Published online 24 October 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D529–D534***
doi: 10.1093/nar/gkw989

proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes

Daniel R. Mende[1,2], Ivica Letunic[3], Jaime Huerta-Cepas[1], Simone S. Li[1,4], Kristoffer Forslund[1], Shinichi Sunagawa[1,5] and Peer Bork[1,6,7,8,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [2]Daniel K. Inouye Center for Microbial Oceanography Research and Education, University of Hawai'i at Manoa, Honolulu, HI 96822, USA, [3]Biobyte solutions GmbH, Bothestrasse 142, 69126 Heidelberg, Germany, [4]School of Biotechnology and Biomolecular Sciences, University of New South Wales, 2052 Sydney, Australia, [5]Institute of Microbiology, Department of Biology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland, [6]Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, 69120 Heidelberg, Germany, [7]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany and [8]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

Received September 5, 2016; Revised October 11, 2016; Editorial Decision October 12, 2016; Accepted October 17, 2016

ABSTRACT

The availability of microbial genomes has opened many new avenues of research within microbiology. This has been driven primarily by comparative genomics approaches, which rely on accurate and consistent characterization of genomic sequences. It is nevertheless difficult to obtain consistent taxonomic and integrated functional annotations for defined prokaryotic clades. Thus, we developed proGenomes, a resource that provides user-friendly access to currently 25 038 high-quality genomes whose sequences and consistent annotations can be retrieved individually or by taxonomic clade. These genomes are assigned to 5306 consistent and accurate taxonomic species clusters based on previously established methodology. proGenomes also contains functional information for almost 80 million protein-coding genes, including a comprehensive set of general annotations and more focused annotations for carbohydrate-active enzymes and antibiotic resistance genes. Additionally, broad habitat information is provided for many genomes. All genomes and associated information can be downloaded by user-selected clade or multiple habitat-specific sets of representative genomes. We expect that the availability of high-quality genomes with comprehensive functional annotations will promote advances in clinical microbial genomics, functional evolution and other subfields of microbiology. proGenomes is available at http://progenomes.embl.de.

INTRODUCTION

Microbes play a major role in shaping the earth and have vast impacts on human health and well-being. Until recently, however, little was known about their diversity, genetics and functional potential. Over the last two decades, this has changed with the availability of sequenced genomes, which has revolutionized our understanding of microbes (1–3). The extensive use of genome sequencing in microbiology has led to an exponential increase in the number of sequenced bacteria and archaea (4) (Figure 1). However, a genome sequence alone, even if perfectly assembled, is of limited value without annotation to reveal interpretable information. The most basic annotation level should provide the taxonomic designation of a genome and the sequences of genes it encodes. Functional annotation of the latter can reveal, for example, the biochemical processes that underlie phenotypic features of a specific microbe.

The field of comparative genomics, through increasing efforts in the characterization of genomes, has led to new advances in our understanding of bacterial and archaeal life (5). Although such studies necessitate the use of consistently annotated genomes, the current state-of-the-art does not yet provide an easy entry point to obtain these. A number of publicly accessible databases provide genomes with basic and more elaborate annotations. The NCBI RefSeq database (6) provides a comprehensive sets of genomes with minimal annotations that include consistently predicted gene models. Other databases such as the DOE's Joint Genome Institute Integrated Microbial Genomes & Microbiomes (JGI IMG/M) database (7), the PATRIC (Pathosystems Resource Integration Center) database (8) and Ensembl Bacteria (9) provide additional layers of information for the deposited genomes through integration

*To whom correspondence should be addressed. Tel: +49 6221 387 8361; Fax: +49 6221 387 517; Email: bork@embl.de

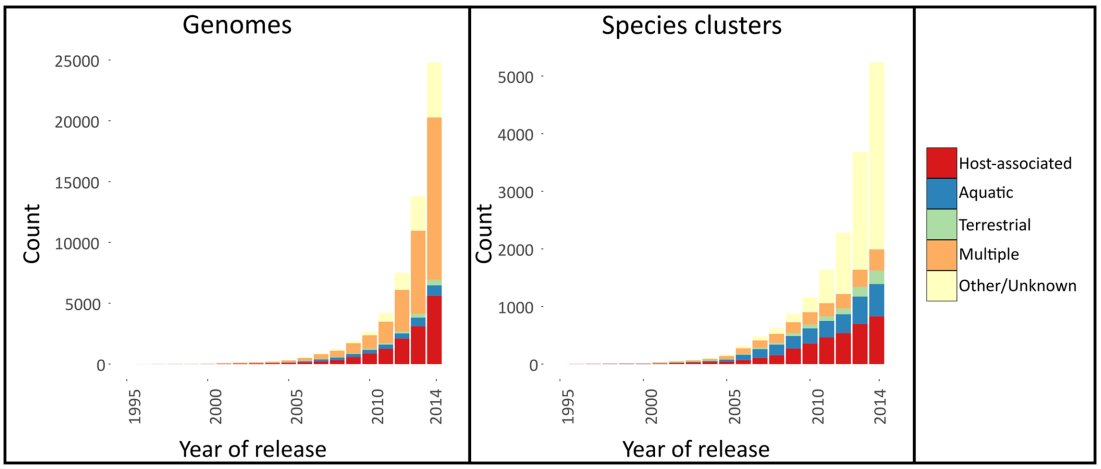

Figure 1. Availability of sequenced genomes and species clusters availability over time. Colors represent the habitat annotation of the genomes/species clusters.

of other data sources. Yet, taxonomic annotations are usually provided by the submitter of the genome sequence. This leads to inconsistencies across different clades of the tree of life, especially at the species level, as the species definition for bacteria and archaea remains a highly debated topic among microbiologists (10). Furthermore, the taxonomic classification of prokaryotes is constantly updated, which hampers efforts to download subsets of species for a desired project. Obtaining a consistent functional ontology for a number of genomes can also be challenging as a number of functional databases exist, each of which covers distinct aspects of functional diversity (e.g. antibiotic resistance (11) or metabolic pathways (12)), and current genome resources are either incomplete or lack cross-referencing information.

To address these issues, we have developed proGenomes (http://progenomes.embl.de), a prokaryotic genome resource that enables direct access to genomes of any taxonomic clade, in conjunction with a number of consistently and hierarchically annotated gene functions for each genome. Additionally, we provide a robust operational species classification in the form of up-to-date species clusters, which perform well compared to the NCBI reference taxonomy (13). In order to minimize redundancy, a representative genome is selected from each specI species cluster reflecting its role in the literature and other criteria (Figure 2). The resulting non-redundant genome sets are thus well suited for metagenomic or large-scale phylogenetic studies.

The sets of genes from each genome were translated into proteomes and consistently annotated using eggNOG, one of the most comprehensive databases for orthologs and their functional annotation, with 1.9 million orthologous groups (14). Using the eggNOG-mapper tool, we were able to annotate almost 79 million protein-coding genes (15) including indirect annotations to KEGG pathways and predicted gene family names.

We also provide more specific annotations of carbohydrate-active enzymes, as well as antibiotic resistance determinants, which are additional features not currently provided together by other databases. The range of such annotations will be extended in future updates.

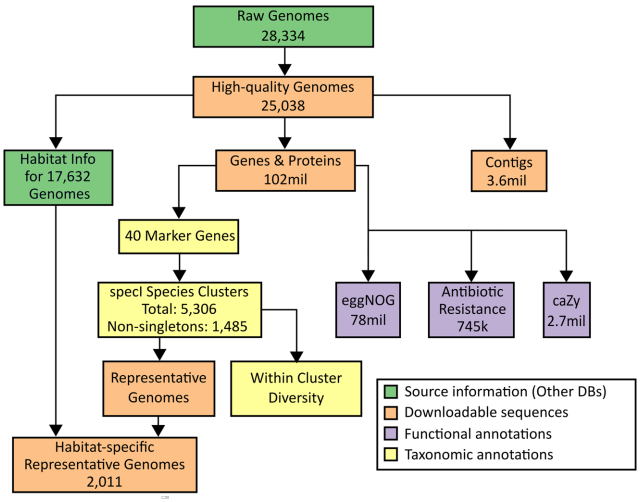

Figure 2. Workflow to generate the underlying data of the database.

We envision that comprehensive functional annotation of high-quality genomes will facilitate research into clinical applications of microbial genomics, as well as studies of functional evolution.

DATABASE CONSTRUCTION AND CHARACTERISTICS

The goal of proGenomes is to provide the available microbial genomes and customizable subsets of them in a readily downloadable and user-friendly manner. Users can find genomes of interest by providing the name of a genome or a taxonomic clade. The website also allows users to explore all the provided information interactively, and genome sequences and annotations of individual genomes or whole taxonomic clades can be easily downloaded. Pre-packaged sets of representative genomes are also available for batch download. The computational pipeline that generates the data presented on the website is summarized in Figure 2. We aim to update the database up to two times per year. Fur-

ther, we plan to perform major biennial updates, that will involve the integration of additional annotation sources or major improvements of existing parts of the workflow.

Genome collection

The genome collection is based on all bacterial and archaeal genomes that were available from the NCBI Nucleotide database on 14 December 2014. Gene predictions were obtained from the deposited genomes, where available. If this information was not provided, genes were predicted using geneMarkS (16). We filtered out low quality assemblies that had an N50 score of <10k bp and/or consisted of more than 300 contigs. Incompletely assembled genomes with <30 of 40 universal, single copy marker genes were also removed (17,18). The detection of a sufficient number of these marker genes in a genome provides a universally applicable measure of genome completeness. Altogether, this resulted in a set of 25 038 high-quality genomes.

Species clusters definitions using the specI approach

As mentioned above, the assignment of genomes to species is contentious (12). The exponentially increasing number of sequenced genomes necessitate the development and use of automatic, unbiased and systematic approaches to tackle this issue. specI species clusters provide an accurate and consistent solution, as they are based solely on genomic sequence (but also largely consistent with consensus from morphological and phenotypic evaluation) and can be applied to any set of sequenced genomes. We calculated specI species clusters using the methodology described in (13), resulted in 5306 specI species clusters for the 25 038 genomes currently in proGenome. This represents a significant advance in comparison to previous efforts, such as the MetaRef database (19), which provided a similar classification for 2818 genomes based on clade specific genes.

The specI approach utilizes a set of 40 universal, single-copy marker gene families (MGs) (17,18) that are provided as part of the resource for each genome. The MGs have been used to reconstruct the tree of life (17) and to study the phylogenetic relations within specific clades (13,20). The fetchMG tool (21) was used to extract the MGs from all high-quality genomes. To generate the updated specI species clusters, all-versus-all alignments were calculated for each of the 40 MGs using vsearch (Rognes *et al.*, https://github.com/torognes/vsearch) and genome-to-genome distances were calculated as gene length weighted mean averages. The genome-to-genome distances were then used as input for average linkage clustering. An average marker gene nucleotide identity cutoff of 96.5% was applied to generate the specI species clusters. This yielded a total of 5306 specI species clusters, of which 1485 contained more than one genome and 3821 were singletons. Non-singleton clusters contained 14.3 genomes on average. The largest cluster (specI_v2_Cluster67: Staphylococcus aureus) contained 4172 genomes. The updated specI species clusters can be easily accessed in proGenomes, either by directly searching for them or by link from any constituent genome. Due to their consistency, these clusters represent an unbiased starting point for pangenomic studies and can also serve as benchmark sets for metagenomic binning approaches.

Selection of representative genomes

Many applications in microbial genomics require non-redundant datasets. This can be due to the detrimental effects of redundancy itself (e.g. when trying to uniquely assign metagenomic reads to reference genomes, as in (22)) or because of significant efficiency gains at comparable accuracies (e.g. (14)). NCBI RefSeq currently provides a set of representative genomes from 4287 species (8), however, many species clusters were not represented in this set. We therefore provide a set of 5510 representative genomes, which are available for bulk download (Figure 3). Additionally, habitat-specific subsets of representative genomes are also available.

Before selecting the representative genomes, we established a small 'whitelist' of genomes of special interest. This chiefly serves to ensure iconic model organism strains are guaranteed to be shown, even where automated measures might have indicated other strains as potential representatives. Users can vote on the website for additional genomes that should be included in this whitelist in future versions of the resource.

To compile these sets of representative genomes at least one genome per specI species cluster was selected. If a specI species cluster contained one or more genomes on the whitelist these were selected. Otherwise, we selected one representative genome from every non-singleton cluster using citation statistics (reflecting the use of a strain for experimental or other model system work) as well as genome quality statistics (N50), whereby completely assembled genomes were selected preferentially. Additionally, all genomes in singleton specI species clusters were included.

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. We focused on the functional annotation of protein-coding genes as they encode for most of the functional repertoire. This was achieved using the eggNOG (14) resource, as it provides a general annotation framework with a broad coverage of different protein functional categories. As mentioned above, proGenomes currently also provides focused annotations of antibiotic resistance and of carbohydrate active enzymes, with further annotations planned in future updates. Antibiotic resistance annotations are provided based on integrated results from the Comprehensive Antibiotic Resistance Database (CARD) (23) and ResFams (24) resources. For CARD, its associated resistance gene identifier tool was run on all proteins in proGenomes, with gene family assignments identified by sequence similarity using the curated CARD cutoffs and sequence (SNP) variation in antibiotic target genes identified using alignment to Hidden Markov Models (HMMs). For each proGenomes protein, the best hit above cutoff was retained in the case of resistance gene family annotation. Similarly, the set of SNPs to the best-scoring model above cutoff was retained for sequence variants. For proteins with no CARD resistance gene annotation, the best ResFams HMM hit above threshold was retained. Since

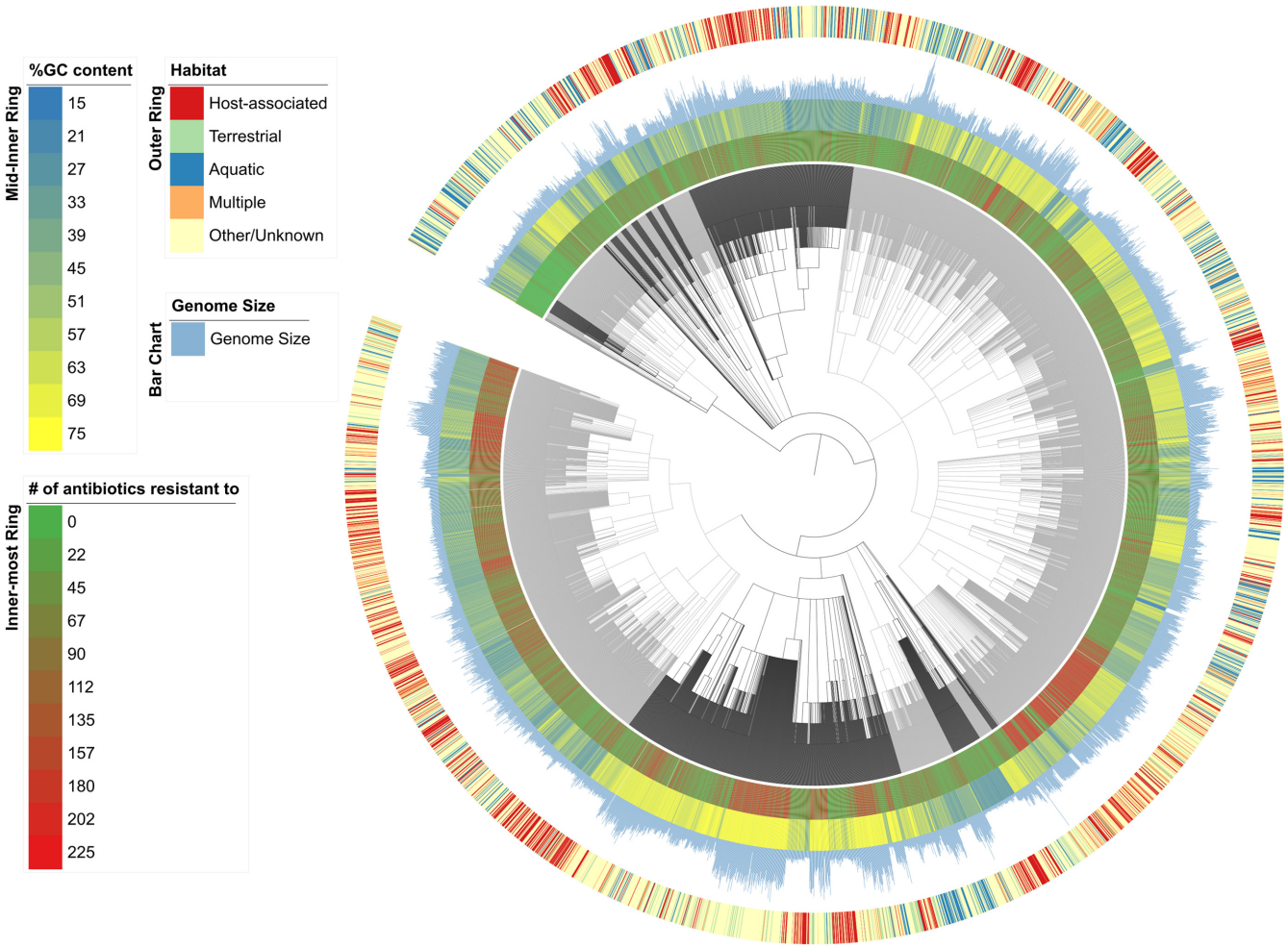

Figure 3. Overview of the representative genome set according to the NCBI Taxonomy. GC content, habitat information, genome size and antibiotic resistance gene carriage are displayed as additional datasets. Different Phyla are displayed as alternating light and dark gray clades within the tree (28).

both databases map to the antibiotic resistance ontology (ARO), the ARO hierarchy (as per CARD version 1.7) 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. Carbohydrate-active enzyme annotations as defined by CAZy (25) were generated using the dbCAN HMM models (26).

Overall, almost 80 Million protein-coding genes were annotated (eggNOG: 78 921 163; CAZy: 2 704 372; CARD + ResFams: 745 070). This information can be examined interactively on the proGenomes website.

Habitat information

Habitat information is provided for most genomes in the proGenomes database. This information can be utilized for in depth studies of selected environments or comparisons between different habitats such as (27). Habitat information was obtained from the manually curated PATRIC database (10) (accessed 15 March 2015). Specifically, the 'Host Name', 'Body Sample Site', and 'Habitat' fields were used. In cases where, for example, different assemblies for the same organism existed, data was collated for each NCBI Taxonomy ID. Information was available for 17 632 of the 25 038 organisms. This enabled us to broadly classify each specI species cluster into one of four different habitat types: host-associated (835), aquatic (566), terrestrial (234) or multiple (376) (Figures 1 and 3). Representative genomes for these subsets of clusters are available for bulk download from the website.

Website

The proGenomes website (http://progenomes.embl.de) can be used to browse the resource and enables direct access to the whole database. It has a searchable interface that can be used to find data from any taxonomic group or specI species cluster (Figure 4). All provided information can be explored interactively at the level of taxonomic groups or individual genomes. For larger taxonomic groups, information about

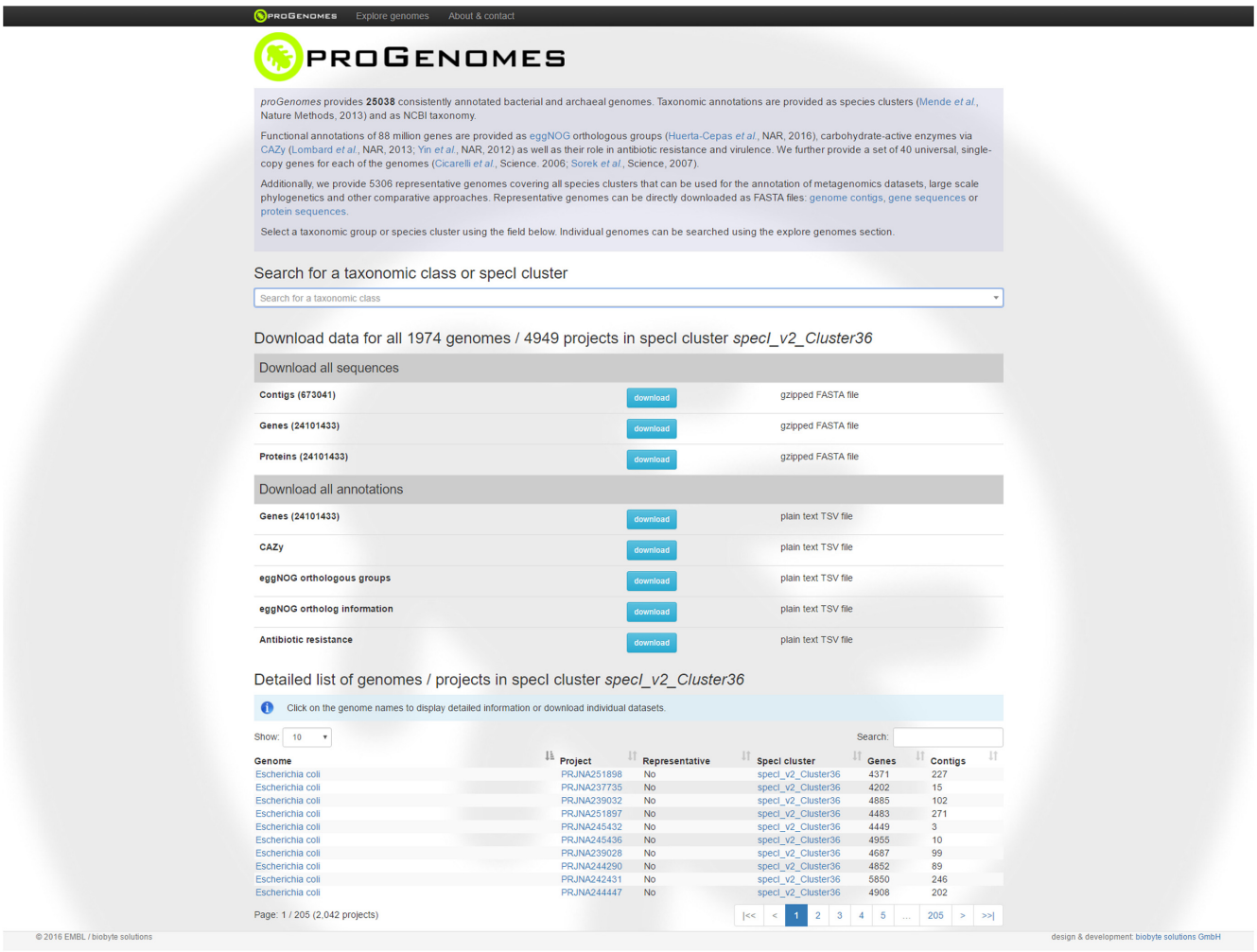

Figure 4. Clade/specI species cluster view on the proGenomes website. All sequences and annotations for the genomes within a clade/specI species cluster can be downloaded directly. Individual member genomes can be accessed at the bottom of the page.

all genomes within that group is displayed, with direct access to the genome, gene and protein sequences and annotations. For individual genomes, we provide all annotations in an interactive environment, which enables users to access additional information through direct links to relevant external database entries.

DISCUSSION

proGenomes provides consistent taxonomic and functional annotations for a large number of 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 Bork group, in particular Yan-Ping Yuan for technical support, as well as Lars Juhl Jensen for providing the initial whitelist. D.R.M. would further like to thank Sebastian Schmidt and Frank Aylward for helpful discussions.

FUNDING

European Commission MetaCardis project [FP7-HEALTH-305312]; International Human Microbiome Standards Consortium [HEALTH-FP7-2010-261376]; European Research Council CancerBiome project [268985]; GALAXY project [Project reference: 668031]; Novo Nordisk Foundation [NNF14CC0001]; European Union Horizon 2020 research and innovation programme [686070]; European Molecular Biology Laboratory (EMBL); Australian Postgraduate Award (to S.S.L.); EMBL Australia International Ph.D. Fellowship (to S.S.L.); European Molecular Biology Organization [ALTF 721–2015 to D.R.M]; LTFCOFUND2013 [PCOFUND-

GA-2013-609409 to D.R.M]. 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. 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. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B., Shukla,M. *et al.* (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.
5. Medini,D., Duccio,M., Davide,S., Julian,P., Relman,D.A., Claudio,D., Richard,M., Stanley,F. and Rino,R. (2008) Microbiology in the post-genomic era. *Nat. Rev. Microbiol.*, **6**, 419–430.
6. Tatusova,T., Ciufo,S., Federhen,S., Fedorov,B., McVeigh,R., O'Neill,K., Tolstoy,I. and Zaslavsky,L. (2014) Update on RefSeq microbial genomes resources. *Nucleic Acids Res.*, **43**, D599–D605.
7. Markowitz,V.M., Chen,I.-M.A., Palaniappan,K., Chu,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Woyke,T., Huntemann,M. *et al.* (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.*, **42**, D560–D567.
8. 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.
9. 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.
10. Rosselló-Mora,R. and Amann,R. (2001) The species concept for prokaryotes. *FEMS Microbiol. Rev.*, **25**, 39–67.
11. Davis,J.J., Boisvert,S., Brettin,T., Kenyon,R.W., Mao,C., Olson,R., Overbeek,R., Santerre,J., Shukla,M., Wattam,A.R. *et al.* (2016) Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci. Rep.*, **6**, 27930.
12. Kanehisa,M., Minoru,K., Yoko,S., Masayuki,K., Miho,F. and Mao,T. (2015) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
13. Mende,D.R., Sunagawa,S., Zeller,G. and Bork,P. (2013) Accurate and universal delineation of prokaryotic species. *Nat. Methods*, **10**, 881–884.
14. Huerta-Cepas,J., Jaime,H.-C., Damian,S., Kristoffer,F., Helen,C., Davide,H., Walter,M.C., Thomas,R., Mende,D.R., Shinichi,S. *et al.* (2015) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286–D293.
15. Huerta-Cepas,J., Forslund,K., Szklarczyk,D., Jensen,L.J., von Mering,C. and Bork,P. (2016) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *bioRxiv*, **76331**, 1–11.
16. Borodovsky,M., Mark,B. and Alex,L. (2014) Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Curr. Protoc. Microbiol.*, **32**, doi:10.1002/0471250953.bi0405s35.
17. 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.
18. 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.
19. Huang,K., Brady,A., Mahurkar,A., White,O., Gevers,D., Huttenhower,C. and Segata,N. (2014) MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Res.*, **42**, D617–D624.
20. Minguez,P., Letunic,I., Parca,L., Garcia-Alonso,L., Dopazo,J., Huerta-Cepas,J. and Bork,P. (2015) PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic Acids Res.*, **43**, D494–D502.
21. Sunagawa,S., Mende,D.R., Zeller,G., Izquierdo-Carrasco,F., Berger,S.A., Kultima,J.R., Coelho,L.P., Arumugam,M., Tap,J., Nielsen,H.B. *et al.* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods*, **10**, 1196–1199.
22. 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.
23. McArthur,A.G., Waglechner,N., Nizam,F., Yan,A., Azad,M.A., Baylay,A.J., Bhullar,K., Canova,M.J., De Pascale,G., Ejim,L. *et al.* (2013) The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.*, **57**, 3348–3357.
24. 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.
25. Lombard,V., Vincent,L., Ramulu,H.G., Elodie,D., Coutinho,P.M. and Bernard,H. (2013) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**, D490–D495.
26. Yin,Y., Mao,X., Yang,J., Chen,X., Mao,F. and Xu,Y. (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.*, **40**, W445–W451.
27. Smillie,C.S., Smith,M.B., Friedman,J., Cordero,O.X., David,L.A. and Alm,E.J. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, **480**, 241–244.
28. Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.