OrthoDB in 2020: evolutionary and functional annotations of orthologs

Evgeny M. Zdobnov*, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Matthew Berkeley and Evgenia V. Kriventseva*

Department of Genetic Medicine and Development, University of Geneva Medical School, rue Michel-Servet 1, 1211 Geneva, Switzerland, and Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland

Received September 15, 2020; Revised October 12, 2020; Editorial Decision October 13, 2020; Accepted October 29, 2020

ABSTRACT

OrthoDB provides evolutionary and functional annotations of orthologs, inferred for a vast number of available organisms. OrthoDB is leading in the coverage and genomic diversity sampling of Eukaryotes, Prokaryotes and Viruses, and the sampling of Bacteria is further set to increase three-fold. The user interface has been enhanced in response to the massive growth in data. OrthoDB provides three views on the data: (i) a list of orthologous groups related to a user query, which are now arranged to visualize their hierarchical relations, (ii) a detailed view of an orthologous group, now featuring a Sankey diagram to facilitate navigation between the levels of orthology, from more finely-resolved to more general groups of orthologs, as well as an arrangement of orthologs into an interactive organism taxonomy structure, and (iii) we added a gene-centric view, showing the gene functional annotations and the pair-wise orthologs in example species. The OrthoDB standalone software for delineation of orthologs, Orthologer, is freely available. Online BUSCO assessments and mapping to OrthoDB of user-uploaded data enable interactive exploration of related annotations and generation of comparative charts. OrthoDB strives to predict orthologs from the broadest coverage of species, as well as to extensively collate available functional annotations, and to compute evolutionary annotations such as evolutionary rate and phyletic profile. OrthoDB data can be assessed via SPARQL RDF, REST API, downloaded or browsed online from https://orthodb.org.

INTRODUCTION

OrthoDB is a leading resource of evolutionary and functional annotations of orthologs (1). Orthology enables formulation of the most specific functional hypothesis for descendant genes (2,3), and therefore it is the preferred way for tentative characterization of genes in newly sequenced organisms. Orthology is also the cornerstone for comparative evolutionary studies. Initially coined for a pair of species, the term orthologs referred to genes that derive from a single gene in the last common ancestor (LCA) of the species (4). Generalization of the term to more than two species, however, required the LCA to be specified and led to the concepts of groups of orthologs or orthogroups (5) and levels of orthology (6,7). Although orthologs are clearly defined evolutionarily as descendant genes from a single gene of the LCA of the considered organisms, practical delineation of orthologs in extant species can be challenging as there are limitations to the accuracy with which gene and species genealogies can be reconstructed. Large-scale delineation of gene orthology is an even more demanding and challenging task, as there is a trade-off between accuracy and scalability of methods for making evolutionary inferences. The demand for automated methods for ortholog prediction is growing, and the challenges are exemplified by the numerous approaches proposed (6,8–16). Orthologer is the software behind the delineation of orthologs in OrthoDB that is freely available for standalone computations.

The concept of orthology is inherently hierarchical, as each phylogenetic clade or subclade of species has a distinct common ancestor, leading to more finely-resolved orthologs for more closely related species. OrthoDB has explicitly emphasized this since its inception (7), defining levels of orthology at each major radiation of the species taxonomy. Overcoming the original pair-wise thinking of orthology however is tangled and many find navigation
to the most relevant level of orthology confusing. This prompted us to enhance the OrthoDB web interface to improve usability and simplify access to the ever-growing database.

Evolutionary and functional annotations of orthologs, predicted across thousands of genomes from all forms of life including viruses, remains a key feature of OrthoDB. Functional annotations of genes were extensively collated from major resources including Uniprot (17) and NCBI (18) gene records, as well as InterPro (19) Gene Ontology (GO) (20). Notably, we collated and made searchable references such as to KEGG (21) pathways and to Online Mendelian Inheritance In Man (OMIM®) (22) linking to human diseases. All data were processed to assemble consolidated and non-redundant per-gene annotation records, presented as short one-line descriptions. Consequent aggregations to the orthogroup-level from the corresponding gene-level annotations allowed us to compile: (a) one-line orthologous group descriptors to concisely but precisely outline functional knowledge in a human-readable language as described in an earlier publication (23) whenever possible assign functional categories of COG (24), GO, KEGG pathways and the enzyme EC number. Such high-level functional descriptors are informative for comparative studies and metagenomics. Functional annotation of genes is complicated and may contain errors. Discordant annotations should be considered with caution, and OrthoDB makes such errors in the underlying data apparent.

Evolutionary annotations were computed for each orthologous group from the available genomics data and sequence alignment metrics. As detailed earlier in (23) metrics include: ‘phyletic profile’ that reflects gene universality (proportion of species with orthologs) and duplicability (proportion of multi-copy versus single-copy orthologs), ‘evolutionary rate’ that reflects the relative constraints on protein sequence conservation or divergence, and ‘sibling groups’ that reflects the sequence uniqueness of the orthologs. The universality of a gene family hints at a functional load that is widely necessary and basal, while lineage-restricted genes may underlie lineage-specific adaptations. Duplicability is also indicative of the type of molecular functions, e.g. members of a signal-transduction pathway or a protein complex may evolve under the single-copy control (25). ‘Evolutionary rate’ is a relative measure of sequence divergence, where faster-than-average genome evolution may indicate positive selection. The ‘sibling groups’ allow navigation to gene families possibly having similar molecular functions. These annotations, which provide an evolutionary perspective, remain unique to OrthoDB.

The OrthoDB resource is public, including both data and data processing software. The optional registration allows the authenticated users to upload their own proteomic data, for example from freshly sequenced genomes, for performing online BUSCO (26) analysis and for mapping to the current OrthoDB data. This enables the user to map existing functional annotations to the new genes, as well as to generate user-tailored comparative charts depicting the total gene count, the fraction of common genes, the fraction of the most conserved single-copy genes, etc.

COVERAGE OF ORGANISMS

In 2020, OrthoDB continues to provide the worldwide leading coverage of organisms. It notably includes Viruses and is set to further increase the sampling of Bacteria three-fold in release 11. The actual statistics of the analyzed organisms, their taxonomic relations, and the genome assembly accession numbers are shown and searchable under the ‘Advanced’ section of the user interface. The orthology levels, referring to the last common ancestors from which extant orthologs evolved, are defined according to the NCBI Taxonomy (27). Protein-coding gene translations are retrieved mostly from RefSeq and NCBI complete genomes (28). The number of sequenced genomes continues to grow exponentially, and it is important to uniformly cover the emerging diversity of sequences. This will not only facilitate studies of these organisms, but will also empower resolution of complex ortholog genealogies, with multiple gene duplications and losses. Sampling of species diversity was shown to be a major factor affecting the accuracy of inferred gene orthology, besides the quality of the underlying data. OrthoDB makes such errors in the underlying data apparent.

WEB INTERFACE

The OrthoDB user interface has been enhanced to respond to the massive growth in available data. There are now three views on the data: (i) a sorted list of ortholog groups resulting from free-text search against all identifiers and textual description records by user query, (ii) a detailed view of an ortholog group and (iii) a gene-centric view.

Since the last OrthoDB publication:

(i) the retrieved list of ortholog groups is now arranged to visualize their hierarchical relations. Previously, when a user searched by a keyword, an identifier or a complex query, OrthoDB returned a list of ortholog groups sorted only by textual relevance (i.e. a count of the number of query matches per group). The list remained unsorted if a single gene identifier pointed one gene in one organism, e.g. if a Uniprot id or NCBI gid was queried. Since even simple queries by a gene identifier return several ortholog groups corresponding to different levels of orthology, for example across Eukaryotes, vertebrates, mammals etc., such an unsorted list can be confusing. In the revised interface we opted for double-sorting, firstly by textual relevance, then by the taxonomic level, from the more general to the more specific. The hierarchy is visualized by a ladder-like shifted list of ortholog group banners. This results in a more biologically intuitive output when OrthoDB is queried by a particular gene identifier, for example via a URL coming from NCBI linkouts (Figure 1B).

(ii) we added a Sankey diagram to the detailed view of an ortholog group. This flow diagram shows the ortholog group within the hierarchy of the
Figure 1. Elements of OrthoDB user interface include: (A) a gene-centric view, providing available gene annotations and a list of pair-wise orthologs in example species; (B) a hierarchically arranged list, sorted by relevance of ortholog groups matching the user query; (C) detailed view of an ortholog group, facilitating the navigation between the levels of orthology with a Sankey flow diagram and (D) arranging the list of orthologs into an interactive organism taxonomy structure to focus only on relevant data; (E) the 'Advanced' section of the web interface enables user-tailored selection of organisms to focus on, specifying explicitly relevant levels of orthology, or phyloprofile filters.

(iii) we added a gene-centric view, providing a list of pair-wise orthologs in example species in addition to the available annotations. Although since its inception OrthoDB explicitly emphasized groups of orthologs and their hierarchical nature (7), the notion of pair-wise orthology (i.e. between two species) is commonly used, and we hope this gene-centric view will facilitate bridging these two concepts. The pair-wise orthology in example species is arranged by the phylogenetic distance to the organism in which the gene matching the query was found. These are consequently linked to the corresponding groups of orthologs at the closest level of orthology to enable navigation to orthologs in the other organisms.

ORTHOLGER SOFTWARE
Orthologer, the software used to delineate orthologs in OrthoDB, is freely available for standalone computations (see more from https://www.orthodb.org/orthodb_userguide.html#standalone-orthologer-software). It is based on pair-wise assessments of protein...
sequence homology between complete genomes using MMseqs2 (31) and subsequent clustering, as previously described in (23). The latest adjustment to Orthologer includes the addition of an optional mode to perform hierarchical computation of orthology guided by a user-provided organism taxonomy.

DATA AVAILABILITY
As for previous versions of OrthoDB we provide data files for bulk download, one file per level of orthology; as well as the underlying amino acid gene translations. To retrieve substantial subsets of data from OrthoDB or to access it programmatically we provide a REST API, documented at https://www.orthodb.org/orthodb_userguide.html#api, that returns data in JSON, FASTA or TAB formats. All data are distributed under the Creative Commons Attribution 3.0 License from http://www.orthodb.org/.

The RDF SPARQL interface was introduced in 2016 and it is gaining momentum. Adopting URIs of UniProt proteins and Ensembl genes, to be compatible with both UniProt and Ensembl SPARQL endpoints, it provides the possibility for very elaborate queries and a number of clickable links to Ensembl Genomes, NCBI, Interpro and GO resources. Users can also navigate to OrthoDB records by following links from FlyBase ‘Orthologs’ section, UniProt ‘Phylogenomic databases’ section, or NCBI ‘Additional links’/ Gene LinkOut’ section.

CONCLUSIONS AND PERSPECTIVES
The growing number of sequenced genomes increases the power of comparative analyses, but it also presents challenges regarding scalability of methods and data presentation to end users. OrthoDB strives to uniformly sample the available genomic space and to refine the accuracy of ortholog delineations.

ACKNOWLEDGEMENTS
We would like to thank all current and former members of the group, and in particular Dr Robert M. Waterhouse for his continuing support of OrthoDB and the user community.

FUNDING
University of Geneva, Swiss National Science Foundation [310030_189062 to E.Z.]; Swiss Institute of Bioinformatics SER1 (to E.Z. and E.K.).

Conflict of interest statement. None declared.

This paper is linked to: doi:10.1093/nar/gkaa1018.

REFERENCES
1. Kriventseva,E.V., Kuznetsov,D., Teglenfeldt,F., Manni,M., Dias,R., Simão,F.A. and Zdobnov,E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res., 47, D807–D811.
2. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. Science, 278, 631–637.
3. Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. Nat. Rev. Genet., 14, 360–366.
4. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. Syst. Biol., 19, 99–113.
5. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet., 39, 309–338.
6. van der Heijden,R.T., Snel,B., van Noort,V. and Huyten,M.A. (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics, 8, 83.
7. Kriventseva,E.V., Rahman,N., Espinosa,O. and Zdobnov,E.M. (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res., 36, D271–D275.
8. Fischer,S., Brunk,B.P., Chen,F., Gao,X., Harb,O.S., Iodice,J.B., Shannugam,D., Roos,D.S. and Stoeckert,C.J. (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr. Protoc. Bioinformatics, doi:10.1002/9780471250953.0801235.
9. Nakaya,A., Katayama,T., Itoh,M., Hiramuka,K., Kawashima,S., Moriya,Y., Okuda,S., Tanaka,M., Tokimatsu,T., Yamanishi,Y. et al. (2013) KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. Nucleic Acids Res., 41, D353–D355.
10. Huerta-Cepas,J., Capella-Gutiérrez,S., Puyet,E., Marquejou-Houben,M. and Gabaldón,F. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res., 42, D897–D902.
11. Sonnhammer,E.L.L. and Ostlund,G. (2015) InParanoid 8: orthology detection for evolutionary and functional annotations of orthologs. Nucleic Acids Res., 43, D232–D239.
12. Luchyanova,I., Miøara,M., Nishide,H. and Chiba,H. (2015) MIBD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Res., 43, D270–D276.
13. 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.
14. Galperin,M.Y., Kristensen,D.M., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2019) Microbial genome analysis: the COG approach. Brief. Bioinform., 20, 1063–1070.
15. Zdobnov,E.M., Teglenfeldt,F., Kuznetsov,D., Waterhouse,R.M., Simão,F.A., Ioannidis,P., Seppey,M., Lototscher,A. and Kriventseva,E.V. (2017) OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res., 45, D744–D749.
16. Altenhoff,A.M., Glover,N.M., Train,C.-M., Kaleb,K., Warwick, Vanetskyy, D., Duls, de Varies,T.M., Zikl,K., Stevenson,C., Long,J. et al. (2018) The OMIA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res., 46, D477–D485.
17. Consortium,T.U. (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res., 47, D506–D515.
18. Sawyer,E.W., Cavanagh,M., Clark,K., Osell,L., Pruitt,K.D. and Karsch-Mizrachi,I. (2019) GenBank. Nucleic Acids Res., 47, D94–D99.
19. Mitchell,A.L., Attwood,T.K., Babbitt,P.C., Blom,M., Bork,P., Bridge,A., Brown,S.D., Chang,H.-Y., El-Gebali,S., Fraser,M.I. et al. (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res., 47, D351–D360.
20. Gene Ontology Consortium, T. (2019) The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res., 47, D330–D338.
21. Kanehisa,M., Sato,Y., Furumichi,M., Morishima,K. and Tanabe,M. (2019) New approach for understanding genome variations in KEGG. Nucleic Acids Res., 47, D590–D595.
22. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res., 43, D789–D798.
23. Kriventseva,E.V., Teglenfeldt,F., Petty,T.J., Waterhouse,R.M., Simão,F.A., Podzynak,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.
24. 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.
25. Waterhouse, R.M., Zdobnov, E.M. and Kriventseva, E.V. (2011) Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. Genome Biol Evol, 3, 75–86.
26. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. and Zdobnov, E.M. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol., 35, 543–548.
27. Federhen, S. (2012) The NCBI Taxonomy database. Nucleic Acids Res., 40, D136–D143.
28. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., 44, D733–D745.
29. 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.
30. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S. and Phillippy, A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol., 17, 132.
31. Steinegger, M. and Söding, J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol., 35, 1026–1028.