proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes

Daniel R. Mende [1,*], Ivica Letunic[2], Oleksandr M. Maistrenko[3], Thomas S.B. Schmidt[3], Alessio Milanese[3], Lucas Paoli[4], Ana Hernández-Plaza[5], Askarbek N. Orakov[3], Sofia K. Forslund[6], Shinichi Sunagawa[4], Georg Zeller[3], Jaime Huerta-Cepas[5], Luis Pedro Coelho[7,8] and Peer Bork[3,6,9,10,*]

[1]Department of Medical Microbiology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands, [2]Biobyte solutions GmbH, Bothestr, 142, 69117 Heidelberg, Germany, [3]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [4]Institute of Microbiology, Department of Biology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland, [5]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Campus de Montegancedo-UPM, 28223, Pozuelo de Alarcón, Madrid, Spain, [6]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany, [7]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China, [8]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, China, [9]Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, 69120 Heidelberg, Germany and [10]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

ABSTRACT

Microbiology depends on the availability of annotated microbial genomes for many applications. Comparative genomics approaches have been a major advance, but consistent and accurate annotations of genomes can be hard to obtain. In addition, newer concepts such as the pan-genome concept are still being implemented to help answer biological questions. Hence, we present proGenomes2, which provides 87 920 high-quality genomes in a user-friendly and interactive manner. Genome sequences and annotations can be retrieved individually or by taxonomic clade. Every genome in the database has been assigned to a species cluster and most genomes could be accurately assigned to one or multiple habitats. In addition, general functional annotations and specific annotations of antibiotic resistance genes and single nucleotide variants are provided. In short, proGenomes2 provides threefold more genomes, enhanced habitat annotations, updated taxonomic and functional annotation and improved linkage to the NCBI BioSample database. The database is available at http://progenomes.embl.de/.

INTRODUCTION

Large-scale genomics has been instrumental for our improved understanding of microbes. Microbiology has developed into a data-intensive field with the availability of thousands of sequenced genomes (1–3). Over the last 20+ years, the number of bacteria and archaea with sequenced genomes has grown exponentially (4,5). To facilitate an understanding of microbes from their genomic data, annotations are essential. These enable researchers to pinpoint potential functions and allow for comparative analyses (6). For this reason, we initially developed proGenomes and are continuing to improve the database. Several publicly accessible databases provide genomes with basic or even more elaborate annotations. For example, the NCBI RefSeq database (7) make a comprehensive set of genomes available to the public (though only minimal annotations are provided). Further, databases such as Ensembl Bacteria (8), the DOE's Joint Genome Institute Integrated Microbial Genomes & Microbiomes (JGI IMG/M) database (9), or the PATRIC (Pathosystems Resource Integration Center) database (10) contain more sophisticated, but often se-

lect information and annotations. For these databases, the taxonomic annotations are selected by the submitter of each genome. This leads to inconsistencies across different clades across the tree of life, especially at the species level, as the species definition for bacteria and archaea remains a highly debated topic among microbiologists (11,12). In general and not only due to user errors, inconsistencies are widespread in genomic databases (13–15). A successful effort to increase the consistency of the taxonomy at higher taxonomic levels is the Genome Taxonomy Database (GTDB) (12), while specI (5) using genomics information to delineate species was used in proGenomes v1 (4).

The pan-genome concept has been an important advance in microbial genomics and microbiology overall (16,17). Due to the availability of many genome sequences within one species, researchers can now explore the pan-genome of many species and study the functional repertoire of species. Still most genomes are studied on an individual basis even in comparative approaches. Dedicated databases for pan-genomes exist, but these are often focused on specific taxonomic clades or lack in-depth functional annotations. Hence, the availability of pan-genomes for many species could facilitate many studies and applications

Here, we present proGenomes2 which was developed to address these issues as an update of the existing proGenomes database. The updated version provides three times as many genome sequences and annotations and a higher phylogenetic coverage while adding information about the pan-genome of every species cluster. A number of workflows were improved for proGenomes2 including enhanced habitat annotations and linkage to the NCBI BioSample database. The database is available at http://progenomes.embl.de/

DATABASE CONSTRUCTION AND CHARACTERISTICS

proGenomes2 provides the available microbial genomes and customizable subsets in a readily downloadable and user-friendly manner. Genomes and sets of genomes can be found and retrieved using the taxonomic name of the organism, species or clade. The provided information can be accessed, explored interactively and downloaded easily. The database will be updated regularly in the future and major upgrades of the underlying computational pipeline are planned every two years. The genomes for proGenomes2 were obtained on 15 May 2017 and the NCBI taxonomy database used was downloaded on 8 January 2019.

Genome collection

We downloaded all bacterial and archaeal genomes that were available from the NCBI Nucleotide database on 15 May 2017. Gene annotations provided with the deposited genomes were used when available. If gene annotations were absent from the deposited genomes, we used geneMarkS to predict genes (18). To exclude low quality genomes, we used the same parameters as used in the original proGenomes version (N50 score >10k bp, <300 contigs and >30 of 40 universal, single copy marker genes (19,20)). We further removed genomes that were since removed from the NCBI Nucleotide database and genomes

that we found to be of low quality by manual quality filtering. After these filtering steps, 87 920 high-quality genomes remained (10 333 genomes were removed). In comparison, proGenomes version 1 contained 25 038 high-quality genomes. proGenomes2 normalizes genome, gene and protein identifiers to a consistent scheme, linking them to NCBI taxonomic and BioSample IDs, to facilitate downstream automated processing. This also ensure access to information about the sequenced sample that is provided by the NCBI BioSample Database (21)

Species clusters definitions using the specI approach

specI species clusters provide an accurate and consistent solution for genomics-based species definition that are largely consistent with consensus from morphological and phenotypic evaluation). We calculated specI species clusters as in the previous proGenomes version using the methodology described in (5), resulted in 12 221 specI species clusters for the 87 920 genomes currently in proGenomes2 (proGenomes version 1: 5306 specI cluster). In short, pairwise genome-to-genome identities were calculated as a length-weighted average of the nucleotide identities of a set of 40 universal, single copy marker genes (19,20) calculated with vsearch (v1.8.0) (22). These were transformed into distances and clustered using average linkage employing a cutoff of 3.5% distance (96.5% nucleotide ID). Genomes are annotated according to the NCBI taxonomy (downloaded on 8 January 2019 and available at https://doi.org/10.5281/zenodo.3357977). Annotation for the specI is derived from the annotation of the genomes that compose the specI clusters.

Selection of representative genomes

Many species and even strains have been sequenced multiple times leading to an increasing amount of redundancy in genomic databases. Non-redundant datasets are increasingly important in microbial genomics (23). Hence, proGenomes provides a non-redundant set of 12 221 representative genomes and habitat-specific subsets. These datasets are precompiled and can be readily downloaded.

For each specI cluster one representative genome was selected. For this purpose, a whitelist containing several highly-important genomes was compiled. For all specI clusters that contained a genome on the whitelist, this genome was selected as representative. For all other specI clusters, the representative genomes were selected using citation counts as well as the N50 measure as a proxy for genome quality, while completely assembled genomes were selected preferentially.

Pan-genomes

In addition to representative genomes, proGenomes provides pan-genomes for the specI clusters. These are nonredundant sets of genes that represent the genetic diversity within a specI (species) cluster. Per specI-cluster pangenomes were generated in two steps. First, genes were dereplicated by sorting the genes and removing identical sequences. Second, these dereplicated gene sets were further
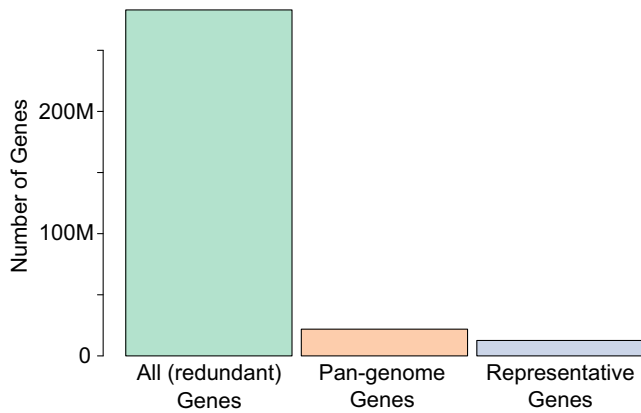
Figure 1. Cumulative number of genes in specI clusters with more than one genome. Number of all redundant genes (left), genes represented in pan-genomes (middle) and genes from representative genomes (right).

clustered with cd-hit-est (24) to produce non-redundant versions, at 95% identity and 90% coverage (exact command used: -c 0.95 -G 0 -g 1 -aS 0.9) (25). The resulting pangenomes can be used for many applications ranging from metagenomic mapping to evolutionary analyses of functional genes across different species. For clusters with more than one genome, this reduced the number of genes from 283 million to 63 million, while providing a far greater coverage of the functional repertoire as the representative genomes alone (21.8 million) (Figure 1).

Functional annotation

The functional repertoire of a microbial genome defines its phenotype, lifestyle and ecological role. Hence, it is pivotal to our understanding of a microorganism that we have a consistent, accurate and comprehensive functional annotation of its genes. One of the main aspects of proGenomes is to provide consistently-computed functional annotation of protein-coding genes. For general annotations, we use the eggNOG-mapper for eggNOG 5.0 (26) software that assigns protein-coding genes to orthologous groups which are in turn assigned to functions and broader functional categories.

proGenomes2 provides dedicated antibiotic resistance annotations of both antimicrobial resistance genes and resistance-conveying single nucleotide variants. The antibiotic resistance annotations in proGenomes2 are provided based on integrated results from the Comprehensive Antibiotic Resistance Database (23,27) and ResFams (28) resources as in proGenomes version 1. Since both databases (CARD v3.0.0 and ResFams v1.2.2) map to the antibiotic resistance ontology (ARO), the ARO hierarchy (as per CARD version 3.0) was used to assess which antibiotics each resistance gene determinant protects against. Proxy terms for 'unspecified beta-lactam' and 'multidrug efflux pump' were added to reconcile ambiguities in some annotations. For complexes listed in the ARO, such as components with disparate subunits, such synergies between hits were counted within each genome, reflecting how the presence of several interacting antibiotic resistance genes can provide further resistance. Overall, almost 288 million protein-

coding genes were annotated using eggNOG (proGenomes version 1: 80 million). This information can be explored interactively on the proGenomes website.

Habitat information

proGenomes2 provides annotations of each genome and species to a habitat. As previously, the habitat annotations are based on the PATRIC database. Information regarding the isolation source was parsed from Patric database version 3.5.28 (accessed on 8 December 2018) (29). Habitat annotations are available for 7218 out of the 12 221 specI clusters (59 344/87 920 isolates). Species were considered associated with a habitat if at least one genome was isolated from that habitat. We expanded the habitat classification from the previous version and provide three broad (soil-associated, aquatic, host-associated) and additionally five more specific categories (mud/sediment, freshwater, disease-associated, food-associated)—biologically meaningful categories (Figure 2). Representative genomes for these subsets of clusters are available for bulk download from the website. We removed the category 'multiple' used in the previous version and allow the annotation to more than one category in proGenomes2. As many genomes were annotated as food-associated, isolated from sediment or freshwater, we introduced these as a new categories. Annotations are provided with each genome and specI cluster as well as one large downloadable file.

Website

To navigate and access the proGenomes2 database, we provide a website (http://progenomes.embl.de). A search function allows users to access data for specific taxonomic groups or specI clusters directly. Information about single genomes is provided interactively with direct links to relevant external database information, while for higher level taxonomic groups concise information about the organisms belonging to that group are displayed.

The website further allows users to provide their own genomes for annotation and contextualization within the proGenomes framework. Taxonomic placement with respect to the above-described specI clusters proceeds in four steps. First, genes and proteins in the uploaded genome are identified using Prodigal (30). Second, among these genes the 40 marker genes on which the specI clusters were defined are identified using the methodology described in (5). In a third step, the extracted marker genes are compared to those from the 12 221 specI clusters using vsearch (22) with the gene-specific thresholds introduced above. Finally, a consistent annotation for the genome is derived from the mapping of individual marker genes (more details are documented on the progenome website). This taxonomic placement tool is also available as stand-alone software. Functional annotations via the eggnog-mapper web server (31) are accessible via a link.

Future outlook

We are planning to upgrade proGenomes further in the near future. The upcoming improvements include the integration
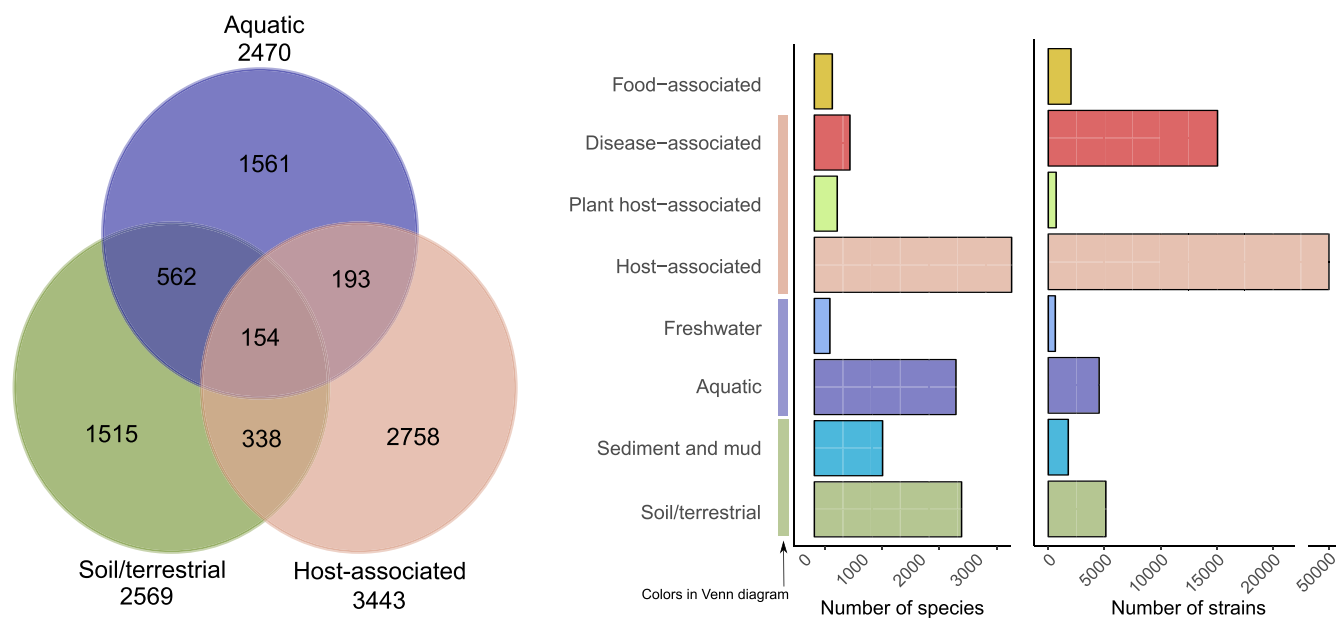
Figure 2. Habitat annotations. In proGenomes2 organisms can be associated with multiple habitats categories. The left panel shows the overlap between specI species cluster habitat annotations. The right panel shows how many species (clusters) and strains are associated with each habitat category.

of the GTDB taxonomy and the inclusion of metagenomics-assembled genomes (MAGs). To enable this we aim to improve our quality measures for genomes in general and with a specific focus on issues related to MAGs.

DISCUSSION

proGenomes provides consistent taxonomic and functional annotations for quality filtered genomes, as well as a non-redundant, habitat-specific sets of representative genomes. The easy-to-use website provides a wide range of information relevant to researchers interested in microbial genomics and allows the customization of subsets of genomes for download, thus facilitating comparative studies that address questions from evolution, population genetics, functional genomics and many other research fields. We intend proGenomes to be a valuable resource for studies ranging from those focusing on one or a few organisms to those analyzing large-scale evolutionary patterns or complex microbial communities.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the Bork group, in particular Yan-Ping Yuan for technical support.

FUNDING

REFERENCES

1. Hall,N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.*, **210**, 1518–1525.
2. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A. and Merrick,J.M. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.
3. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) The minimal gene complement of Mycoplasma genitalium. *Science*, **270**, 397–403.
4. 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.
5. Mende,D.R., Sunagawa,S., Zeller,G. and Bork,P. (2013) Accurate and universal delineation of prokaryotic species. *Nat. Methods*, **10**, 881–884.
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–D580.

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–D591.

11. Rosselló-Mora,R. and Amann,R. (2001) The species concept for prokaryotes. *FEMS Microbiol. Rev.*, **25**, 39–67.

12. 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.

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. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.

17. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13950–13955.

18. Borodovsky,M. and Lomsadze,A. (2014) Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Curr. Protoc. Microbiol.*, **32**, Unit 1E.7.

19. 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.

20. 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.

21. Federhen,S., Clark,K., Barrett,T., Parkinson,H., Ostell,J., Kodama,Y., Mashima,J., Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2014) Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand. Genomic Sci.*, **9**, 1275–1277.

22. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ.*, **4**, e2584.

23. Schloissnig,S., Arumugam,M., Sunagawa,S., Mitreva,M., Tap,J., Zhu,A., Waller,A., Mende,D.R., Kultima,J.R., Martin,J. *et al.* (2013) Genomic variation landscape of the human gut microbiome. *Nature*, **493**, 45–50.

24. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

25. Jain,C., Rodriguez-R,L.M., Phillippy,A.M., Konstantinidis,K.T. and Aluru,S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 5114.

26. 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.

27. Jia,B., Raphenya,A.R., Alcock,B., Waglechner,N., Guo,P., Tsang,K.K., Lago,B.A., Dave,B.M., Pereira,S., Sharma,A.N. *et al.* (2017) CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **45**, D566–D573.

28. Gibson,M.K., Forsberg,K.J. and Dantas,G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 207–216.

29. Wattam,A.R., Davis,J.J., Assaf,R., Boisvert,S., Brettin,T., Bun,C., Conrad,N., Dietrich,E.M., Disz,T., Gabbard,J.L. *et al.* (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.*, **45**, D535–D542.

30. Hyatt,D., Chen,G.-L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

31. Huerta-Cepas,J., Forslund,K., Coelho,L.P., Szklarczyk,D., Jensen,L.J., von Mering,C. and Bork,P. (2017) Fast Genome-Wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, **34**, 2115–2122.