Nucleic Acids Research Advance Access published November 17, 2015

eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences

Jaime Huerta-Cepas[1], Damian Szklarczyk[2,3], Kristoffer Forslund[1], Helen Cook[4], Davide Heller[2,3], Mathias C. Walter[5], Thomas Rattei[6], Daniel R. Mende[7], Shinichi Sunagawa[1], Michael Kuhn[8], Lars Juhl Jensen[4], Christian von Mering[2,3,*] and Peer Bork[1,9,10,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [2]Institute of Molecular Life Sciences, University of Zurich, Zurich 8057, Switzerland, [3]Bioinformatics/Systems Biology Group, Swiss Institute of Bioinformatics (SIB), Zurich 8057, Switzerland, [4]The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen N 2200, Denmark, [5]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg 85764, Germany, [6]CUBE—Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna 1090, Austria, [7]Daniel K. Inouye Center for Microbial Oceanography: Research and Education, University of Hawaii, Honolulu, HI 96822, USA, [8]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany, [9]Germany Molecular Medicine Partnership Unit (MMPU), University Hospital Heidelberg and European Molecular Biology Laboratory, Heidelberg 69117, Germany and [10]Max Delbrück Centre for Molecular Medicine, Berlin 13125, Germany

Received September 14, 2015; Revised October 30, 2015; Accepted November 02, 2015

ABSTRACT

eggNOG is a public resource that provides Orthologous Groups (OGs) of proteins at different taxonomic levels, each with integrated and summarized functional annotations. Developments since the latest public release include changes to the algorithm for creating OGs across taxonomic levels, making nested groups hierarchically consistent. This allows for a better propagation of functional terms across nested OGs and led to the novel annotation of 95 890 previously uncharacterized OGs, increasing overall annotation coverage from 67% to 72%. The functional annotations of OGs have been expanded to also provide Gene Ontology terms, KEGG pathways and SMART/Pfam domains for each group. Moreover, eggNOG now provides pairwise orthology relationships within OGs based on analysis of phylogenetic trees. We have also incorporated a framework for quickly mapping novel sequences to OGs based on precomputed HMM profiles. Finally, eggNOG version 4.5 incorporates a novel data set spanning 2605 viral OGs, covering 5228 proteins from 352 viral proteomes. All data are accessible for bulk downloading, as a web-service, and through a completely redesigned web interface. The new access points provide faster searches and a number of new browsing and visualization capabilities, facilitating the needs of both experts and less experienced users. eggNOG v4.5 is available at http://eggnog.embl.de.

INTRODUCTION

Orthology and paralogy are central concepts in evolutionary biology. They allow distinguishing between molecular sequences that, despite sharing a common ancestry, evolved by different mechanisms: orthologs are the result of speciation events, whereas paralogs originate from gene duplications. This distinction is widely used in molecular biology, since the evolutionary forces shaping the respective classes of sequences are profoundly different and impact the analysis of functional divergence (1). It is generally assumed that orthologous genes are more likely to conserve their function than paralogs, which, in contrast to orthologs, are partially released from selective pressures after duplication. This idea is commonly referred as the Ortholog Conjecture and, although recently questioned (2,3), it is still considered generally valid and represents the basis of most functional an-

[*]To whom correspondence should be addressed. Tel: +49 6221 387 85 26; Email: bork@embl.de
Correspondence may also be addressed to Christian von Mering. Tel: +41 44 635 31 47; Email: mering@imls.uzh.ch

notation methods (4). Consequently, precise orthology assignments are crucial in many fields such as phylogenetics, pharmacology and comparative genomics. However, due to the intricate evolution of most gene families, which often involves multiple nested duplications, genomic rearrangements and horizontal gene transfers, orthology prediction remains as a highly challenging task (4,5), both analytically and computationally.

Therefore multiple orthology resources have been developed that provide precomputed predictions, each based on a different methodology and organism range, and all having different strengths and weaknesses (6,7). The inference approaches fall into two main categories, namely graph-based (8–15) and tree-based (16–19) methods. Graph-based algorithms allow analysis of more species at once and produce groups of orthologous sequences with the common ancestor defined by the set of species considered at the taxonomic level. Tree-based approaches, by contrast, provide finer resolution (i.e. using tree topology to identify specific speciation and duplication events), but they require heavier computations and are more sensitive to methodological artifacts (20).

We maintain a database of Orthologous Groups (OGs) and functional annotations called eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) (21). eggNOG uses a graph-based unsupervised clustering algorithm extending the COG methodology (22) to produce genome wide orthology inferences, which are further adjusted to provide lineage specific resolution. The database currently covers 2031 eukaryotic and prokaryotic organisms, as well as precomputed mappings for 1655 additional prokaryotes (12). The present manuscript describes the most recent release of eggNOG (*v4.5*, 2015), featuring a number of improvements over its previous release. The most notable ones include (i) modifications to the clustering algorithm in order to make OGs hierarchically consistent across taxonomic levels, (ii) improved annotation of OGs, (iii) the availability of HMM-based tools for fast protein sequence assignment to OGs, (iv) the addition of viral OGs*,* (v) the availability of fine-grained orthology inferences derived from phylogenetic analysis, (vi) a completely re-designed web interface and (vii) programmatic access through a RESTful Application Programming Interface (API). eggNOG v4.5 is available at http://eggnog.embl.de.

OVERVIEW OF THE COMPUTATIONAL PIPELINE

Apart from the central graph-based clustering algorithm, the eggNOG production pipeline involves a number of quality controls as well as pre- and post-processing steps, which have evolved over the last 8 years since its first publication (21). Given the amount of change accumulated at present and previous versions, we describe here the current status of the complete pipeline (Figure 1), highlighting the most recent updates and additions since release 4.0 (12).

Data set preparation and pairwise sequence comparison

The workflow starts by collecting genomes from public databases (19,23–26). Genomes and proteomes are downloaded, parsed and subjected to quality controls that prevent the inclusion of partial or draft genomes (Figure

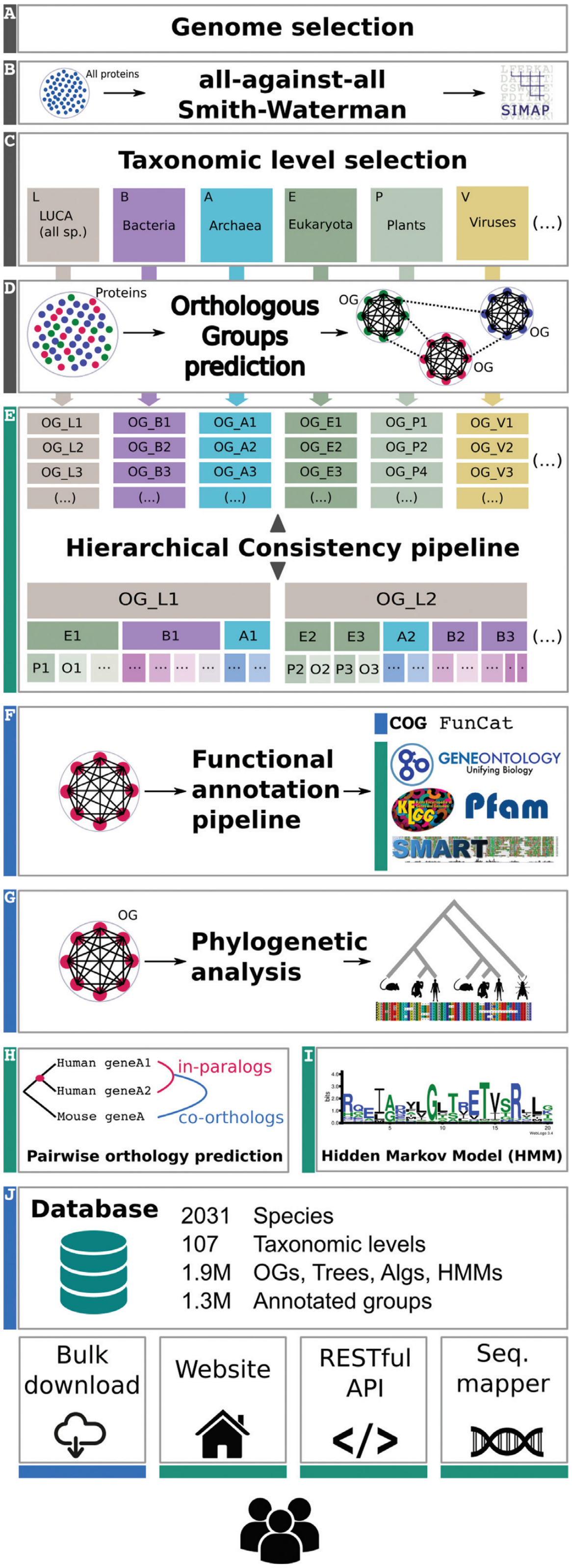


Figure 1 Schematic representation of the eggNOG pipeline: Boxes labelled in green indicate new data and/or methods added in this version. Blue labels represent updated methodology and/or data with respect to previous versions. Grey boxes indicate unchanged steps in version 4.5.

1A). This step is coordinated with the STRING (27) and STITCH (28) databases so that the underlying set of protein sequences and names is shared among all three resources. The new viral proteins included in eggNOG v4.5 were retrieved by selecting all reference viral proteomes in Uniprot via XML download on 31 August 2015. These proteomes were filtered by a series of quality controls, which removed 50 proteomes. Eight additional proteomes were included following manual review. Viral proteins are often translated as a single polyprotein, which is cleaved to form functional proteins. Prior to inclusion in eggNOG, such polyproteins were cleaved *in silico* following the 'chain' annotations present in Uniprot entries, and only the smallest units were retained, so that the protein sequences are non-redundant.

Pairwise sequence comparison

Protein sequences from the selected organisms and viruses are extracted and used to compute an *all-against-all* pairwise similarity matrix (Figure 1B), a task that is currently carried out by the SIMAP project (29). The comparison uses Smith–Waterman alignments and compositional adjustment of the scores, as in BLAST, to prevent spurious hits between low-complexity sequence regions. Hits with bit-scores above 50 are stored and indexed in a relational database, which forms the input to the next stage of the algorithm.

Definition of taxonomic levels

Because the resolution of OGs depends on the taxonomic level, the eggNOG clustering pipeline is independently executed at different predefined taxonomic levels, each spanning a different clade in the overall tree of life. Levels are manually chosen to cover evolutionarily relevant groups as well as to maximally make use of well-studied model organisms (Figure 1C). This gives rise to the hierarchical structure of the data in eggNOG (Figure 2A), where, for example, a set of mammalian sequences with a common ortholog at the base of vertebrates could be part of a single mammal-specific OG (OG:0UIPS in Figure 2A), but constitute two separate supraprimate-specific groups (OG:1AVEH and OG:1AU76 in Figure 2A). In addition, eggNOG v4.5 uses 16 predefined taxonomic levels to classify the 352 viral proteomes (Figure 2B).

Building Orthologous Groups

eggNOG's clustering algorithm (Figure 1D) takes its basis from the manually curated Clusters of Orthologous Groups covering the three domains of life: COGs (universal with best coverage for Bacteria) (30), KOGs (Eukaryotes) (8) and arKOGs (Archaea) (31). These groups are conserved at their corresponding taxonomic level in eggNOG, and are extended with additional proteomes. For each of the predefined taxonomic levels, first, groups of in-paralogous proteins are created. Then, closely related groups of in-paralogs and single genes are merged creating clusters of homologous proteins. Such clusters can also later be split again if there is a reciprocal best hit between proteins from clusters from separate lineages. The eggNOG algorithm used to build OGs has been benchmarked and compared to similar approaches in the past using OrthoBench (6,7). Moreover, the OrthoBench test suite is regularly used to evaluate the quality of OGs every time eggNOG receives an update. Note that, although other benchmarking frameworks are available, they usually require pairwise orthology predictions, therefore preventing the correct evaluation of OGs. We have, however, incorporated such type of benchmarks to test eggNOG's new capacity to produce fine-grained predictions (see sections bellow).

Hierarchical consistency of nested groups

The eggNOG pipeline is run independently for each of the predefined taxonomic levels considered. The imperfect quality of the proteomes used and the heuristics of the pipeline are factors that can lead to minor inconsistencies between levels, such as disagreements on when duplication events occurred for a given set of homologous proteins. This sometimes prevented the correct propagation of annotations across nested groups in previous versions, and occasionally also caused inconsistencies when applying third-party analysis pipelines to eggNOG. Version 4.5 of the eggNOG algorithm resolves this by incorporating a post-clustering step that scans all groups in all levels and eliminates inconsistencies by splitting and merging the OGs (Figure 1E). The scanning is performed in root-ward direction starting from each of the leaves in predetermined sequence. The algorithm checks for any OG that underwent the division at the parental level. If such split was found, the algorithm determines, in sequence (from largest to smallest), the species overlap between each pair of resulting groups. If no overlap is detected, the split OGs are merged together at the parental level, otherwise the proteins of the smaller group are separated from the proteins of the larger group downstream at every child level. Some apparent inconsistencies remain which reflect gene fusion events; these however represent the true mosaic history of the affected proteins and have therefore been retained. In order to examine whether the consistency pipeline had affected the quality of the groups, we have benchmarked the new, consistent, eggNOG using OrthoBench 2 (7). The benchmark results show a slight increase in the F-measure for the bilateria level (from 71.2% to 72.4%) and a small decrease for gammaproteobacteria (from 94.6% to 93.2%), indicating no major impact on the quality of groups.

Phylogenetic analysis

Amino acid sequences from each OG at each taxonomic level are further analysed using phylogenetic methods (Figure 1G). For this release, 1.9 million phylogenetic trees were built using a slightly modified version of a previously described methodology (32,33). The currently used approach includes reconstructing multiple sequence alignments based on the consensus of several aligning and gap cleaning programs (34–38), evolutionary model testing and maximum likelihood inference (39,40). The inferred speciation events are used to derive a list of pairwise orthology predictions to make the group concept of eggNOG more comparable with pair-based orthology methods (Figure 1H); the list is provided with this eggNOG version. Finally, a Hidden Markov

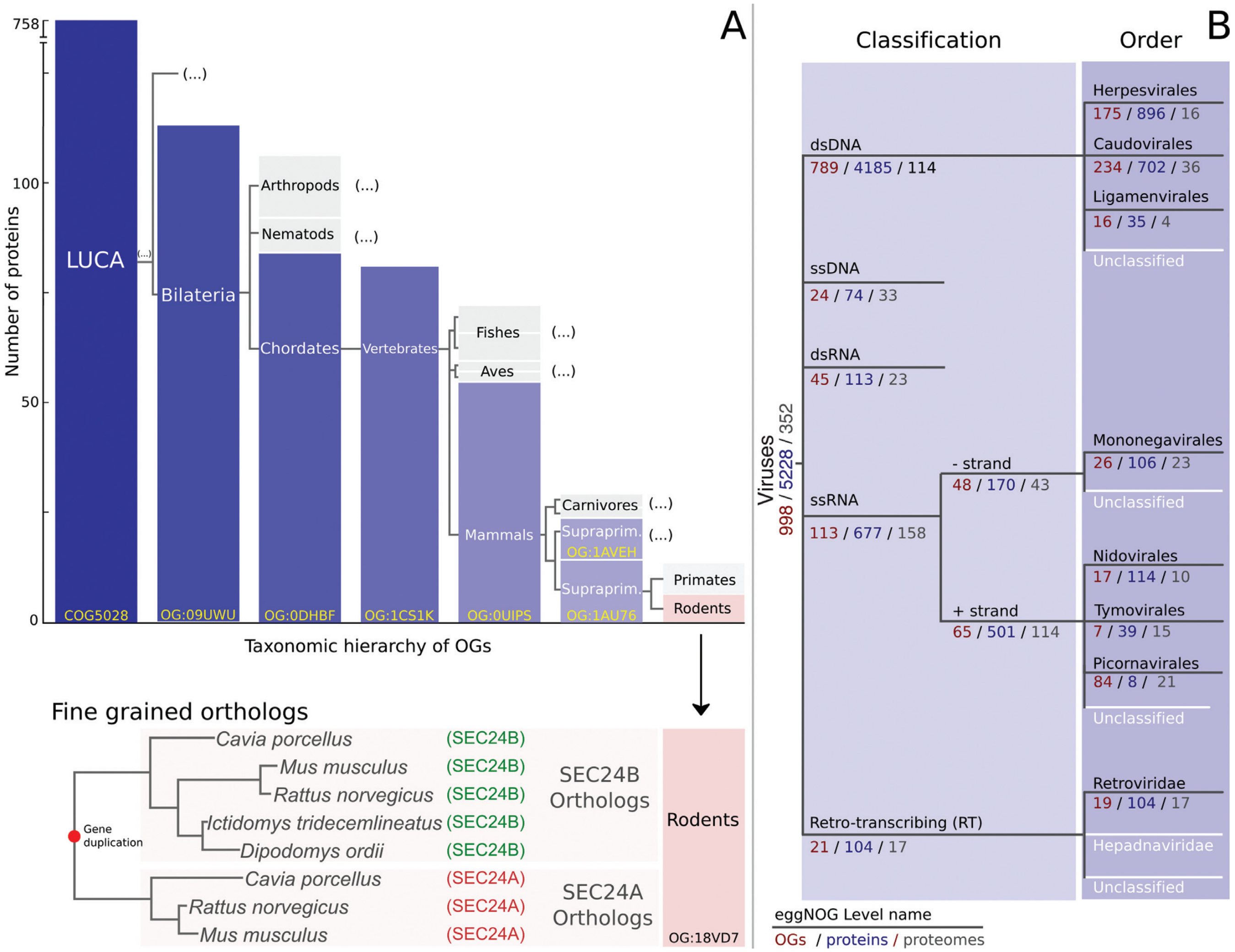
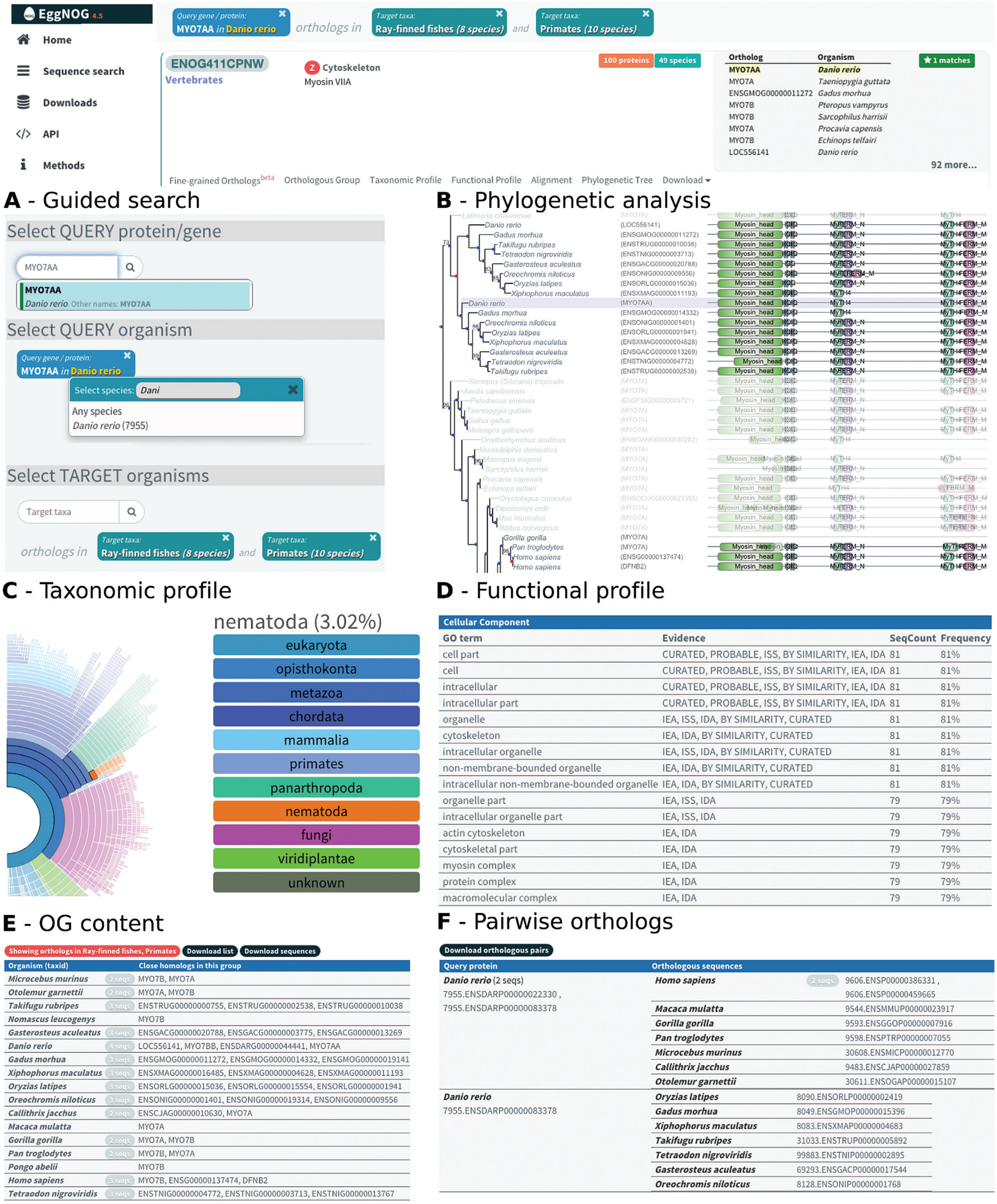


Figure 2 (**A**) Hierarchically consistent structure of OGs including genes from the SEC24 protein family, from the root taxonomic level (Last Universal Common Ancestor, LUCA) to the rodents specific level. Each OG is represented by a box labelled with the lineage name it belongs to, and whose size is proportional to the number of proteins grouped. Boxes filled with a blue gradient represent the nested hierarchy of OGs specifically containing the mouse SEC24A protein. Grey boxes indicate collapsed branches in the OG hierarchy. Note that another Bilateria-specific OG exist, but has been collapsed for readability reasons. The most lineage-specific OG containing the mouse SEC24A protein is at the rodents taxonomic level, which is coloured in pink. Fine grained orthology for SEC24 genes, based on the phylogenetic analysis of the 18VD7 rodent-specific group, is shown in the bottom part, with tree branches indicating a lineage specific duplication. (**B**) Viral taxonomic tree. Black branches indicate levels for which OGs were calculated, whereas white branches indicate no OG was calculated at this level. Numbers indicate the number of OGs at this level, the number of proteins contained in all OGs at this level and the number of proteomes represented by the proteins within all OGs at this level, respectively.

Model (HMM) profile is built for each group based on the untrimmed version of the multiple sequence alignment using HMMER (41) (Figure 1I), which can be used for specific OG assignments in external data sets.

Functional annotation of orthologous groups

Once the consistency has been ensured, functional descriptions are assigned to each OG using an automated procedure (Figure 1F). At the taxonomic level of each OG, available functional annotations are collected from many sources including free-text descriptions in source genome databases, COG functional categories (8), Gene Ontology terms (42), KEGG pathways (43) and SMART/Pfam protein domains (44,45). From these, a heuristic procedure aims to identify the most descriptive shared description substring among annotated members of the group. This integrated description line is then provided together with the group as a bare-bone descriptor of what is known in terms of its role and function. While this text summary is human-readable, it cannot be used for statistical analysis, and groups are therefore also classified into the single-letter functional categories used by the COG database. The individual assignments are made by a Support Vector Machine (SVM) classifier trained on proteins within COGs, KOGs and arKOGs, using as features text description words and substrings, protein domain and Gene Ontology term assignments, as well as KEGG pathway membership information. Further technical details regarding the functional annotation pipeline are available at the methods section of the main website: http://eggnog.embl.de/#/app/methods.



EggNOG 4.5

- Home
- Sequence search
- Downloads
- API
- Methods

Query gene / protein: MYO7AA in Danio rerio ✖ orthologs in Target taxa: Ray-finned fishes (8 species) ✖ and Target taxa: Primates (10 species) ✖

ENOG411CPNW
Vertebrates

Z Cytoskeleton
Myosin VIIA

100 proteins 49 species

Ortholog	Organism
MYO7AA	*Danio rerio*
MYO7A	*Taeniopygia guttata*
ENSGMOG00000011272	*Gadus morhua*
MYO7B	*Pteropus vampyrus*
MYO7B	*Sarcophilus harrisii*
MYO7A	*Procavia capensis*
MYO7B	*Echinops telfairi*
LOC556141	*Danio rerio*

★ 1 matches

92 more...

Fine-grained Orthologs beta · Orthologous Group · Taxonomic Profile · Functional Profile · Alignment · Phylogenetic Tree · Download ▾

A - Guided search

Select QUERY protein/gene

MYO7AA

MYO7AA
Danio rerio. Other names: MYO7AA

Select QUERY organism

Query gene / protein: MYO7AA in Danio rerio ✖

Select species: Dani ✖

Any species
Danio rerio (7955)

Select TARGET organisms

Target taxa

orthologs in Target taxa: Ray-finned fishes (8 species) ✖ and Target taxa: Primates (10 species) ✖

B - Phylogenetic analysis

Latimeria chalumnae (MYO7A)
Danio rerio (LOC556141)
Gadus morhua (ENSGMOG00000011272)
Takifugu rubripes (ENSTRUG00000010038)
Tetraodon nigroviridis (ENSTNIG00000003713)
Gasterosteus aculeatus (ENSGACG00000020788)
Oreochromis niloticus (ENSONIG00000009556)
Oryzias latipes (ENSORLG00000015036)
Xiphophorus maculatus (ENSXMAG00000011193)
Danio rerio (MYO7AA)
Gadus morhua (ENSGMOG00000014332)
Oreochromis niloticus (ENSONIG00000001401)
Oryzias latipes (ENSORLG00000001941)
Xiphophorus maculatus (ENSXMAG00000004628)
Gasterosteus aculeatus (ENSGACG00000013269)
Tetraodon nigroviridis (ENSTNIG00000004772)
Takifugu rubripes (ENSTRUG00000002538)
Xenopus (Silurana) tropicalis (MYO7A)
Anolis carolinensis (MYO7A)
Pelodiscus sinensis (ENSPSIG00000009721)
Taeniopygia guttata (MYO7A)
Gallus gallus (MYO7A)
Meleagris gallopavo (MYO7A)
Ornithorhynchus anatinus (ENSOANG00000030202)
Monodelphis domestica (MYO7A)
Macropus eugenii (MYO7A)
Sarcophilus harrisii (MYO7A)
Procavia capensis (MYO7A)
Echinops telfairi (MYO7A)
Oryctolagus cuniculus (ENSOCUG00000023355)
Dipodomys ordii (MYO7A)
Mus musculus (MYO7A)
Rattus norvegicus (MYO7A)
Gorilla gorilla (MYO7A)
Pan troglodytes (MYO7A)
Homo sapiens (ENSG00000137474)
Homo sapiens (DFNB2)
73
83
85
96
99
96
96
98
Myosin_head
MyTH4
FERM_N
FERM_M

C - Taxonomic profile

D - Functional profile

Cellular Component

GO term	Evidence	SeqCount	Frequency
cell part	CURATED, PROBABLE, ISS, BY SIMILARITY, IEA, IDA	81	81%
cell	CURATED, PROBABLE, ISS, BY SIMILARITY, IEA, IDA	81	81%
intracellular	CURATED, PROBABLE, ISS, BY SIMILARITY, IEA, IDA	81	81%
intracellular part	CURATED, PROBABLE, ISS, BY SIMILARITY, IEA, IDA	81	81%
organelle	IEA, ISS, IDA, BY SIMILARITY, CURATED	81	81%
cytoskeleton	IEA, IDA, BY SIMILARITY, CURATED	81	81%
intracellular organelle	IEA, ISS, IDA, BY SIMILARITY, CURATED	81	81%
non-membrane-bounded organelle	IEA, IDA, BY SIMILARITY, CURATED	81	81%
intracellular non-membrane-bounded organelle	IEA, IDA, BY SIMILARITY, CURATED	81	81%
organelle part	IEA, ISS, IDA	79	79%
intracellular organelle part	IEA, ISS, IDA	79	79%
actin cytoskeleton	IEA, IDA	79	79%
cytoskeletal part	IEA, IDA	79	79%
myosin complex	IEA, IDA	79	79%
protein complex	IEA, IDA	79	79%
macromolecular complex	IEA, IDA	79	79%

E - OG content

Showing orthologs in Ray-finned fishes, Primates · Download list · Download sequences

Organism (taxid)		Close homologs in this group
Microcebus murinus	2 seqs	MYO7B, MYO7A
Otolemur garnettii	2 seqs	MYO7A, MYO7B
Takifugu rubripes	3 seqs	ENSTRUG00000000755, ENSTRUG00000002538, ENSTRUG00000010038
Nomascus leucogenys		MYO7B
Gasterosteus aculeatus	3 seqs	ENSGACG00000020788, ENSGACG00000003775, ENSGACG00000013269
Danio rerio	4 seqs	LOC556141, MYO7BB, ENSDARG00000044441, MYO7AA
Gadus morhua	3 seqs	ENSGMOG00000011272, ENSGMOG00000014332, ENSGMOG00000019141
Xiphophorus maculatus	3 seqs	ENSXMAG00000016485, ENSXMAG00000004628, ENSXMAG00000011193
Oryzias latipes	3 seqs	ENSORLG00000015036, ENSORLG00000015554, ENSORLG00000001941
Oreochromis niloticus	3 seqs	ENSONIG00000001401, ENSONIG00000019314, ENSONIG00000009556
Callithrix jacchus	2 seqs	ENSCJAG00000010630, MYO7A
Macaca mulatta		MYO7A
Gorilla gorilla	2 seqs	MYO7A, MYO7B
Pan troglodytes	2 seqs	MYO7B, MYO7A
Pongo abelii		MYO7B
Homo sapiens	3 seqs	MYO7B, ENSG00000137474, DFNB2
Tetraodon nigroviridis	3 seqs	ENSTNIG00000004772, ENSTNIG00000003713, ENSTNIG00000013767

F - Pairwise orthologs

Download orthologous pairs

Query protein	Orthologous sequences		
Danio rerio (2 seqs) 7955.ENSDARP00000022330 , 7955.ENSDARP00000083378	***Homo sapiens***	2 seqs	9606.ENSP00000386331 , 9606.ENSP00000459665
	Macaca mulatta		9544.ENSMMUP00000023917
	Gorilla gorilla		9593.ENSGGOP00000007916
	Pan troglodytes		9598.ENSPTRP00000007055
	Microcebus murinus		30608.ENSMICP00000012770
	Callithrix jacchus		9483.ENSCJAP00000027859
	Otolemur garnettii		30611.ENSOGAP00000015107
Danio rerio 7955.ENSDARP00000083378	***Oryzias latipes***		8090.ENSORLP00000002419
	Gadus morhua		8049.ENSGMOP00000015396
	Xiphophorus maculatus		8083.ENSXMAP00000004683
	Takifugu rubripes		31033.ENSTRUP00000005892
	Tetraodon nigroviridis		99883.ENSTNIP00000002895
	Gasterosteus aculeatus		69293.ENSGACP00000017544
	Oreochromis niloticus		8128.ENSONIP00000001768

Figure 3 Website screenshots showing fish and primate orthologs for the myosin protein MYO7AA. (**A**) The guided search dialog used to retrieve the orthologs. (**B**) Partial tree representation of the associated phylogenetic tree. Blue nodes in the tree represent speciation events. Red nodes indicate duplication events (in-paralogs). Pfam domains are shown in-line for all the orthologous sequences. Note that tree visualization is adapted to the query, highlighting

UPDATES AND ADDITIONS SINCE PREVIOUS RELEASE

During the last two years, the development work on the above computational pipeline has resulted in a series of improvements, changes and additions, aiming to make the procedure more stringent in preparation for the next major update to the underlying data set. In parallel, work on a revised web front-end and programmatic access has been undertaken, aiming to more directly address the needs of different strata of eggNOG users.

Improved and extended functional annotations

The most common application of eggNOG remains the functional characterization of novel genes or proteins by mapping into the space of OGs for which annotations are available. Such annotations, as described above, include human-readable functional summaries as well as single-letter functional codes as defined for the COGs. With the reconciliation of nested groups at the predefined taxonomic levels, clades closer to the tips of the trees that lack annotations may now inherit from their parent groups closer to the root. In comparison to previous versions, 95 890 (5%) previously uncharacterized groups were annotated with text descriptions using this strategy, yielding 1 368 357 (72%) annotated groups in total. COG functional categories were assigned to 143 683 (7.5%) groups previously lacking them, yielding 936 917 (49%) annotated OGs in total. These cases are specifically flagged in case any application wants to exclude them. Due to the increasing need for a controlled vocabulary of functional annotations, eggNOG v4.5 provides now access to Gene Ontology, KEGG, SMART and Pfam mappings. Functional terms, as well as their relative frequencies within each group of orthologs, can be browsed interactively or queried programmatically using the API.

Faster and more sensitive sequence annotation based on HMM models

Many applications of eggNOG build on determining which OG a novel gene falls within. Although pairwise sequence similarity tools such as BLAST (46) are extensively applied for that purpose, the use of profile Hidden Markov Models (HMMs) can provide higher sensitivity for detecting remote similarities and overall performs better in large data sets (41). Moreover, the structure of eggNOG is particularly suitable for HMM analysis, as the hierarchical taxonomic structure of nested OGs allows adjusting the search to use the most appropriate level for each analysis. Users can choose to increase the resolution of mappings and annotations by restricting searches to lineage-specific levels, thus maximizing sequence similarity within groups, and therefore have access to better-quality multiple sequence alignments and HMMs. In eggNOG v4.5, 1.9 million HMMs have been reconstructed based on the complete set of OGs at all taxonomic levels. A collection of raw HMM files is available for download for each level to annotate external data sets. Furthermore, three optimized databases are provided that cover the three domains of life: Archaea, Bacteria and Eukaryota. These three databases have been designed to contain a selection of HMMs where larger OGs at the deepest taxonomic levels have been split into their corresponding lineage-specific, but more fine-grained, OGs.

Annotation of viral orthologous groups

Viruses have not been taxonomically or functionally annotated in other orthology resources. In this version of eggNOG, we cover non-cellular life for the first time, by the addition of viral orthologous groups constructed analogously but in parallel to the rest of the genomes covered by eggNOG. The viral proteins were processed by the standard eggNOG pipeline, which was seeded with phage orthology groups published in Kristensen et al. (47), analogous to how COGs, KOGs and arKOGs are used to seed each cellular domain. An additional step to merge orthology groups was performed on those viral OGs where a majority of the included proteins had the same Pfam domain architectures at the clan level. This resulted in 2605 final viral orthology groups at the top level covering 5228 proteins from 352 proteomes. Viral OGs have been calculated at 16 separate levels within the virus taxonomy (Figure 2B). Viral OGs, along with their associated HMMs, multiple sequence alignments, phylogenetic trees and functional annotations are available for downloading.

Pairwise orthology predictions

Although eggNOG currently has more than a hundred predefined taxonomic levels of orthology resolution, it is not infrequent that OGs contain in-paralogs and masked co-orthology relationships at some levels, the more the closer to the root of the tree of life, particularly at the deepest taxonomic levels. This is irrelevant when the intended analysis focuses on using the functional description of OGs, such as for the annotation of genomic and metagenomic data (48–50). However, accurate distinction among one-to-one, one-to-many and many-to-many relationships is often needed to address evolutionary questions such as the reconstruction of the tree of life (51), estimating the relative age of sequences (52), or studying gene duplication (53). For this reason, eggNOG v4.5 allows refining the content of each OG through the automated analysis of precomputed phylogenetic trees. This allows us to extract fine-grained orthology relationships among the protein members of each OG, even when the maximum level of taxonomic resolution is reached in the OGs hierarchy (Figure 2A). The performance of pairwise orthology predictions from eggNOG was recently evaluated as part of the Quest for Orthologs Benchmarking initiative, showing comparable results to other pairwise orthology resources (http://orthology.benchmarkservice.org).

←

the seed and target species and graying out the rest. (**C**) Taxonomic profile representation showing the distribution of orthologs in the tree of life. (**D**) Functional profile based on Gene Ontology terms associated to the OG. (**E**) Filtered content of the OG (protein names and sequences), restricted to the query and target species. (**F**) Pairwise orthology predictions adapted to the query protein and the target species. In-paralogy and co-orthology relationships are resolved according to the speciation and duplication events inferred from the phylogenetic tree.

New web interface: faster searches and advanced data browsing

All the described improvements have been fully integrated into the new eggNOG back-end database and the completely redesigned front-end web interface (Figure 3), which enables much faster searches and provides many new browsing and visualization capabilities. Special efforts have been made to facilitate the access to eggNOG data for less experienced users. The default search panel (Figure 3A) allows for guided queries in three simple steps. First, users enter a protein or gene name, which is instantly searched and autocompleted based on an in-house ID translation database covering all major sequence providers. Second, users are asked to disambiguate the source organism for the selected protein, which allows distinguishing between genes having the same name in different species (i.e. CDK1 in human versus CDK1 in chimp) and enables eggNOG to infer pairwise orthology mappings. Finally, users can provide a list of target organisms, or complete lineages, from which they would like to retrieve orthologs. This ensures an interpretable output and allows eggNOG to automatically select the most appropriate taxonomic level for the given query. For example, if a query is set to find rat orthologs of mouse CDK1, eggNOG will automatically retrieve the corresponding OG at the *rodents* taxonomic level and limit the displayed results to rat and mouse proteins only.

Finally, five new information channels have been added to the interface, which permit users to browse the extended data associated with each OG: (i) *Phylogenetic trees* and *Alignments* provide an integrative overview of the evolutionary relationships of all member proteins within each OG together with their functional annotations. The phylogenetic tree image (Figure 3B) highlights the query and target sequences, aligned domain regions, and the inferred duplication and speciation events within the group. (ii) The *Taxonomic Profile* channel (Figure 3C) offers a visual overview in the form of a *sunburst* representation about the distribution of orthologs across different taxonomic subdivisions. (iii) The *Functional Profiles* allow to inspect the frequency of functional terms, domains and pathways found within each group (Figure 3D). (iv) The *Orthologous Group* channel (Figure 3E) displays the complete list of members in a group, filtering out any species that are not present in the query and allowing users to download both ortholog names and sequences in FASTA format. (v) The *Pairwise Orthology* provides a refined list of orthologs to the specific queried protein, including fine-grained delineation of one-to-one, one-to-many and many-to-many relationships (Figure 3F).

Programmatic access

A scalable RESTful web service has been implemented that permits programmatic access to all eggNOG data, as well as their integration in third party resources. It currently supports queries to retrieve complete OGs, protein sequences, alignments, phylogenetic trees, HMM models and functional profiles in text and JSON formats. When a particular protein name is fixed as a query, pairwise orthology predictions can also be fetched using the web service API.

CONCLUSIONS

With the changes, updates and additions described above, eggNOG v4.5 provides one of the most complete and scalable databases for orthology prediction and functional annotation publicly available. The introduction of hierarchical consistency between groups, and the ability to stringently derive pairwise orthology relationships, brings the possibility of using eggNOG data both for large-scale sequence annotation projects and for evolutionary analyses requiring finer resolution. The redesign of the website frontend and backend databases offers fast and seamless integration with all eggNOG data, which ultimately aims at covering a variety of use-cases and users. Finally, the extensive changes described here enable more efficient and regular incorporation of newly sequenced high quality genomes to keep comprehensive species coverage.

ACKNOWLEDGEMENTS

Valuable feedback helping to guide some of the improvements to eggNOG was received within the Quest for Orthologs initiative.

FUNDING

European Commission MetaCardis [FP7-HEALTH-305312]; International Human Microbiome [HEALTH-FP7-2010-261376]; LTFCOFUND2013 [PCOFUND-GA-2013-609409]; European Research Council CancerBiome project [contract number 268985]; European Molecular Biology Organization [ALTF 721-2015]; CellNetworks (Excellence Initiative of the University of Heidelberg); Novo Nordisk Foundation [NNF14CC0001]; European Molecular Biology Laboratory (EMBL). Funding for open access charge: European Molecular Biology Laboratory (EMBL).

Conflict of interest statement. None declared.

REFERENCES

1. Ohno,S. (2013) *Evolution by Gene Duplication*. Springer Science & Business Media, NY.
2. Studer,R.A. and Robinson-Rechavi,M. (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.*, **25**, 210–216.
3. Nehrt,N.L., Clark,W.T., Radivojac,P. and Hahn,M.W. (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.*, **7**, e1002073.
4. Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.
5. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
6. 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.
7. Trachana,K., Forslund,K., Larsson,T., Powell,S., Doerks,T., Mering,C. and von Bork,P. (2014) A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS One*, **9**, e111122.
8. Tatusov,R., Fedorova,N., Jackson,J., Jacobs,A., Kiryutin,B., Koonin,E., Krylov,D., Mazumder,R., Mekhedov,S., Nikolskaya,A. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
9. Chen,F., Mackey,A.J., Stoeckert,C.J. and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.

10. Schreiber,F. and Sonnhammer,E.L.L. (2013) Hieranoid: hierarchical orthology inference. *J. Mol. Biol.*, **425**, 2072–2081.
11. Altenhoff,A.M., Kunca,N., Glover,N., Train,C.-M., Sueki,A., Pili ota,I., Gori,K., Tomiczek,B., Muller,S., Redestig,H. *et al.* (2014) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.*, **43**, D240–D249.
12. Powell,S., Forslund,K., Szklarczyk,D., Trachana,K., Roth,A., Huerta-Cepas,J., Gabaldón,T., Rattei,T., Creevey,C., Kuhn,M. *et al.* (2014) EggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.
13. Sonnhammer,E.L.L. and Östlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–D239.
14. Linard,B., Allot,A., Schneider,R., Morel,C., Ripp,R., Bigler,M., Thompson,J.D., Poch,O. and Lecompte,O. (2015) OrthoInspector 2.0: software and database updates. *Bioinformatics*, **31**, 447–448.
15. Kriventseva,E.V., Tegenfeldt,F., Petty,T.J., Waterhouse,R.M., Simão,F.A., Pozdnyakov,I.A., Ioannidis,P. and Zdobnov,E.M. (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.*, **43**, D250–D256.
16. Mi,H., Guo,N., Kejariwal,A. and Thomas,P.D. (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.*, **35**, D247–D252.
17. Pryszcz,L.P., Huerta-Cepas,J. and Gabaldón,T. (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, **39**, e32.
18. Huerta-Cepas,J., Capella-Gutiérrez,S., Pryszcz,L.P., Marcet-Houben,M. and Gabaldón,T. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, 897–902.
19. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
20. Rokas,A. and Carroll,S.B. (2006) Bushes in the tree of life. *PLoS Biol.*, **4**, 1899–1904.
21. Jensen,L.J., Julien,P., Kuhn,M., Mering,C., von Muller,J., Doerks,T. and Bork,P. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, **36**, D250–D254.
22. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
23. Kersey,P.J., Allen,J.E., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C., Hughes,D.S.T., Humphrey,J., Kerhornou,A., Khobova,J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.
24. Nordberg,H., Cantor,M., Dusheyko,S., Hua,S., Poliakov,A., Shabalov,I., Smirnova,T., Grigoriev,I.V. and Dubchak,I. (2014) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.*, **42**, D26–D31.
25. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
26. Tatusova,T., Ciufo,S., Fedorov,B., O'Neill,K. and Tolstoy,I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553-D559.
27. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2014) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
28. Kuhn,M., Szklarczyk,D., Pletscher-Frankild,S., Blicher,T.H., Mering,C., von Jensen,L.J. and Bork,P. (2014) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.*, **42**, D401–D407.
29. 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.
30. 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.
31. 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.
32. Huerta-Cepas,J., Dopazo,H., Dopazo,J. and Gabaldón,T. (2007) The human phylome. *Genome Biol.*, **8**, R109.
33. Huerta-Cepas,J., Dopazo,J. and Gabaldón,T. (2010) ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, **11**, 24.
34. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
35. Wallace,I.M., O'Sullivan,O., Higgins,D.G. and Notredame,C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.
36. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
37. 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.
38. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
39. Guindon,S., Dufayard,J.-F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
40. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
41. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
42. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
43. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
44. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
45. Letunic,I., Doerks,T. and Bork,P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.*, **43**, D257–D260.
46. Lipman,D.J., Lipman,D.J., Myers,E.W., Myers,E.W., Gish,W., Gish,W., Miller,W., Miller,W., Altschul,S.F. and Altschul,S.F. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
47. Kristensen,D.M., Waller,A.S., Yamada,T., Bork,P., Mushegian,A.R. and Koonin,E.V. (2013) Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.*, **195**, 941–950.
48. Li,J., Jia,H., Cai,X., Zhong,H., Feng,Q., Sunagawa,S., Arumugam,M., Kultima,J.R., Prifti,E., Nielsen,T. *et al.* (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, **32**, 834–841.
49. Sunagawa,S., Coelho,L.P., Chaffron,S., Kultima,J.R., Labadie,K., Salazar,G., Djahanschiri,B., Zeller,G., Mende,D.R., Alberti,A. *et al.* (2015) Structure and function of the global ocean microbiome. *Science*, **348**, 1261359.
50. Ku,C., Nelson-Sathi,S., Roettger,M., Sousa,F.L., Lockhart,P.J., Bryant,D., Hazkani-Covo,E., McInerney,J.O., Landan,G. and Martin,W.F. (2015) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*, **524**, 427–432.
51. Ciccarelli,F.D., Doerks,T., Mering,C., von Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
52. Huerta-Cepas,J. and Gabaldón,T. (2011) Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics*, **27**, 38–45.
53. Baker,C.R., Hanson-Smith,V. and Johnson,A.D. (2013) Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science*, **342**, 104–108.