Published online 12 November 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D309–D314***
doi: 10.1093/nar/gky1085

eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses

Jaime Huerta-Cepas[1,2,*,†], Damian Szklarczyk[3,†], Davide Heller[3], Ana Hernández-Plaza[2], Sofia K. Forslund[1,4], Helen Cook[5], Daniel R. Mende[6], Ivica Letunic[7], Thomas Rattei[8], Lars J. Jensen[5], Christian von Mering[3] and Peer Bork[1,9,10,11,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [2]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain, [3]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland, [4]Experimental and Clinical Research Center, a cooperation of Charité-Universitätsmedizin Berlin and Max Delbruck Center for Molecular Medicine, 13125 Berlin, Germany, [5]The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen N 2200, Denmark, [6]Daniel K. Inouye Center for Microbial Oceanography: Research and Education (C-MORE), University of Hawaii, Honolulu, HI 96822, USA, [7]Biobyte solutions GmbH, Bothestr 142, 69126 Heidelberg, Germany, [8]CUBE-Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna 1090, Austria, [9]Germany Molecular Medicine Partnership Unit (MMPU), University Hospital Heidelberg and European Molecular Biology Laboratory, Heidelberg, Germany, [10]Max Delbrück Centre for Molecular Medicine, Berlin, Germany and [11]Department of Bioinformatics, Biocenter University of Würzburg, Würzburg, Germany

Received September 15, 2018; Editorial Decision October 10, 2018; Accepted October 26, 2018

ABSTRACT

eggNOG is a public database of orthology relationships, gene evolutionary histories and functional annotations. Here, we present version 5.0, featuring a major update of the underlying genome sets, which have been expanded to 4445 representative bacteria and 168 archaea derived from 25 038 genomes, as well as 477 eukaryotic organisms and 2502 viral proteomes that were selected for diversity and filtered by genome quality. In total, 4.4M orthologous groups (OGs) distributed across 379 taxonomic levels were computed together with their associated sequence alignments, phylogenies, HMM models and functional descriptors. Precomputed evolutionary analysis provides fine-grained resolution of duplication/speciation events within each OG. Our benchmarks show that, despite doubling the amount of genomes, the quality of orthology assignments and functional annotations (80% coverage) has persisted without significant changes across this update. Finally, we improved eggNOG online services for fast functional annotation and orthology prediction of custom genomics or metagenomics datasets. All precomputed data are publicly available for downloading or via API queries at http://eggnog.embl.de

INTRODUCTION

Identifying orthologs, those sequences diverging from a common ancestry after a speciation event, constitutes a fundamental task in molecular and evolutionary biology. Compared to paralogs, which are sequences diverged after a duplication event, orthologs are more prone to retain their ancestral function (1,2), even at long evolutionary timescales (3). Therefore, differentiating between these two subtypes of homology relationships is crucial to produce accurate functional predictions (2,4,5). It is also essential for proper analysis in, for example, phylogenetics and comparative genomics (6) or the study of cell-type evolution (7). Hence, several databases have been developed over the years that provide precomputed orthology predictions using different approaches and operational definitions (8–13). Most of those resources, including eggNOG, are part of the international consortium Quest for Orthologs (14), were standard-

*To whom correspondence should be addressed. Tel: +34 913364556; Email: j.huerta@upm.es
Correspondence may also be addressed to Peer Bork. Tel: +49 6221 387 85 26; Email: bork@embl.de
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

ized benchmarking approaches (15) and reference datasets are developed and shared.

eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) is a public resource in which thousands of genomes are analyzed at once to establish orthology relationships between all their genes. Compared to similar databases, eggNOG focuses on providing: (i) comprehensive functional annotations for the inferred orthologs, (ii) predictions across thousands of genomes covering the three domains of life and viruses, and iii) hierarchical resolution of orthology assignments and fine-grained relationships (i.e. in-paralogies) based on phylogenetic analysis. For that, a species-aware clustering algorithm based on the concept of triangulation of best reciprocal hits (16) is applied to identify Orthologous Groups (OGs): sets of homologous sequences that started diverging from the same speciation event. As orthology relationships vary depending on the assumed reference speciation event (outgroup)—with increasing resolution toward the tips of the tree of life—since its inception in 2008 (17), eggNOG computes orthology predictions at different taxonomic levels. All OGs from all taxonomic levels are then functionally annotated and analyzed using phylogenetic methods, which allows users to further explore the history of speciation and duplication events within each OG, infer pairwise orthology relationships between specific species, or trace functional changes therein.

Here, we describe eggNOG v5.0, including the following improvements over previous versions: (i) a major upgrade of the underlying databases, featuring one of the most comprehensive selection of prokaryotic, eukaryotic and viral genomes available; (ii) updates in the online service for custom (meta-)genome annotation, now including options for fast orthology prediction and improved computational power via cloud computing and (iii) better visualization options of OGs and their associated functional data.

UPDATES AND ADDITIONS SINCE PREVIOUS RELEASE

Genomes update

eggNOG 5.0 has increased the number of genomes used for inferring orthology from 2031 core organisms to 5090**.** Viral proteomes have also been upgraded, increasing from 352 to 2502 proteomes collected from Uniprot and filtered by completeness (those with less than three proteins after in silico cleaving of polyproteins were discarded). In order to select best representative prokaryotic genomes, we used the SpecI species delineation method (18) against a total set of 25 038 genomes retrieved from RefSeq (19), obtaining 4445 reference species. Similarly, 477 eukaryotic genomes were collected from Ensembl (11) and other project-oriented resources (see online methods at http://eggnog.embl.de/). In all cases, genomes and proteomes were standardized and checked for completeness and minimum quality before inclusion into the database. For instance, incomplete prokaryotic genomes missing more than 4 out of 40 universal, single copy, marker genes (20) were excluded, as well as genomes that could not be assembled to fewer than 300 contigs or genomes with an N50 of <10 000.

Taxonomic levels and non-supervised Orthologous Groups

An Orthologous Group (OG) is defined as a cluster of three or more homologous sequences that diverge from the same speciation event (16,17). Different OGs could therefore be inferred depending on the speciation split considered, that is, implicitly, the taxonomic resolution one considers. Older speciation events lead to larger OGs with more in-paralogs (duplication events occurred after the speciation) and higher functional divergence among their members. By contrast, recent speciations lead to smaller and usually more functionally specific sets of orthologs. For example, this implies that vertebrate-specific OGs would yield more fine-grained functional differentiation than OGs built using all eukaryotic species.

In order to better reflect this taxonomic range and improve the precision of eggNOG functional predictions, in this version we have largely increased the number of predefined taxonomic levels (speciation splits) for which OGs are independently computed. In total, we applied the non-supervised eggNOG clustering method described in Jensen *et al.* (17) on 379 taxonomic levels, leading to 4.4M OGs (compared to 107 levels and 1.9M OGs in the previous version (21)). OGs were built using best reciprocal hits information derived from an all-against-all Smith-Waterman matrix provided by the SIMAP project (22). In addition, manually curated OGs available for the three domains of life were integrated into the corresponding levels in eggNOG, namely bacterial subset of COGs (23), archaeal arCOGs (24) and eukaryotic KOGs (25). Similarly, viral OGs were updated using deeper taxonomic categories, now descending to the family level. The taxonomic distribution in eggNOG v5.0, as well as the number of organisms, OGs inferred, and functional annotation coverage per level is shown in Figure 1.

Hierarchical consistency of OGs

Relationships between more rootward OGs and their nested children OGs at more specific taxonomic levels were explicitly tracked and ascertained to be consistent, with exceptions only for mosic proteins with multi-domain combinations, where individual domains might have evolved independently (26,27). Hierarchical inconsistencies are the inevitable product of executing eggNOG's clustering algorithm independently at each taxonomic level. Given that the set of species vary at each level, nested OGs might describe slightly incompatible evolutionarily histories for the same set of proteins. Solving those cases is particularly important for third-party applications (e.g. STRING (28)), in which information needs to be propagated across the hierarchy of taxonomic levels. Therefore, from version 4.5, we apply a post-processing step to ensure hierarchical consistency of all nested OGs.

In this database update, we have improved our methodology by implementing a more accurate strategy based on gene-tree reconciliation. Briefly, for each hierarchical inconsistency found, we subsample the proteins spanning the affected OGs and perform gene-tree to species-tree reconciliation. Each reconciled tree sample represents a vote towards one of the conflicting evolutionary hypotheses. We combine the reconciliations by majority voting to decide

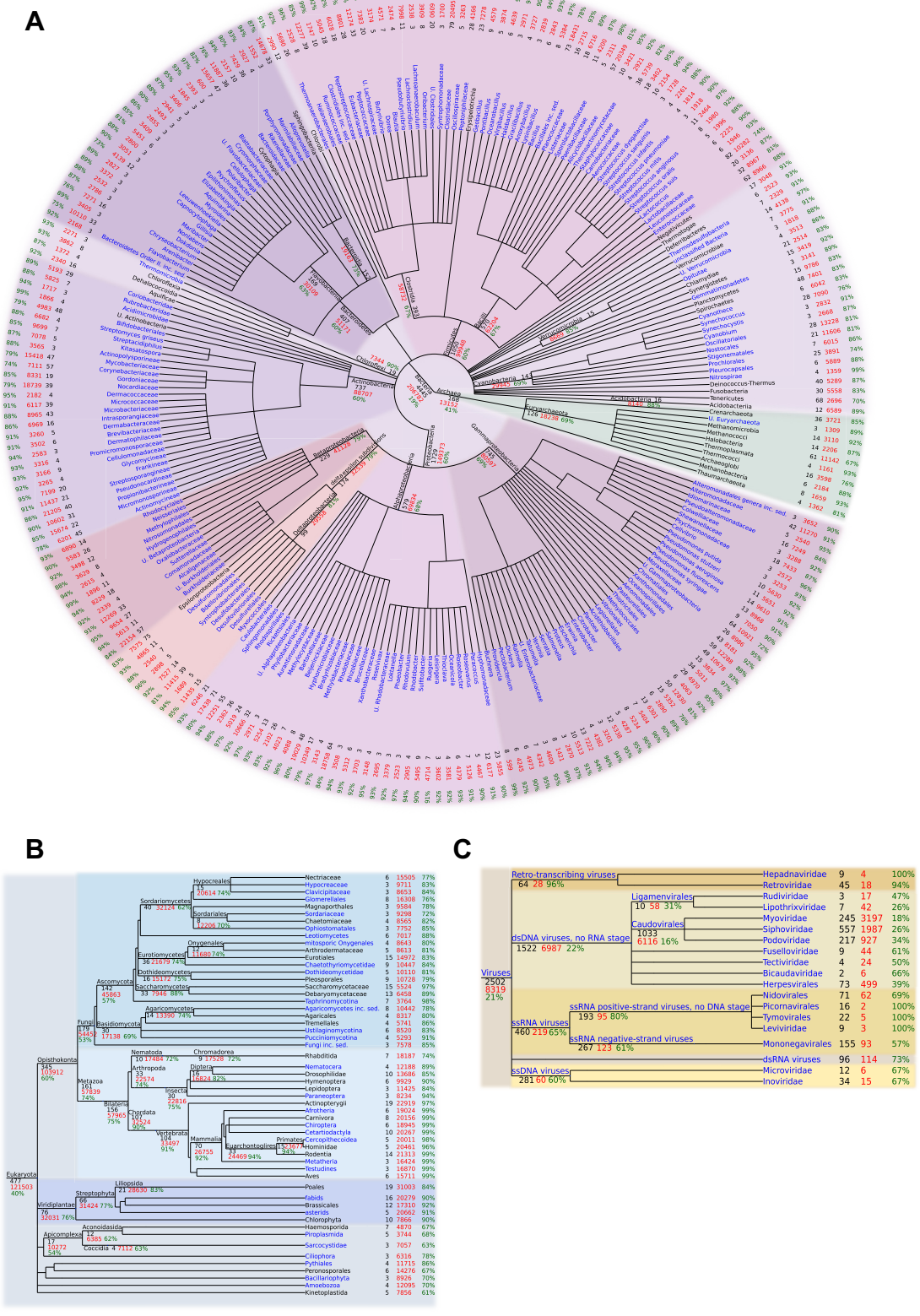

Figure 1. Taxonomic levels for which OGs have been independently computed based on (**A**) prokaryotic, (**B**) eukaryotic and (**C**) viral genomes. Names in blue indicate new taxonomic levels with respect to previous eggNOG versions. Numbers indicate the the amount of OGs per level (red), number of species covered (black) and functional annotation coverage (green).

how to resolve the inconsistency. Given the large number of species in this version of eggNOG, we have however retained some size control heuristics, such as the rule that COGs should not be merged. A full description of the reconciliation method is available at https://github.com/meringlab/og_consistency_pipeline.

Phylogenetics analysis

As in previous releases, all OGs in eggNOG v5.0 were analyzed using a comprehensive phylogenetic approach. Based on recent benchmarks (29), we adapted our phylogenomic strategy to the following steps: multiple sequence alignments inferred with Clustal Omega (30), soft alignment trimming by removing columns with less than five aligned residues, model testing using ModelFinder (31), maximum likelihood trees computed with IqTree (32) and branch supports calculated using the ultrafast bootstrap method (33). The full workflow was executed using the ETE toolkit v3.1.1 (34), which integrates the complete pipeline as a built in gene-tree workflow (code name 'eggnog50_full'). For ∼57 000 OGs**,** due to the increasing gene family sizes, computation was not possible in this pipeline, so a fall-back method was used where IqTree was executed with the less-sensitive option '–fast'. All 4.4M trees were analyzed to infer speciation and duplication events (i.e. in-paralogy relationships) using the species overlap algorithm described in (35), leading to pairwise orthology tables (differentiating one-to-one versus many-to-many relationships) for each OG.

Functional annotations

Orthologous Groups were functionally annotated using updated versions of Gene Ontology (36), KEGG pathways (37), SMART/PFAM domains (38) and expanded to CAZy (39) and KEGG modules. Moreover, general free text descriptions and COG functional categories were updated for each OG using the automated text-mining and machine learning-based pipeline described in (21). In short, OGs were assigned text descriptions based on a heuristic to find the most informative text substring from either names of assigned SMART domains, assigned Gene Ontology terms, or common substrings in free text annotations from the source gene databases. In total, 80% of all OGs were annotated using at least one functional source. Finally, we improved the online visualization of functional annotations, which can now be explored from an evolutionary point of view by plotting functional descriptors together with the phylogenetic tree and the duplication/speciation events inferred for each OG (Figure 2).

Fast functional and orthology assignments for custom user data

eggNOG v5.0 has also improved the underlying precomputed data used by the online version of eggNOG-mapper (40), a tool for rapid annotation of custom (meta-)genomes. Moreover, our online services are now cloud-enabled, permitting intensive computations required by functional annotation of massive datasets to run on dedicated servers with hundreds of CPUs available. We have also introduced a new option for fast batch orthology assignments of custom sets of sequences, which allows users to assign orthology relationships between novel genes and all genomes represented in eggNOG.

BENCHMARK

The average quality of orthology predictions and functional annotations was benchmarked in order to estimate the effect of adding novel genomes. Both orthobench2 (41) and the Quest For Orthologs (QFO) benchmark (15) were used. Compared to eggNOG v4.5, we improved the performance in the orthobench's Bilaterian (from 72.1% to 73.1% *F*-measure) and Gammaproteobacteria test (from 93.2% to 94.7% *F*-measure). On the other hand, the QFO benchmark allowed us to evaluate the performance of both OG-based predictions and fine-grained predictions. Results show a clear tradeoff in the precision-recall ratio depending on the strategy selected, which in turn reflects different use cases of orthology assignments. OG-based predictions produced results with high recall values, predicting more than twice the number of orthologous pairs with <10.6% drop in average Schlicker similarity compared to the benchmark average in the Enzyme Classification and Gene Onthology Conservation tests. This high recall pattern is in general preferred by probabilistic prediction methods such as interolog inference in the STRING database (28). By contrast, fine-grained predictions showed higher precision values, while maintaining a similar recall as the previous EggNOG versions, which is usually preferred for accurate functional transfers. In general, for the majority of QFO benchmark tests, the performance of eggNOG 5.0 was slightly better or stayed at the Pareto line compared to previous eggNOG version (detailed plots and results are available at http://orthology.benchmarkservice.org). Taken together, this indicates that the large increase of genomes had no major impact on the quality of the inferred orthologous groups, suggesting the eggNOG approach continues to scale well.

CONCLUSIONS AND PERSPECTIVES

By further streamlining and modernizing the automated approach for the construction of eggNOG orthologous groups, as well as synchronizing with improved or newly developed source databases (e.g. proGenomes for the classification of high quality prokaryotic genomes, (42)), we have been able to more than double core genome coverage for eggNOG, including extensive expansion of viral gene families, largely without loss of quality of orthology reconstruction or functional annotation. Due to a supervised increase of pre-defined taxonomic levels as basis for OG calculation, we almost tripled to number of OGs to 4.4M. Version 5 of eggNOG should thus be a useful resource for ecological, evolutionary or medical -omics analysis, also serving as an entry point for fast functional annotation of newly sequenced genes, genomes and metagenomes. We are currently working on conceptual and algorithmic improvements to be able to continue to keep pace with a vastly expanding number of organisms and meta-genomes sequenced.

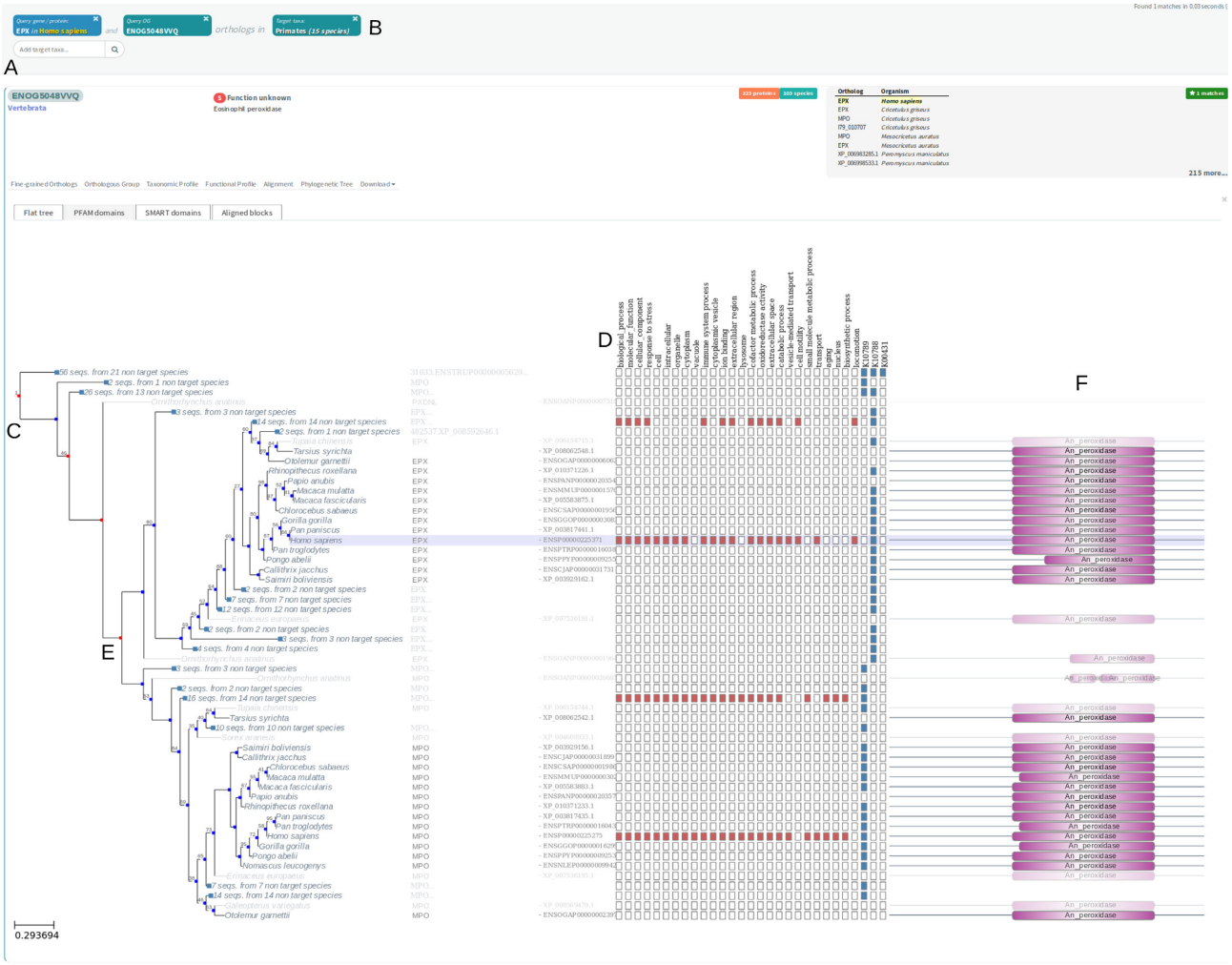

Figure 2. Visualization of the phylogeny associated to the OG ENOG5048VVQ at the vertebrate level (**A**) extracted from the eggNOG website. Target orthologs were restricted to primates in the phylogenetic tree to facilitate exploration (**B**). Duplication nodes (in-paralogies) are labeled in red, and speciation events in blue (**C**). The functional profile of each orthologous sequence is shown in the presence/absence matrix (**D**). Functional differences can be noticed at both sides of the duplication event separating EPX from MPO sequences (**E**) in both GO Slim terms (red squares in matrix D) and KEGG Modules (blue squares in matrix D), while having similar domain architectures (**F**).

FUNDING

European Union's Horizon 2020 research and innovation programme [686070]; Ramón y Cajal Programme [RYC-2016-20621]; Consejería de Educación, Juventud y Deporte de la Comunidad de Madrid and Fondo Social Europeo [PEJ-2017-AI/TIC-7514]; FP7 METACARDIS [HEALTH-F4–2012-305312]; Novo Nordisk Foundation (Copenhagen) [NNF14CC0001]; European Molecular Biology Laboratory; Cloud computing was supported by BMBF [de.NBI network #031A537B]; Swiss National Science Foundation [31003A-160095]. Funding for open access charge: European Molecular Biology Laboratory.
Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Fitch,W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
3. Kachroo,A.H., Laurent,J.M., Yellman,C.M., Meyer,A.G., Wilke,C.O. and Marcotte,E.M. (2015) Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, **348**, 921–925.
4. Zhang,J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.*, **18**, 292–298.
5. Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.
6. Moreira,D. and Philippe,H. (2000) Molecular phylogeny: pitfalls and progress. *Int. Microbiol.*, **3**, 9–16.
7. Arendt,D. (2008) The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.*, **9**, 868–882.
8. Altenhoff,A.M., Glover,N.M., Train,C.-M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., de Farias,T.M., Zile,K., Stevenson,C., Long,J. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.
9. Huerta-Cepas,J., Capella-Gutierrez,S., Pryszcz,L.P., Denisov,I., Kormes,D., Marcet-Houben,M. and Gabaldón,T. (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.*, **39**, D556–D560.

10. Sonnhammer,E.L.L. and Östlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–D239.
11. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
12. Mi,H., Huang,X., Muruganujan,A., Tang,H., Mills,C., Kang,D. and Thomas,P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.
13. Van Bel,M., Diels,T., Vancaester,E., Kreft,L., Botzki,A., Van de Peer,Y., Coppens,F. and Vandepoele,K. (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.*, **46**, D1190–D1196.
14. Forslund,K., Pereira,C., Capella-Gutierrez,S., Sousa da Silva,A., Altenhoff,A., Huerta-Cepas,J., Muffato,M., Patricio,M., Vandepoele,K., Ebersberger,I. *et al.* (2017) Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics*, **34**, 323–329.
15. Altenhoff,A.M., Boeckmann,B., Capella-Gutierrez,S., Dalquen,D.A., DeLuca,T., Forslund,K., Huerta-Cepas,J., Linard,B., Pereira,C., Pryszcz,L.P. *et al.* (2016) Standardized benchmarking in the quest for orthologs. *Nat. Methods*, **13**, 425–430.
16. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
17. Jensen,L.J., Julien,P., Kuhn,M., von Mering,C., Muller,J., Doerks,T. and Bork,P. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, **36**, D250–D254.
18. Mende,D.R., Sunagawa,S., Zeller,G. and Bork,P. (2013) Accurate and universal delineation of prokaryotic species. *Nat. Methods*, **10**, 881–884.
19. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
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. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M. *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286–D293.
22. Arnold,R., Goldenberg,F., Mewes,H.W. and Rattei,T. (2014) SIMAP - The database of all-against-all protein sequence similarities and annotations with new interfaces and increased coverage. *Nucleic Acids Res.*, **42**, D279–D284
23. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
24. Makarova,K., Wolf,Y. and Koonin,E. (2015) Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. *Life*, **5**, 818–840.
25. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
26. Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences—where are the bottlenecks? *Nat. Genet.*, **18**, 313.
27. Sjölander,K., Datta,R.S., Shen,Y. and Shoffner,G.M. (2011) Ortholog identification in the presence of domain architecture rearrangement. *Brief. Bioinform.*, **12**, 413–422.
28. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. *et al.* (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
29. Zhou,X., Shen,X., Hittinger,C.T. and Rokas,A. (2017) Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol. Biol. Evol.*, **35**, 486–503.
30. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539
31. Kalyaanamoorthy,S., Minh,B.Q., Wong,T.K.F., von Haeseler,A. and Jermiin,L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, **14**, 587–589.
32. Nguyen,L.-T., Schmidt,H.A., von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
33. Minh,B.Q., Nguyen,M.A.T. and von Haeseler,A. (2013) Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.*, **30**, 1188–1195.
34. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.
35. Huerta-Cepas,J., Dopazo,H., Dopazo,J. and Gabaldón,T. (2007) The human phylome. *Genome Biol.*, **8**, R109.
36. The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
37. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
38. Letunic,I. and Bork,P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.*, **46**, D493–D496.
39. Levasseur,A., Drula,E., Lombard,V., Coutinho,P.M. and Henrissat,B. (2013) Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol. Biofuels*, **6**, 41.
40. Huerta-Cepas,J., Forslund,K., Pedro Coelho,L., Szklarczyk,D., Juhl Jensen,L., von Mering,C. and Bork,P. (2016) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol,* **34**, 2115–2122.
41. Trachana,K., Forslund,K., Larsson,T., Powell,S., Doerks,T., von Mering,C. and Bork,P. (2014) A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS One*, **9**, e111122.
42. 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.