Published online 18 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D271–D275***
doi:10.1093/nar/gkm845

OrthoDB: the hierarchical catalog of eukaryotic orthologs

Evgenia V. Kriventseva[3], Nazim Rahman[1], Octavio Espinosa[1] and Evgeny M. Zdobnov[1,2,4,*]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, [2]Swiss Institute of Bioinformatics, 1 rue Michel-Servet, [3]Department of Structural Biology and Bioinformatics, University of Geneva Medical School, 1 rue Michel-Servet, 1211 Geneva, Switzerland and [4]Imperial College London, South Kensington Campus, SW7 2AZ London, UK

Received August 20, 2007; Revised September 24, 2007; Accepted September 25, 2007

ABSTRACT

The concept of orthology is widely used to relate genes across different species using comparative genomics, and it provides the basis for inferring gene function. Here we present the web accessible OrthoDB database that catalogs groups of orthologous genes in a hierarchical manner, at each radiation of the species phylogeny, from more general groups to more fine-grained delineations between closely related species. We used a COG-like and Inparanoid-like ortholog delineation procedure on the basis of all-against-all Smith-Waterman sequence comparisons to analyze 58 eukaryotic genomes, focusing on vertebrates, insects and fungi to facilitate further comparative studies. The database is freely available at http://cegg.unige.ch/orthodb

INTRODUCTION

Identification of orthologous genes is the cornerstone of comparative genomics, which is increasingly becoming an essential part of modern molecular biology. Functions of orthologous genes are often preserved through evolution, as by definition, orthologous genes descend by speciation from the common ancestor gene (1–3). Although the conservation of ortholog functions is not required or guaranteed, it is the most likely evolutionary scenario and provides a strong working hypothesis, particularly when the ortholog copy-number is preserved over a long period of time. Identification of orthologs is intricate as it assumes knowledge of ancestral state of the genes, and it requires knowledge of the complete gene repertoires. It is also complicated by gene duplication, fusion and exon shuffling, as well as pseudonization and loss, which make the problem particularly challenging with complex eukaryotic genomes. The fast growing number of available complete genomes facilitates a much better resolution of the gene genealogies, while at the same time greatly increasing the computational challenges.

There are two main approaches to delineate orthologous genes: (i) from reconciliation of gene trees with the species phylogeny and (ii) from classification of all-against-all sequence comparisons of complete genomes. The phylogeny approach takes advantage of well-studied evolution of conserved cores of globular proteins using quantitative models of amino acid substitutions (4–6). The notable examples of the tree-based approach to delineate orthologous genes are HOVERGEN (7) and TreeFam (8). The expert curation of the phylogenetic trees and the underlying multiple sequence alignments is both, advantageous, providing better accuracy and disadvantageous, limiting the comprehensiveness, homogeneity of quality and expandability to new species. Although given the appropriate data phylogenetic methods are likely to give more accurate models of ancestral sequences and therefore to yield more accurate orthology prediction, their applicability to current genomic data is hindered by several factors, most importantly: (i) they require substantially more computational resources, (ii) the reconciliation of gene and species trees relies on poorly quantified models of gene duplication and loss, and (iii) they are sensitive to completeness of predicted genes as the evolutionary models are designed for only well-conserved globular cores of proteins and missing data (gaps) render the approach inapplicable. The tree-based approaches also require the knowledge of the species phylogeny, and although the consensus on animal phylogeny seems to be close, it is still constantly challenged. The alternative approach of clustering orthologous genes on the basis of their whole-length similarity around Best-Reciprocal-Hits (BRHs, also known as SymBets, bi-directional BeTs and best–best hits, denoting sequences most similar to each other in between-genome comparisons) was first

*To whom correspondence should be addressed. Tel: +41 22 379 59 73; Fax: +41 22 379 57 06; Email: evgeny.zdobnov@medecine.unige.ch

introduced by the database of Clusters of Orthologous Groups (COGs) (9). Triggered by the earlier availability of much smaller and simpler bacterial genomes the database has quickly gained wide recognition and was later extrapolated to eukaryotic genomes (KOGs) (9). The identification of BRHs is widely adopted currently in the field of comparative genomics for its simplicity and feasibility of application to large-scale data. In terms of phylogenetic trees, BRHs could be interpreted as genes from different species with the shortest connecting path over the distance-based tree. The simplest application of this approach using BLAST (10) for interspecies comparisons suffers from inaccuracies of sequence distance estimates and ignores many gene duplications after the speciation that are, in fact, co-orthologs that are difficult to differentiate functionally. However, using these genes as anchors of orthologous groups in different species, additional co-orthologs can be identified as genes that are more similar to them in intra-genome comparisons than to any other gene in the other genomes, as popularized by the pairwise Inparanoid approach (11). There are also a few alternative clustering heuristics with varying compromises between specificity and selectivity (12) that focus on the growing number of available eukaryotic genomes such as the probabilistic approach of OrthoMCL (13) and the vertebrate-centric Ensembl-Compara (14).

Another important feature of orthology and paralogy classification, which is currently underappreciated, is that it is relative to a particular ancestor, as orthology of genes is defined by their descent from a common ancestor gene by speciation (1–3). Therefore, the more distantly related species are considered the more general (inclusive) orthologous groups become, because all lineage-specific duplications since this last common ancestor should be considered as co-orthologs. Inversely, orthologous groups become more fine-grained (more 1 : 1 relations) when closely related species are considered, as there was less time for gene duplications to occur. The concept of hierarchical orthologous groups has already prompted development of Levels of Orthology From Trees (LOFT) (15) tool to interpret the gene-trees in the context of species tree, COrrelation COefficient-based CLustering (COCO-CL) (16) methodology to refine clusters of homologous genes, and PHOG approach (17) to resolve orthology at each taxonomy node using explicit modeling of the ancestral sequences and relying on PHOG-BLAST (18) profile–profile comparisons.

Aiming to fuel comparative genomic studies we focused here on the most represented eukaryotic phyla, namely, we analyzed 23 fungi, 19 insect (plus one crustacean) and 15 vertebrate species with complete proteomes available (Table 1). For this analysis, we employed our own implementation of COG-like and Inparanoid-like ortholog identification procedures from all-against-all sequence comparisons across multiple species (19–22), and here we explicitly delineate the hierarchy of the orthologous groups, consistently applying the procedure to the sets of species with varying levels of relatedness according to the species tree (Figure 1).

METHODS

Orthology delineation

Groups of orthologous genes were automatically identified using a strategy employed previously (19–22) that is based on all-against-all protein sequence comparisons using the Smith-Waterman algorithm as implemented in ParAlign (23) with default parameters, followed by clustering of best reciprocal hits from highest scoring ones to 10^{-6} *e*-value cutoff for triangulating BRH or 10^{-10} cutoff for unsupported BRH, and requiring a sequence alignment overlap of at least 30 amino acids across all members of a group. Furthermore, the orthologous groups were expanded by genes that are more similar to each other within a proteome than to any gene in any of the other species, and by very similar copies that share over 97% sequence identity, which were identified initially using CD-Hit (24). We considered only the longest transcript per gene or the most common as specified in UniProt (25). The outlined procedure was first applied to all species considered, and then to each subset of species according to the radiation of the phylogenetic tree.

Phylogeny reconstruction

To guide computation of the ortholog hierarchy we produced the multiple alignment of concatenated single-copy orthologs, using well-aligned regions extracted with Gblocks (26) from individually aligned orthologous sequences using Muscle (27). This was used to compute the phylogenetic trees using the maximum-likelihood method as implemented in PHYML (28), employing the JTT model, a gamma correction with four discrete classes, and an estimated alpha parameter and proportion of invariable sites.

DATABASE CONTENT

Overview statistics

As detailed in Table 1 we analyzed 23 complete proteomes of fungal species, 19 insects and 15 vertebrates at different levels of the species phylogeny. Overall, this effort spans 870 737 genes, 82% of which have been classified into 10 876 orthologous groups in fungi, 19 835 in insects and 23 940 in vertebrates, providing the first systematic classification of the wealth of data that will provide the basis for further comparative evolutionary analyses.

WEB INTERFACE

The database is freely accessible from http://cegg.unige.ch/orthodb

Hierarchy of the orthologous groups

Orthology and paralogy classification is relative to the set of species considered [namely, to the particular ancestor (1–3)] and is more general (inclusive) for distantly related species, and more fine-grained (specific) for closely related species. We therefore delineated orthologous groups at each radiation node of the species phylogeny. To clearly

Table 1. Sets of covered complete proteomes

Lineage	Species name	Abbreviation	No. of genes*	Classified (%)	Source
Vertebrates	*Bos taurus*	Btar	21755	88	Ensembl v45 Jun2007
	Canis familiaris	Cfam	19305	94	Ensembl v45 Jun2007
	Danio rerio	Drer	24961	82	Ensembl v45 Jun2007
	Gasterosteus aculeatus	Gacu	20791	88	Ensembl v45 Jun2007
	Gallus gallus	Ggal	16736	83	Ensembl v45 Jun2007
	Homo sapiens	Hsap	22937	96	Ensembl v45 Jun2007
	Monodelphis domestica	Mdom	19520	91	Ensembl v45 Jun2007
	Macaca mulatta	Mmul	21944	90	Ensembl v45 Jun2007
	Mus musculus	Mmus	24496	87	Ensembl v45 Jun2007
	Ornithorhynchus anatinus	Oana	15723	87	Ensembl v45 Jun2007
	Pan troglodytes	Ptro	20965	97	Ensembl v45 Jun2007
	Rattus norvegicus	Rnov	22993	89	Ensembl v45 Jun2007
	Tetraodon nigroviridis	Tnig	28005	71	Ensembl v45 Jun2007
	Takifugu rubripes	Trub	21880	91	Ensembl v45 Jun2007
	Xenopus tropicalis	Xtro	18025	81	Ensembl v45 Jun2007
Insects**	*Aedes aegypti*	Aaeg	16789	89	AaegL1.1
	Anopheles gambiae	Agam	13133	87	AgamP3.45
	Apis mellifera	Amel	10330	87	GLEAN + curated_set
	Bombyx mori	Bmor	21302	48	SW_ge2k_BGF
	Culex pipiens	Cpip	23165	66	JCVI.CpipJ1.0_5
	Drosophila ananassae	Dana	22551	74	CAP freeze_20061030
	Drosophila erecta	Dere	16880	91	CAP freeze_20061030
	Drosophila grimshawi	Dgri	16901	87	CAP freeze_20061030
	Drosophila melanogaster	Dmel	13733	98	CAP freeze r4.3.FB
	Drosophila mojavensis	Dmoj	17738	84	CAP freeze_20061030
	Drosophila persimilis	Dper	23029	77	CAP freeze_20061030
	Drosophila pseudoobscura	Dpse	17328	91	CAP freeze_20061030
	Daphnia pulex	Dpul	30940	42	FrozenGC_2007_07_03
	Drosophila sechellia	Dsec	21332	81	CAP freeze_20061030
	Drosophila simulans	Dsim	17049	89	CAP freeze_20061030
	Drosophila virilis	Dvir	17679	86	CAP freeze_20061030
	Drosophila willistoni	Dwil	20211	77	CAP freeze_20061030
	Drosophila yakuba	Dyak	18816	86	CAP freeze_20061030
	Pediculus humanus	Phum	11206	82	TIGR.061807
	Tribolium castaneum	Tcas	16616	67	GLEAN + curated_set
Fungi	*Ashbya gossypii*	ASHG	4720	97	UniProt v12 Jul2007
	Aspergillus clavatus	ASPC	9120	95	UniProt v12 Jul2007
	Aspergillus fumigatus	ASPF	9629	95	UniProt v12 Jul2007
	Aspergillus oryzae	ASPO	12055	84	UniProt v12 Jul2007
	Aspergillus terreus	ASPT	10405	90	UniProt v12 Jul2007
	Candida glabrata	CANG	5180	95	UniProt v12 Jul2007
	Chaetomium globosum	CHAG	11040	78	UniProt v12 Jul2007
	Coccidioides immitis	COCI	10435	69	UniProt v12 Jul2007
	Cryptococcus neoformans	CRYN	6438	83	UniProt v12 Jul2007
	Debaryomyces hansenii	DEBH	6311	89	UniProt v12 Jul2007
	Encephalitozoon cuniculi	ENCC	1909	62	UniProt v12 Jul2007
	Kluyveromyces lactis	KLUL	5326	92	UniProt v12 Jul2007
	Lodderomyces elongisporus	LODE	5781	91	UniProt v12 Jul2007
	Magnaporthe grisea	MAGG	12685	71	UniProt v12 Jul2007
	Neosartorya fischeri	NEOF	10403	95	UniProt v12 Jul2007
	Neurospora crassa	NEUC	10076	75	UniProt v12 Jul2007
	Phaeosphaeria nodorum	PHAN	16451	59	UniProt v12 Jul2007
	Pichia guilliermondii	PICG	5919	93	UniProt v12 Jul2007
	Pichia stipitis	PICS	5797	96	UniProt v12 Jul2007
	Schizosaccharomyces pombe	SCHP	5008	88	UniProt v12 Jul2007
	Ustilago maydis	USTM	6546	78	UniProt v12 Jul2007
	Yarrowia lipolytica	YARL	6525	81	UniProt v12 Jul2007
	Saccharomyces cerevisiae	YEAS	6214	88	UniProt v12 Jul2007

*Only the longest transcript per gene was considered.
**Including one crustacean.

show the hierarchy level of the classifications and to allow easy navigation along the hierarchy we display the interactive species tree (Figure 1). The default level for an initial user query is set to fungi, arthropods or vertebrates and the level can be adjusted afterwards by selecting a radiation of interest on the phylogeny. Each result page provides a precompiled Bookmarklet, a snippet of JavaScript code that can be easily bookmarked in the user browser, to allow direct query to a particular phylogeny level.

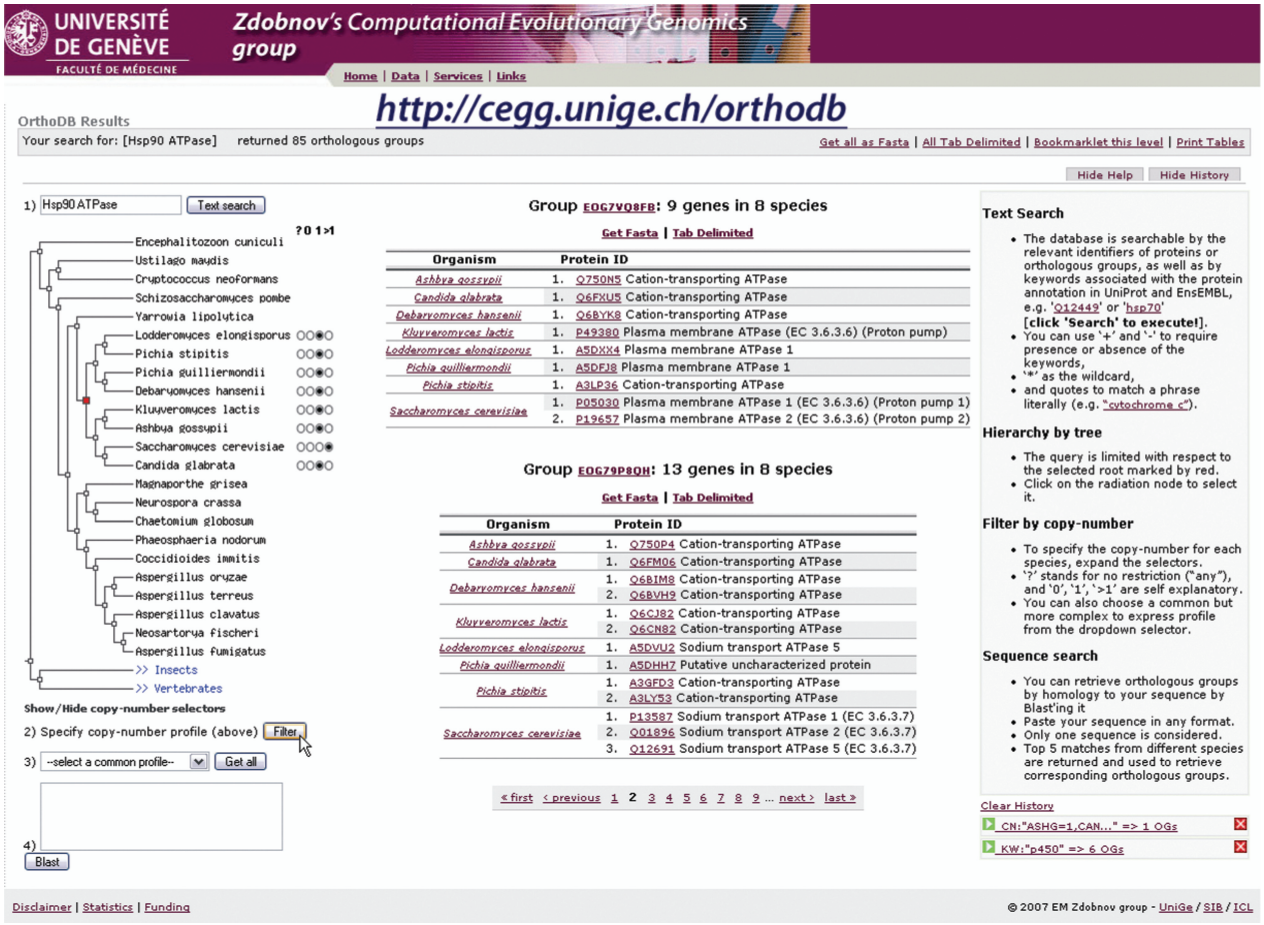

Figure 1. Example screenshot of the OrthoDB web interface (http://cegg.unige.ch/orthodb). The left panel enumerates the modes to query and browse the database by: (1) a keyword, (2) a user specified phylogenetic gene copy-number profile, (3) a common phylogenetic profile, or (4) sequence homology search; the middle panel is reserved for displaying results; and the right panel accommodates help and query history messages.

Stable identifiers

We assigned short identifiers using Noid utility to the generated orthologous groups that we will maintain unique across subsequent updates of the database to allow stable references to the data.

Querying by keywords

The database is searchable by the relevant identifiers used for the proteins or orthologous groups, as well as by keywords associated with the protein annotation in UniProt (25) or Ensembl (14). Currently the search is implemented as mySQL full-text index and the query is interpreted in Boolean mode that allows use of '+' and '−'operators to indicate that a word is required to be present or absent, respectively, for a match to occur; parentheses used to group words into subexpressions; '*' serves as the wildcard operator; and a phrase is matched literally if it is enclosed within quotes (e.g. 'cytochrome c'). The results always refer to the relevant orthologous groups, not separate genes.

Filtering by phylogenic profile

Another feature of the database interface is filtering orthologous groups by a phylogenic profile. This can be done by activating the set of selectors next to the phylogenetic tree and specifying the ortholog copy-number requirements in the species of interest, where '?' notation stands for no restriction ('any number') and '0', '1', '>1' are self explanatory. The 'Filter' button in the 'Specify copy-number profile' section will execute the corresponding query. In addition, we provide a set of precompiled queries for phylogenic profiles of common interest (via the selection list) that are more complicated to express, e.g. 'all but one' type: all single-copy orthologs but allowing for a loss or run-away in one of the species, or multigene orthologs in all but one species, etc. This allows viewing of the gene clusters that have undergone expansions or losses in the specific lineages, which is informative in the evolutionary context (29). These queries, as well as text search, are performed with respect to the selected speciation root, marked by red on the phylogenetic tree.

Query by sequence homology

Not all protein identifiers are widely known, particularly for automatically annotated genomes, and functional annotations for many genes are still anticipated. We therefore provide data querying by sample sequences,

e.g. a user submitted sequence is matched using Blast against the collected proteomes, and the top five matches from distinct proteomes are shown to the user and used to retrieve the associated orthologous groups, ranking by the number of hits to each group. Please note that if a sequence of an as yet unanalyzed species is used, the query will return the best matching ortholog cluster, however, this may not be sufficient to assume orthology.

Export of data

All results or particular groups can be retrieved as tab-delimited text or as Fasta formatted protein sequences with annotation of the orthologous group.

FUTURE PERSPECTIVES

All current approaches to identify orthologous groups of genes have different deficiencies and there are ways to improve their sensitivity and specificity. The implemented infrastructure in principal does not depend on the particular choice of the method, although our own implementation of a COG-like and Inparanoid-like ortholog identification procedure seems to produce reliable results according to extensive checks in the frame of our previous research projects. We plan also to test other available orthology delineation procedures.

ACKNOWLEDGEMENTS

We thank Thomas Junier for technical support, Dr Stefan Wyder, Dr Ivo Pedruzzi and Robert M Waterhouse for feedback and quality checks. We would like to acknowledge UniProt, Ensembl, Vectorbase, Flybase, the AAA fly genomes initiative, genome analysis consortiums and sequencing centers Agencourt, Baylor, BGI, Broad Institute, JCVI and JGI for the data. Foundation Giorgi-Cavaglieri and Swiss National Science Foundation are acknowledged for funding (SNF 3100A0-112588 to EMZ). Funding to pay the Open Access publication charges for this article was provided by SNF 3100A0-112588.

Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
3. Sonnhammer,E.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
4. Dayhoff,M.O. (1976) The origin and evolution of protein superfamilies. *Fed. Proc.*, **35**, 2132–2138.
5. Jones,C.H., Tatti,K.M. and Moran,C.P. Jr (1992) Effects of amino acid substitutions in the −10 binding region of sigma E from *Bacillus subtilis*. *J. Bacteriol.*, **174**, 6815–6821.
6. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
7. Duret,L., Mouchiroud,D. and Gouy,M. (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.*, **22**, 2360–2365.
8. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Heriche,J.K., Osmotherly,L., Li,R., Liu,T., Zhang,Z. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.
9. 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. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
10. 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.
11. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, 476–480.
12. 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.
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. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
15. 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.
16. Jothi,R., Zotenko,E., Tasneem,A. and Przytycka,T.M. (2006) COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, **22**, 779–788.
17. Merkeev,I.V., Novichkov,P.S. and Mironov,A.A. (2006) PHOG: a database of supergenomes built from proteome complements. *BMC Evol. Biol.*, **6**, 52.
18. Merkeev,I.V. and Mironov,A.A. (2006) PHOG-BLAST–a new generation tool for fast similarity search of protein families. *BMC Evol. Biol.*, **6**, 51.
19. Waterhouse,R.M., Kriventseva,E.V., Meister,S., Xi,Z., Alvarez,K.S., Bartholomay,L.C., Barillas-Mury,C., Bian,G., Blandin,S. *et al.* (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, **316**, 1738–1743.
20. Zdobnov,E.M. and Bork,P. (2007) Quantification of insect genome divergence. *Trends Genet.*, **23**, 16–20.
21. Zdobnov,E.M., von Mering,C., Letunic,I., Torrents,D., Suyama,M., Copley,R.R., Christophides,G.K., Thomasova,D., Holt,R.A. *et al.* (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science*, **298**, 149–159.
22. Consortium. (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, **443**, 931–949.
23. Saebo,P.E., Andersen,S.M., Myrseth,J., Laerdahl,J.K. and Rognes,T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.*, **33**, 535–539.
24. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
25. Consortium. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
26. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
27. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
28. Guindon,S., Lethiec,F., Duroux,P. and Gascuel,O. (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.*, **33**, W557–559.
29. Wyder,S., Kriventseva,E.V., Schröder,R., Kadowaki,T. and Zdobnov,E.M. (2007) Quantification of ortholog losses in insects and vertebrates. *Genome Biol.*, (in press).