP OTUs with habitat information. In proGenomes3, we link to the MAP website which also enables the visualization of the world-wide distribution of MAP OTUs (Figure 2B).

As before we compiled sets of representative genomes for different habitats which can be downloaded directly from the proGenomes website.

Links to outside databases

Dedicated databases often provide very detailed information which is not mirrored in proGenomes3. To accommodate easier access to this information, we added additional links to outside databases such as NCBI Genome (48), BacDive (49), GTDB (12) and MAP (17).

Database design

The core of proGenomes is a relational database system powered by PostgreSQL, which stores all relevant information on the included genomes and their features which are available through the web user interface. Due to its size (close to 8 Tb), the sequence information (genomes, gene and protein sequences) is stored in custom indexed FASTA flatfiles. This allows the retrieval and download of user requested individual sequences with acceptable response times.

Website

proGenomes3 can be accessed via its dedicated website (http://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.

As in previous versions, user-supplied genomes can be taxonomically annotated using the same placement algorithm as described previously for proGenomes2.

Future outlook

We are constantly improving proGenomes and will continue to do so in the future. Our goal is to provide even richer annotation sets as well as datasets that can be used for data science applications for microbial genomes. One major focus will be on the ever-growing number of MAGs has motivated plans for their inclusion in future proGenomes releases.

DISCUSSION

proGenomes3 provides nearly one million high-quality genomes with consistent taxonomic, functional, and habitat annotations. These data can be accessed via a dedicated website that also provides additional information such as links to other relevant databases or by direct download of sets of representative genomes (general and habitat specific). proGenomes continues to facilitate comparative studies addressing questions from evolution, population genetics, functional genomics and many other research fields for researchers at all levels of experience in genomics. Previous versions have been used to establish important resources such as eggNOG (20), mOTUs (50,51) and the Global Microbial Gene Catalog (52), while being used in research projects that led to impactful discoveries (4,53–55)

Hence, we expect proGenomes3 to be a valuable resource for many upcoming studies ranging from those focusing on one or a few organisms to those analyzing large-scale evolutionary patterns or complex microbial communities.

DATA AVAILABILITY

No new data were generated or analysed in support of this research. proGenomes3 is available at https://progenomes.embl.de/.

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.

FUNDING

Amsterdam UMC; European Molecular Biology Laboratory (EMBL); Swiss National Science Foundation (SNSF) [205321_184955 to S.S.]; NCCR Microbiomes [51NF40_180575 to S.S. and C.v.M.]; German Federal Ministry of Education and Research (BMBF); de.NBI network [031A537B to P.B., 031L0181A to G.Z.]; German Research Foundation (DFG) [395357507 – SFB 1371 to G.Z., 'NFDI4Microbiota' to P.B.]; European Grant with code [PGC2018-098073-A-I00MCIU/AEI/FEDER to J.H.-C.]; Spanish Ministry of Universities [FPU-19/06635 to C.S.-L.]. Funding for open access charge: EMBL.
Conflict of interest statement. None declared.

REFERENCES

1. Hall,N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.*, **210**, 1518–1525.
2. Mende,D.R., Letunic,I., Huerta-Cepas,J., Li,S.S., Forslund,K., Sunagawa,S. and Bork,P. (2017) proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res.*, **45**, D529–D534.
3. Mende,D.R., Sunagawa,S., Zeller,G. and Bork,P. (2013) Accurate and universal delineation of prokaryotic species. *Nat. Methods*, **10**, 881–884.
4. Maistrenko,O.M., Mende,D.R., Luetge,M., Hildebrand,F., Schmidt,T.S.B., Li,S.S., Rodrigues,J.F.M., von Mering,C., Pedro Coelho,L., Huerta-Cepas,J. *et al.* (2020) Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.*, **14**, 1247–1259.
5. Hug,L.A., Baker,B.J., Anantharaman,K., Brown,C.T., Probst,A.J., Castelle,C.J., Butterfield,C.N., Hernsdorf,A.W., Amano,Y., Ise,K. *et al.* (2016) A new view of the tree of life. *Nat. Microbiol.*, **1**, 16048.
6. Medini,D., Serruto,D., Parkhill,J., Relman,D.A., Donati,C., Moxon,R., Falkow,S. and Rappuoli,R. (2008) Microbiology in the post-genomic era. *Nat. Rev. Microbiol.*, **6**, 419–430.

7. Tatusova,T., Ciufo,S., Federhen,S., Fedorov,B., McVeigh,R., O'Neill,K., Tolstoy,I. and Zaslavsky,L. (2015) Update on refseq microbial genomes resources. *Nucleic Acids Res.*, **43**, D599–D605.
8. Kersey,P.J., Allen,J.E., Armean,I., Boddu,S., Bolt,B.J., Carvalho-Silva,D., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C. *et al.* (2016) Ensembl genomes 2016: more genomes, more complexity. *Nucleic Acids Res.*, **44**, D574–D80.
9. Chen,I.-M.A., Chu,K., Palaniappan,K., Pillay,M., Ratner,A., Huang,J., Huntemann,M., Varghese,N., White,J.R., Seshadri,R. *et al.* (2019) IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.*, **47**, D666–D677.
10. Wattam,A.R., Abraham,D., Dalay,O., Disz,T.L., Driscoll,T., Gabbard,J.L., Gillespie,J.J., Gough,R., Hix,D., Kenyon,R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D91.
11. Parks,D.H., Chuvochina,M., Waite,D.W., Rinke,C., Skarshewski,A., Chaumeil,P.-A. and Hugenholtz,P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004.
12. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumeil,P.-A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.
13. Beaz-Hidalgo,R., Hossain,M.J., Liles,M.R. and Figueras,M.-J. (2015) Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for aeromonas genomes in the genbank database. *PLoS One*, **10**, e0115813.
14. Chen,Q., Zobel,J. and Verspoor,K. (2017) Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database*, **2017**, baw163.
15. Vilgalys,R. (2003) Taxonomic misidentification in public DNA databases. *New Phytol.*, **160**, 4–5.
16. Buttigieg,P.L., Pafilis,E., Lewis,S.E., Schildhauer,M.P., Walls,R.L. and Mungall,C.J. (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semantics*, **7**, 57.
17. Matias Rodrigues,J.F., Schmidt,T.S.B., Tackmann,J. and von Mering,C. (2017) MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics*, **33**, 3808–3810.
18. Mukherjee,S., Stamatis,D., Bertsch,J., Ovchinnikova,G., Sundaramurthi,J.C., Lee,J., Kandimalla,M., Chen,I.-M.A., Kyrpides,N.C. and Reddy,T.B.K. (2021) Genomes online database (GOLD) v.8: overview and updates. *Nucleic Acids Res.*, **49**, D723–D733.
19. Thompson,L.R., Sanders,J.G., McDonald,D., Amir,A., Ladau,J., Locey,K.J., Prill,R.J., Tripathi,A., Gibbons,S.M., Ackermann,G. *et al.* (2017) A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, **551**, 457–463.
20. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernández-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
21. Cantalapiedra,C.P., Hernández-Plaza,A., Letunic,I., Bork,P. and Huerta-Cepas,J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, **38**, 5825–5829.
22. Siguier,P., Perochon,J., Lestrade,L., Mahillon,J. and Chandler,M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**, D32–D36.
23. Leplae,R., Lima-Mendez,G. and Toussaint,A. (2010) ACLAME: a CLAssification of mobile genetic elements, update 2010. *Nucleic Acids Res.*, **38**, D57–D61.
24. Bi,D., Xu,Z., Harrison,E.M., Tai,C., Wei,Y., He,X., Jia,S., Deng,Z., Rajakumar,K. and Ou,H.-Y. (2012) ICEberg: a web-based resource for integrative and conjugative elements found in bacteria. *Nucleic Acids Res.*, **40**, D621–D626.
25. Khedkar,S., Smyshlyaev,G., Letunic,I., Maistrenko,O.M., Coelho,L.P., Orakov,A., Forslund,S.K., Hildebrand,F., Luetge,M., Schmidt,T.S.B. *et al.* (2022) Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res.*, **50**, 3155–3168.
26. Orakov,A., Fullam,A., Coelho,L.P., Khedkar,S., Szklarczyk,D., Mende,D.R., Schmidt,T.S.B. and Bork,P. (2021) GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.*, **22**, 178.
27. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
28. Chklovski,A., Parks,D.H., Woodcroft,B.J. and Tyson,G.W. (2022) CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. bioRxiv doi: http://dx.doi.org/10.1101/2022.07.11.499243, 11 July 2022, preprint: not peer reviewed.
29. Mende,D.R., Sunagawa,S., Zeller,G. and Bork,P. (2013) Accurate and universal delineation of prokaryotic species. *Nat. Methods*, **10**, 881–884.
30. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol.*, **17**, 132.
31. Sorek,R., Zhu,Y., Creevey,C.J., Francino,M.P., Bork,P. and Rubin,E.M. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, **318**, 1449–1452.
32. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
33. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.
34. Mende,D.R., Letunic,I., Maistrenko,O.M., Schmidt,T.S.B., Milanese,A., Paoli,L., Hernández-Plaza,A., Orakov,A.N., Forslund,S.K., Sunagawa,S. *et al.* (2020) proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic. Acids. Res.*, **48**, D621–D625.
35. Van Rossum,T., Costea,P.I., Paoli,L., Alves,R., Thielemann,R., Sunagawa,S. and Bork,P. (2021)metaSNV v2: detection of SNVs and subspecies in prokaryotic metagenomes. *Bioinformatics*, **38**, 1162–1164.
36. Olm,M.R., Crits-Christoph,A., Bouma-Gregson,K., Firek,B.A., Morowitz,M.J. and Banfield,J.F. (2021) inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol*, **39**, 727–736.
37. Pafilis,E., Frankild,S.P., Fanini,L., Faulwetter,S., Pavloudi,C., Vasileiadou,A., Arvanitidis,C. and Jensen,L.J. (2013) The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS One*, **8**, e65390.
38. Deorowicz,S., Debudaj-Grabysz,A. and Gudyś,A. (2016) FAMSA: fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.*, **6**, 33964.
39. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
40. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.
41. Zhang,H., Yohe,T., Huang,L., Entwistle,S., Wu,P., Yang,Z., Busk,P.K., Xu,Y. and Yin,Y. (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.*, **46**, W95–W101.
42. Liu,B., Zheng,D., Jin,Q., Chen,L. and Yang,J. (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.
43. Doster,E., Lakin,S.M., Dean,C.J., Wolfe,C., Young,J.G., Boucher,C., Belk,K.E., Noyes,N.R. and Morley,P.S. (2020) MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res.*, **48**, D561–D569.
44. Arango-Argoty,G., Garner,E., Pruden,A., Heath,L.S., Vikesland,P. and Zhang,L. (2018) DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, **6**, 23.

45. Carroll,L.M., Larralde,M., Fleck,J.S., Ponnudurai,R., Milanese,A., Cappio,E. and Zeller,G. (2021) Accurate de novo identification of biosynthetic gene clusters with GECCO. bioRxiv doi: http://dx.doi.org/10.1101/2021.05.03.442509, 04 May 2021, preprint: not peer reviewed.
46. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
47. Davis,J.J., Wattam,A.R., Aziz,R.K., Brettin,T., Butler,R., Butler,R.M., Chlenski,P., Conrad,N., Dickerman,A., Dietrich,E.M. *et al.* (2020) The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Res.*, **48**, D606–D612.
48. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Connor,R., Funk,K., Kelly,C., Kim,S. *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **50**, D20–D26.
49. Reimer,L.C., Sardà Carbasse,J., Koblitz,J., Ebeling,C., Podstawka,A. and Overmann,J. (2022) BacDive in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res.*, **50**, D741–D746.
50. Milanese,A., Mende,D.R., Paoli,L., Salazar,G., Ruscheweyh,H.-J., Cuenca,M., Hingamp,P., Alves,R., Costea,P.I., Coelho,L.P. *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.*, **10**, 1014.
51. Ruscheweyh,H.-J., Milanese,A., Paoli,L., Karcher,N., Clayssen,Q., Metzger,M.I., Wirbel,J., Bork,P., Mende,D.R., Zeller,G. *et al.* (2021) Reference genome-independent taxonomic profiling of microbiomes with mOTUs3. bioRxiv doi: http://dx.doi.org/10.1101/2021.04.20.440600, 08 April 2022, preprint: not peer reviewed.
52. Coelho,L.P., Alves,R., Del Río,Á.R., Myers,P.N., Cantalapiedra,C.P., Giner-Lamia,J., Schmidt,T.S., Mende,D.R., Orakov,A., Letunic,I. *et al.* (2022) Towards the biogeography of prokaryotic genes. *Nature*, **601**, 252–256.
53. Paoli,L., Ruscheweyh,H.-J., Forneris,C.C., Hubrich,F., Kautsar,S., Bhushan,A., Lotti,A., Clayssen,Q., Salazar,G., Milanese,A. *et al.* (2022) Biosynthetic potential of the global ocean microbiome. *Nature*, **607**, 111–118.
54. Schmidt,T.S.B., Li,S.S., Maistrenko,O.M., Akanni,W., Coelho,L.P., Dolai,S., Fullam,A., Glazek,A.M., Hercog,R., Herrema,H. *et al.* (2022) Drivers and determinants of strain dynamics following fecal microbiota transplantation. *Nat. Med.*, **28**, 1902–1912.
55. Nocedal,I. and Laub,M.T. (2022) Ancestral reconstruction of duplicated signaling proteins reveals the evolution of signaling specificity. *Elife*, **11**, e77346.