

OPEN ACCESS

Full open access to this and thousands of other papers at <http://www.la-press.com>.

PhyloTreePruner: A Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for Phylogenomics

Kevin M. Kocot^{1,†}, Mathew R. Citarella^{2,3,†}, Leonid L. Moroz²⁻⁴ and Kenneth M. Halanych¹

¹Department of Biological Sciences, 101 Rouse Life Sciences, Auburn University, Auburn, USA. ²Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, Florida, USA. ³Department of Neuroscience and ⁴McKnight Brain Institute, University of Florida, Gainesville, Florida, USA. [†]These authors contributed equally to this work. Corresponding author email: kmkocot@auburn.edu

Abstract: Molecular phylogenetics relies on accurate identification of orthologous sequences among the taxa of interest. Most orthology inference programs available for use in phylogenomics rely on small sets of pre-defined orthologs from model organisms or phenetic approaches such as all-versus-all sequence comparisons followed by Markov graph-based clustering. Such approaches have high sensitivity but may erroneously include paralogous sequences. We developed PhyloTreePruner, a software utility that uses a phylogenetic approach to refine orthology inferences made using phenetic methods. PhyloTreePruner checks single-gene trees for evidence of paralogy and generates a new alignment for each group containing only sequences inferred to be orthologs. Importantly, PhyloTreePruner takes into account support values on the tree and avoids unnecessarily deleting sequences in cases where a weakly supported tree topology incorrectly indicates paralogy. A test of PhyloTreePruner on a dataset generated from 11 completely sequenced arthropod genomes identified 2,027 orthologous groups sampled for all taxa. Phylogenetic analysis of the concatenated supermatrix yielded a generally well-supported topology that was consistent with the current understanding of arthropod phylogeny. PhyloTreePruner is freely available from <http://sourceforge.net/projects/phyloreepruner/>.

Keywords: phylogenomic, orthology, paralogy, gene tree

Evolutionary Bioinformatics 2013:9 429–435

doi: [10.4137/EBO.S12813](https://doi.org/10.4137/EBO.S12813)

This article is available from <http://www.la-press.com>.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article published under the Creative Commons CC-BY-NC 3.0 license.



Introduction

Phylogenomics has been termed to describe the use of many concatenated orthologous sequences derived from genome or transcriptome data for phylogeny reconstruction.^{1,2} As the cost of high-throughput sequencing continues to decrease, phylogenomics is rapidly replacing target-gene (PCR-based) methods, as larger sets of data can be brought to bear on evolutionary questions while requiring less time in the laboratory, often at a lower cost. Several recent studies have demonstrated the utility of phylogenomics for addressing outstanding questions in organismal phylogeny.^{3–11}

As in any phylogenetic analysis, sequences being compared in a phylogenomic analysis must reflect the evolutionary history of the taxa of interest.¹² In other words, sequences employed in species tree reconstruction must be orthologous (see Reference 13 for review of terms related to orthology and paralogy). Several programs are available to *de novo* parse sequences into putatively orthologous groups for phylogenomics.^{13–19} Most of these programs rely on all-versus-all Basic Local Alignment Search Tool (BLAST) searches followed by one of several clustering methods. For example, OrthoMCL¹³ employs an all-versus-all BLAST²⁰ followed by a Markov clustering (MCL) algorithm in an effort to group orthologs separately from paralogs. Such graph-based methods are useful for identifying and grouping related sequences (ie, those with high sensitivity) but may erroneously group paralogs (ie, those with low specificity). However, comparative studies have shown that phylogenetic (tree-based) approaches have much higher specificity, particularly for detecting orthologs among distantly related taxa for which the phylogenomic approach is most often employed.^{21–23} Several studies^{24–26} have demonstrated that paralogs have the potential to mislead phylogenetic reconstruction.

The utility of a phylogenetic tree-based approach to refine orthology inferences made using graph-based or other methods was demonstrated in some recent phylogenomic studies. Briefly, in addition to other steps to help exclude paralogous sequences, Dunn et al³ made parsimony trees for all putatively orthologous groups identified by their initial orthology inference method (all-vs.-all BLASTP²⁰ followed by clustering with TribeMCL¹⁹) that contained 2 or more sequences per taxon for 1 or more taxa. All but 1 of the sequences

from the same taxon were deleted if they formed a monophyletic clade with bootstrap support of 80% or greater. Remaining groups that still had more than 1 sequence per taxon were then visually inspected and groups with evidence of paralogy were excluded. Otherwise, all sequences from the problematic taxa were excluded and the remainder of the group was retained for concatenation and analysis. Hejnlol et al⁴ also applied multiple filters to help exclude groups containing paralogous sequences followed by a phylogenetic tree-based screening approach. As in Dunn et al,³ in clades of sequences from the same taxon, all sequences but 1 were deleted (although in the case of Hejnlol et al⁴ this was done without consideration of the bootstrap support value). Next, “paralog pruning” was performed, in which the largest subtree that had no more than 1 sequence per taxon was identified and pruned away for further analysis. The remaining tree was retained for additional rounds of pruning and isolation of subtrees containing orthologous sequences until the remaining tree had no more than one sequence per taxon. Subtrees identified by this approach were then required to pass additional filtering criteria before being retained.

Given the greater accuracy of ortholog sets inferred by workflows including a phylogenetic approach versus a phenetic-only approach,^{21–23} we feel that a phylogenetic approach should be routinely incorporated into orthology inference for phylogenomics. However, no easy-to-use, standalone software that implements such an approach is currently available (but see the “treeprune” component of the unpublished Agalma pipeline; <https://bitbucket.org/caseywdunn/agalma>). Our purpose here is to provide a platform that allows automation of orthology inference using a phylogenetic approach.

Overview of PhyloTreePruner

We have developed PhyloTreePruner, an automated, phylogenetic tree-based utility that refines orthology inferences made using phenetic approaches (eg, all-versus-all BLAST and MCL clustering) following the general approaches of Dunn et al³ and Hejnlol et al.⁴ One important novel aspect of PhyloTreePruner is that it collapses poorly-supported nodes into polytomies in order to avoid unnecessarily discarding sequences in cases where a weakly-supported tree topology incorrectly suggests paralogy.

PhyloTreePruner screens single-gene trees and corresponding alignments for evidence of paralogy and produces a reduced alignment containing only sequences inferred to be strictly orthologous (Fig. 1). The user provides a fasta file containing the multi-sequence alignment and a Newick-format tree file generated from that alignment (Fig. 1A). First, nodes with support values below a user-selected cutoff value (0.50 in the example shown in Fig. 1) are collapsed into polytomies (Fig. 1B and C). Next, the largest subtree that meets the following criteria is identified and retained: if a taxon is represented by more than 1 sequence, all sequences from that taxon must form a monophyletic clade or be part of the same polytomy. One important difference between our approach and previous phylogenetic tree-based approaches is that polytomies with

sequences from 2 or more taxa are permitted. Tests of PhyloTreePruner showed that collapsing weakly supported nodes decreased the number of sequences unnecessarily deleted because a weakly supported tree topology incorrectly recovered orthologs as paralogs (eg, Fig. 2). Putative paralogs (sequences falling outside of the maximally inclusive subtree identified above) were then deleted from a copy of the input alignment produced by PhyloTreePruner (Fig. 1D and E).

Importantly, multiple sequences from the same taxon that form a clade in a gene tree are commonly observed in phylogenomic datasets. These sequences may represent a special case of paralogy often referred to as in-paralogy.¹⁴ In a gene tree, 2 or more sequences from the same taxon are in-paralogs if they were produced by 1 or more gene duplication

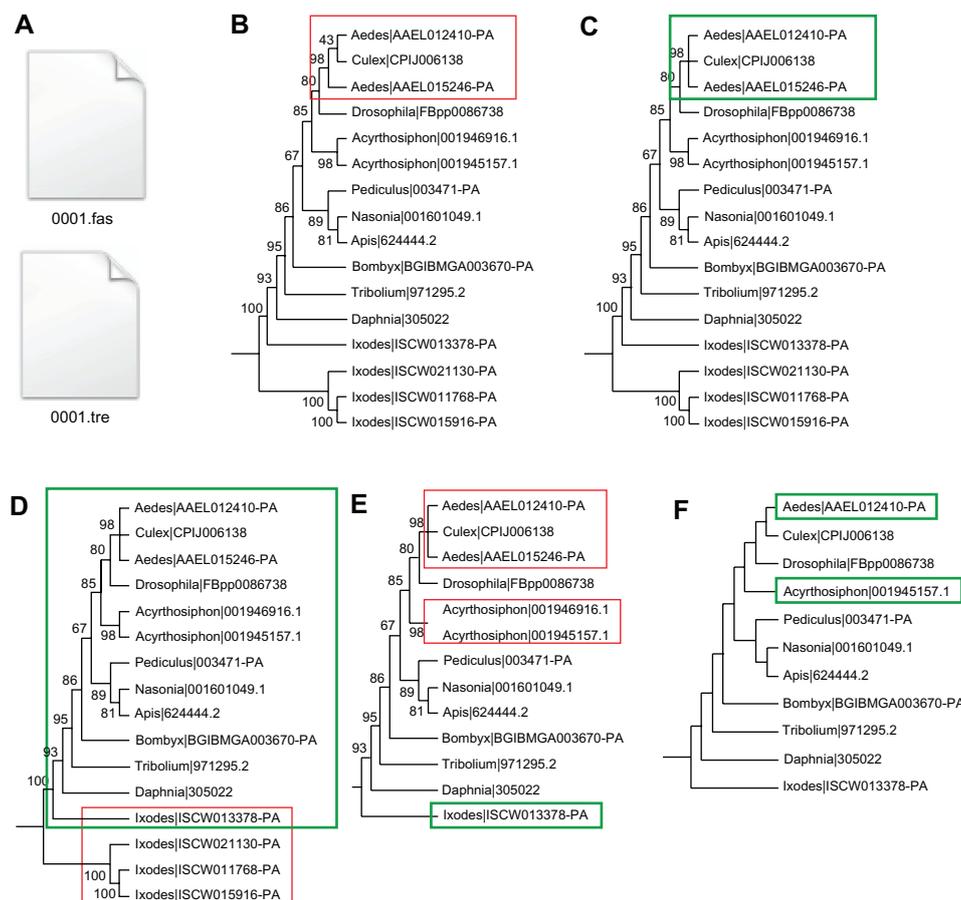


Figure 1. Illustration of the PhyloTreePruner tree-pruning algorithm. **(A)** PhyloTreePruner reads the single-gene tree and corresponding alignment file. **(B)** Nodes in the single-gene tree with support values below the user-defined threshold are identified (red box) and **(C)** collapsed into polytomies (green box). **(D)** PhyloTreePruner identifies the maximally inclusive subtree in which all taxa are represented by exactly one sequence, or, if there is more than one sequence from a taxon, these sequences form a monophyletic clade or are part of the same polytomy (green box). In this example, PhyloTreePruner identifies a potential paralogy issue with the Ixodes sequences (red box). This example shows the necessity of correct single-gene tree rooting. **(E)** PhyloTreePruner deletes sequences inferred to be paralogs from the tree and the corresponding sequence alignment file (red boxes). **(F)** In cases where more than one sequence remains from the same taxon, PhyloTreePruner selects the longest sequence and deletes all others (green boxes). This step can be skipped if preferred and another method (eg, SCaFoS) can be used to select the best sequence for each taxon.

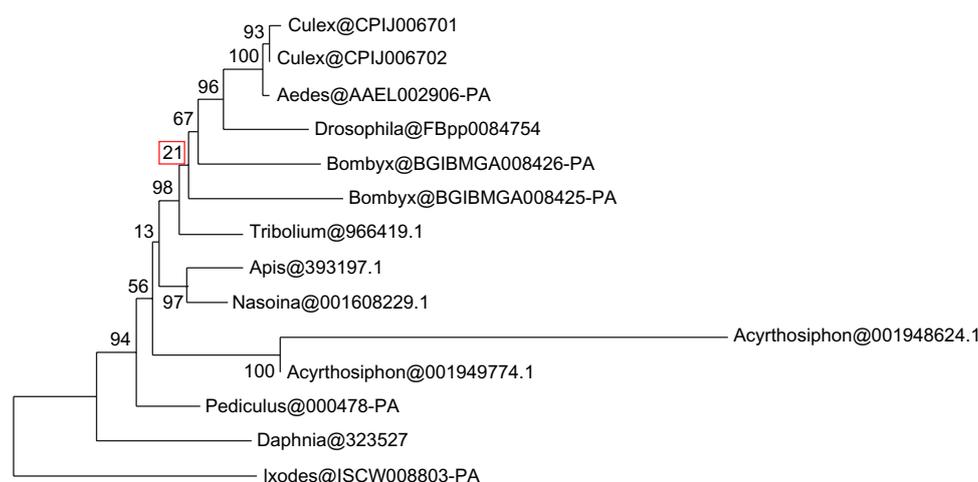


Figure 2. Example of a single-gene tree showing a weakly supported node (red box) that incorrectly recovers two sequences from the same taxon as paralogs. PhyloTreePruner collapses nodes with support values below a user-defined threshold and allows sequences from multiple taxa to be part of the same polytomy. Thus, PhyloTreePruner would “rescue” this group from being discarded if a minimum support value above 21 was used.

events that occurred after all speciation events documented on that tree. Because the gene duplication event occurred after all relevant speciation events, any in-paralog retained in the dataset used to generate the final species tree should result in the same reconstructed phylogeny (compare the trees in Fig. 1E and F). Alternatively, transcriptome data are commonly employed in phylogenomics and it is not uncommon to recover multiple splice variants of the same gene in a given transcriptome assembly. Therefore, in cases where multiple sequences from the same taxon formed a clade, all sequences but the longest (presumably the most complete splice variant) were deleted. Notably, this feature can be disabled and another program (eg, SCaFoS³¹) can be used to select the best sequence for each taxon using another metric (eg, pairwise distance). Because taxa with completely sequenced genomes were used in this example, in order to minimize missing data in the final matrix, any groups not retaining a sequence from all eleven taxa were discarded.

The final output produced by PhyloTreePruner is a reduced version of the input fasta file with “_pruned.fa” appended to its name. This file contains only sequences inferred to be orthologous among the sampled taxa (although in-paralogous sequences may be retained at the user’s discretion). Output files containing fewer than the minimum specified number of orthologous sequences are not produced. The original input trees and fasta files are not modified or deleted by PhyloTreePruner.

In order to demonstrate the utility of PhyloTreePruner for selection of orthologous groups of sequences, we assembled a dataset of protein-coding gene sequences derived from eleven arthropod taxa with completely sequenced genomes, made single-gene trees for each group, and applied the PhyloTreePruner pruning algorithm to these trees to identify and remove paralogous sequences.

Case Study: Arthropod Phylogenomics Approach

Translated gene models (predicted transcripts) from 11 arthropod genomes (Table 1) were downloaded from the InParanoid 7.0 database¹⁴ and an all-versus-all

Table 1. Taxon sampling.

Taxon	Species	Predicted transcripts
Chelicerata	<i>Ixodes scapularis</i>	20,486
Branchiopoda	<i>Daphnia pulex</i>	30,930
Insecta	<i>Acyrtosiphon pisum</i>	10,248
	<i>Aedes aegypti</i>	15,419
	<i>Apis mellifera</i>	9,093
	<i>Bombyx mori</i>	14,623
	<i>Culex pipiens</i>	18,883
	<i>Drosophila melanogaster</i>	14,076
	<i>Nasonia vitripennis</i>	9,163
	<i>Pediculus humanus</i>	11,194
	<i>Tribolium castaneum</i>	9,761

Note: All sequences were downloaded from InParanoid 7.0¹⁴ from the “processed sequences” directory.

BLASTP comparison was performed with an e-value cut-off of 0.00001. In order to cluster putative orthologs, OrthoMCL 2.0¹³ was employed using an inflation parameter of 2.1. We selected OrthoMCL because this software identifies more orthologous groups than most other graph-based orthology determination algorithms but suffers from low specificity (high false positive rate)²³ thus necessitating further refinement of the orthology inferences made by this method. PhyloTreePruner is also compatible with the popular orthology inference program HaMStR²⁷ but we caution that the “representative” option cannot be used as only one sequence will be selected for each taxon.

Resulting fasta files were then processed with a modified version of the bioinformatics pipeline used by Kocot et al.⁶ Here, groups that did not have at least one sequence from each taxon were discarded. Remaining groups were then aligned with MAFFT²⁸ (mafft—auto—localpair—maxiterate 1000). In order to remove ambiguously aligned and uninformative positions in the resulting alignments, trimming was performed with Gblocks²⁹ (Gblocks -t = p -p = n -b1 = number of sequences/2 -b2 = b1 -b3 = 8 -b4 = 2 -b5 = h). Any resulting alignments or sequences shorter than 100 AAs were then deleted. Finally, an “approximately maximum likelihood” tree was inferred for each group using FastTree 2³⁰ (FastTreeMP -slow -gamma).

Resulting single-gene trees and alignments were screened for paralogy with PhyloTreePruner as described above. The minimum number of sequences per file was set to 11 and nodes with support values below 70 were collapsed. In cases where 2 or more sequences were present for a taxon, only the longest splice variant or in-paralog was retained.

Groups of orthologous sequences identified by PhyloTreePruner were concatenated using FASconCAT³² and analyzed using maximum likelihood in RAxML 7.2.7³³ under the WAG+GAMMA+F model on the Auburn University Molette Lab SkyNet server. Nodal support was assessed using 100 bootstrap replicates. The tick *Ixodes* (Chelicerata) was used to root the resulting trees.

Results

OrthoMCL identified a set of 19,007 putatively orthologous groups. After all groups that did not have at least 1 sequence from each taxon were discarded, the

dataset was reduced to only 2,553 groups. After each alignment was trimmed with Gblocks and both alignments and individual sequences shorter than 100 AAs were deleted, 2,514 groups longer than 100 AAs that fulfilled the criterion of having at least one sequence from all taxa remained.

PhyloTreePruner further reduced this set to 2,027 orthologous groups with sequences from all 11 taxa. Of the 2,027 orthologous groups (OGs), 751 (37%) required pruning to exclude paralogs (including in-paralogs). The average number of sequences pruned from an OG was 1.81. Most (518) OGs only had 1 paralogous sequence removed, followed by 132 that had 2 paralogs removed, 51 that had 3 paralogs removed, 20 that had 4 paralogs removed, 15 that had 5 paralogs removed and 15 that had 6 or more paralogs removed. After concatenation, this data matrix was 863,121 amino acid positions in length with 458,480 distinct alignment patterns and only 6.24% missing data.

Phylogenetic analysis yielded a topology (Fig. 3) consistent with the current understanding of arthropod phylogeny.³⁴ Notably, support for Paraneoptera (*Acyrtosiphon* + *Pediculus*), a group generally recognized by morphologists, was very weak. However, difficulty in placing the louse *Pediculus* has been reported in previous molecular studies.³⁵

Applicability to Transcriptome Data

Our example employed completely sequenced genomes. However, taxa represented by significantly less data (eg,

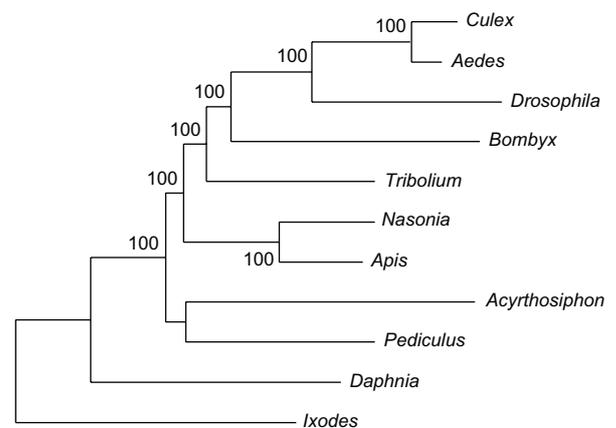


Figure 3. Phylogram of the most likely tree recovered in the RAxML analysis of the concatenated data matrix. The tick *Ixodes* was used to root the tree. Bootstrap support values above 50 are shown at each node. Scale bar = 0.05 substitutions per site. Notably, bootstrap support for Paraneoptera (*Acyrtosiphon* + *Pediculus*) was weak, consistent with the results of other phylogenomic studies.^{34,35}



moderately-sized Sanger expressed sequence tag libraries) may also be used. We caution that only sequences overlapping significantly with all other sequences in a single-gene alignment should be used for gene tree construction because very short sequences may only slightly or not at all overlap. A simple script deleting sequences that span less than 50% of the multiple sequence alignment (`remove_short_seqs.sh`, bundled with PhyloTreePruner) before tree reconstruction could be implemented to help avoid this problem.

Additionally, the minimum number of taxa sampled per gene can be decreased to allow as few taxa per orthology group as desired. In typical cases where transcriptome data are employed, most genes will not be sampled for all taxa. However, caution should be exercised when permitting very few taxa per group as paralogy is more likely to go undetected. As the cost of high-throughput sequencing technologies continues to decrease and more sequence data become available from a broader sampling of the tree of life, the severity of these problems will undoubtedly decrease.

Conclusions

PhyloTreePruner is a utility for improving orthology inferences made with graph-based or other methods using a phylogenetic tree-based approach. Wrapper scripts that automate multiple sequence alignment in MAFFT,²⁸ alignment trimming in Gblocks,²⁹ and the construction of single-gene trees using either FastTree²³⁰ or RAxML³³ are bundled with PhyloTreePruner. Gene trees and the corresponding alignments are analyzed by PhyloTreePruner and reduced alignments containing only sequences inferred as orthologs are produced. PhyloTreePruner can be configured to select the longest in-paralog/splice variant for each taxon or another program such as SCaFoS³¹ can be employed to select the best sequence for each taxon using another criterion (eg, lowest average pairwise distance). PhyloTreePruner can also be configured to automatically exclude groups with sequences from fewer than the desired minimum number of taxa.

PhyloTreePruner is implemented in Java and works on virtually all Unix-based operating systems. Source code, documentation, the example dataset described herein, and wrapper scripts to help automate dataset assembly and tree reconstruction are available from <http://sourceforge.net/projects/phyloreepruner/>.

Acknowledgements

We thank Damien Waits for help testing PhyloTreePruner and members of the Auburn University (AU) Molette Laboratory for providing comments that helped to improve an earlier version of this manuscript. We especially thank Dr. Scott Santos for his advice and access to the Molette Laboratory SkyNet server. This is AU Molette Laboratory contribution #19 and AU Marine Biology contribution #111.

Author Contributions

Conceived and designed the experiments: KMK. Wrote the software: MRC. Analyzed the data: KMK and MRC. Wrote the first draft of the manuscript: KMK. Contributed to the writing of the manuscript: KMK, MRC, KMH, and LLM. Agree with manuscript results and conclusions: KMK, MRC, KMH, and LLM. Jointly developed the structure and arguments for the paper: KMK, MRC, KMH, and LLM. Made critical revisions and approved final version: KMK, MRC, KMH, and LLM.

Funding

We acknowledge funding from NSF DEB-1210518 to KMK, NSF DEB-1036537 and NSF IOS-0843473 to KMH, McKnight Brain Research Foundation, University of Florida Opportunity Funds, NSF IOS-1146575, NSF CNS-0821622, NIH 5R01GM097502, R01MH097062, and NIH 5R21DA030118 to LLM, and NASA 11-EXO11-0127 to all authors.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Telford MJ. Phylogenomics. *Curr Biol*. 2007;17(22):R945–6.
2. Telford MJ. Resolving animal phylogeny: a sledgehammer for a tough nut? *Dev Cell*. 2008;14(4):457–9.
3. Dunn CW, Hejnol A, Matus DQ, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 2008;452(7188):745–9.
4. Hejnol A, Obst M, Stamatakis A, et al. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci*. 2009;276(1677):4261–70.
5. Struck TH, Paul C, Hill N, et al. Phylogenomic analyses unravel annelid evolution. *Nature*. 2011;471(7336):95–8.
6. Kocot KM, Cannon JT, Todt C, et al. Phylogenomics reveals deep molluscan relationships. *Nature*. 2011;477(7365):452–6.
7. Smith SA, Wilson NG, Goetz FE, et al. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*. 2011;480(7377):364–7.
8. Torruella G, Derelle R, Paps J, et al. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol Biol Evol*. 2012;29(2):531–44.
9. Capella-Gutiérrez S, Marcet-Houben M, Gabaldón T. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biol*. 2012;10:47.
10. Timme RE, Bachvaroff TR, Delwiche CF. Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One*. 2012;7(1):e29696.
11. Pick KS, Philippe H, Schreiber F, et al. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol*. 2010;27(9):1983–7.
12. Philippe H, Brinkmann H, Lavrov DV, et al. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 2011;9(3):e1000602.
13. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13(9):2178–89.
14. Ostlund G, Schmitt T, Forslund K, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res*. 2010;38(Database issue):D196–203.
15. Kim K, Kim W, Kim S. ReMark: an automatic program for clustering orthologs flexibly combining a Recursive and a Markov clustering algorithms. *Bioinformatics*. 2011;27(12):1731–3.
16. Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*. 2011;12:11.
17. Mahmood K, Webb GI, Song J, Whisstock JC, Konagurthu AS. Efficient large-scale protein sequence comparison and gene matching to identify orthologs and co-orthologs. *Nucleic Acids Res*. 2012;40(6):e44.
18. DeLuca TF, Cui J, Jung JY, St Gabriel KC, Wall DP. Roundup 2.0: enabling comparative genomics for over 1800 genomes. *Bioinformatics*. 2012;28(5):715–6.
19. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30(7):1575–84.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
21. Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*. 2007;2(4):e383.
22. Gabaldón T. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol*. 2008;9(10):235.
23. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*. 2009;5(1):e1000262.
24. Martin AP, Burg, TM. 2002. Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. *Syst Biol*. 2002;51(4):570–87.
25. Pirie MD, Vargas MP, Botermans M, Bakker FT, Chatrou LW. Ancient paralogy in the cpDNA trnL-F region in Annonaceae: implications for plant molecular systematics. *Am J Bot*. 2007;94(6):1003–16.
26. Struck TH. The impact of paralogy on phylogenomic studies—a case study on annelid relationships. *PLoS One*. 2013;8(5):e62892.
27. Ebersberger I, Strauss S, von Haeseler A. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol*. 2009;9:157.
28. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 2005;33(2):511–8.
29. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*. 2007;56(4):564–77.
30. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010;5(3):e9490.
31. Roure B, Rodriguez-Ezpeleta N, Philippe H. SCAFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol*. 2007;7 Suppl 1:S2.
32. Kück P, Meusemann K. FASconCAT: Convenient handling of data matrices. *Mol Phylogenet Evol*. 2010;56(3):1115–8.
33. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22(21):2688–90.
34. Trautwein MD, Wiegmann BM, Beutel R, Kjer KM, Yeates DK. Advances in insect phylogeny at the dawn of the postgenomic era. *Annu Rev Entomol*. 2012;57:449–68.
35. Letsch HO, Meusemann K, Wipfler B, Schütte K, Beutel R, Misof B. Insect phylogenomics: results, problems and the impact of matrix composition. *Proc Biol Sci*. 2012;279(1741):3282–90.