Published online 5 November 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D807–D811***
doi: 10.1093/nar/gky1053

OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs

Evgenia V. Kriventseva[1,2,*], Dmitry Kuznetsov[1,2], Fredrik Tegenfeldt[1,2], Mosè Manni[1,2], Renata Dias[1,2], Felipe A. Simão[1,2] and Evgeny M. Zdobnov[1,2,*]

[1] Department of Genetic Medicine and Development, University of Geneva Medical School, rue Michel-Servet 1, 1211 Geneva, Switzerland and [2] Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland

Received September 15, 2018; Revised October 16, 2018; Editorial Decision October 17, 2018; Accepted October 29, 2018

ABSTRACT

OrthoDB (https://www.orthodb.org) provides evolutionary and functional annotations of orthologs. This update features a major scaling up of the resource coverage, sampling the genomic diversity of 1271 eukaryotes, 6013 prokaryotes and 6488 viruses. These include putative orthologs among 448 metazoan, 117 plant, 549 fungal, 148 protist, 5609 bacterial, and 404 archaeal genomes, picking up the best sequenced and annotated representatives for each species or operational taxonomic unit. OrthoDB relies on a concept of hierarchy of levels-of-orthology to enable more finely resolved gene orthologies for more closely related species. Since orthologs are the most likely candidates to retain functions of their ancestor gene, OrthoDB is aimed at narrowing down hypotheses about gene functions and enabling comparative evolutionary studies. Optional registered-user sessions allow on-line BUSCO assessments of gene set completeness and mapping of the uploaded data to OrthoDB to enable further interactive exploration of related annotations and generation of comparative charts. The accelerating expansion of genomics data continues to add valuable information, and OrthoDB strives to provide orthologs from the broadest coverage of species, as well as to extensively collate available functional annotations and to compute evolutionary annotations. The data can be browsed online, downloaded or assessed via REST API or SPARQL RDF compatible with both UniProt and Ensembl.

INTRODUCTION

Genomic sequencing is the most comprehensive method for the molecular interrogation of organisms, with the potential to reveal the complete repertoire of genes and enable the study of cellular processes at the molecular level. Homology, the recognition of gene sequence similarities as evidence of shared ancestry, allows for hypotheses on a gene's function when biological roles of related genes in other species are characterized. Homologs with a reference to a specific phylogeny radiation, i.e. descendants from a single gene of the last common ancestor, are termed orthologs and referred to below as ortholog groups or OGs (1,2). Such gene genealogies, pinned to particular ancestor genes, enable the most specific functional hypothesis for the descendant genes (3,4). Orthology is also the cornerstone for comparative evolutionary studies. The large-scale delineation of gene orthology is a popular but challenging task as evidenced by numerous proposed approaches (5–14).

OrthoDB is one of the largest resources of orthologs (15). Beyond the benchmarks of the underlying algorithm (15,16), the accuracy of our methodology earned its reputation through many comparative genomic studies (e.g. 17–19), particularly in the i5K initiative (20). The concept of orthologous groups is inherently hierarchical, as each phylogenetic clade or subclade of species has a distinct common ancestor; OrthoDB has explicitly emphasized this aspect since its inception (21). The ortholog delineation procedure is applied at each major radiation of the species taxonomy to produce more finely resolved groups of closely related species and to allow users to select the most relevant level.

OrthoDB provides tentative functional annotations of groups of orthologs and mapping to functional categories by summarizing functional gene annotations, extensively collected from other public resources. Annotation of genes is complicated and contains errors. Although in many cases OrthoDB makes such errors in the underlying data apparent, discordant annotations should be considered with caution. The evolutionary annotations of the orthologs remain another distinguishing feature of OrthoDB (Figure 1). In

[*] To whom correspondence should be addressed. Tel: +41 22 379 54 32; Fax:+41 22 379 57 06; Email: evgenia.kriventseva@unige.ch
Correspondence may also be addressed to Evgeny Zdobnov. Tel: +41 22 379 59 73; Email: evgeny.zdobnov@unige.ch

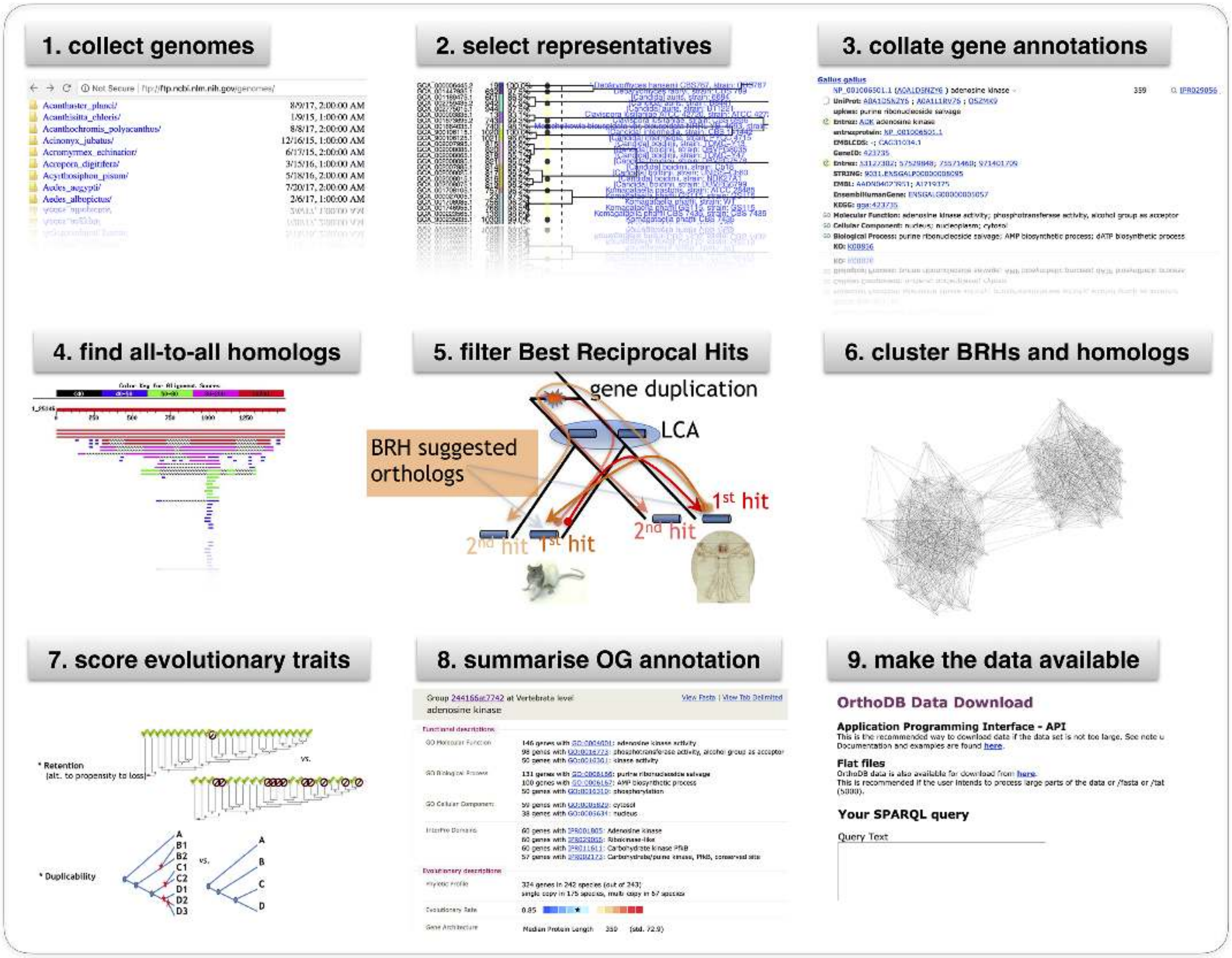
1. collect genomes
Not Secure | ftp://ftp.ncbi.nlm.nih.gov/genomes/
Acanthaster_planci/
Acanthisitta_chloris/
Acanthochromis_polyacanthus/
Acinonyx_jubatus/
Acromyrmex_echinatior/
Acropora_digitifera/
Acyrthosiphon_pisum/
Aedes_aegypti/
Aedes_albopictus/
2. select representatives
3. collate gene annotations
Gallus gallus
adenosine kinase
4. find all-to-all homologs
5. filter Best Reciprocal Hits
gene duplication
LCA
BRH suggested orthologs
1st hit
2nd hit
2nd hit
1st hit
6. cluster BRHs and homologs
7. score evolutionary traits
* Retention (alt. to propensity to loss)
vs.
* Duplicability
A
B1
B2
C1
C2
D1
D2
D3
vs.
A
B
C
D
8. summarise OG annotation
Group 244166at7742 at Vertebrata level
adenosine kinase
View Fasta | View Tab Delimited
Functional descriptions
GO Molecular Function
146 genes with GO:0004001: adenosine kinase activity
98 genes with GO:0016773: phosphotransferase activity, alcohol group as acceptor
50 genes with GO:0016301: kinase activity
GO Biological Process
131 genes with GO:0006166: purine ribonucleoside salvage
100 genes with GO:0006167: AMP biosynthetic process
50 genes with GO:0016310: phosphorylation
GO Cellular Component
59 genes with GO:0005829: cytosol
38 genes with GO:0005634: nucleus
InterPro Domains
60 genes with IPR001805: Adenosine kinase
60 genes with IPR029056: Ribokinase-like
60 genes with IPR011611: Carbohydrate kinase PfkB
57 genes with IPR002173: Carbohydrate/purine kinase, PfkB, conserved site
Evolutionary descriptions
Phyletic Profile
324 genes in 242 species (out of 243)
single copy in 175 species, multi copy in 67 species
Evolutionary Rate
0.85
Gene Architecture
Median Protein Length 359 (std. 72.9)
9. make the data available
OrthoDB Data Download
Application Programming Interface - API
This is the recommended way to download data if the data set is not too large. See note u
Documentation and examples are found here.
Flat files
OrthoDB data is also available for download from here.
This is recommended if the user intends to process large parts of the data or /fasta or /tat
(5000).
Your SPARQL query
Query Text

Figure 1. OrthoDB graphical abstract depicting the data processing pipeline.

this update (v10) we further increased the coverage of organisms, adjusted the underlying algorithm, and improved usability of the web interface.

COVERAGE OF ORGANISMS

A substantial fraction of widely inherited genes evolve under the single-copy control (22). These are the easiest to predict and are the basis for our BUSCO tool (23). The BUSCO acronym stands for Benchmarking Universal Single-Copy Orthologs. The software aims at quantitative assessment of completeness of genome assemblies, gene sets or transcriptomes, based on evolutionarily informed expectations of gene content, complementing the technical metrics like N50. Although we derive these BUSCO marker genes from OrthoDB, OrthoDB also strives to resolve the challenging cases of gene duplications and losses. Sampling of species diversity was shown to be a major factor affecting accuracy of inferred gene orthology, besides the quality of the underlying genomes and their annotations (16). OrthoDB thus strives to cover as much sequence diversity as practical with our computational resources. When there are multiple genomes available with greater than about 96% identity using MASH estimates (24), we sample the best annotated representatives with the most complete gene sets according to BUSCO metrics (25).

OrthoDB v10 now covers 1271 eukaryotes, 5609 bacteria, 404 archaea and 6488 viruses, as detailed in Table 1 (with the figures from the other resources as of September 2018). Overall, OrthoDB v10 covers over 37 million genes, classifying them into over 8.5 million tentative groups of orthologs at 624 levels of granularity. The orthology-levels, referring to the last common ancestors from which extant orthologs evolved, are defined according to the NCBI Taxonomy (26). Protein-coding gene translations for this release were retrieved mostly from RefSeq and NCBI complete genomes and the genome assembly ID is referenced now in the browsable taxonomy of organisms.

Table 1. Coverage of genomic diversity by the largest orthology resources

	OrthoDB	KEGG-OC	eggNOG	OMA
Eukaryota	1271	394	238	383
- Metazoa	448	n.a.	89	156
— Vertebrata	243	n.a.	51	71
— Arthropoda	170	n.a.	22	52
- Viridiplantae	117	n.a.	23	54
- Fungi	549	n.a.	85	107
Bacteria	5609	4301	1678[a]	1635
Archaea	404	253	115	149
Viruses	6488	0	352	0

[a] Plus 1655 additional bacteria were subsequently mapped.

THE ALGORITHM AND SOFTWARE

The OrthoDB computational pipeline for delineation of orthologs is based on assessments of pairwise gene homology between complete genomes and their subsequent clustering (Figure 1). The pipeline has previously been described (15), and our software is freely available from https://orthodb.org/software.

The latest adjustments to the OrthoDB algorithm include: (i) optimizations of the underlying data structures to reduce the memory usage, enabling us to increase the number of organisms covered, (ii) use of MMseqs2 (27) for homology searches to speed up the computations growing as square of the total number of genes, (iii) introduction of an additional species overlap criteria (6) for merging seed clusters to improve accuracy of inferred orthology and (iv) splitting of frequently mispredicted gene fusions to avoid spurious association of different orthologs.

FUNCTIONAL AND EVOLUTIONARY ANNOTATIONS

Gene functional annotations were collated from the major resources including Uniprot and NCBI gene records, as well as InterPro and Gene Ontology (GO). All data were processed to assemble consolidated and non-redundant per-gene annotation records, presented as short one-line descriptions that are click-expandable to immediately access the complete annotation record. The relative amount of available annotation data per gene is indicated by the size (one to five chevrons) of the click-expandable widget. Notably, we collated and made searchable references such as to KEGG pathways and to Online Mendelian Inheritance in Man (OMIM®) linking to human diseases.

OG functional annotations are consequent aggregations from their corresponding gene-level annotations, aiming to provide the user with an overview of the possible functions of the member orthologs. The compilation of one-line orthologous group descriptors to briefly but precisely outline functional knowledge in a human-readable language is a non-trivial task. We achieved this by identifying the best scoring single phrase found in any part of available annotation for all genes in the orthologous group. All these phrases were matched using a full-text search engine educated with a list of biological stop words, against a body of all annotation records of all genes in the group. This querying was performed separately for subsets of the annotations, partitioned according to data provenance. An empirically evaluated weight factor was used to choose the best phrases from each source, with the data source precedence as follows: Uniprot to ENSEMBL to NCBI to Interpro and then GO.

Functional categories of COG, GO, and KEGG pathways were assigned to OGs whenever possible. Such high-level functional descriptors are informative for comparative studies, e.g. in metagenomics.

Evolutionary annotations are computed for each OG from the available genomic data and sequence alignment scores. As detailed earlier (15), these 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 on a function that is widely necessary and basal, for example Orco is an essential co-receptor for insect odorant-sensing, while a lineage-restricted genes may underlie the lineage-specific adaptations, for example the *Drosophila* gland-specific peptide 26Ab. 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 (22), both of the examples above show this pattern (and the handful of 'duplications' or 'missing' of Orco actually point to deficiencies of the underlying genomes or their annotations). '*Evolutionary rate*' is a relative measure, where slower than genome average evolution may indicate stronger purifying selection, like in the case of the Orco gene, while the faster than average evolutionary rate may indicate positive selection, like in the case of the gland-specific peptide 26Ab. Although specific function of the 26Ab peptide is still unknown, this protein is transferred from male to female during mating and may act as key player in species-specific male reproductive success and thus, may be subject to rapid evolution resulting from sexual conflict and competition. The '*sibling groups*' allow navigation to gene families possibly having similar molecular functions. These annotations providing an evolutionary perspective remain unique to OrthoDB.

WEB INTERFACE

The OrthoDB resource is public. The optional registration allows the authenticated users to upload their own data for performing online BUSCO analysis and for mapping to current OrthoDB OGs, enabling the user to explore mapped functional annotations and 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. The growing number of available genomes may hamper the user experience while browsing the data. Therefore, by default, orthologs from only user-selected or reference species are shown, and users may choose to toggle a check-box to view orthologs from all available species.

CONCLUSION 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 informatively sample the available genomic space and to refine the accuracy of ortholog delineations.

DATA AVAILABILITY

As for the 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/v10/?page=api, that returns data in *JSON*, *FASTA* or *TAB* formats. All data are distributed under the Creative Commons Attribution 3.0 License from https://www.orthodb.org/.

The RDF SPARQL interface was introduced in the previous OrthoDB v9.1 and it is gaining momentum, being compatible with both UniProt and Ensembl SPARQL endpoints. Adopting Uniform Resource Identifier (URI) of UniProt proteins and Ensembl genes, 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.

ACKNOWLEDGEMENTS

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

FUNDING

University of Geneva and Swiss National Science Foundation Grant [31003A_166483 to E.Z.]; Swiss Institute of Bioinformatics SERI (to E.Z., E.K.); University of Geneva COINF (to E.Z., E.K.); São Paulo Research Foundation [17/15195–2 to R.O.D.]. Funding for open access charge: SIB.
Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
2. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
3. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
4. Gabaldon,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.
5. 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 Bioinformatics*, **8**, 83–95.
6. Fischer,S., Brunk,B.P., Chen,F., Gao,X., Harb,O.S., Iodice,J.B., Shanmugam,D., Roos,D.S. and Stoeckert,C.J. Jr (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinform.*, doi:10.1002/0471250953.bi0612s35.
7. Nakaya,A., Katayama,T., Itoh,M., Hiranuka,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–D357.
8. Huerta-Cepas,J., Capella-Gutierrez,S., Pryszcz,L.P., Marcet-Houben,M. and Gabaldon,T. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, D897–D902.
9. Sonnhammer,E.L. and Ostlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–D239.
10. Uchiyama,I., Mihara,M., Nishide,H. and Chiba,H. (2015) MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. *Nucleic Acids Res.*, **43**, D270–D276.
11. 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.
12. Galperin,M.Y., Kristensen,D.M., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2017) Microbial genome analysis: the COG approach. *Brief. Bioinform.*, doi:10.1093/bib/bbx117.
13. Zdobnov,E.M., Tegenfeldt,F., Kuznetsov,D., Waterhouse,R.M., Simao,F.A., Ioannidis,P., Seppey,M., Loetscher,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.
14. Altenhoff,A.M., Glover,N.M., Train,C.M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., de Farias,T.M., Zile,K., Stevenson,C., Long,J. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.
15. Kriventseva,E.V., Tegenfeldt,F., Petty,T.J., Waterhouse,R.M., Simao,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. 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.
17. Waterhouse,R.M., Kriventseva,E.V., Meister,S., Xi,Z., Alvarez,K.S., Bartholomay,L.C., Barillas-Mury,C., Bian,G., Blandin,S., Christensen,B.M. *et al.* (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, **316**, 1738–1743.
18. Bovine Genome,S., Analysis,C., Elsik,C.G., Tellam,R.L., Worley,K.C., Gibbs,R.A., Muzny,D.M., Weinstock,G.M., Adelson,D.L., Eichler,E.E. *et al.* (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, **324**, 522–528.
19. Hoy,M.A., Waterhouse,R.M., Wu,K., Estep,A.S., Ioannidis,P., Palmer,W.J., Pomerantz,A.F., Simao,F.A., Thomas,J., Jiggins,F.M. *et al.* (2016) Genome sequencing of the phytoseiid predatory mite metaseiulus occidentalis reveals completely atomized hox genes and superdynamic intron evolution. *Genome Biol. Evol.*, **8**, 1762–1775.
20. i, K.C. (2013) The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.*, **104**, 595–600.

21. 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.
22. 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.
23. Waterhouse,R.M., Seppey,M., Simao,F.A., Manni,M., Ioannidis,P., Klioutchnikov,G., Kriventseva,E.V. and Zdobnov,E.M. (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, **35**, 543–548.
24. 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**, 132–146.
25. Simao,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
26. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
27. Steinegger,M. and Soding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.