D284–D289 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 16 November 2011*
doi:10.1093/nar/gkr1060

eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges

Sean Powell[1], Damian Szklarczyk[2], Kalliopi Trachana[1], Alexander Roth[3], Michael Kuhn[4], Jean Muller[5,6], Roland Arnold[7], Thomas Rattei[8], Ivica Letunic[1], Tobias Doerks[1], Lars J. Jensen[2,*], Christian von Mering[3,*] and Peer Bork[1,9,*]

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark, [3]University of Zurich and Swiss Institute of Bioinformatics, Winterthurerstrasse 190, 8057 Zurich, Switzerland, [4]Biotechnology Center, TU Dresden, 01062 Dresden, Germany, [5]Institute of Genetics and Molecular and Cellular Biology, CNRS, INSERM, University of Strasbourg, [6]Genetic Diagnostics Laboratory, CHU Strasbourg Nouvel Hôpital Civil, Strasbourg, France, [7]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario M55 3E1, Canada, [8]University of Vienna, Department of Computational Systems Biology, Althanstrasse 14, 1090 Vienna, Austria and [9]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, 13092 Berlin, Germany

Received September 15, 2011; Revised and Accepted October 26, 2011

ABSTRACT

Orthologous relationships form the basis of most comparative genomic and metagenomic studies and are essential for proper phylogenetic and functional analyses. The third version of the eggNOG database (http://eggnog.embl.de) contains non-supervised orthologous groups constructed from 1133 organisms, doubling the number of genes with orthology assignment compared to eggNOG v2. The new release is the result of a number of improvements and expansions: (i) the underlying homology searches are now based on the SIMAP database; (ii) the orthologous groups have been extended to 41 levels of selected taxonomic ranges enabling much more fine-grained orthology assignments; and (iii) the newly designed web page is considerably faster with more functionality. In total, eggNOG v3 contains 721 801 orthologous groups, encompassing a total of 4 396 591 genes. Additionally, we updated 4873 and 4850 original COGs and KOGs, respectively, to include all 1133 organisms. At the universal level, covering all three domains of life, 101 208 orthologous groups are available, while the others are applicable at 40 more limited taxonomic ranges. Each group is amended by multiple sequence alignments and maximum-likelihood trees and broad functional descriptions are provided for 450 904 orthologous groups (62.5%).

INTRODUCTION

Orthology, defined as homology via speciation (1), is a crucial concept in evolutionary biology and is essential for disciplines such as comparative genomics, metagenomics and phylogenomics. The concepts of orthology and paralogy, with the latter being defined as homology via duplication (1), have been used as a foundation to introduce the concept of clusters of orthologous groups: proteins that have evolved from a single ancestral sequence existing in the last common ancestor (LCA) of the species that are being compared, through a series of speciation and duplication events (2). Orthologous groups (OGs) have proven useful for functional analyses and the annotation of newly sequenced genomes (3–5) as orthologs tend to have equivalent functions (6).

A number of orthology prediction methods have been recently introduced that can be classified into (i) graph-based methods, from the reciprocal-best-hit approach (7) to more sophisticated methods, such as the identification of best-hit triangles (2,8–11) and other clustering-based approaches (12–15) or (ii) tree-based methods that can be further classified into methods that use tree reconciliation to infer orthologs (16–19) and those that do not

*To whom correspondence should be addressed. Tel: +49 6221 387 8361; Fax: +49 6221 387 8517; Email: bork@embl.de
Correspondence may also be addressed to Lars J. Jensen. Tel: +45 35 32 50 25; Fax: +45 35 32 50 01; Email: lars.juhl.jensen@cpr.ku.dk
Correspondence may also be addressed to Christian von Mering. Tel: +41 44 6353147; Fax: +41 44 6356864; Email: mering@imls.uzh.ch

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

(20,21). Their methodological advantages and disadvantages have been reviewed in refs (22–24).

An important point is that OGs depend on their taxonomic context. The broader the taxonomic range, the deeper the LCA is placed, resulting in larger OGs with lower resolution of the orthologous relationships. Thus, the smaller taxonomical range results in more fine-grained groups. Therefore, the first and most successful resource, COG (2), provided OGs for certain taxonomic ranges, namely COGs for all three domains of life, KOGs for Eukaryotes (8) and arCOGs for Archaea (9). Some automatic orthology prediction methods also provide distinct sets of OGs for an increasing number of taxonomic groups [e.g. OrthoDB (10), eggNOG (11) and OMA (12)].

The functional annotation of OGs is particularly necessary, as functional insights from well-studied proteins/species can be transferred to uncharacterized orthologs. Moreover, several genome annotation tools [e.g. (25)] use the functional annotations of OGs to automatically map function information to large-scale genomic data. The most common form of orthologous group annotation is a consensus-based (longest common string) approach (9,12,18,21,26) in which the description of the OG is derived from available annotations of the member proteins. Only a few available resources conduct a more robust manual annotation of the groups (8) or incorporate multiple annotation sources for the description and annotate the groups with functional categories (8,11).

Here, we describe the third version of eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups), a database that provides orthologous groups for 943 Bacteria, 69 Archaea and 121 Eukaryotes. In total, 721 801 OGs have been computed including about twice as many orthologous relations for genes compared to the previous version. Most importantly, it contains considerably more taxonomically restricted OGs with higher resolution, covering 41 taxonomically relevant ranges such as Proteobacteria or Metazoans.

SELECTION OF GENOMES

We downloaded complete proteomes from RefSeq (27), Ensembl (28), UniProt (29), GiardiaDB (30), JGI (http://genome.jgi-psf.org/) and TAIR (31). This particular set of genomes also forms the basis for the most recent STRING (32) and STITCH (33) database, allowing for easy integration across these databases.

The analyses were performed on 1133 complete genomes, encoding 5 214 234 proteins. The genomes were selected based on pertinence and quality. Except for the many model organisms that were included in the database, the species were selected based on their taxonomic position to ensure a dense sampling of 41 selected taxonomical ranges (see below) as well as a broad coverage of the tree of life. As genome quality significantly affects the accuracy of orthology assignment (34,35) all genomes in eggNOG v3 were manually selected for genomic quality based on sequencing coverage and genome completeness judged by the coverage of 40 phylogenetic marker genes (36,37).

CONSTRUCTION OF ORTHOLOGOUS GROUPS AT DIFFERENT TAXONOMIC LEVELS

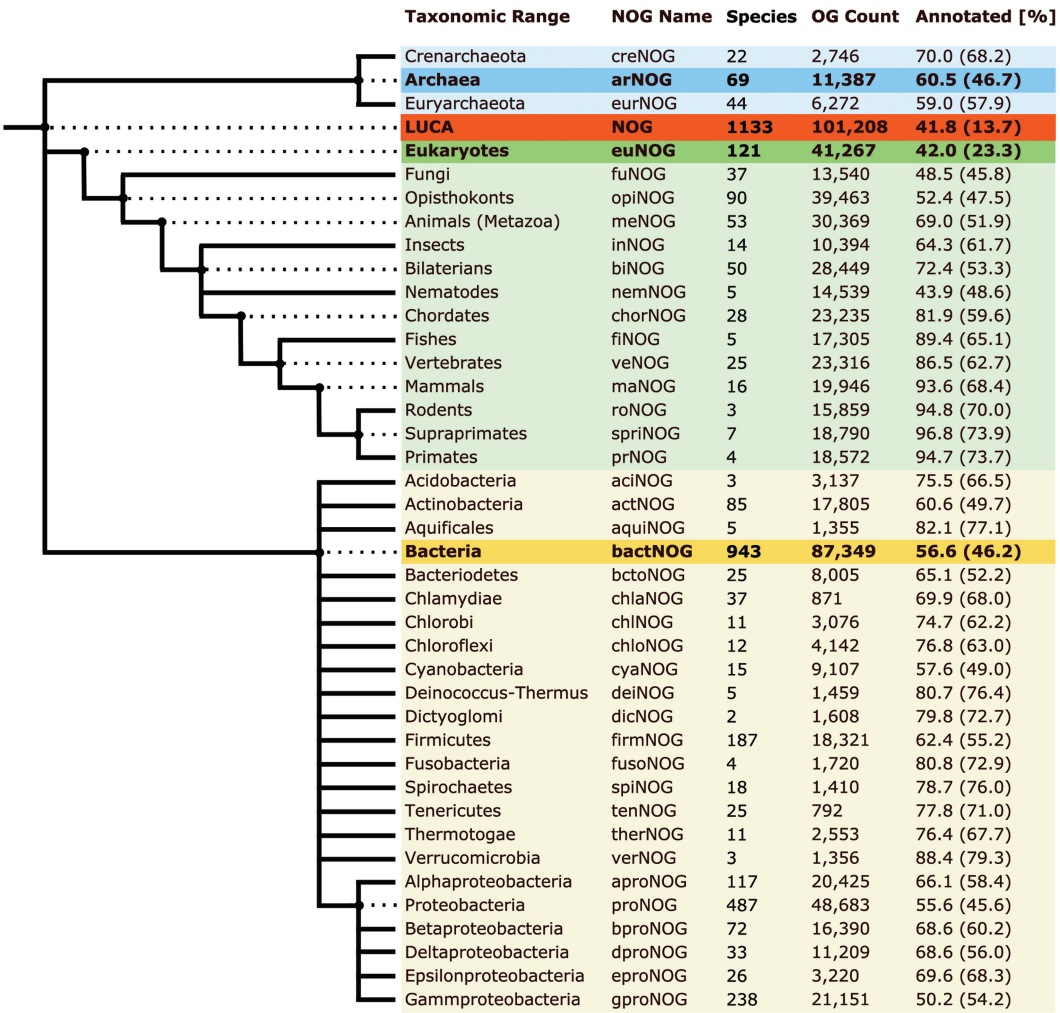
The first step of the eggNOG pipeline is an all-against-all similarity search. Due to the quadratic escalation of computational power necessary for such an all-against-all search, eggNOG v3 now uses the SIMAP database (38) for the required homology comparisons. SIMAP uses the FASTA heuristics (39), which are better at capturing sequences with a lower degree of similarity than BLAST (40), which was previously used in eggNOG, at the cost of reduced performance.

After the homology searches and the subsequent clustering step (11), 4 396 591 (84%) of all proteins investigated were assigned to at least one of the 721 801 orthologous groups generated by eggNOG (Figure 1). We extended the COGs, KOGs and arCOGs (8,9) to include the 1133 organisms, 121 eukaryotic and 69 archaeal species, respectively. As an enhancement to the 4873 COGs, 4850 KOGs and 7538 arCOGs, additional groups have been created as non-supervised OGs (NOGs), eukaryote-specific NOGs (euNOGs) and archaea-specific NOGs (arNOGs), extending those original COGs/KOGs/arNOGs by 101 208 NOGs, 41 267 euNOGs and 11 387 arNOGs. To provide a higher resolution of orthologous groups in frequently used taxonomic ranks, we applied our procedure to several subsets of organisms separately. Apart from the level of Eukaryotes (euNOGs) and Archaea (arNOGs), to provide information for all three domains of life, we provide newly derived bacteria-specific NOGs (bactNOGs). Subsequently, the orthology for 22 bacterial levels such as Firmicutes (firmNOGS), Proteobacteria (proNOGs) and Actinobacteria (actNOGs) (Figure 1) is further resolved, as well as for 14 major levels in the eukaryotic clade including Animals (meNOGs) and Fungi (fuNOGs).

AUTOMATED ANNOTATION OF PROTEIN FUNCTION

An important feature of eggNOG v3 is the automatic functional annotation of the OGs. The groups are annotated with a function description based on the functional annotations of each protein member within the group (26) and in parallel with one of 25 functional categories (11) compatible with those provided by the COG and KOG databases (8).

In eggNOG v3, the functional annotation pipeline has similarly been optimized to scale to the large amount of data. This has led to a significant improvement in computation time while simultaneously increasing the total number of functionally annotated OGs. Between eggNOG v2 and eggNOG v3, for corresponding taxonomic levels, the total number of annotated OGs increased by 28.8% and 10.0% for function description and functional category, respectively. In summary, of the 721 801 OGs in eggNOG v3, 62.5% have a functional annotation and



Taxonomic Range	NOG Name	Species	OG Count	Annotated [%]
Crenarchaeota	creNOG	22	2,746	70.0 (68.2)
Archaea	**arNOG**	**69**	**11,387**	**60.5 (46.7)**
Euryarchaeota	eurNOG	44	6,272	59.0 (57.9)
LUCA	**NOG**	**1133**	**101,208**	**41.8 (13.7)**
Eukaryotes	**euNOG**	**121**	**41,267**	**42.0 (23.3)**
Fungi	fuNOG	37	13,540	48.5 (45.8)
Opisthokonts	opiNOG	90	39,463	52.4 (47.5)
Animals (Metazoa)	meNOG	53	30,369	69.0 (51.9)
Insects	inNOG	14	10,394	64.3 (61.7)
Bilaterians	biNOG	50	28,449	72.4 (53.3)
Nematodes	nemNOG	5	14,539	43.9 (48.6)
Chordates	chorNOG	28	23,235	81.9 (59.6)
Fishes	fiNOG	5	17,305	89.4 (65.1)
Vertebrates	veNOG	25	23,316	86.5 (62.7)
Mammals	maNOG	16	19,946	93.6 (68.4)
Rodents	roNOG	3	15,859	94.8 (70.0)
Supraprimates	spriNOG	7	18,790	96.8 (73.9)
Primates	prNOG	4	18,572	94.7 (73.7)
Acidobacteria	aciNOG	3	3,137	75.5 (66.5)
Actinobacteria	actNOG	85	17,805	60.6 (49.7)
Aquificales	aquiNOG	5	1,355	82.1 (77.1)
Bacteria	**bactNOG**	**943**	**87,349**	**56.6 (46.2)**
Bacteriodetes	bctoNOG	25	8,005	65.1 (52.2)
Chlamydiae	chlaNOG	37	871	69.9 (68.0)
Chlorobi	chlNOG	11	3,076	74.7 (62.2)
Chloroflexi	chloNOG	12	4,142	76.8 (63.0)
Cyanobacteria	cyaNOG	15	9,107	57.6 (49.0)
Deinococcus-Thermus	deiNOG	5	1,459	80.7 (76.4)
Dictyoglomi	dicNOG	2	1,608	79.8 (72.7)
Firmicutes	firmNOG	187	18,321	62.4 (55.2)
Fusobacteria	fusoNOG	4	1,720	80.8 (72.9)
Spirochaetes	spiNOG	18	1,410	78.7 (76.0)
Tenericutes	tenNOG	25	792	77.8 (71.0)
Thermotogae	therNOG	11	2,553	76.4 (67.7)
Verrucomicrobia	verNOG	3	1,356	88.4 (79.3)
Alphaproteobacteria	aproNOG	117	20,425	66.1 (58.4)
Proteobacteria	proNOG	487	48,683	55.6 (45.6)
Betaproteobacteria	bproNOG	72	16,390	68.6 (60.2)
Deltaproteobacteria	dproNOG	33	11,209	68.6 (56.0)
Epsilonproteobacteria	eproNOG	26	3,220	69.6 (68.3)
Gammproteobacteria	gproNOG	238	21,151	50.2 (54.2)

Figure 1. In addition to the over 100 000 orthologous groups in the last universal common ancestor (LUCA), eggNOG v3 also provides orthologous groups and functional annotation for an additional 40 taxonomic levels. Here we display each level with its abbreviated name, species count, orthologous group count and annotation coverage. The annotation coverage for both the functional description of the groups as well as the functional category (in parentheses) is given.

47.6% have been classified into a functional category (for details see Figure 1).

FURTHER IMPROVEMENTS

As the exponential growth of genomes and genes therein leads to considerable issues regarding performance, a number of technical improvements and speedups have been introduced; for example the parallelization of some key aspects of the OG pipeline have contributed to the performance enhancement.

One important step in the eggNOG pipeline is the inference of in-paralogs. Proteins that belong to a given subset of species and are more similar to each other than to proteins belonging to species outside that subset are defined as in-paralogs. In this release, we determined the aforementioned subsets automatically: for the universal, domain- and phylum-specific OGs, we grouped organisms within the same taxonomic order. For taxonomical ranges between the phylum and class, we used the taxonomical family, while for ranges below the class level we grouped given species together.

QUALITY ASSESSMENT OF eggNOG v3.0

So far, the majority of quality assessment tests are based on the functional conservation of predicted orthologs (41–44); however, it has been acknowledged that a phylogeny-based benchmarking approach would be more appropriate (44,45). We therefore manually curated a set of orthologous groups exemplifying multiple caveats of orthology prediction (35), named Reference OGs (RefOGs), which were used to assess the quality between this release and eggNOG v2. As many as 95% of the reference orthologs can be detected in the new release compared to only 75% in the previous version (Figure 2). This is mainly due to the updated genome annotations in eggNOG v3. We estimated the impact of four error sources: (i) false assignments, (ii) missing orthologs, (iii) fusions and (iv) fissions (for details

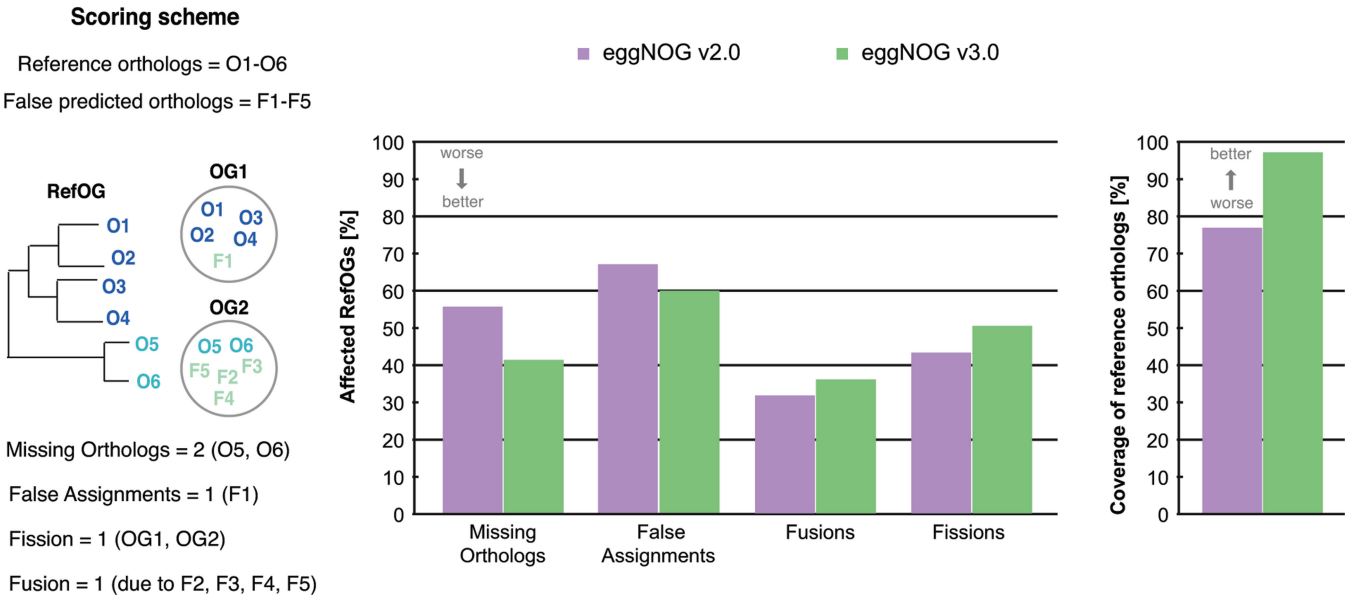


Figure 2. Quality assessment of eggNOG v3. We used 70 manually curated families (RefOGs) to test the accuracy of orthology prediction of the new release compared to eggNOG v2. For each release, we identified the orthologous group (OG) with the largest overlap of each RefOG and calculated how many genes were not predicted in the OG (missing orthologs) and how many genes were over-predicted in the OG (false assignments). Additionally, we checked if members of the same RefOG have been separated into multiple OGs (RefOG fission) and how many of those OGs include more than three false assignments (RefOG fusion). Missing orthologs influence 41% of the RefOGs; however, this is significantly less than the 57% in eggNOG v2. Similarly, less RefOGs include false assignments in eggNOG v3 (60%) compared to version 2 (66%). However, there are slightly less artificial OG fusions and fissions in eggNOG v2. Given that an addition of species can introduce false assignments, our results suggest that the eggNOG methodology can tolerate a large number of species, and at the same time improve its coverage against the tested benchmark dataset.

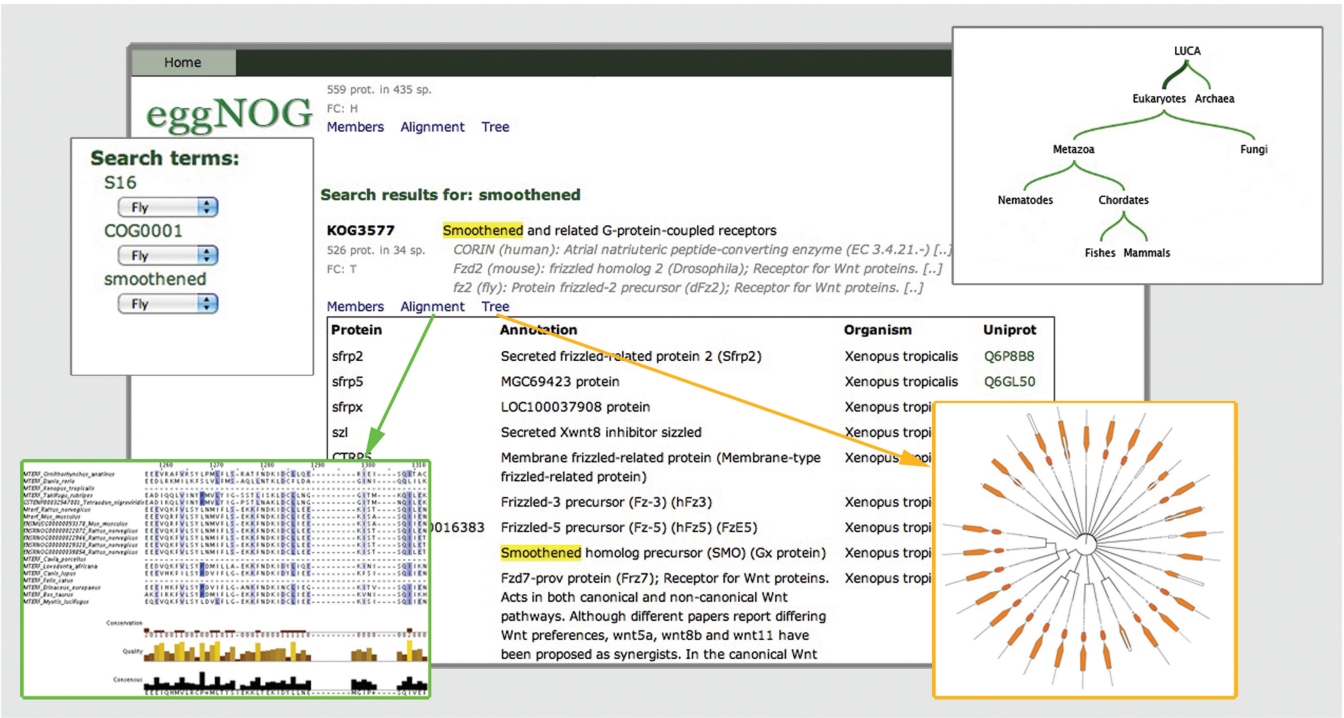


Figure 3. Screenshot of a results page. The eggNOG database was queried for the term 'smoothened'. The top left picture demonstrates the simplified navigation of multiple search terms and species selection. The navigation tree at the top right of the page allows the user to change the view to more coarse-grained orthologous groups, for example, the mammalian orthologous groups. The group features, such as member proteins, alignments (green arrow) and phylogenetic trees with SMART domains (orange arrow), can be accessed inline and do not require a page refresh.

see Figure 2). eggNOG v3 is less influenced than eggNOG v2 by false assignments and missing orthologs. Especially, for the missing orthologs, only 41% of the RefOGs are affected in this release compared to 57% in previous one. The high coverage of the benchmark set (95%) due to new genome annotations is the major contributor to this observation, highlighting the importance of frequent database updates, which is one of our goals. On the other hand, the previous release contains slightly fewer artificial fusions and fissions. As coverage of compared species affect the accuracy of orthology assignment (35), it can be expected that the addition of more species does not always improve all benchmark parameters.

ACCESS OPTIONS

To improve the usability of eggNOG v3, a new, modernized web interface was developed. As with the previous versions, the new interface provides data that can be downloaded under the Creative Commons Attribution 3.0 License at http://eggnog.embl.de. The available data include the OGs, protein sequences, multiple sequence alignments, precomputed gene trees (Figure 3) as well as the annotation of 62% of the OGs. Possible queries include multiple OG names, gene names and/or protein names. One goal of the new interface is to simplify the navigation of the various OGs by (i) a cleaner, more intuitive interface as well as (ii) an interactive species tree on the right side of the search results. The interactive species tree facilitates the navigation across different hierarchical levels by following the orthologs through the taxonomic levels. *Homo sapiens* serves as the default species for protein name queries; however, this can be changed to a multiple of common species within the search results. The multiple sequence alignments can be displayed using the Jalview applet (46) or downloaded in aligned or unaligned form. Precomputed phylogenetic trees are also provided and can be viewed together with any assigned PFAM (47) and SMART (48) domain via iTOL (49) or downloaded in Newick format.

CONCLUSIONS/PERSPECTIVES

With eggNOG v3, we provide one of the most comprehensive and up-to-date databases of orthologous groups available that delivers protein function annotation for 1133 genomes across the three domains of life. Not only does eggNOG v3 cover a broad taxonomic spectrum, but it also supplies orthologous groups for 41 manually selected taxonomic ranges. The modern, easy-to-use web interface facilitates the usage of the database with novel extended functionalities, such as an interactive species tree to assist the navigation through the increased number of hierarchical levels. Our future plans include the ongoing improvement of the quality of orthology and functional assignments, a further increase of taxonomic ranges and technical improvements to manage the computational challenges that come along with the expected exponential increase of available genomes.

ACKNOWLEDGEMENTS

We would like to thank Yan Yuan for all his help and support on all technical and infrastructure issues we encountered during this project.

FUNDING

EMBL; MetaHit RTD EC (201052); Novo Nordisk Foundation Center for Protein Research; Swiss Institute of Bioinformatics; and the University of Zurich through its Research Priority Program 'Systems Biology and Functional Genomics'. Funding for open access charge: EMBL (internal).

Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
3. Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.
4. Huynen,M.A., Snel,B., von Mering,C. and Bork,P. (2003) Function prediction and protein networks. *CuCrr. Opin. Cell. Biol*, **15**, 191–198.
5. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.
6. Koonin,E.M. (2005) Orthologs, paralogs and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
7. Östlund,G., Schmitt,T., Forslund,K., Köstler,T., Messina,D.N., Roopra,S., Frings,O. and Sonnhammer,E.L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.
8. 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.
9. Makarova,K.S., Sorokin,A.V., Novichkov,P.S., Wolf,Y.I. and Koonin,E.V. (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct.*, **2**, 33.
10. Waterhouse,R.M., Zdobnov,E.M., Tegenfeldt,F., Li,J. and Kriventseva,E.V. (2011) OrthoDBL the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.*, **39**, D283–D288.
11. Muller,J., Szklarczyk,D., Julien,P., Letunic,I., Roth,A., Kuhn,M., Powell,S., von Mering,C., Doerks,T., Jensen,L.J. *et al.* (2010) eggNOG v2.0. extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*, **38**, D190–D195.
12. Altenhoff,A.M., Schneider,A., Gonnet,G.H. and Dessimoz,C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.
13. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr and Roos,D.S. (2006) OrthoMCL-DB. Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
14. Uchiyama,I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res.*, **35**, D343–D346.
15. Linard,B., Thompson,J.D., Poch,O. and Lecompte,O. (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinform.*, **12**, 11.

16. Wapinski,I., Pfeffer,A., Friedman,N. and Regev,A. (2007) Automaticgenome- wide reconstruction of phylogenetic gene trees. *Bioinformatics*, **23**, i549–i58.
17. Huerta-Cepas,J., Bueno,A., Dopazo,J. and Gabaldón,T. (2008) PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.*, **36**, D491–D496.
18. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees. Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–35.
19. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Hériché,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam. 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
20. van der Heijden,R.T., Snel,B., van Noort,V. and Huynen,M.A. (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinform.*, **8**, 83.
21. Datta,R.S., Meacham,C., Samad,B., Neyer,C. and Sjölander,K. (2009) Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res.*, **37**, W84–W89.
22. Kuzniar,A., van Ham,R.C., Pongor,S. and Leunissen,J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.*, **24**, 539–551.
23. Gabaldon,T. (2008) Large-scale assignment of orthology Back to phylogenetics? *Genome Biol.*, **9**, 235.
24. Kristensen,D.M., Wolf,Y.I., Mushegian,A.R. and Koonin,E.V. (2011) Computational methods for Gene Orthology inference. *Brief Bioinform.*, **12**, 379–391.
25. Kuzniar,A., Lin,K., He,Y., Nijveen,H., Pongor,S. and Leunissen,J.A. (2009) ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Res.*, **37**, W428–W434.
26. 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**, D250–254.
27. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
28. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **36**, D491–496.
29. The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
30. Aurrecoechea,C., Brestelli,J., Brunk,B.P., Carlton,J.M., Dommer,J., Fischer,S., Gajria,B., Gao,X., Gingle,A., Grant,G. *et al.* (2009) GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res.*, **37**, D526–D530.
31. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
32. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
33. Kuhn,M., Szklarczyk,D., Franceschini,A., Campillos,M., von Mering,C., Jensen,L.J., Beyer,A. and Bork,P. (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.
34. Milinkovitch,M.C., Helaers,R., Depiereux,E., Tzika,A.C. and Gabaldón,T. (2010) 2x genomes–depth does matter. *Genome Biol.*, **11**, R16.
35. Trachana,K., Larsson,T.A., Powell,S., Chen,W.H., Doerks,T., Muller,J. and Bork,P. (2011) Orthology prediction methods: a quality assessment using curated protein families. *Bioessays*, **33**, 769–780.
36. 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.
37. Creevey,C.J., Doerks,T., Fitzpatrick,D.A., Raes,J. and Bork,P. (2011) Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS One*, **6**, e22099.
38. Rattei,T., Tischler,P., Götz,S., Jehl,M.A., Hoser,J., Arnold,R., Conesa,A. and Mewes,H.W. (2010) SIMAP–a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res.*, **38**, D223–D226.
39. Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.
40. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
41. Pryszcz,L.P., Huerta-Cepas,J. and Gabaldon,T. (2010) MetaPhOrs. Orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, **39**, e32.
42. Hulsen,T., Huynen,M.A., de Vlieg,J. and Groenen,P.M. (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.*, **7**, R31.
43. Chen,F., Mackey,A.J., Vermunt,J.K. and Roos,D.S. (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**, e383.
44. Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.
45. Boeckmann,B., Robinson-Rechavi,M., Xenarios,I. and Dessimoz,C. (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief Bioinform.*, **12**, 423–435.
46. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
47. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
48. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6. Recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
49. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–W78.