LETTERS

Broad phylogenomic sampling improves resolution of the animal tree of life

Casey W. Dunn[1]†, Andreas Hejnol[1], David Q. Matus[1], Kevin Pang[1], William E. Browne[1], Stephen A. Smith[2], Elaine Seaver[1], Greg W. Rouse[3], Matthias Obst[4], Gregory D. Edgecombe[5], Martin V. Sørensen[6], Steven H. D. Haddock[7], Andreas Schmidt-Rhaesa[8], Akiko Okusu[9], Reinhardt Møbjerg Kristensen[10], Ward C. Wheeler[11], Mark Q. Martindale[1] & Gonzalo Giribet[12,13]

Long-held ideas regarding the evolutionary relationships among animals have recently been upended by sometimes controversial hypotheses based largely on insights from molecular data[1,2]. These new hypotheses include a clade of moulting animals (Ecdysozoa)[3] and the close relationship of the lophophorates to molluscs and annelids (Lophotrochozoa)[4]. Many relationships remain disputed, including those that are required to polarize key features of character evolution, and support for deep nodes is often low. Phylogenomic approaches, which use data from many genes, have shown promise for resolving deep animal relationships, but are hindered by a lack of data from many important groups. Here we report a total of 39.9 Mb of expressed sequence tags from 29 animals belonging to 21 phyla, including 11 phyla previously lacking genomic or expressed-sequence-tag data. Analysed in combination with existing sequences, our data reinforce several previously identified clades that split deeply in the animal tree (including Protostomia, Ecdysozoa and Lophotrochozoa), unambiguously resolve multiple long-standing issues for which there was strong conflicting support in earlier studies with less data (such as velvet worms rather than tardigrades as the sister group of arthropods[5]), and provide molecular support for the monophyly of molluscs, a group long recognized by morphologists. In addition, we find strong support for several new hypotheses. These include a clade that unites annelids (including sipunculans and echiurans) with nemerteans, phoronids and brachiopods, molluscs as sister to that assemblage, and the placement of ctenophores as the earliest diverging extant multicellular animals. A single origin of spiral cleavage (with subsequent losses) is inferred from well-supported nodes. Many relationships between a stable subset of taxa find strong support, and a diminishing number of lineages remain recalcitrant to placement on the tree.

Expressed sequence tags (ESTs) provide opportunities to sample diverse genes from a large number of taxa[6]. Several recent phylogenomic studies, based largely on EST data, analysed matrices containing more than 140 genes from up to 34 metazoans (multicellular animals)[7–9]. However, the included species were not well sampled across extant metazoan diversity. These analyses also relied on either ribosomal proteins or a list of target genes identified from a small (1,152 ESTs) choanoflagellate data set[10], limiting the possibilities of EST studies to inform gene selection and homology assignment. Rather than look for predefined sets of genes in our data, we present an explicit procedure for gene selection (see Methods and Supplementary Fig. 2).

Our complete matrix includes data from 77 taxa (of which 71 are metazoans) and 150 genes. On average, taxa in our matrix include 50.9% of the 150 genes, and overall matrix completeness is 44.5%. Maximum likelihood (WAG model of sequence evolution; Figs 1 and 2) and bayesian (CAT[11] and WAG models of sequence evolution; Fig. 2) analyses of our matrix support the major groups of the 'new animal phylogeny'[2]. These groups have also been supported by other EST-based analyses[9], but not by phylogenomic studies that consider a small number of animal taxa[12]. Primary analyses of the 77-taxon matrix recover Metazoa, Bilateria and Protostomia with strong bootstrap support (>90%). This is an improvement compared to some previous phylogenomic studies that did not recover Protostomia, which in part led one study to conclude that it may not be possible to reconstruct the relationships of several major clades of animals because the metazoan radiation was too rapid[13]. It now seems that those findings were largely caused by limited taxon sampling, a result consistent with reanalyses[14]. Bootstrap support for Lophotrochozoa and Ecdysozoa is low in the 77-taxon consensus tree, but this is caused by the instability of a relatively small number of taxa (see below). Whereas Deuterostomia had poor support in recent phylogenomic analyses[15], in analyses of our 77-taxon matrix maximum likelihood bootstrap support for Deuterostomia is >80%. Within Deuterostomia, *Xenoturbella* was found to be sister to Ambulacraria (echinoderms and hemichordates) in a study that included 1,372 *Xenoturbella* ESTs[7]. Our inclusion of 3,840 additional *Xenoturbella* ESTs is consistent with this previous analysis (Figs 1, 2). None of our results are congruent with Coelomata, a group consisting of taxa that have a coelomic body cavity, which was favoured before molecular data became available. Coelomata has been recovered in some studies using many genes from a very small number of taxa[12,16], but it now seems clear that this is an artefact of poor taxon sampling.

Low-support values on consensus trees can be caused by large-scale structural rearrangements or by the instability of particular taxa. If, for instance, a taxon is only placed within a particular clade 50% of

[1]Kewalo Marine Laboratory, PBRC, University of Hawaii, 41 Ahui Street, Honolulu, Hawaii 96813, USA. [2]Department of Ecology and Evolutionary Biology, Yale University, PO Box 208105, New Haven, Connecticut 06520, USA. [3]Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive 0202, La Jolla, California 92093, USA. [4]Kristineberg Marine Research Station, Kristineberg 566, 450 34 Fiskebäckskil, Sweden. [5]Department of Palaeontology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK. [6]Ancient DNA and Evolution Group, Biological Institute, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark. [7]Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, California 95039, USA. [8]Zoological Museum, University of Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany. [9]Biology Department, Simmons College, The Fenway, Boston, Massachusetts 02115, USA. [10]Zoological Museum, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark. [11]Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024, USA. [12]Department of Organismic and Evolutionary Biology, [13]Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA. †Present address: Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman Street, Providence, Rhode Island 02912, USA.

the time, the support for that clade will be 50%, even if all other features of the tree are identical. This can obscure strongly supported relationships among stable taxa. We therefore used quantitative criteria to remove unstable taxa by calculating leaf stability indices[17], which measure the consistency of a taxon's position relative to other taxa across replicates, for all ingroup taxa (Fig. 1) and generated a new 64-taxon data set including only the most stable taxa (leaf stability, >90%). Some of the 13 unstable taxa (Entoprocta, Myzostomida, the sponge *Suberites domuncula* and the acoels) had poor gene sampling (Supplementary Tables 1 and 2, and Supplementary Fig. 3), which may simply provide too few informative characters for phylogenetic reconstruction. Acoels have also been found to be unstable in other phylogenomic studies[15]. Other unstable taxa (for example, Rotifera, Bryozoa and Gnathostomulida) had good gene sampling, suggesting that improved taxon sampling may be the most promising strategy for resolving their positions. Most unstable taxa moved between only a few positions (Supplementary Fig. 8), with most placed closer to Platyhelminthes than to other stable taxa, recovering with poor support a group known as Platyzoa[18]. Platyhelminths have relatively long branches, and it may

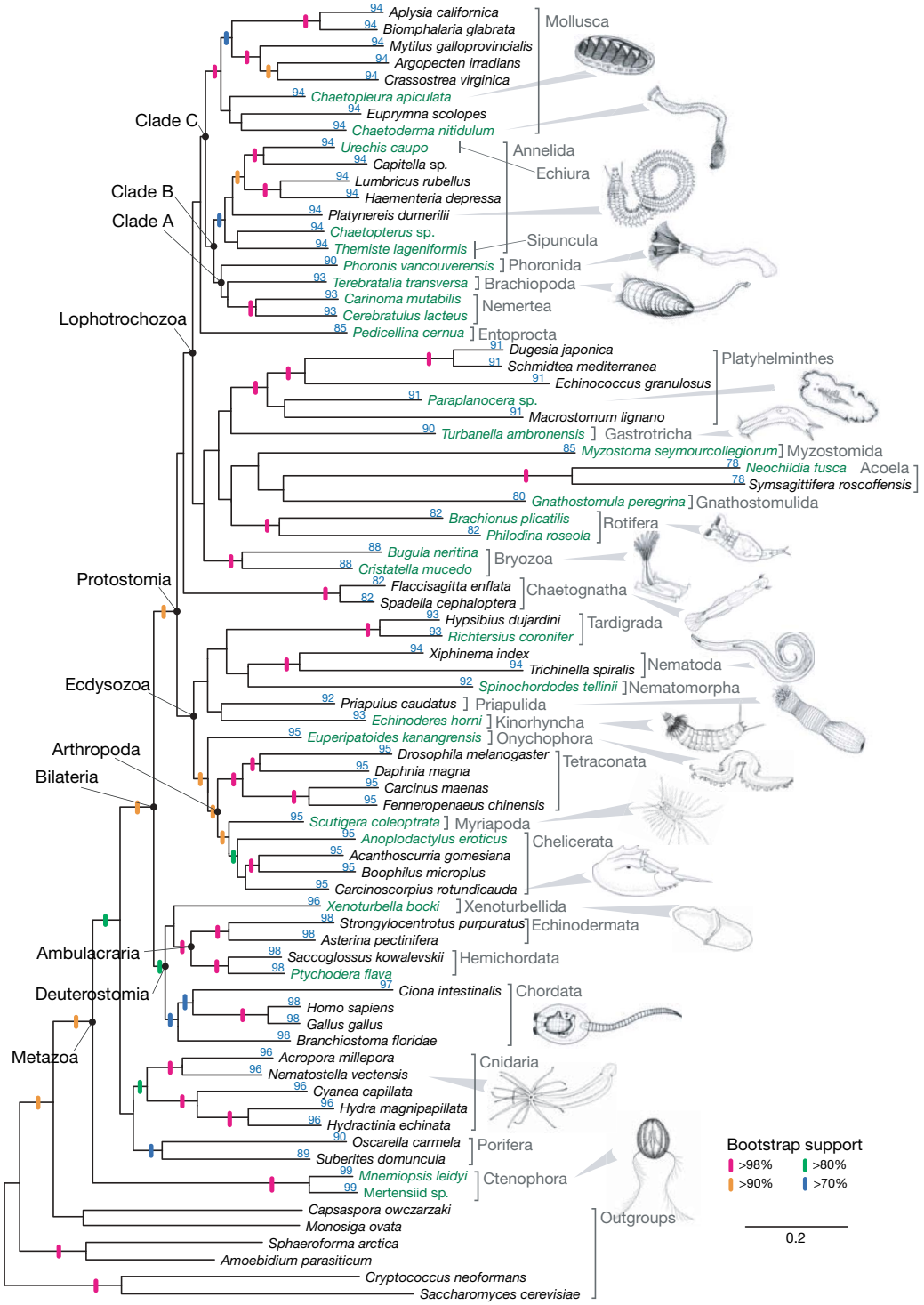


Figure 1 | Phylogram of the 77-taxon RaxML maximum likelihood analyses conducted under the WAG model. The figured topology and branch lengths are for the sampled tree with the highest likelihood (1,000 searches, log likelihood = –796,399.2). Support values are derived from 1,000 bootstrap replicates. Leaf stabilities are shown in blue above each branch. Taxa for which we collected new data are shown in green.

be that Platyzoa is an artefact of attracting unstable long-branch species to their vicinity.

Analyses of the 64-taxon matrix (Fig. 2 and Supplementary Fig. 9) show strong support for several important clades. To test if confidence in the relationships between stable taxa is overestimated in the absence of unstable taxa, we pruned away the 13 unstable taxa from each of the 1,000 bootstrap trees inferred from the 77-taxon matrix. This generated a set of trees containing only stable taxa, but for which relationships had been inferred in the presence of unstable taxa. Clade frequencies were calculated from this pruned tree set and mapped onto the most probable 64-taxon tree (Fig. 2). These reduced-tree support values are very similar to bootstrap support values calculated from the 64-taxon matrix, indicating that unstable taxa do not affect the inference of most relationships between stable taxa, only obscure these affinities.

The 64-taxon matrix strongly supports a sister-group relationship between Platyhelminthes and the remaining lophotrochozoans. A similar result, although uniting gastrotrichs with platyhelminths, was proposed recently[19]. Consistent with recent findings[20], *Urechis caupo*, an echiuran, is placed as sister to the annelid *Capitella* sp., and the sipunculan *Themiste lageniformis* is allied with annelids rather than molluscs. All analyses place Annelida as sister to a novel group that we call Clade A (Fig. 2), consisting of the nemerteans, a phoronid and a brachiopod, with variable support across analyses. Bayesian support for a group consisting of Annelida + Clade A (Clade B, Fig. 2) is strong (100% posterior probability in CAT and WAG analyses), whereas bootstrap support is moderate (84%). Although a brachiopod–annelid relationship is supported by the shared presence of chitinous chaetae, this new relationship implies that chaetae have been lost in nemerteans and phoronids (as in sipunculans, leeches

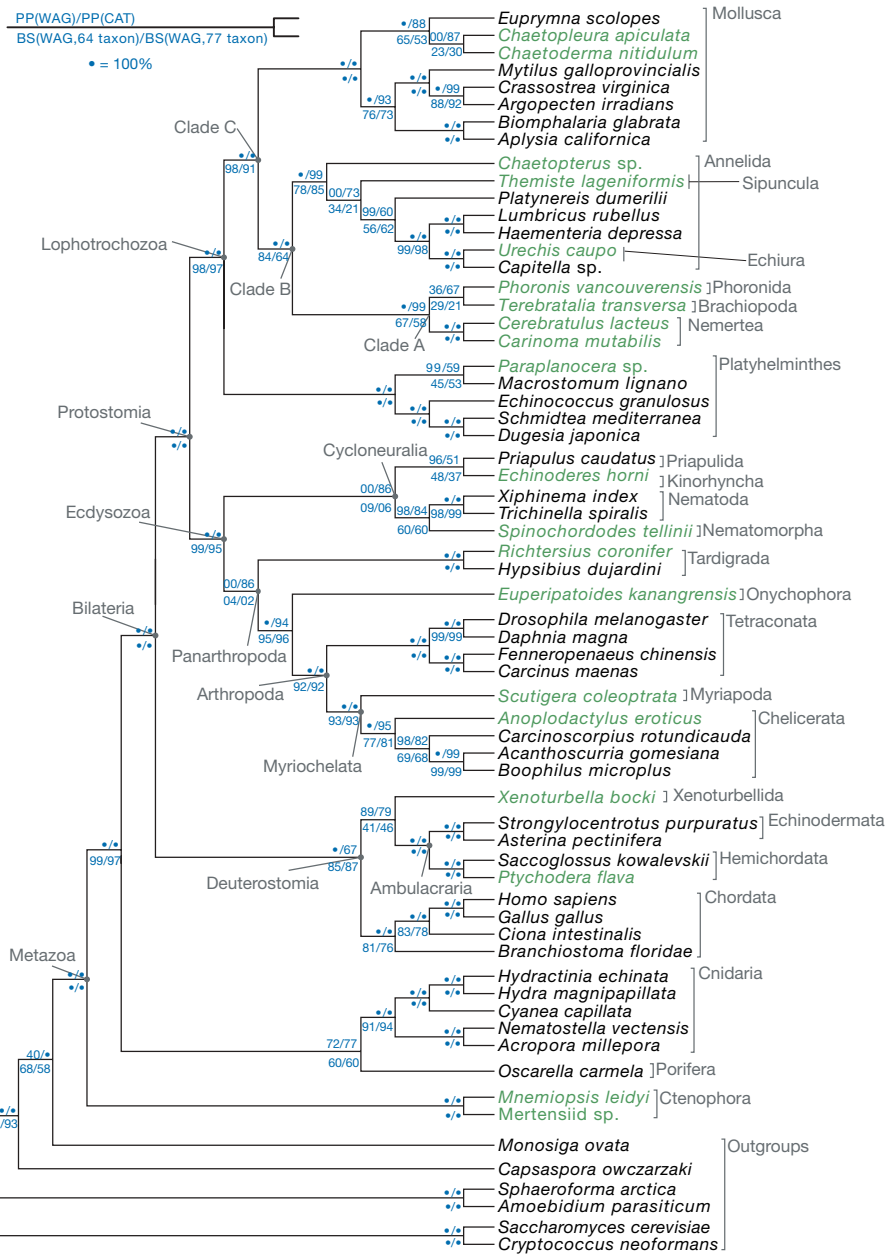


Figure 2 | Cladogram of the 64-taxon PhyloBayes bayesian analyses conducted under the CAT model. Posterior probabilities (PP) estimated under the CAT (15 PhyloBayes runs of 6,000 generations each; 1,200 generation burn-in) and WAG+I+Γ (8 MrBayes runs of two-million generations each; 125,000 generation burn-in; 4 chains per run) models. Maximum likelihood bootstrap support was calculated for the 64-taxon data set (2,000 replicate RaxML runs) and for the relationship of these 64 taxa in the 77-taxon analysis (by pruning all other taxa from the bootstrap replicates summarized in Fig. 1). Taxa for which we collected new data are shown in green. Support values, as specified at the top-left of the figure, are shown in blue.

and some other annelids). A monophyletic Mollusca, recovered here with significant support for the first time[21], is found to be sister to Clade B. Mollusca + Clade B (Clade C, Fig. 2) unites animals that produce chitinous chaetae with those that secrete $CaCO_3$ spicules and/or shells (that is, epidermal extracelluar formations for which secretory cells develop into a cup/follicle with microvilli at their base). A palaeontological scenario[22] identifies mollusc spicules and annelid/brachiopod chaetae as having been derived from distinctive fossil 'coelosclerites'. This scenario and a single origin of these epidermal formations are consistent with our cladogram.

The inclusion for the first time of nematomorphs, onychophorans and kinorhynchs in a phylogenomic analysis provides important insight into the structure of Ecdysozoa. Maximum likelihood bootstrap support for relationships within Ecdysozoa are similar in the 64- and 77-taxon analyses. The onychophoran is unambiguously placed as sister to arthropods in a clade of coelomate ecdysozoans that excludes Tardigrada, resolving a long-standing issue about the arthropods' sister group[5]. Tardigrades have traditionally been hypothesized to be allied with arthropods and onychophorans (together forming Panarthropoda)[23], but recent molecular data have suggested an alternative grouping of tardigrades with nematodes[9]. We find that the CAT model favours the former hypothesis (with Tardigrada sister to Onychophora + Arthropoda) whereas WAG favours the latter, indicating that at least one of these models is prone to systematic error for this particular problem (see Supplementary Information for further discussion of this issue).

We find strong support at all key internal arthropod nodes, and several contentious relationships of central interest are well resolved for the first time. Pycnogonids (sea spiders) group with chelicerates, rejecting placement of sea spiders as the earliest branching arthropod lineage[24]. Our results reject Mandibulata (Myriapoda, Crustacea and Hexapoda) in favour of myriapods being sister to chelicerates plus pycnogonids[25,26].

The spiral cleavage programme, a complex and highly stereotyped mode of early embryonic development, is present in at least Annelida, Entoprocta, Mollusca, Nemertea and Platyhelminthes[23], constituting a synapomorphy of at least the lophotrochozoan taxa included in the 64-taxon analysis. The placement of the lophophorate taxa Phoronida and Brachiopoda, which have radial cleavage and lie well within this assemblage, implies that they have lost spiral cleavage and also that their larvae are derived from the trochophore found in annelids, nemerteans and molluscs. Although phoronids do not show spiral cleavage, their mesoderm has a dual ecto/endodermal origin[27]—an important characteristic of spiralian embryology. Spiral cleavage has also been lost in cephalopod molluscs and in some neoophoran platyhelminths[23], establishing that this major shift has occurred repeatedly. Spiral cleavage may also have been lost or extensively modified in some of the unstable taxa not considered in the 64-taxon analysis (for example, gastrotrichs).

The placement of ctenophores (comb jellies) as the sister group to all other sampled metazoans is strongly supported in all our analyses. This result, which has not been postulated before, should be viewed as provisional until more data are considered from placozoans and additional sponges. If corroborated by further analyses, it would have major implications for early animal evolution, indicating either that sponges have been greatly simplified or that the complex morphology of ctenophores has arisen independently from that of other metazoans. Independent analyses of ribosomal and non-ribosomal proteins (Supplementary Information and Supplementary Fig. 10) indicate that support for this hypothesis (and for others presented for the first time here, such as Clade A and Clade B) is much greater in the combined analyses than in partitioned analyses with fewer genes. This may explain why these novel clades have not been recovered before, because support requires very broad gene sampling.

A few other principal groups have yet to be incorporated into phylogenomic studies, including Nemertodermatida, Loricifera, Cycliophora and Micrognathozoa. On the basis of our present findings, we predict that resolution across the metazoan tree will continue to improve as phylogenomic data from these additional taxa are collected and sampling is improved within clades already represented.

METHODS SUMMARY

Complementary DNA libraries were prepared for 29 species, and about 3,000 clones 5′ sequenced from each (Supplementary Table 1). All of our original sequence data have been deposited in the NCBI Trace Archive. These ESTs were assembled into a set of unique transcripts for each species, which were then translated into proteins using similarity and extension. Data from 48 additional species were downloaded from public archives (Supplementary Table 2). We present a new approach to identification of orthologous genes in animal phylogenomic studies (Supplementary Fig. 2) that relies on a Markov cluster algorithm[28,29] to analyse the structure of BLAST hits to a subset of the NCBI HomoloGene Database. The stringency of clustering is adjusted by means of the inflation parameter to best recapitulate the orthology groupings of HomoloGene.

Phylogenetic trees were inferred with bayesian and maximum likelihood approaches. The stabilities of taxa were assessed with leaf stabilities[17], as calculated by Phyutility[30] (available at http://code.google.com/p/phyutility/). Unstable taxa were removed from both sequence matrices and tree sets to assess the relationships of a stable subset of taxa to each other.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 September; accepted 20 December 2007.
Published online 5 March 2008.

1. Giribet, G. Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion? *Mol. Phylogenet. Evol.* **24,** 345–357 (2002).
2. Halanych, K. M. The new view of animal phylogeny. *Ann. Rev. Ecol. Evol. Sys.* **35,** 229–256 (2004).
3. Aguinaldo, A. M. A. *et al.* Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387,** 489–493 (1997).
4. Halanych, K. M. *et al.* Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science* **267,** 1641–1643 (1995).
5. Schmidt-Rhaesa, A. Tardigrades - Are they really miniaturized dwarfs? *Zool. Anz.* **240,** 549–555 (2001).
6. Philippe, H. & Telford, M. J. Large-scale sequencing and the new animal phylogeny. *Trends Ecol. Evol.* **21,** 614–620 (2006).
7. Bourlat, S. J. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444,** 85–88 (2006).
8. Delsuc, F. *et al.* Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439,** 965–968 (2006).
9. Philippe, H., Lartillot, N. & Brinkmann, H. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.* **22,** 1246–1253 (2005).
10. Philippe, H. *et al.* Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol. Biol. Evol.* **21,** 1740–1752 (2004).
11. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21,** 1095–1109 (2004).
12. Philip, G. K., Creevey, C. J. & McInerney, J. O. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. *Mol. Biol. Evol.* **22,** 1175–1184 (2005).
13. Rokas, A., Kruger, D. & Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. *Science* **310,** 1933–1938 (2005).
14. Baurain, D., Brinkmann, H. & Philippe, H. Lack of resolution in the animal phylogeny: closely spaced cladogenesis or undetected systematic errors? *Mol. Biol. Evol.* **24,** 6–9 (2006).
15. Philippe, H. *et al.* Acoel flatworms are not Platyhelminthes: evidence from phylogenomics. *PLoS One* **2,** e717 (2007).
16. Blair, J. E. *et al.* The evolutionary position of nematodes. *BMC Evol. Biol.* **2,** 1–7 (2002).
17. Thorley, J. L. & Wilkinson, M. Testing the phylogenetic stability of early tetrapods. *J. Theor. Biol.* **200,** 343–344 (1999).
18. Giribet, G., Distel, D. L., Polz, M., Sterrer, W. & Wheeler, W. C. Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. *Syst. Biol.* **49,** 539–562 (2000).
19. Telford, M. J., Wise, M. J. & Gowri-Shankar, V. Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the Bilateria. *Mol. Biol. Evol.* **22,** 1129–1136 (2005).
20. Struck, T. H. *et al.* Annelid phylogeny and the status of Sipuncula and Echiura. *BMC Evol. Biol.* **7,** 57 (2007).

21. Giribet, G. *et al.* Evidence for a clade composed of molluscs with serially repeated structures: monoplacophorans are related to chitons. *Proc. Natl Acad. Sci. USA* **103,** 7723–7728 (2006).
22. Conway Morris, S. & Peel, J. S. Articulated Halkieriids from the Lower Cambrian of North Greenland and their role in early protostome evolution. *Phil. Trans. R. Soc. Lond. B* **347,** 305–358 (1995).
23. Nielsen, C. *Animal Evolution, Interrelationships of the Living Phyla* 2nd edn (Oxford Univ. Press, Oxford, 2001).
24. Giribet, G., Edgecombe, G. D. & Wheeler, W. C. Arthropod phylogeny based on eight molecular loci and morphology. *Nature* **413,** 157–161 (2001).
25. Mallatt, J. M., Garey, J. R. & Shultz, J. W. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol. Phylogenet. Evol.* **31,** 178–191 (2004).
26. Hwang, U. W. *et al.* Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature* **413,** 154–157 (2001).
27. Freeman, G. & Martindale, M. Q. The origin of mesoderm in phoronids. *Dev. Biol.* **252,** 301–311 (2002).
28. van Dongen, S. A cluster algorithm for graphs. *National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam.* Technical Report INS-R0010 (Stichting Mathematisch Centrum, Amsterdam, 2000).
29. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
30. Smith, S. A. & Dunn, C. W. Phyutility: a phyloinformatics tool for trees, alignments, and molecular data. *Bioinformatics* doi:10.1093/bioinformatics/btm619 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank all participants in the Protostome Assembling the Tree of Life (AToL) Project as well as E. J. Edwards, T. Dubuc, A. Stamatakis, J. Q. Henry and S. Maslakova. A.H. received support from the Deutsche Forschungsgemeinschaft, and M.O. received support from the Swedish Taxonomy Initiative and the Royal Swedish Academy of Sciences. The *Capitella* sp. EST data were produced by the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/Capitella), as were the *Mnemiopsis* dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) data. This work was funded by two consecutive collaborative grants from the AToL program from the US National Science Foundation. Ctenophore sequencing was supported by NASA.

Author Information The concatenated sequence matrix has been deposited at TreeBase (http://www.treebase.org). The raw sequence data are available at the NCBI Trace Archives (http://www.ncbi.nlm.nih.gov/Traces), and can be retrieved with the query 'center_name='KML-UH''. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.W.D. (casey_dunn@brown.edu).

METHODS

Molecular techniques. Total RNA was prepared using TRIzol (Molecular Research Center), the RNeasy Mini Kit (Qiagen), the RNAqueous-micro kit (Ambion) or Dynabeads (Invitrogen) from fresh specimens or tissue that had been stored in RNA*later* (Ambion) at –20 °C. First-strand cDNA was synthesized using the GeneRacer Kit (Invitrogen), which selects for full-length mRNA. Twenty cycles of PCR with the GeneRacer 5′ and 3′ primers were then performed (94 °C for 30 s, 69 °C for 30 s, and 72 °C for 4 min, with an initial denaturation of 94 °C for 5 min and a final extension of 72 °C for 10 min; BD Advantage 2 Polymerase Mix, Clontech). The PCR products of most taxa were enriched for larger fragments using ChromaSpin TE400 columns (Clontech). PCR products were concentrated with the MinElute PCR Purification Kit (Qiagen) and ligated into pGEM-T Easy (Promega). The ligations were sent to Macrogen Ltd for transformation, plating, colony picking, minipreping, and 5′ sequencing with the GeneRacer 5′ primer. All of our original sequence data have been deposited in the NCBI Trace Archive.

Sequence preprocessing. The PartiGene Pipeline v3.0 (ref. 31) was used to preprocess EST data, with several modifications (Supplementary Fig. 2). The option to use quality data for assembly was enabled. Partigene outputs multiple contiguous sequences for a given transcript when PHRAP (http://www.phrap.org/) does not fully assemble the sequences assigned to a transcript. Low-quality ends were trimmed from these partially assembled sequences, which were then aligned with ClustalW[32] and the highest-quality bases chosen for the consensus. Transcripts were translated by similarity and extension (using the SwissProt database).

The 2,137 *Xenoturbella bocki* sequences from dbEST were assembled along with the 3,840 new sequences that we generated. The 3,360 ESTs we prepared from *Mnemiopsis leidyi* were also combined with data from dbEST that had been generated by the US Department of Energy Joint Genome Institute. In addition, we considered 48 taxa from other publicly available sources (Supplementary Table 2).

Orthology assignment. We developed an explicit method for selecting genes from EST data sets to maximise gene intersection across taxa and to minimise problems with orthology and paralogy (Supplementary Fig. 2). Promiscuous domains (Conserved Domain Database[33] accession numbers pfam01535, pfam00400, pfam00047, smart00407, cd00099, pfam00076, pfam00023, pfam01576, pfam00041, cd00031, smart00112, cd00096, cd00204, pfam00023, smart00248, pfam01344, pfam00018, pfam00038, pfam00096, pfam00595, pfam00651, pfam00169, pfam00105, pfam00435, pfam00084, pfam00017, smart00225, smart00367, smart00135, cd00020, pfam00514, cd00020, smart00185, cd00014, pfam00307 and smart00033) were identified by RPSBLAST and masked before orthology assignment. These domains are a subset of those masked in the construction of NCBI KOG database of eukaryotic orthologues[34]. We constructed a local database of all *Homo sapiens*, *Canis familiaris*, *Gallus gallus*, *Drosophila melanogaster* and *Anopheles gambiae* sequences that have orthology assignments in the National Center for Biotechnology Information (NCBI) HomoloGene database, and the masked sequences were queried against these sequences with BLASTP. BLASTP hits were then passed to TribeMCL (the version bundled with mcl v6.58) for Markov Chain Clustering (MCL)[29,35]. The MCL inflation parameter was varied in intervals of 0.1 to identify the value that generated the maximum number of clusters with sequences from one HomoloGene group.

Groups with sequences from fewer than 25 taxa were discarded. We also discarded groups with sequences from fewer than 5 of the taxa we collected original EST data for to prevent gene selection from being dominated by some of the much larger EST and genomic data sets included from public archives. The number of sequences for each taxon represented within each group was then enumerated, and groups with a median of greater than one or a mean greater than 2.5 were discarded. This eliminated many groups that had a high rate of lineage-specific duplication. Two features of the cluster graph were then evaluated for properties potentially indicative of paralogy problems. First, the group was rejected if it included no Homologene sequences. Second, the TribeMCL group was rejected if it included any Homologene sequences belonging to a Homologene group with sequences in another TribeMCL group.

Most TribeMCL groups contained multiple sequences for some taxa, which could be paralogues, splice variants or the result of EST assembly errors. The sequences for each of these problematic TribeMCL groups were aligned with ClustalW v1.83 (ref. 32), and parsimony trees (100 bootstrap replications) were inferred with PAUP* v4.0b10 (ref. 36). All but one of the sequences from the same taxon were automatically excluded from the group if they were monophyletic with a bootstrap score of >80%. The retained sequence was selected to have a stop codon if possible. Trees for TribeMCL groups that still had taxa with multiple sequences were then visually inspected. If there were strongly supported deep nodes indicating the existence of multiple paralogues shared by multiple taxa the entire group was excluded. Otherwise, all sequences for the problematic taxa were excluded from the group and sequences from nonproblematic taxa retained.

All groups that passed the above criteria were prepared for tree building. 5′ untranslated regions were removed by blasting each sequence against the other sequences in the same group and trimming ends that were not included in the resulting HSPs (10^{-4} *e*-value threshold). The sequences of each TribeMCL group were aligned with Muscle v3.6 (ref. 37) and trimmed with Gblocks v0.91b[38] (settings: –b2 = [65% of the number of sequences] –b3 = 10 –b4 = 5 –b5 = a). These trimmed alignments for each gene were then concatenated into a single alignment (21,152 positions long), which has been deposited in TreeBase.

To compare matrix construction methods between studies, sequences were queried by BLASTP (10^{-20} *e*-value threshold) against the sequences of the most frequently used matrix of genes in metazoan EST studies[9]. The identity of the top-scoring hit, if any hits were found, was putatively assigned to the query sequence. Alignment and trimming were executed as described above, and the least-divergent sequences were assembled into a matrix (24,708 positions long) with SCaFoS[39].

Phylogenetic analyses. Phylogenetic analysis of our large matrix was computationally intensive and took several months on more than 120 processors spread across multiple modern computer clusters. A preliminary matrix was evaluated under a mixed model with MrBayes v.3.1.2 (ref. 40), which selected WAG with 100% posterior probability. Maximum likelihood analyses were performed with RAxML-VI-HPC v.2.2.1 (ref. 41). All searches were completed with the PROTMIXWAG option. PhyloBayes v.2.1 (ref. 11) was used for bayesian analyses conducted under the CAT model, and MrBayes v.3.1.2 for bayesian analyses under the WAG model (with Gamma approximation of among site rate variation and allowing for invariable sites). Burn-ins were determined by plotting parameters across all runs for a given analysis. Leaf stabilities[17] were calculated with the tree analysis program Phyutility[30] (available at http://code.google.com/p/phyutility/), which was also used to determine where unstable taxa wandered across the bootstrap replicates (Supplementary Fig. 8).

31. Parkinson, J. *et al.* PartiGene — constructing partial genomes. *Bioinformatics* **20,** 1398–1404 (2004).
32. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
33. Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.* **33** (Database issue), D192–D196 (2005).
34. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4,** 41 (2003).
35. van Dongen, S. *Graph Clustering by Flow Simulation*. PhD thesis, Univ. Utrecht (2000).
36. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods)* Version 4 (Sinauer Associates, Sunderland, Massachusetts, 2003).
37. Edgar, R. C. & Journals, O. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
38. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17,** 540–552 (2000).
39. Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. *BMC Evol. Biol.* **7,** S2 (2007).
40. Huelsenbeck, J. P. & Ronquist, F. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17,** 754–755 (2001).
41. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).

Supplementary Discussion

Comparisons of gene selection strategies

There is relatively little overlap between the 150 genes selected by our approach, the manually curated list of 146 target genes used in most metazoan phylogenomic studies[9], and the 43 genes amplified by directed PCR (via 50 primer combinations) in another recent report[13] (Supplementary Tables 3,4; Supplementary Fig. 4). The union of the two previously published sets of genes listed above includes markers that are also commonly sequenced in smaller-scale phylogenetic studies (e.g. *elongation factor-2*, *ATPase alpha-subunit*, *70 kDa heat shock proteins*, and *DNA-directed RNA polymerase II largest subunit*) but that are rejected by our gene selection strategy. The matrix assembly results presented here therefore have the potential to inform gene selection both in future phylogenomic studies and more traditional phylogenetic investigations based on directed PCR of a small number of genes.

Examination of relevant selection metrics (Supplementary Table 4) indicates that most of the 126 genes from these other datasets that are rejected by our selection strategy pass the taxon sampling criteria imposed here (79 genes), but that most (68 genes) fail the criteria imposed on the median (must equal one) and mean (must be less than 2.5) number of sequences per taxon. Twenty-eight genes that passed these criteria were found to have clustering properties that indicate potential paralogy issues or were found to have paralogy problems when evaluated phylogenetically (see Methods for explanations of these criteria). Although 40 ribosomal proteins are included in our final matrix, 30 ribosomal proteins used in these previous studies were rejected (Supplementary Table 4), many due to paralogy problems observed following phylogenetic analysis. This indicates that the assembly of phylogenomic datasets exclusively from the 70 or more ribosomal proteins that may be recovered in a typical EST screen, which has been done in some studies in part because of their abundance in EST surveys[41], may lead to

systematic error due to paralogy problems. Ribosomal proteins should be carefully evaluated for paralogy, the same as for any other type of gene.

All of the genes shared between our matrix and one or both of the other matrices have one-to-one mapping across studies (Supplementary Table 4). There is not always a one-to-one mapping, however, between genes rejected from our matrix and the corresponding genes included in other matrices. Mapping can be many-to-one, as where clusters 275, 561, and 970 all map to *vata* from the Philippe et al. matrix[9]. Mapping can also be one-to-many or many-to-many, as for the *psma* genes. Mappings other than one-to-one indicate genes for which different studies disagree on paralogy assignment, with, for instance, one study grouping together all sequences as a single gene that another study assigned to multiple genes. The acceptance of only those genes with one-to-one mapping across studies, even though this was not part of the actual selection criteria, is one indication that our gene selection strategy is quite conservative.

We assembled a matrix of the genes used in most other animal phylogenomic studies[9] for the 64 stable taxa. Phylogenetic analyses of this alternative matrix (Supplementary Fig. 5) produce trees that are largely congruent with analyses of our own matrix. There are several differences in topology (e.g. neither Annelida nor Chordata are monophyletic in the alternative tree), but the general agreement between two matrices of similar size that differ by more than 100 genes shows that estimates of the animal phylogeny converge with large amounts of molecular sequence data, and that the major findings of the present study are not artefacts of a particular set of genes. The similarity of trees inferred from the two matrices also indicates that many of the genes rejected by our approach may not have paralogy problems, and that the gene selection method presented here is probably quite conservative. Our criteria for the mean and median number of sequences per taxon, for example, may by more restrictive than is absolutely necessary.

Gene accumulation plots based on our dataset (Supplementary Fig. 6) indicate that slightly more genes are obtained from the standard predefined list[9] when sequencing fewer than 1600 ESTs. Beyond this, in the range of most EST studies, more genes are obtained by our method.

Rejection of genes without paralogy problems could be reduced in future studies by applying less stringent phenetic gene-selection criteria (such as the cutoffs for mean and median number of sequences per taxon, which are computationally innexpensive) and evaluating proportionally more genes with phylogenetic tools (which is more computationally expensive). Like other animal phylogenomic studies, our original assignment of genes to orthologous groups is itself phenetic, relying on sequence similarity, though it takes into account more information by clustering with a graph theory approach rather than relying on BLAST ranks. One could avoid rejecting paralogs that diverged prior to the radiation of the group of interest (and hence could be informative for the problem at hand) by making the initial clustering less stringent, generating a smaller number of larger clusters. Each cluster would then be evaluated phylogenetically, and informative sub-trees pruned away as their own clusters.

Ribosomal vs. Non-ribosomal proteins

As noted above, the 150 genes in our final matrix include 40 ribosomal proteins for which no paralogy problems were identified. All 40 of the ribosomal proteins fall within the top 44 best-sampled genes across taxa (Supplementary Table 3), so they constitute a disproportionally large fraction of the character data. We therefore partitioned the 64-taxon matrix into two matrices, one consisting of the 40 ribosomal proteins (5526 aa long, 30.7% missing characters) and the remaining 110 non-ribosomal proteins (15626 aa long, 61.0% missing characters).

Comparisons between ML bootstrap support values derived from the non-ribosomal, ribosomal, and combined matrices (Supplementary Fig. 10) show that bootstrap support from the combined matrix is greater than or equal to the support derived from either sub-matrix for most nodes. These nodes indicate features of the tree for which there is no conflict between the ribosomal and non-ribosomal partitions. At many such nodes support is much greater when all genes are analyzed in combination. These include Clade A, Clade B, and the node that places ctenophores as the earliest branching metazoans. This is an encouraging result, as other nodes that remain recalcitrant in the present study of 150 genes may be resolved as even more genes are considered (through deeper EST sequencing or improved gene selection methods).

Ribosomal support is greater than combined support at nine ingroup nodes (Supplementary Fig. 10). At these nodes there is conflict between the partitions but signal from the ribosomal proteins contributed more strongly to the combined analysis than did signal from non-ribosomal proteins. With the exception of the node uniting *Drosophila melanogaster* and *Daphnia magna*, where the combined and ribosomal analyses differ in support by only 1%, none of these nodes had greater than 76% bootstrap support in the combined analyses and are not relevant to the major conclusions of the paper. Bias from ribosomal proteins that conflict with other genes therefore does not appear to be a problem in our study. Non-ribosomal support is greater than combined support at only two ingroup nodes, Cnidaria and Nematoda.

Gene lengths

Genes selected for phylogenetic analysis by our method tend to be slightly shorter than those in the population as a whole (Supplementary Fig. 7). A similar bias towards shorter genes is also apparent in the matrix of the genes used in most other animal phylogenomic studies[9], though it is less pronounced. The selection of shorter proteins may be due to an enrichment for highly expressed genes that maximize gene intersection across taxa. It has previously been found

that gene expression levels vary inversely with gene length[42], perhaps due to stronger selection on highly expressed genes to be shorter so as to reduce overall amino acid use.

Since all ESTs were sequenced from the 5' end and mRNA was enriched for complete transcripts, most gene sequences presented here will be complete at the 5' end. 59.65% of genes derived from our EST data are complete at the 3' end, as determined by the presence of a stop codon. Since the length of genes without a stop codon is underestimated, we also considered only those genes with a stop codon, and found a similar pattern in gene length (Supplementary Fig. 7).

Tardigrada

WAG analyses (both ML and Bayesian) do not recover Panarthropoda and favour a close relationship between tardigrades, nematodes, and a nematomorph (Fig. 1, Supplementary Fig. 9). However, of the 15 independent 64-taxon Bayesian runs based on the CAT model of evolution, 13 recover the morphologically-founded groups Cycloneuralia (i.e., Priapulida, Kinorhyncha, Nematomorpha, and Nematoda) and Panarthropoda (with Tardigrada as sister to Onychophora +Arthropoda), each with a posterior probability of 98% or greater. The other two CAT Bayesian runs place the clade composed of the kinorhynch and priapulid as sister to the remaining ecdysozoans (posterior probability of 100%), and place the tardigrades in a clade with the nematomorph and nematodes (also with a posterior probability of 100%). This leads to insignificant posterior probabilities of 86% for Panarthropoda and Cycloneuralia when the posterior distributions of trees are combined across runs (Fig. 2) and illustrates the large computational burden of phylogenomic analyses, since a small number of runs may not have revealed this lack of convergence. The strong dependence on the model of molecular evolution for the placement of the tardigrades indicates that at least one of these models is prone to systematic error for this particular problem. The two alternative placements of tardigrades

determine whether paired panarthropod appendages have a single or dual origin, but both topologies identify unique onychophoran-arthropod synapomorphies such as a dorsal heart with segmental ostia and open, haemocoelic circulation as shared derived characters.

Supplementary Tables

Supplementary Table 1 | Specimen data for sequenced taxa. RL- tissue was stored in RNA*later* prior to preparation. CVBS- Connecticut Valley Biological Supply (USA). dbEST- indicates that our data were supplemented with sequences from dbEST.

Species	Number of ESTs	Number of Matrix Genes	Tissue	Collection Location	Extraction method
Anoplodactylus eroticus	3744	81	embryos, larvae	Honolulu, HI, USA	TRI REAGENT
Brachionus plicatilis	3552	94	whole animals	Culture	Dynabeads
Bugula neritina	3360	92	hatched larvae	Honolulu, HI, USA	Dynabeads
Carinoma mutabilis	3168	62	whole animal	Friday Harbor, WA, USA	TRI REAGENT
Cerebratulus lacteus	6144	80	embryos (cleavage, gastrula)	Woods Hole, MA, USA	TRI REAGENT
Chaetoderma nitidulum	1632	47	parts of whole animal (RL)	Kristineberg Marine Station, Fiskebackskil, Sweden	Rneasy Micro
Chaetopleura apiculata	2304	45	gills	Woods Hole, MA, USA	TRI REAGENT
Chaetopterus sp.	3360	79	embryos	Woods Hole, MA, USA	TRI REAGENT
Cristatella mucedo	3264	85	statoblasts	Kristineberg, Sweden	TRI REAGENT
Echinoderes horni	3264	74	whole adults	Honolulu, HI, USA	RNAqueous-Micro
Euperipatoides kanangrensis	3360	81	two brains and muscle tissue	Kanangra-Boyd National Park, NSW, Australia	RNAqueous-Micro
Gnathostomula peregrina	3552	73	whole animals	Bermuda	RNAqueous-Micro
Mertensiid sp	3072	62	adult	Monterey, CA, USA	Rneasy Micro
Mnemiopsis leidyi	3360 (+dbEST)	110	early cleavage to gastrula	Woods Hole, MA, USA	TRI REAGENT
Myzostoma seymourcollegiorum	1056	46	whole animals (RL)	Encounter Bay, Australia	Rneasy Micro
Neochildia fusca	1728	21	whole animals (RL)	Woods Hole, MA, USA	Rneasy Micro
Paraplanocera sp.	3744	85	whole animal	Honolulu, HI, USA	TRI REAGENT
Pedicellina cernua	5184	33	whole animals	Kristineberg, Sweden	RNAqueous-Micro

Species	Number of ESTs	Number of Matrix Genes	Tissue	Collection Location	Extraction method
Philodina roseola	3168	82	whole animals	culture (CVBS)	RNeasy micro
Phoronis vancouverensis	2208	27	"heads" (RL)	Friday Harbor, WA, USA	Rneasy Micro
Ptychodera flava	3360	89	adult colar	Honolulu, HI, USA	TRI REAGENT
Richtersius coronifer	3360	66	whole adults	Öland, Sweden	TRI REAGENT
Scutigera coleoptrata	2400	66	part of whole animal (RL)	Cambridge, MA, USA	Dynabeads
Spinochordodes tellinii	2208	25	part of whole animal (RL)	Montpellier, France	TRI REAGENT
Terebratalia transversa	3552	91	embryos/larvae (RL)	Friday Harbor, WA, USA	Rneasy Micro
Themiste lageniformis	2640	70	embryos (cleavage, gastrula)	Honolulu, HI, USA	TRI REAGENT
Turbanella ambronensis	3264	61	whole animals (RL)	Wilhelmshaven, Germany	RNAqueos-Micro
Urechis caupo	2208	78	internal tissue (gland, nervous)	Santa Barbara, CA, USA	TRI REAGENT
Xenoturbella bocki	3840 (+2137 dbEST)	71	part of whole animal (RL)	Strömstad, Sweden	Rneasy Micro

Supplementary Table 2 | Taxa from previously published EST projects and genomic data used in this analysis. dbEST—http://www.ncbi.nlm.nih.gov/dbEST/, HG— http://www.ncbi.nlm.nih.gov/HomoloGene/, JGI—http://www.jgi.doe.gov/.

Taxon	Source	Number of Genes
Acanthoscurria gomesiana	dbEST	83
Acropora millepora	dbEST	101
Amoebidium parasiticum	dbEST	59
Aplysia californica	dbEST	22
Argopecten irradians	dbEST	81
Asterina pectinifera	dbEST	123
Biomphalaria glabrata	dbEST	79
Boophilus microplus	dbEST	105
Branchiostoma floridae	dbEST	101
Capitella sp.	JGI	116
Capsaspora owczarzaki	dbEST	98
Carcinoscorpius rotundicauda	dbEST	18
Carcinus maenas	dbEST	63
Ciona intestinalis	JGI	125
Crassostrea virginica	dbEST	82
Cryptococcus neoformans	http://www-sequence.stanford.edu/group/C.neoformans/download.html	114
Cyanea capillata	from authors[43]	75
Daphnia magna	dbEST	83
Drosophila melanogaster	HG	141
Dugesia japonica	dbEST	59
Echinococcus granulosus	dbEST	92
Euprymna scolopes	dbEST	87
Fenneropenaeus chinensis	dbEST	74
Flacisagitta enflata	assembled from original EST traces[44]	66
Gallus gallus	HG	116
Haementeria depressa	dbEST	42
Homo sapiens	HG	125
Hydra magnipapillata	http://mpc.uci.edu/hampson/public_html/blast/jf9/	93
Hydractinia echinata	dbEST	77
Hypsibius dujardini	dbEST	90
Lumbricus rubellus	dbEST	106
Macrostomum lignano	http://macest.biology.ucla.edu/macest/	56
Monosiga ovata	dbEST	80
Mytilus galloprovincialis	dbEST	66
Nematostella vectensis	JGI	137
Oscarella carmela	dbEST	35
Platynereis dumerilii	Genbank	36
Priapulus caudatus	dbEST	24
Saccharomyces cerevisiae	http://mips.gsf.de/genre/proj/yeast/About/FTP_sites.html	101

Taxon	Source	Number of Genes
Saccoglossus kowalevskii	dbEST	51
Schmidtea mediterranea	dbEST	129
Spadella cephaloptera	EMBL	35
Sphaeroforma arctica	dbEST	88
Strongylocentrotus purpuratus	RefSeq	124
Suberites domuncula	Genbank	45
Symsagittifera roscoffensis	dbEST	33
Trichinella spiralis	dbEST	78
Xiphinema index	dbEST	86

Supplementary Table 3 | Genes selected for phylogenetic analysis. ID- the unique numerical identifier assigned to the gene during the clustering process (this number corresponds to the partition names within the nexus file), Description- the name of the gene as determined from one of the HomoloGene IDs, HomoloGene IDs- identifiers for HomoloGene groups, PG- the identifier of the gene in the matrix of Philippe et al.[9] (if the gene is in both matrices), Number of Taxa- the number of taxa in each cluster after paralog processing.

ID	Description	HomoloGene IDs	PG	Number of taxa
143	ribosomal protein L9	37328, 68697	rpl9	70
243	ribosomal protein S24 isoform c	68148, 82521, 82583, 83665, 74661	-	68
268	ribosomal protein L35a	6994	rpl33a	66
199	ribosomal protein S15	37414, 69555, 54806, 79584	rps15	64
232	ribosomal protein L18a	68104, 757, 66212, 81127	rpl20	64
273	ribosomal protein S11	789	rps11	63
184	ribosomal protein L8	32141	rpl2	63
242	ribosomal protein S16	794, 74778, 73355	rps16	62
200	ribosomal protein L17	81526, 81780, 83344, 67073, 78559, 83922, 66863, 83820	rpl17	62
225	ribosomal protein S12	36049, 53343, 54313, 54294	l12e-A	62
211	ribosomal protein S18	5747, 74651, 78628	rps18	61
279	ribosomal protein L26	764, 74758, 53471, 83250, 67813, 79103, 55206	rpl26	61
169	ribosomal protein L7a	39625, 79798, 79964, 72659, 83856, 83973	l12e-D	61
299	ribosomal protein S17	68133, 54243, 68663, 50231	rps17	60
193	ribosomal protein L27a	81527, 73905	rpl27	60
179	ribosomal protein L18	756, 66225, 83186	rpl18	60
287	ribosomal protein S27	803, 69197, 67034, 68802	rps27	59
260	ribosomal protein S19	74380, 37416, 79606	rps19	59
272	ribosomal protein S13	38660	rps13a	59
241	ribosomal protein L37	68110, 81642, 82208, 78170, 82895	rpl37a	59
305	ribosomal protein L35	31432, 66665, 78302	rpl35	59
213	ribosomal protein L12	68673, 70329, 54987, 55157, 79120, 82909, 82997	rpl12b	59
271	ribosomal protein P2	68111, 68655	-	59
186	ribosomal protein S8	786, 83687, 67837	rps8	58
355	ribosomal protein S29 isoform 1	83197, 83391	rps29	58
236	ribosomal protein S20	37417, 76418, 54224	rps20	58
340	ribosomal protein S21	37418, 54814	-	58
351	ribosomal protein L28	768	-	58
325	ribosomal protein L14	68375, 2956, 42819	rpl14a	57
288	ribosomal protein S25	68149, 48911	rps25	56

ID	Description	HomoloGene IDs	PG	Number of taxa
356	ribosomal protein L34	68109, 79505	rpl34	56
235	ribosomal protein L23	68103, 67103	rpl23a	55
233	ribosomal protein L22 proprotein	37378, 69416, 46103, 67852, 82005	rpl22	55
320	ubiquitin-like protein fubi and ribosomal protein S30 precursor	37562, 54440	-	55
226	ribosomal protein P1 isoform 1	37388, 52692, 66221, 83538	rla2-B	54
319	ribosomal protein L36	41038, 65169, 78365, 17549	-	52
345	ribosomal protein L30	766, 79124, 83359	rpl30	51
156	ribosomal protein L3 isoform a	747, 68434, 45874, 83790	rpl3	51
174	ribosomal protein L19	68105, 82497, 67865, 79127, 79129	rpl19a	50
330	cytochrome c oxidase subunit III	5014, 55063	-	48
435	signal sequence receptor gamma subunit	5154, 82965	-	48
281	cytochrome c oxidase subunit Va precursor	37905	-	47
593	ATP synthase, H+ transporting, mitochondrial F1 complex, delta subunit precursor	37514	-	47
321	ribosomal protein S28	68150, 49064	rps28a	46
333	cytochrome c	68675	-	46
460	tumor protein, translationally-controlled 1	55730, 69044	-	45
352	eukaryotic translation initiation factor 5A	1490, 38886, 56219	eif5a	44
542	eukaryotic translation initiation factor 1	48375, 22219, 83852	-	44
572	ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	4378	-	44
619	Sec61 beta subunit	38229, 80032	-	44
495	ATP synthase, mitochondrial F1 complex, gamma subunit isoform H precursor	3792	-	43
400	cytochrome c oxidase subunit II	5017	-	42
532	mitochondrial ATP synthase, O subunit precursor	1283	-	42
550	cytochrome c oxidase subunit IV isoform 1 precursor	37537, 13082	-	42
586	signal sequence receptor, beta precursor	2369	-	41
700	endothelial differentiation-related factor 1 isoform alpha	2809	-	39
740	defender against cell death 1	1027	-	39
686	cytochrome c oxidase subunit VIa polypeptide 1 precursor	3219, 38020, 66386	-	38
716	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit b isoform 1 precursor	1275	-	38
764	signal sequence receptor, delta	4573	-	38
263	NADH dehydrogenase subunit 1	5011	-	37
646	NADH dehydrogenase (ubiquinone) Fe-S protein 6	37935	-	37

ID	Description	HomoloGene IDs	PG	Number of taxa
717	heat shock 10kDa protein 1 (chaperonin 10)	20500, 68540	-	37
703	proteasome beta 4 subunit	2090	psmb-N	36
507	Sec61 gamma subunit	40767, 83268	-	36
508	triosephosphate isomerase 1	311, 82609	-	36
739	cytochrome c oxidase subunit Vb precursor	37538, 44294	-	36
696	X-linked eukaryotic translation initiation factor 1A	20364, 81626	if1a	35
369	ATP synthase F0 subunit 6	5012	-	35
726	cytochrome c oxidase subunit VIb	39658, 16948	-	35
831	signal sequence receptor, alpha	2368	-	35
585	peroxiredoxin 6	3606, 71226	-	34
629	small nuclear ribonucleoprotein polypeptide D3	3078, 82194	-	34
632	skpA CG16983-PA, isoform A	76877, 64484, 38775	-	34
647	clathrin, light polypeptide A isoform a	1384, 37532	-	34
832	proliferating cell nuclear antigen	1945, 77171	-	34
905	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit g	21294	-	34
961	G10 protein	2906	-	34
805	actin related protein 2/3 complex subunit 3	4178, 82221	ar21	33
852	NADH dehydrogenase (ubiquinone) Fe-S protein 4	1866	-	33
896	NADH dehydrogenase (ubiquinone) Fe-S protein 3	3346	-	33
782	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit f isoform 2a	3594, 80715	-	32
1006	succinate dehydrogenase complex, subunit D precursor	37718, 80882	-	32
1050	cell death-regulatory protein GRIM19	41083, 43992	-	32
766	UV excision repair protein RAD23 homolog B	37704, 48322	rad23	31
613	ubiquinol-cytochrome c reductase binding protein	38164, 78989	-	31
765	small nuclear ribonucleoprotein polypeptide E	37729, 40287	-	31
850	quinoid dihydropteridine reductase	271	-	31
927	beta-tubulin cofactor A	3388	-	31
942	NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa	10884	-	31
368	NADH dehydrogenase subunit 5	36212	-	30
380	NADH dehydrogenase subunit 4	38240	-	30
683	heat-responsive protein 12	4261	-	30
768	iron-sulfur cluster assembly enzyme isoform ISCU1	6991	-	30

ID	Description	HomoloGene IDs	PG	Number of taxa
783	small nuclear ribonucleoprotein polypeptide G	37730, 82718	-	30
801	NADH-ubiquinone oxidoreductase Fe-S protein 7	11535, 56989	-	30
861	hypothetical protein LOC746	40931	-	30
895	eukaryotic translation initiation factor 3, subunit 12	8292, 52710	-	30
965	13kDa differentiation-associated protein	10314	-	30
1041	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 2, 8kDa	37628	-	30
1046	succinate dehydrogenase complex, subunit C precursor	2256	-	30
1118	hypothetical protein LOC55831	10201	-	30
1162	proteasome beta 1 subunit	2087	psma-J	29
1051	actin related protein 2/3 complex subunit 4 isoform a	4177	arc20	29
742	t-complex-associated-testis-expressed 1-like	21304, 4754	-	29
893	small nuclear ribonucleoprotein polypeptide D2	3381, 77955, 76944	-	29
906	translocase of outer mitochondrial membrane 20 homolog	44649, 52617, 47747	-	29
1012	actin related protein 2/3 complex subunit 5	4176, 36463, 52415	-	29
1015	eukaryotic translation initiation factor 3, subunit 4 delta	2784, 70155	-	29
1165	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 5	3664	-	29
1181	hypothetical protein LOC51234	5879	-	29
833	proteasome beta 3 subunit	2089, 74663	psma-I	28
622	electron transfer flavoprotein, alpha polypeptide	100	-	28
793	15 kDa selenoprotein isoform 1 precursor	3145	-	28
916	protein (peptidyl-prolyl cis/trans isomerase) NIMA-interacting 1	4531	-	28
1020	low molecular mass ubiquinone-binding protein	40942, 44637	-	28
1026	esterase D/formylglutathione hydrolase	55623	-	28
1077	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit F6 isoform a precursor	1272, 43209	-	28
1132	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 8, 19kDa	40932, 74890	-	28
1170	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 9, 22kDa	3669	-	28
1292	ATPase, H+ transporting, lysosomal 21kDa, V0 subunit c	2986	-	28
868	proteasome beta 2 subunit	2088, 60427	psma-H	27

ID	Description	HomoloGene IDs	PG	Number of taxa
1007	F-actin capping protein beta subunit	3620	-	27
1054	stromal cell-derived factor 2 precursor	5045, 11101	-	27
1099	ATPase, H+ transporting, lysosomal 14kD, V1 subunit F	3119	-	27
1121	SF3b10	41825	-	27
1136	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6	1861	-	27
1187	prefoldin 4	37645, 82938	-	27
1200	glycine cleavage system protein H (aminomethyl carrier)	12239, 67129	-	27
1245	unactive progesterone receptor, 23 kD	81751, 44698, 57061	-	27
756	adenosine kinase isoform b	4891, 51621	-	26
1009	growth hormone inducible transmembrane protein	8667	-	26
1013	signal peptidase complex subunit 2 homolog	8842, 72649, 55112	-	26
1081	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9, 39kDa	3666	-	26
1129	hypothetical protein LOC51398	13954	-	26
1190	elongin B isoform a	38275, 52996	-	26
1204	prefoldin 5 isoform alpha	1972	-	26
1294	von Hippel-Lindau binding protein 1	2531	-	26
1335	signal peptidase complex subunit 3 homolog	41454	-	26
1381	DNA directed RNA polymerase II polypeptide C	2017	-	26
1409	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 16kDa precursor	31093, 81299	-	26
1489	programmed cell death 5	10506	pace6	25
858	NADH dehydrogenase (ubiquinone) 1, alpha/ beta subcomplex, 1, 8kDa	80336	-	25
930	elongin C	38083	-	25
1047	testis enhanced gene transcript (BAX inhibitor 1)	2419	-	25
1169	B-cell receptor-associated protein 31	38095, 22411	-	25
1244	vacuolar protein sorting 29 isoform 2	9433	-	25
1374	Apg3p	6836	-	25
1378	chromosome 15 open reading frame 24	10597	-	25
1646	translocase of inner mitochondrial membrane 13	40846	-	25

Supplementary Table 4 | Table of metrics relevant to the gene selection approach presented here for all genes used in two recent phylogenomic studies. Gene- the name assigned to the gene in the previous study. Dataset- p designates genes considered by Philippe et al.[9], r designates genes considered by Rokas et al.[13], and r&p designates genes considered by both studies (homology across studies assessed by BLASTP with an e-value cutoff of $1x10^{-20}$ and shared ontology). Exemplar- the GenBank accession number for the sequence used as an Exemplar for each gene in BLASTP searches (usually the longest sequence in the original study). ID- the unique numerical identifier for each cluster generated by our gene selection approach. This ID corresponds to the ID indicated in Supplementary Table 3. If a gene from the previous studies had significant hits to multiple clusters from the present study, each cluster hit is shown on its own row. N_{seq}- number of sequences assigned to the cluster indicated by ID. N_{taxa}- number of taxa with sequences assigned to the cluster. N_{focal}- number of these taxa for which we collected new EST data. Mean- mean number of sequences per taxon in the cluster. Median- median number of sequences per taxon. Other- status of cluster according to other criteria. A dash indicates that there were no other problems or that the cluster was not evaluated for other problems because it failed according to taxon sampling, median, or mean. 1 indicates that features of the graph indicative of paralogy problems were noted (see Methods). 2 indicates that the cluster was evaluated phylogenetically and that the resulting topology indicated paralogy problems (see Methods). Pass- 1 indicates that the cluster passed all selection criteria and is one of the 150 genes included in our phylogenomic analysis.

Gene	Study	Exemplar	ID	N_{seq}	N_{taxa}	N_{focal}	Mean	Median	Other	Pass
ar21	p	NP_650498.2	805	45	38	6	1.18	1	-	1
arc20	p	EAA47728	1051	37	34	5	1.09	1	-	1
arp23	p	AAH02988.1	425	75	47	6	1.60	1	2	0
cct-A	p	EAA08611	46	335	61	17	5.49	5	-	0
cct-A	p	EAA08611	295	98	30	2	3.27	2	-	0
cct-E	p	AAB39290	46	335	61	17	5.49	5	-	0
cct-E	p	AAB39290	295	98	30	2	3.27	2	-	0
cct-G	p	AAH06501	46	335	61	17	5.49	5	-	0
cct-G	p	AAH06501	295	98	30	2	3.27	2	-	0
cct-N	p	CAB08778	46	335	61	17	5.49	5	-	0
cct-T	p	CAA89300	46	335	61	17	5.49	5	-	0
cct-T	p	CAA89300	295	98	30	2	3.27	2	-	0
cct-Z	p	CAA86694	46	335	61	17	5.49	5	-	0
cpn60-mt	p	AAN71181	567	60	36	5	1.67	1.5	-	0
crfg	p	AAK39281	595	58	35	4	1.66	1	-	0
ef2-U5	p	CAA82015	76	232	58	10	4.00	3	-	0
eif5a	p	CAE65142	352	88	54	15	1.63	1	-	1
fibri	p	Q22053	883	42	28	1	1.50	1	-	0
fpps	p	CAA08918	1949	24	23	3	1.04	1	-	0
glcn	p	AAL78196	4926	10	8	0	1.25	1	-	0
if1a	p	AAK29845	696	50	42	9	1.19	1	-	1

Gene	Study	Exemplar	ID	N_{seq}	N_{taxa}	N_{focal}	Mean	Median	Other	Pass
if2b	p	EAA04210	1138	35	33	4	1.06	1	-	0
if2p	p	O60841	1770	26	18	0	1.44	1	-	0
if2p	p	O60841	2241	22	17	0	1.29	1	-	0
if6	p	EAA01111	1226	33	31	4	1.06	1	-	0
l12e-A	p	CAA61806.1	225	112	73	26	1.53	1	-	1
l12e-B	p	EAA52871	365	84	51	15	1.65	2	-	0
l12e-C	p	AAH50495	365	84	51	15	1.65	2	-	0
l12e-D	p	EAA11704	169	133	72	24	1.85	1	-	1
mcm-A	p	XP_226316	303	97	33	4	2.94	2	-	0
mcm-A	p	XP_226316	308	96	26	0	3.69	2	-	0
mcm-B	p	BAA04642	303	97	33	4	2.94	2	-	0
mcm-B	p	BAA04642	308	96	26	0	3.69	2	-	0
metk	p	AAN10507	428	75	50	7	1.50	1	2	0
mra1	p	CAA92394	1491	29	24	1	1.21	1	-	0
nsf1-C	p	XP_341108	29	436	59	15	7.39	5	-	0
nsf1-C	p	XP_341108	63	264	46	7	5.74	3	-	0
nsf1-E	p	EAA13918	29	436	59	15	7.39	5	-	0
nsf1-E	p	EAA13918	63	264	46	7	5.74	3	-	0
nsf1-G	p	AAM48537	29	436	59	15	7.39	5	-	0
nsf1-G	p	AAM48537	63	264	46	7	5.74	3	-	0
nsf1-H	p	EAA49954	29	436	59	15	7.39	5	-	0
nsf1-I	p	BAC36516	29	436	59	15	7.39	5	-	0
nsf1-I	p	BAC36516	63	264	46	7	5.74	3	-	0
nsf1-J	p	EAA01092	29	436	59	15	7.39	5	-	0
nsf1-J	p	EAA01092	63	264	46	7	5.74	3	-	0
nsf1-K	p	AAH08713	29	436	59	15	7.39	5	-	0
nsf1-K	p	AAH08713	63	264	46	7	5.74	3	-	0
nsf1-L	p	NP_010682.1	29	436	59	15	7.39	5	-	0
nsf1-L	p	NP_010682.1	63	264	46	7	5.74	3	-	0
nsf1-M	p	EAA54685	29	436	59	15	7.39	5	-	0
nsf1-M	p	EAA54685	63	264	46	7	5.74	3	-	0
nsf2-A	p	CAA40276	29	436	59	15	7.39	5	-	0
nsf2-A	p	CAA40276	63	264	46	7	5.74	3	-	0
nsf2-F	p	EAA27598	29	436	59	15	7.39	5	-	0
nsf2-F	p	EAA27598	63	264	46	7	5.74	3	-	0
orf2	p	NP_496099	2420	20	18	7	1.11	1	-	0
pace4	p	CAE73168	2155	22	21	1	1.05	1	-	0
pace6	p	CAA88799	1489	29	28	6	1.04	1	-	1
psma-A	p	CAA67615	64	258	66	15	3.91	3	-	0
psma-A	p	CAA67615	73	240	59	16	4.07	3	-	0
psma-A	p	CAA67615	2252	22	10	1	2.20	2	-	0
psma-B	p	EAA28095	73	240	59	16	4.07	3	-	0

Gene	Study	Exemplar	ID	N_{seq}	N_{taxa}	N_{focal}	Mean	Median	Other	Pass
psma-C	p	AAL89878	73	240	59	16	4.07	3	-	0
psma-C	p	AAL89878	447	73	24	1	3.04	2	-	0
psma-D	p	EAA56810	73	240	59	16	4.07	3	-	0
psma-E	p	AAN63095	73	240	59	16	4.07	3	-	0
psma-E	p	AAN63095	2252	22	10	1	2.20	2	-	0
psma-F	p	EAA53450	73	240	59	16	4.07	3	-	0
psma-F	p	EAA53450	447	73	24	1	3.04	2	-	0
psma-F	p	EAA53450	2252	22	10	1	2.20	2	-	0
psma-G	p	EAA13600	73	240	59	16	4.07	3	-	0
psma-G	p	EAA13600	447	73	24	1	3.04	2	-	0
psma-G	p	EAA13600	2252	22	10	1	2.20	2	-	0
psma-H	p	XP_569789.1	868	42	35	5	1.20	1	-	1
psma-I	p	NP_649858.1	833	44	36	6	1.22	1	-	1
psma-J	p	NP_498806.1	1162	34	33	8	1.03	1	-	1
psmb-K	p	AAF52066	264	103	55	14	1.87	1	2	0
psmb-L	p	EAA28906	264	103	55	14	1.87	1	2	0
psmb-M	p	AAF46978	264	103	55	14	1.87	1	2	0
psmb-M	p	AAF46978	667	53	38	5	1.39	1	2	0
psmb-N	p	NP_649529.1	703	50	41	5	1.22	1	-	1
rad23	p	AAH27747	766	47	37	5	1.27	1	-	1
rad51-A	p	AAB64650	1438	30	20	3	1.50	1	-	0
rf1	p	AAM46702	1902	24	20	1	1.20	1	-	0
rf1	p	AAM46702	6389	7	6	0	1.17	1	-	0
rla2-B	p	AAH58685	226	112	72	26	1.56	1	-	1
rpl1	p	EAA05156	113	179	74	23	2.42	1	2	0
rpl12b	p	EAA13967	213	115	71	24	1.62	1	-	1
rpl13	p	AAK92155	101	188	70	21	2.69	1	-	0
rpl14a	p	XP_224414	325	92	70	22	1.31	1	-	1
rpl15a	p	EAA10485	103	186	69	23	2.70	1	-	0
rpl15a	p	EAA10485	2132	22	17	2	1.29	1	-	0
rpl16b	p	EAA14246	191	121	73	22	1.66	1	1	0
rpl17	p	EAA32243	200	118	74	26	1.59	1	-	1
rpl18	p	EAA04761	179	128	71	23	1.80	1	-	1
rpl19a	p	EAA09119	174	130	62	18	2.10	1	-	1
rpl2	p	AAD47076.1	184	124	76	23	1.63	1	-	1
rpl20	p	AAB92041	232	110	73	23	1.51	1	-	1
rpl21	p	EAA00465	298	97	64	22	1.52	1	1	0
rpl22	p	AAF46972	233	110	69	25	1.59	1	-	1
rpl23a	p	AAH49038	235	109	70	25	1.56	1	-	1
rpl24-A	p	AAK18907	160	138	71	25	1.94	2	-	0
rpl24-B	p	EAA32763	160	138	71	25	1.94	2	-	0
rpl25	p	EAA11004	331	91	65	21	1.40	1	1	0

Gene	Study	Exemplar	ID	N_{seq}	N_{taxa}	N_{focal}	Mean	Median	Other	Pass
rpl26	p	CAD37159	279	100	71	25	1.41	1	-	1
rpl27	p	EAA00079	193	120	69	23	1.74	1	-	1
rpl30	p	AAH62278	345	89	63	23	1.41	1	-	1
rpl31	p	EAA00150	389	80	60	22	1.33	1	1	0
rpl32	p	AAN14210	313	95	69	25	1.38	1	1	0
rpl33a	p	AAK92169	268	102	73	25	1.40	1	-	1
rpl34	p	AAN13422	356	86	64	25	1.34	1	-	1
rpl35	p	XP_965481.1	305	96	70	24	1.37	1	-	1
rpl37a	p	AAL99981	241	108	71	26	1.52	1	-	1
rpl39	p	AAR10259	316	94	60	23	1.57	1	1	0
rpl42	p	AAB68420	277	100	65	24	1.54	1	2	0
rpl4B	p	BAB79458	128	163	69	20	2.36	1	2	0
rpl6	p	CAA60588.1	107	182	73	23	2.49	1	2	0
rpl7-A	p	EAA14847	95	192	77	23	2.49	1	2	0
rpl9	p	EAA51069	143	149	79	28	1.89	1	-	1
rps1	p	EAA08803	97	190	76	24	2.50	1	2	0
rps10	p	EAA49455	140	152	75	27	2.03	1	1	0
rps11	p	EAA52085	273	101	71	26	1.42	1	-	1
rps13a	p	AAN52387	272	101	73	24	1.38	1	-	1
rps14	p	EAA06897	145	148	75	24	1.97	1	2	0
rps15	p	EAA01741	199	118	76	27	1.55	1	-	1
rps16	p	AAL26583	242	108	76	26	1.42	1	-	1
rps17	p	EAA50355	299	97	73	24	1.33	1	-	1
rps18	p	EAA54870	211	115	75	26	1.53	1	-	1
rps19	p	EAA05616	260	104	72	25	1.44	1	-	1
rps20	p	EAA51777	236	109	72	25	1.51	1	-	1
rps22a	p	AAH51205	278	100	70	24	1.43	1	1	0
rps23	p	EAA01135	115	177	81	27	2.19	1	2	0
rps25	p	XP_236606	288	99	69	24	1.43	1	-	1
rps26	p	CAB57819	346	89	59	24	1.51	1	1	0
rps27	p	CAC44218	287	99	71	27	1.39	1	-	1
rps28a	p	XP_344014	321	93	61	27	1.52	1	-	1
rps29	p	AAL68340	355	86	67	24	1.28	1	-	1
rps4	p	AAP06482	72	241	73	22	3.30	1	-	0
rps5	p	XP_341789	116	177	72	22	2.46	1	2	0
rps6	p	EAA07587	104	185	66	19	2.80	1	-	0
sap40	p	XP_001177924.1	147	146	63	18	2.32	1	1	0
sra	p	EAA47420	1233	33	25	1	1.32	1	-	0
sra	p	EAA47420	1439	30	24	3	1.25	1	-	0
srp54	p	AAB68136	1233	33	25	1	1.32	1	-	0
srp54	p	AAB68136	1439	30	24	3	1.25	1	-	0

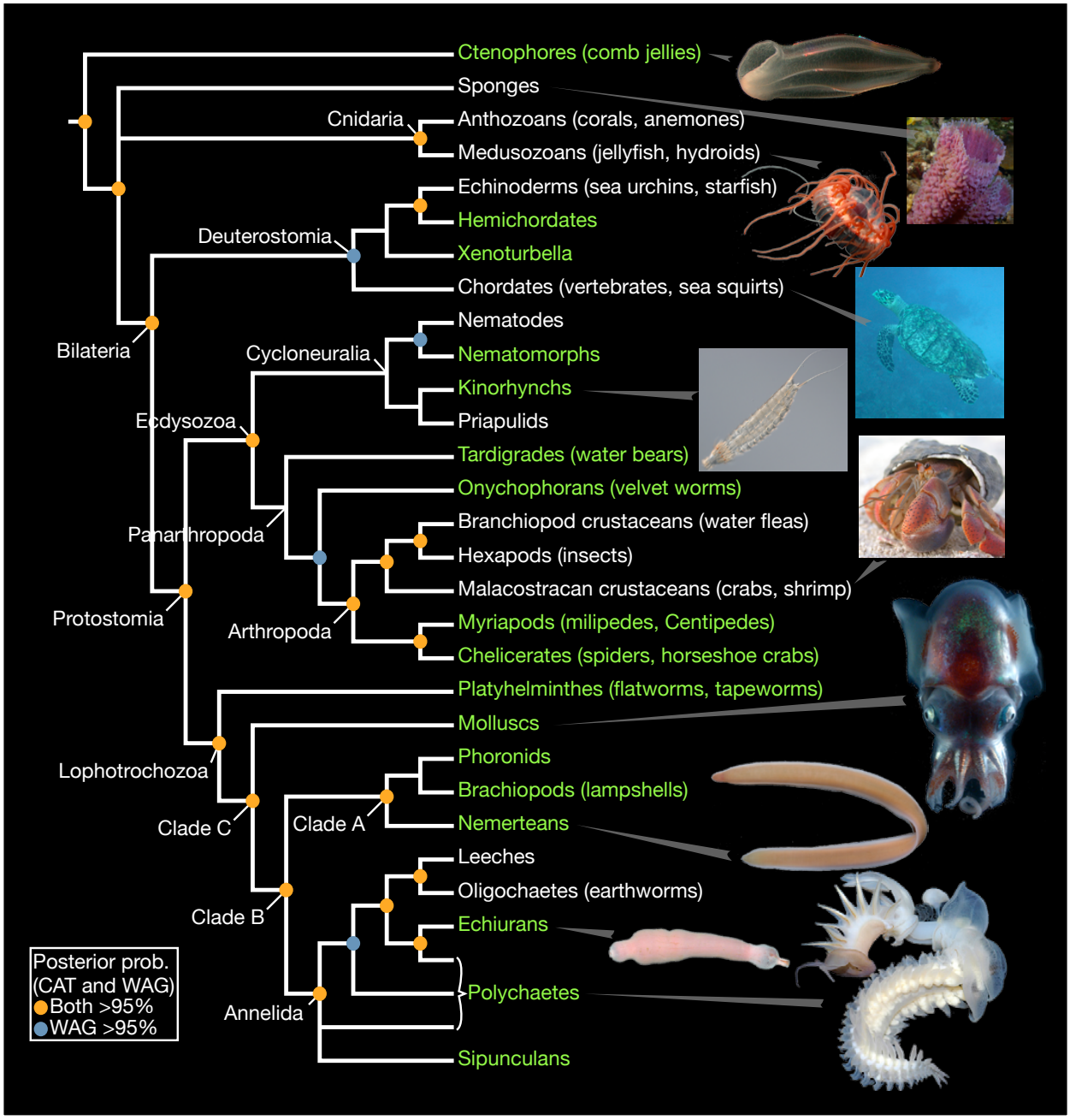
Gene	Study	Exemplar	ID	N_{seq}	N_{taxa}	N_{focal}	Mean	Median	Other	Pass
srs	p	AAH00716	574	59	39	8	1.51	1	2	0
suca	p	EAA54949	744	48	26	1	1.85	2	-	0
suca	p	EAA54949	838	44	33	4	1.33	1	-	0
tfiid	p	CAE64435	1176	34	22	0	1.55	1	-	0
topo1	p	EAA05377	1651	27	16	0	1.69	1	-	0
topo1	p	EAA05377	5484	9	7	0	1.29	1	-	0
vata	p	NP_609595	275	101	51	7	1.98	2	-	0
vata	p	NP_609595	561	61	35	4	1.74	1	-	0
vata	p	NP_609595	970	39	25	1	1.56	1	-	0
vatb	p	EAA08175	275	101	51	7	1.98	2	-	0
vatb	p	EAA08175	458	71	43	8	1.65	1	2	0
vatb	p	EAA08175	561	61	35	4	1.74	1	-	0
vatc	p	AAH56636	1529	28	27	1	1.04	1	-	0
vate	p	AAA35209	753	48	35	3	1.37	1	-	0
w09c	p	EAA43915	624	56	36	5	1.56	1	2	0
wrs	p	BAB23357	1011	38	29	3	1.31	1	-	0
xpb	p	EAA28093	2953	17	14	1	1.21	1	-	0
yif1p	p	AAF56617	3366	15	15	0	1.00	1	-	0
chaperonin complex component TCP-1 beta subunit	r&p	ABB29711.1	46	335	61	17	5.49	5	-	0
chaperonin complex component TCP-1 beta subunit	r&p	ABB29711.1	295	98	30	2	3.27	2	-	0
chaperonin complex component TCP-1 delta subunit	r&p	ABB29710.1	46	335	61	17	5.49	5	-	0
chaperonin complex component TCP-1 delta subunit	r&p	ABB29710.1	295	98	30	2	3.27	2	-	0
DNA-directed RNA polymerase II largest subunit	r&p	ABB29696.1	354	87	19	2	4.58	2	-	0
DNA-directed RNA polymerase II largest subunit	r&p	ABB29696.1	760	48	16	1	3.00	2.5	-	0
elongation factor-2	r&p	ABB29633.1	76	232	58	10	4.00	3	-	0
endoplasmic reticulum heat shock 70 kDa protein	r&p	ABB29693.1	23	494	62	14	7.97	4	-	0
endoplasmic reticulum heat shock 70 kDa protein	r&p	ABB29693.1	349	89	22	1	4.05	1.5	-	0
eukaryotic translation initiation factor 2	r&p	ABB29716.1	1414	30	24	3	1.25	1	-	0

Gene	Study	Exemplar	ID	N_{seq}	N_{taxa}	N_{focal}	Mean	Median	Other	Pass
eukaryotic translation initiation factor 2	r&p	ABB29716.1	2190	22	15	0	1.47	1	-	0
mitochondrial heat shock 70 kDa protein	r&p	ABB29640.1	23	494	62	14	7.97	4	-	0
ribosomal protein 10 large subunit	r&p	ABB29718.1	78	218	75	24	2.91	1	-	0
ribosomal protein 11 large subunit	r&p	ABB29719.1	185	123	74	24	1.66	1	1	0
ribosomal protein 2 small subunit	r&p	ABB29724.1	137	154	68	23	2.26	1	1	0
ribosomal protein 2 small subunit	r&p	ABB29724.1	725	49	27	2	1.81	1	-	0
ribosomal protein 3 large subunit	r&p	ABB29720.1	156	139	72	18	1.93	1	-	1
ribosomal protein 3 small subunit	r&p	ABB29725.1	84	210	77	24	2.73	1	-	0
ribosomal protein 5 large subunit	r&p	ABB29721.1	100	188	75	21	2.51	1	-	0
ribosomal protein 8 small subunit	r&p	ABB29726.1	186	123	75	21	1.64	1	-	1
ribosomal protein P0 large subunit	r&p	ABB29594.1	91	202	70	22	2.89	1	-	0
RNA polymerase I large subunit	r&p	ABB29704.1	760	48	16	1	3.00	2.5	-	0
actin-related protein Arp2 3 complex subunit ARPC2	r	ABB29728.1	1207	34	29	1	1.17	1	-	0
alpha-tubulin	r	ABB29581.1	6	876	74	24	11.84	6	-	0
ATPase alpha-subunit	r	ABB29609.1	99	189	39	2	4.85	3	-	0
beta-tubulin	r	ABB29632.1	6	876	74	24	11.84	6	-	0
cell division control protein 42	r	ABB29714.1	1	1247	70	22	17.81	10	-	0
cytoplasmic heat shock 70 kDa protein	r	ABB29617.1	23	494	62	14	7.97	4	-	0
cytoplasmic heat shock 70 kDa protein	r	ABB29617.1	7802	4	4	0	1.00	1	-	0
DNA replication licensing factor MCM3 component	r	ABB29658.1	303	97	33	4	2.94	2	-	0
DNA replication licensing factor MCM3 component	r	ABB29658.1	308	96	26	0	3.69	2	-	0
DNA replication licensing factor MCM7 component	r	ABB29709.1	303	97	33	4	2.94	2	-	0
DNA replication licensing factor MCM7 component	r	ABB29709.1	308	96	26	0	3.69	2	-	0

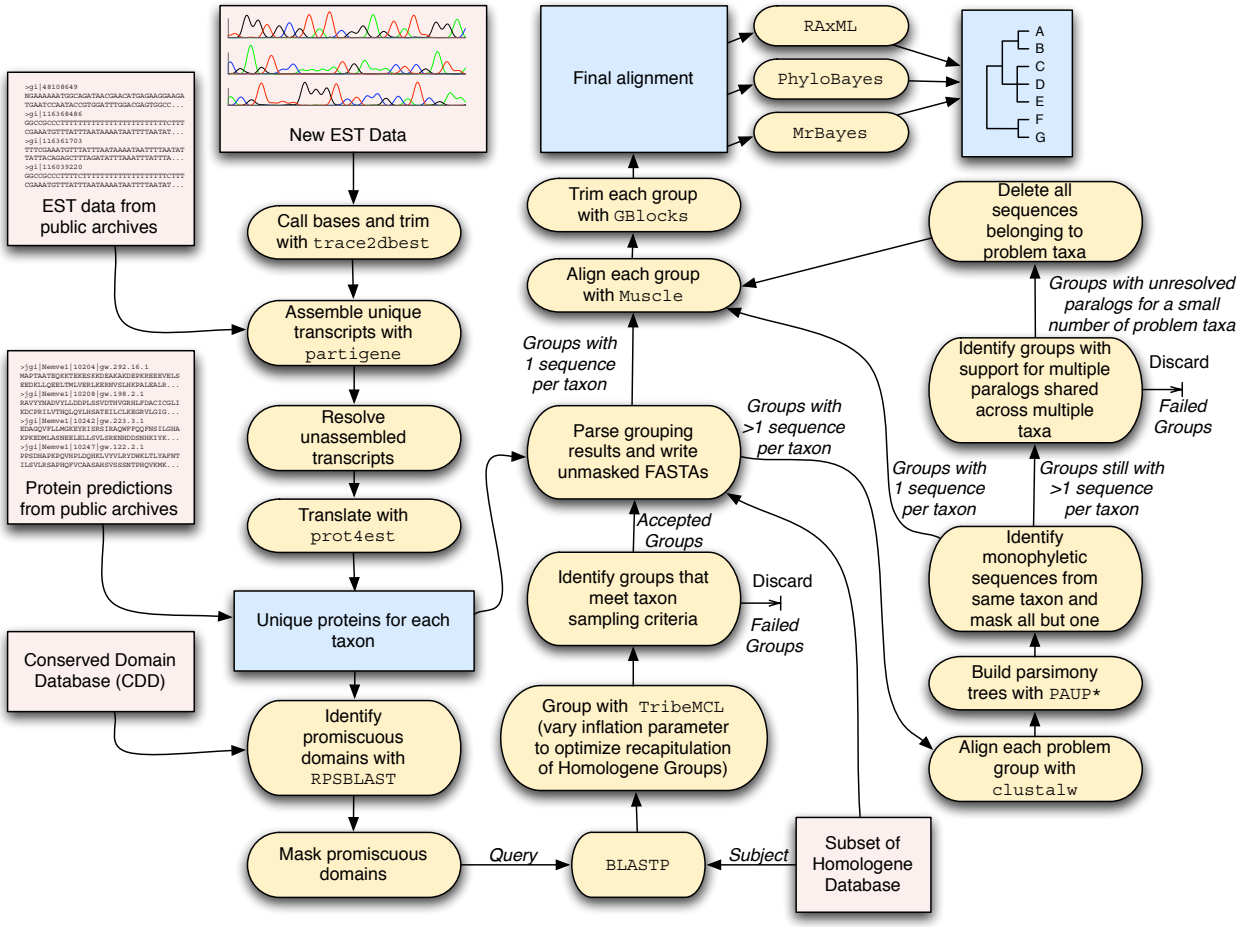
Gene	Study	Exemplar	ID	N_{seq}	N_{taxa}	N_{focal}	Mean	Median	Other	Pass
gamma-glutamylcysteine synthetase	r	ABB29599.1	1627	27	18	0	1.50	1	-	0
glutamyl-tRNA synthetase	r	ABB29702.1	626	56	32	3	1.75	1.5	-	0
glutamyl-tRNA synthetase	r	ABB29702.1	1355	31	24	0	1.29	1	-	0
Gpi-anchor transamidase	r	ABB29607.1	2513	20	18	1	1.11	1	-	0
heat shock 90 kDa protein	r	ABB29634.1	64	258	66	15	3.91	3	-	0
methylthioadenosine phosphorylase MTAP	r	ABB29729.1	1951	24	22	1	1.09	1	-	0
phenylalanyl-tRNA synthetase beta subunit	r	ABB29732.1	1503	29	22	2	1.32	1	-	0
P-type ATPase	r	ABB29630.1	1135	35	17	0	2.06	2	-	0
putative metalloprotease	r	ABB29660.1	1542	28	19	0	1.47	1	-	0
pyruvate carboxylase	r	ABB29597.1	466	71	23	1	3.09	3	-	0
pyruvate carboxylase	r	ABB29597.1	472	70	31	2	2.26	1	-	0
Ras-related nuclear protein	r	ABB29715.1	1	1247	70	22	17.81	10	-	0
Ras-related nuclear protein	r	ABB29715.1	31	422	30	4	14.07	2	-	0
ribosomal protein 8 large subunit	r	ABB29722.1	184	124	76	23	1.63	1	-	1
RNA polymerase I second largest subunit	r	ABB29657.1	578	59	22	1	2.68	2	-	0
RNA polymerase I second largest subunit	r	ABB29657.1	2461	20	10	0	2.00	2	-	0
RNA polymerase II transcription initiation nucleotide excision repair factor TFIIH	r	ABB29708.1	1021	38	18	2	2.11	1	-	0
RNA polymerase II transcription initiation nucleotide excision repair factor TFIIH	r	ABB29708.1	3502	15	10	0	1.50	1	-	0
RNA polymerase III large subunit	r	ABB29654.1	354	87	19	2	4.58	2	-	0
RNA polymerase III large subunit	r	ABB29654.1	760	48	16	1	3.00	2.5	-	0
RNA polymerase III second largest subunit	r	ABB29628.1	578	59	22	1	2.68	2	-	0
RNA polymerase III second largest subunit	r	ABB29628.1	2461	20	10	0	2.00	2	-	0

Gene	Study	Exemplar	ID	N_{seq}	N_{taxa}	N_{focal}	Mean	Median	Other	Pass
SNF2 family DNA-dependent ATPase domain-containing protein	r	ABB29652.1	62	265	37	4	7.16	2	-	0
SNF2 family DNA-dependent ATPase domain-containing protein	r	ABB29652.1	111	180	27	2	6.67	2	-	0
splicing factor 3b subunit 1	r	ABB29653.1	1231	33	25	0	1.32	1	-	0
succinate dehydrogenase iron-sulfur protein	r	ABB29717.1	705	50	39	4	1.28	1	-	0
SWI SNF-related matrix-associated regulator of chromatin a5	r	ABB29626.1	62	265	37	4	7.16	2	-	0

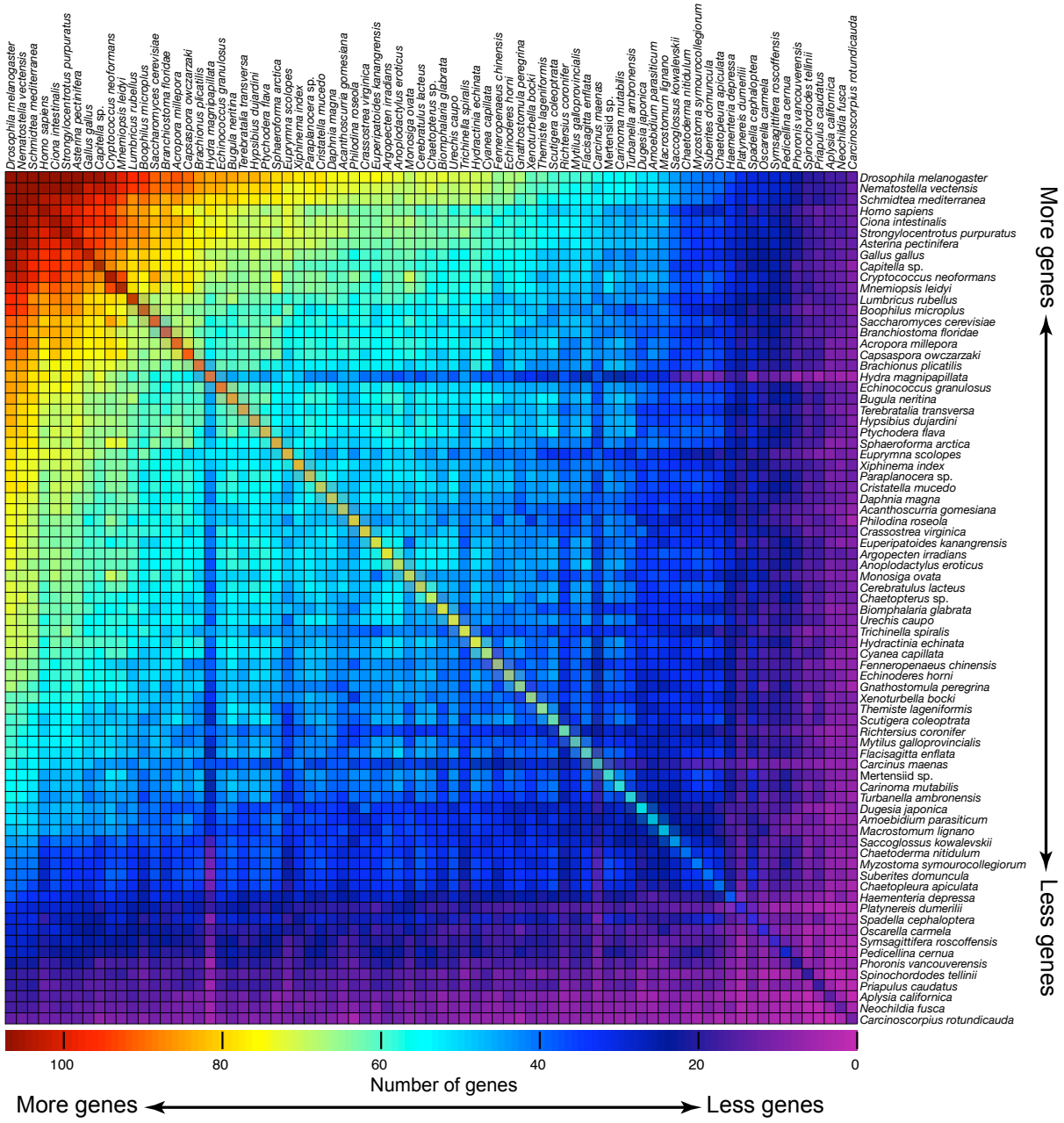
Supplementary Figures



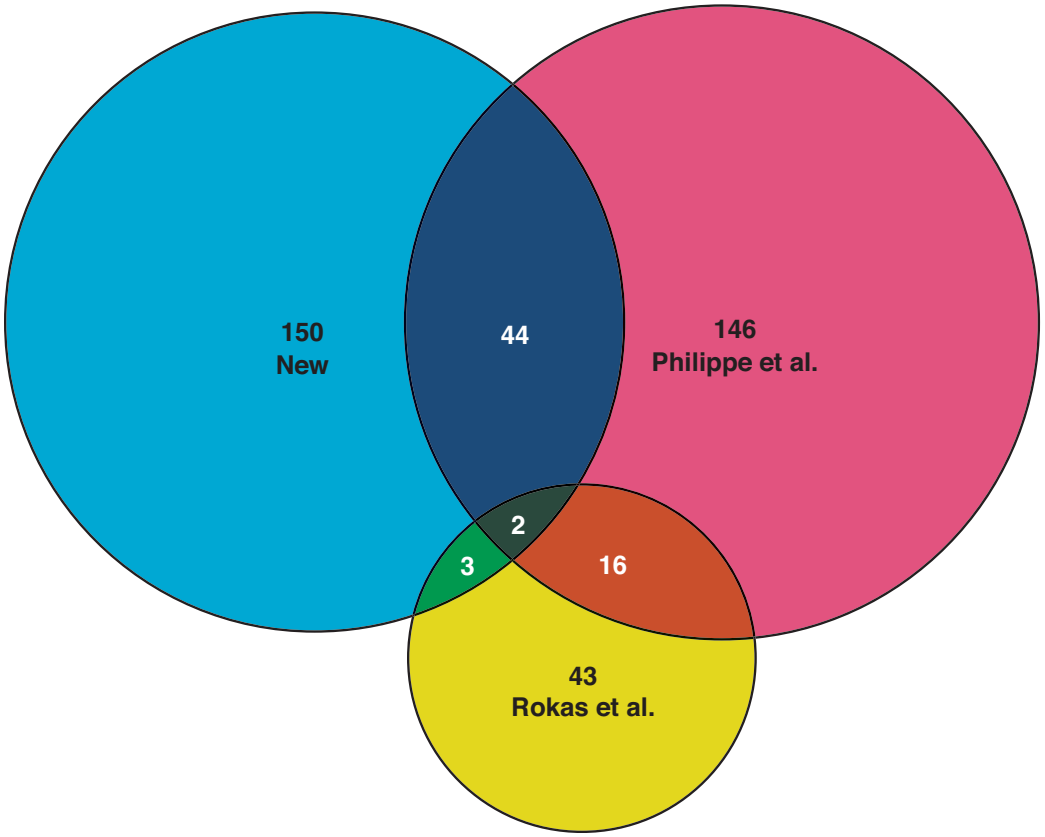
Supplementary Figure 1 | Summary of major findings—the evolutionary relationships among animals as inferred in the present study. Based on Fig. 2, with several clades collapsed for clarity.



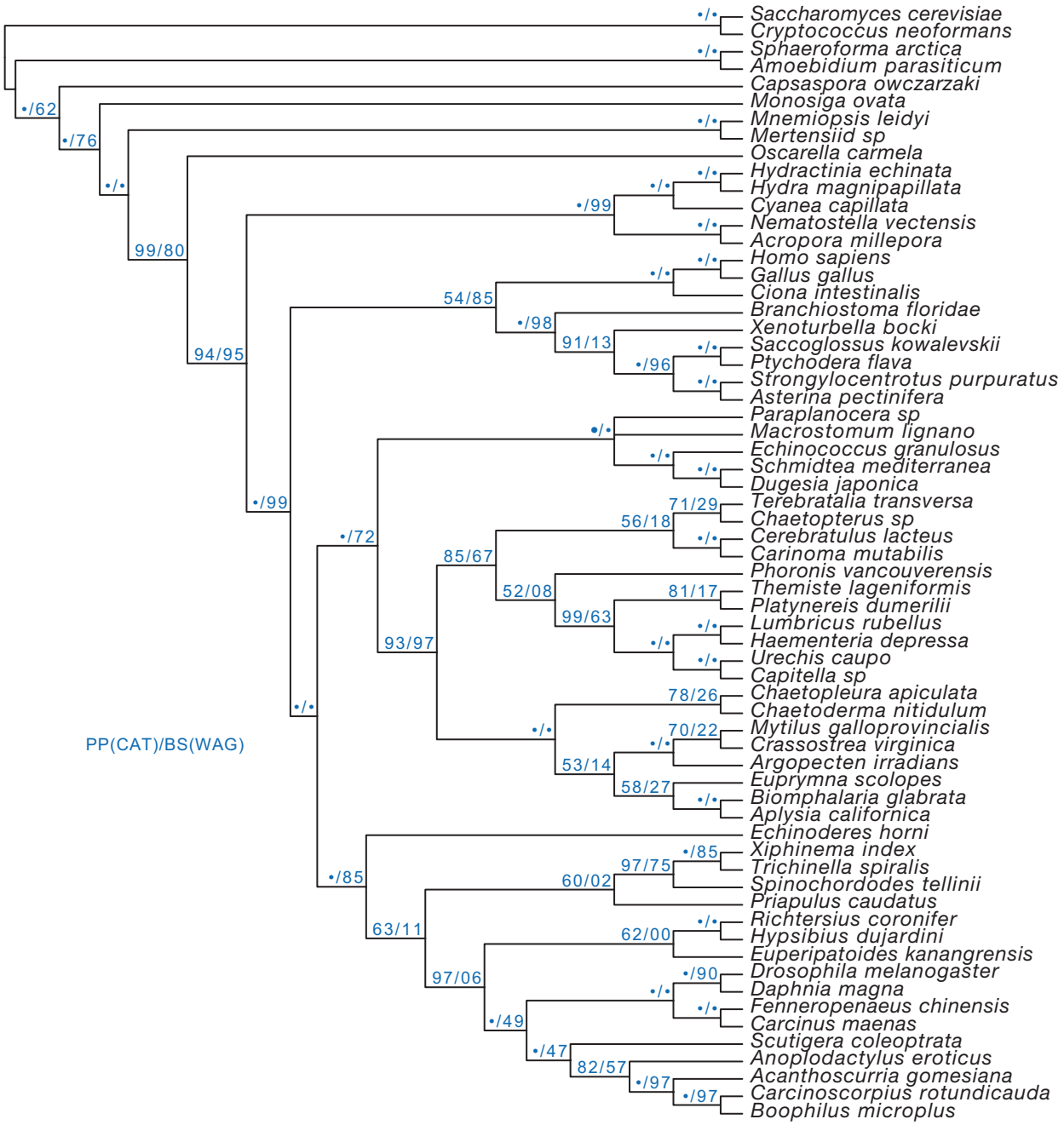
Supplementary Figure 2 | Flow chart of data analysis. "Groups" are sets of genes that are hypothesized to be homologous to each other.



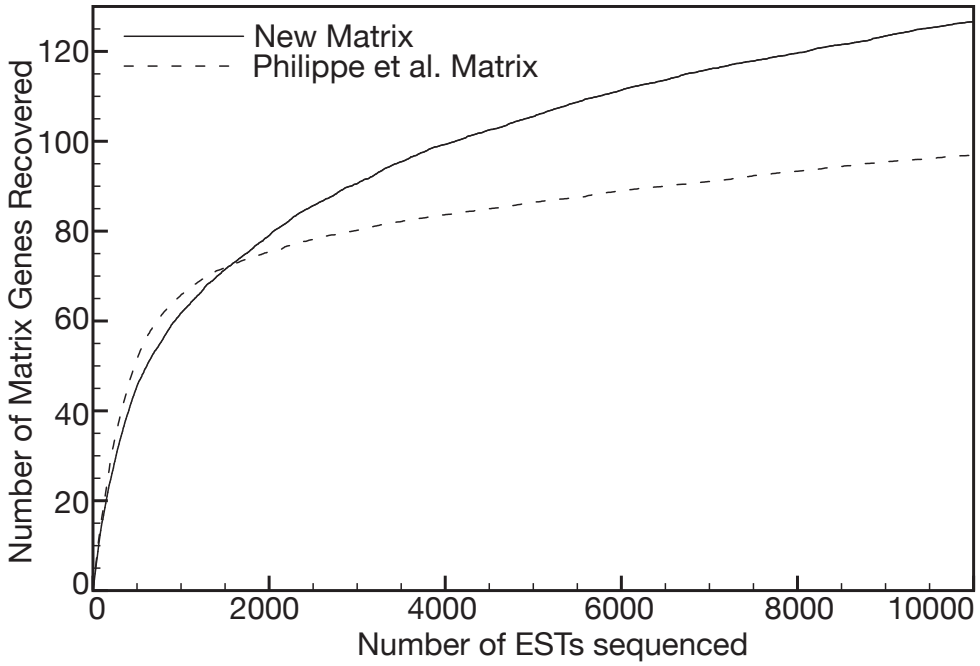
Supplementary Figure 3 | Diagram of gene sampling. Each cell is colour coded to indicate how many of the 150 genes selected for phylogenetic analysis are shared between two corresponding taxa. The diagonal shows how many selected genes were found in a given species. Species are ordered with respect to the number of selected genes.



Supplementary Figure 4 | Venn diagram of gene overlap between the new matrix assembled here, the matrix assembled by Philippe et al.[9], and the matrix assembled by Rokas et al.[13]

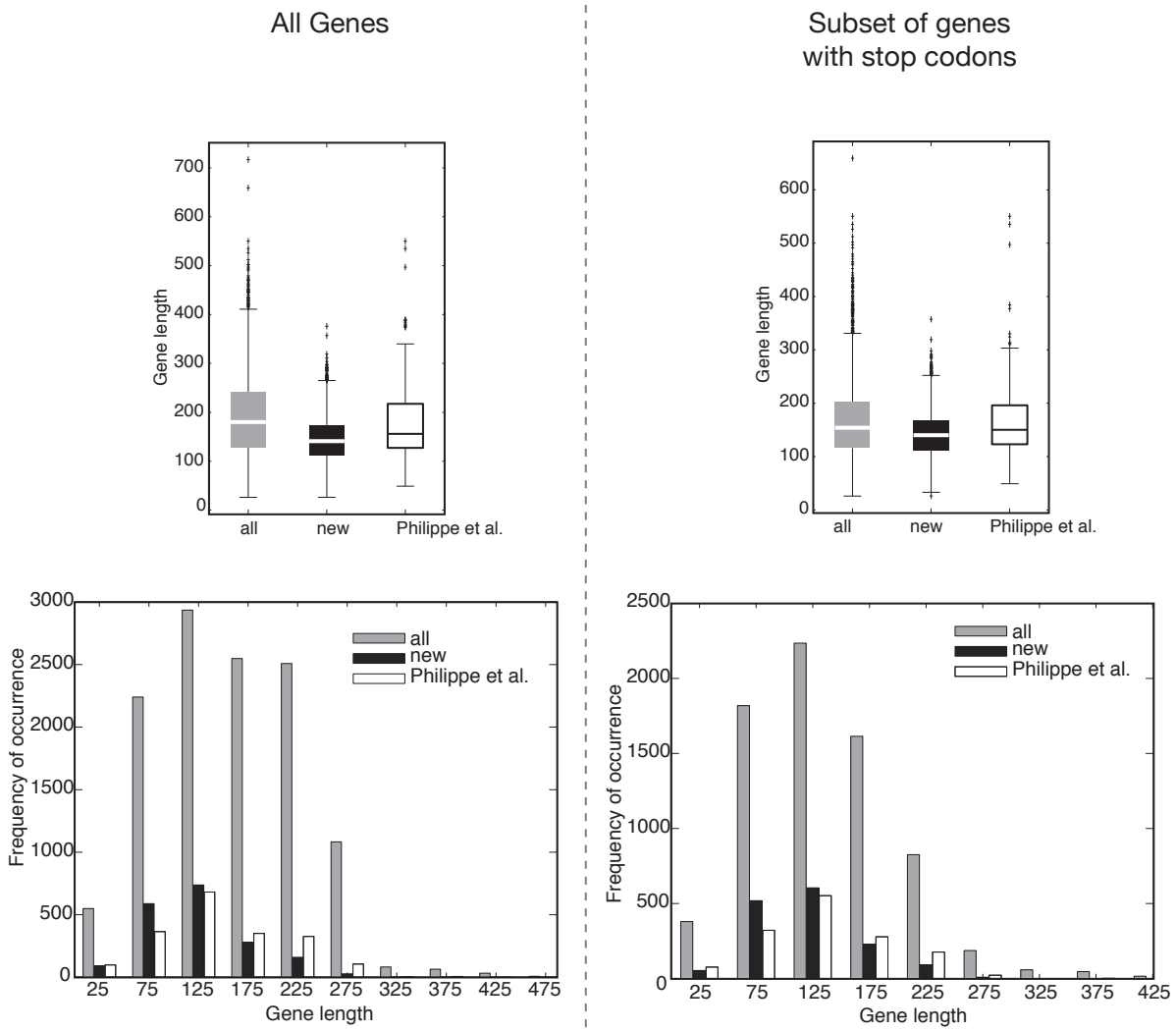


Supplementary Figure 5 | Cladogram of 64-taxon analyses of the predefined list of genes used in most other animal phylogenomic studies[9]. The figured topology was recovered from the Bayesian analysis conducted under the CAT model. Posterior probabilities (PP) were estimated under the CAT model (12 Phylobayes runs of 5000

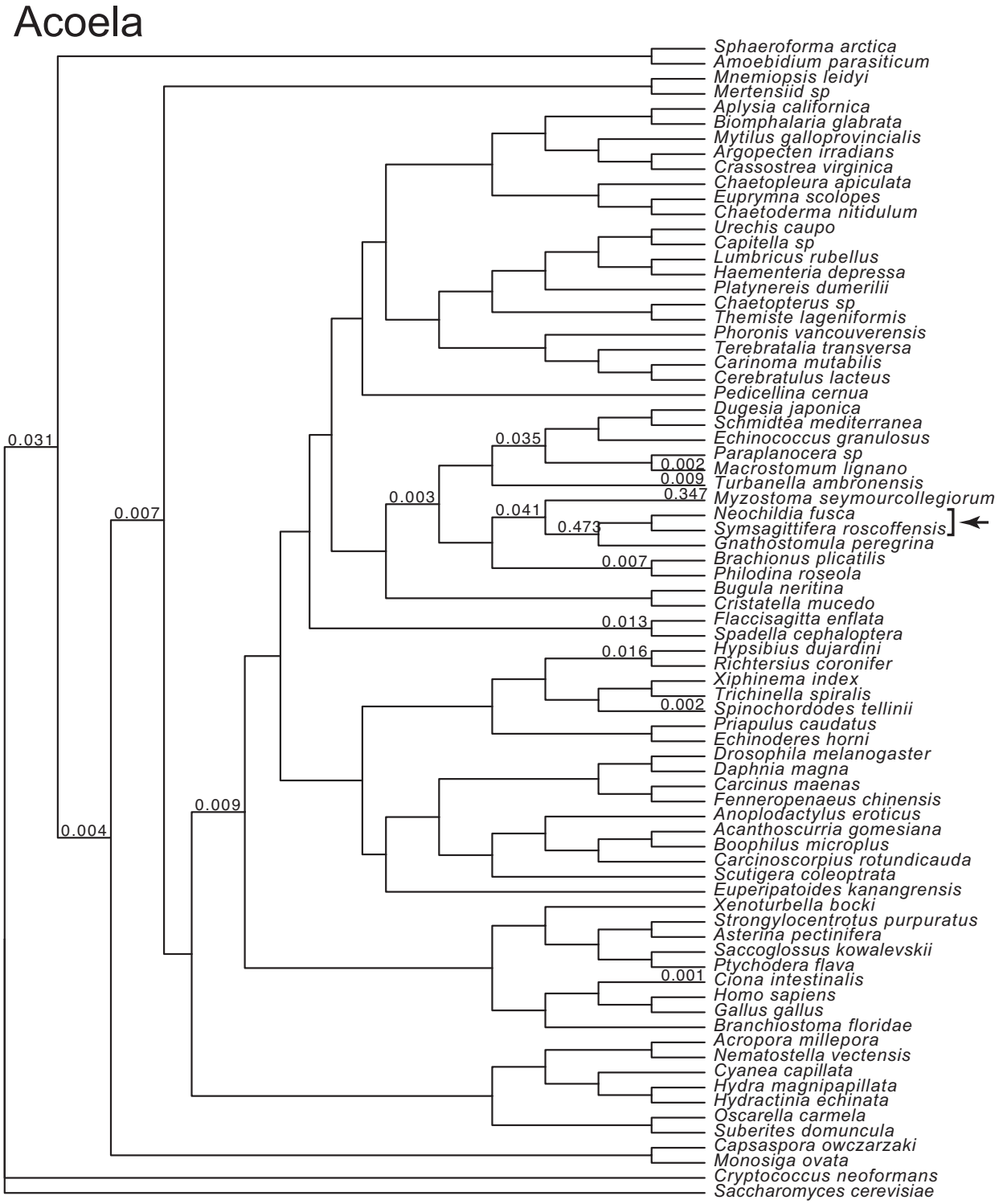


generations each; 1000 generation burnin). ML bootstrap support (BS) was calculated from 1000 bootstrap replicate analyses (RaxML, WAG+Mixed rates model).

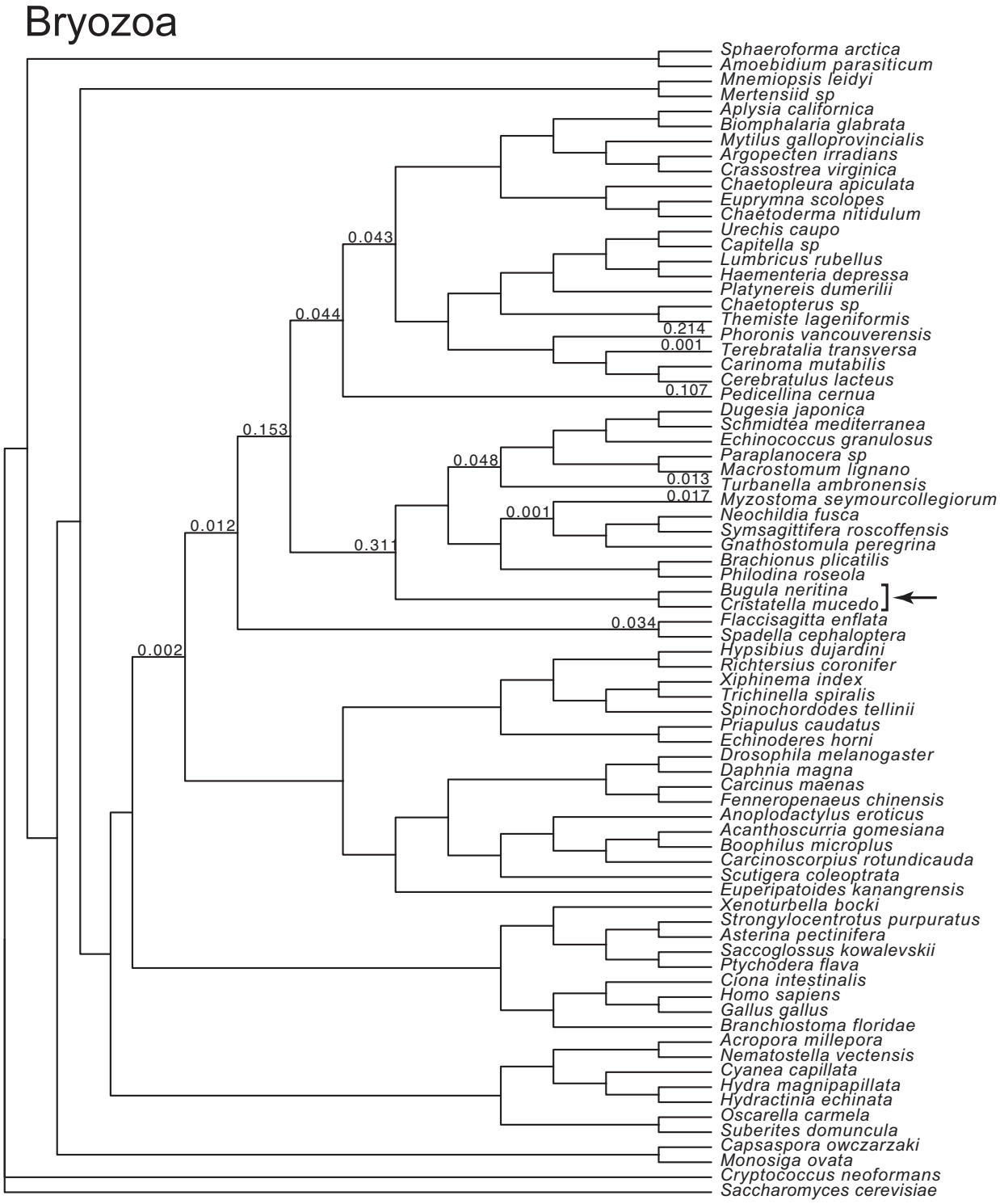
Supplementary Figure 6 | Comparison of gene accumulation curves for our data matrix in relation to that of Philippe et al.[9] ESTs were pooled across the 29 species for which new data were collected. Accumulations were averaged over 50 rarefied replicates.



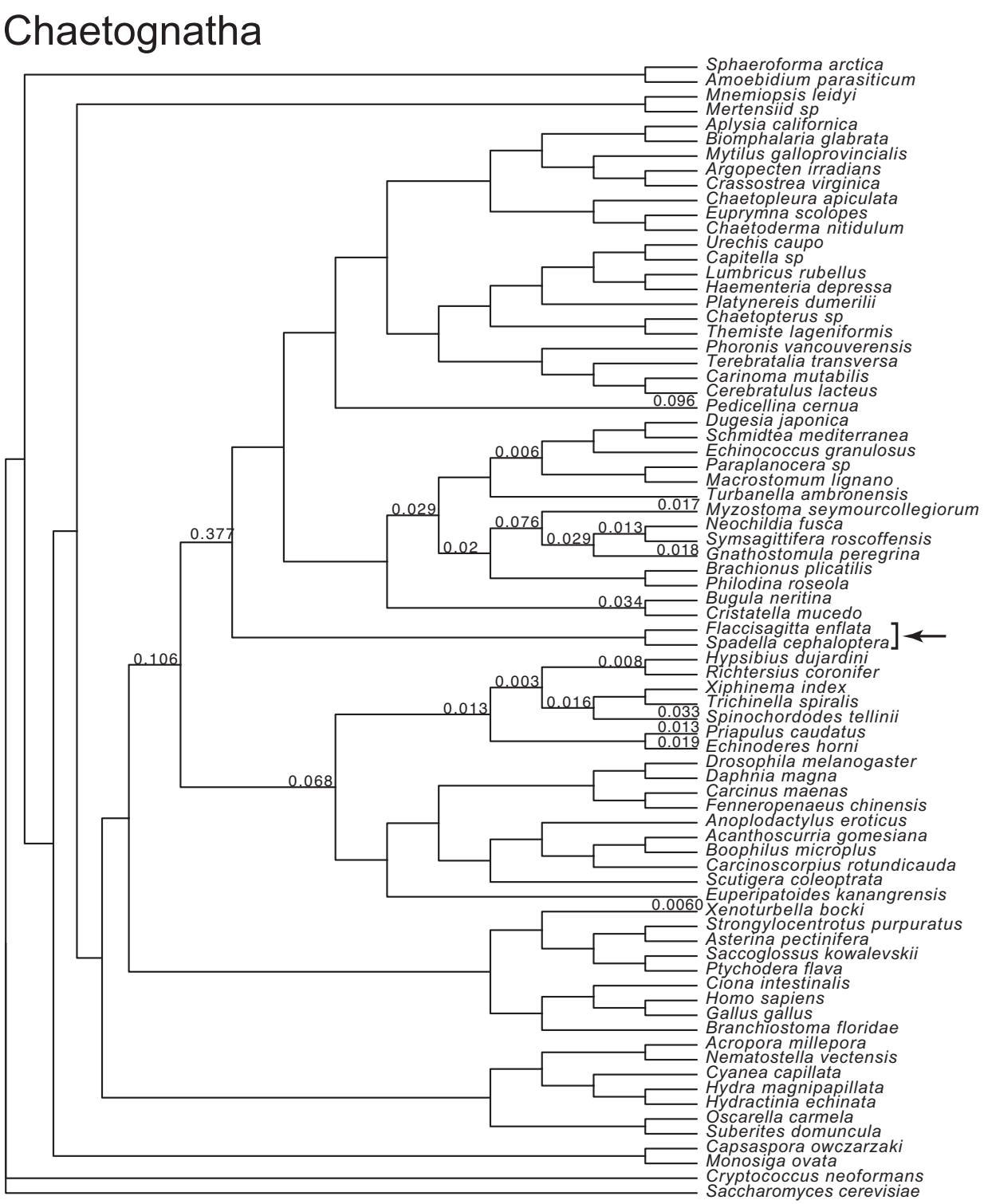
Supplementary Figure 7 | Distributions of gene lengths (number of amino acids) for newly sequenced data. "Philippe et al." designates genes that are orthologs of those in the matrix compiled by Philippe et al.[9], while "new" designates genes assigned to the 150-gene matrix generated here. The distributions of all genes are shown on the left half of the figure, while only the subset of genes with stop codons are shown on the right half. Only those genes translated by similarity (with BLASTX e-value < $1x10^{-8}$ to SWISSPROT, http://www.expasy.org/sprot/) and extension are considered. The boxes indicate the lower quartile, median, and upper quartile; the whiskers indicate the most extreme values within 1.5 times the interquartile range.



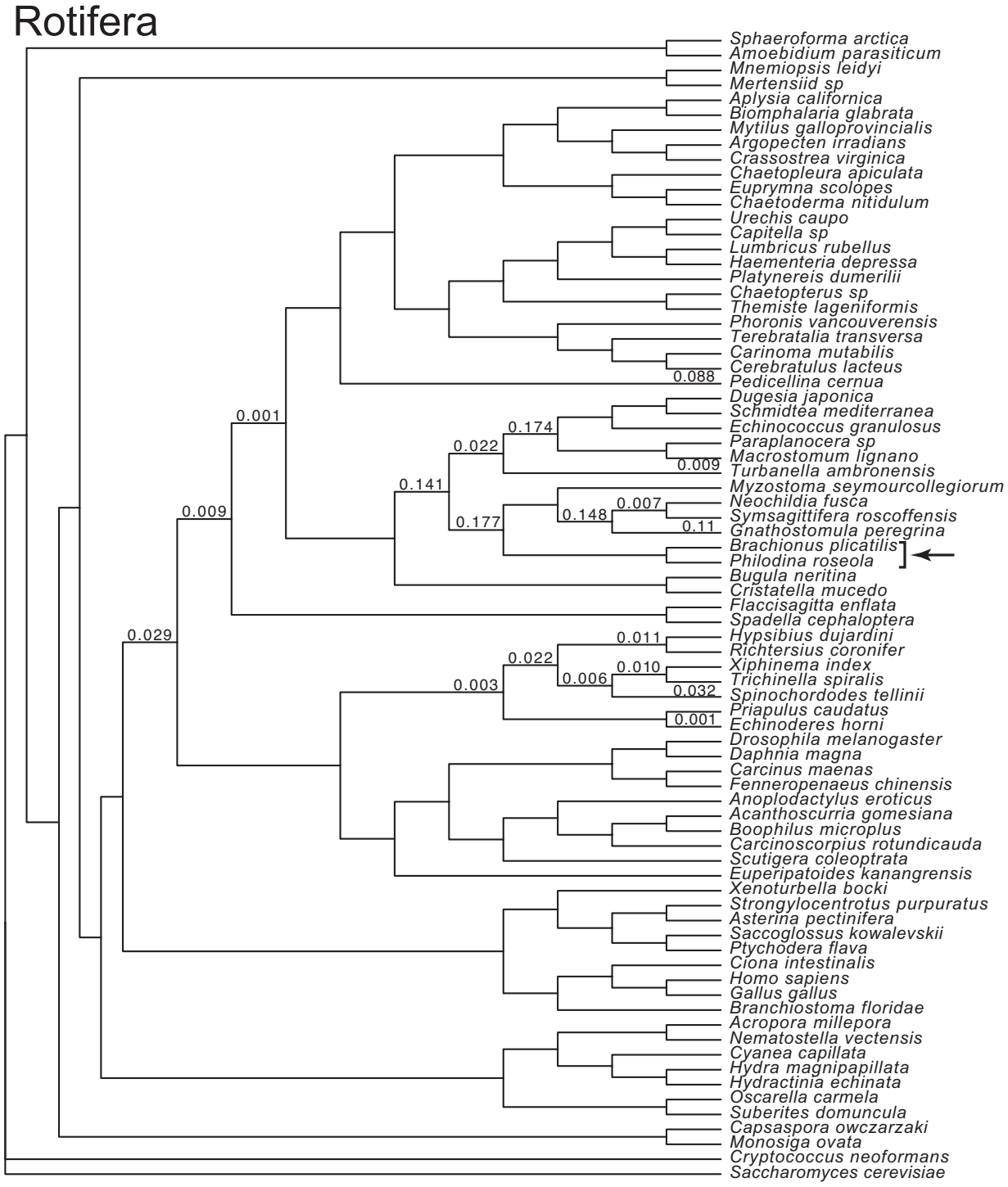
Supplementary Figure 8a



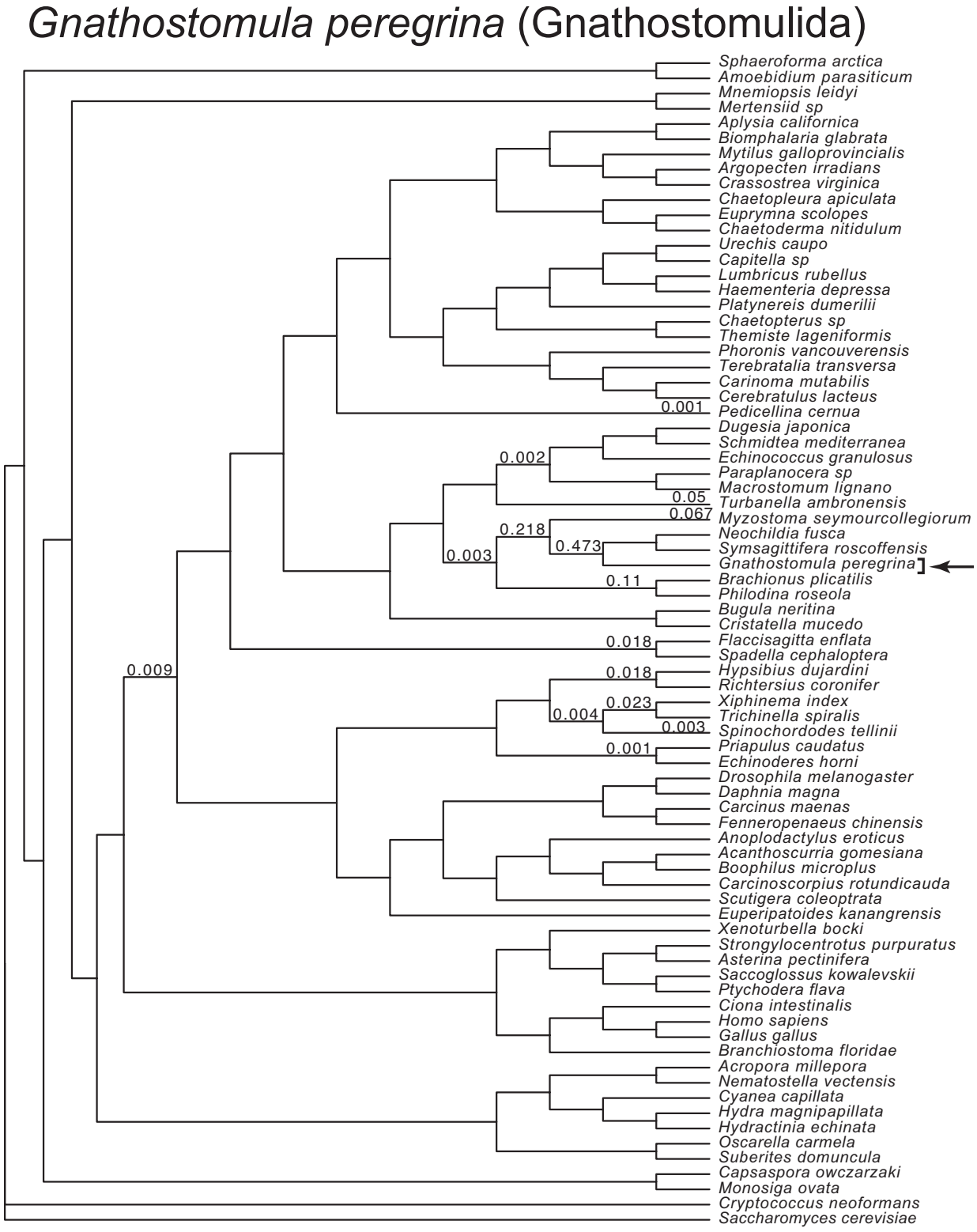
Supplementary Figure 8b



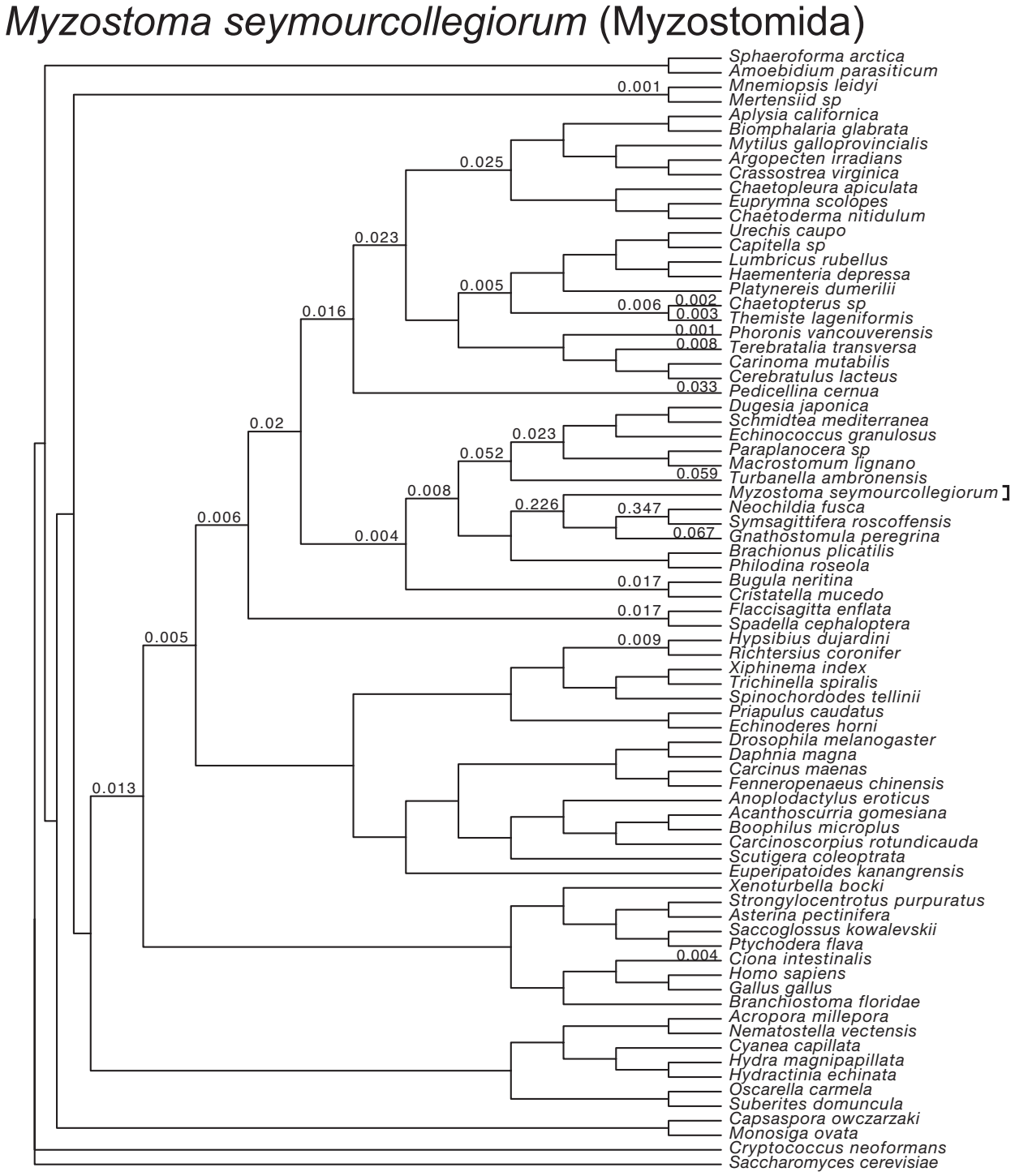
Supplementary Figure 8c



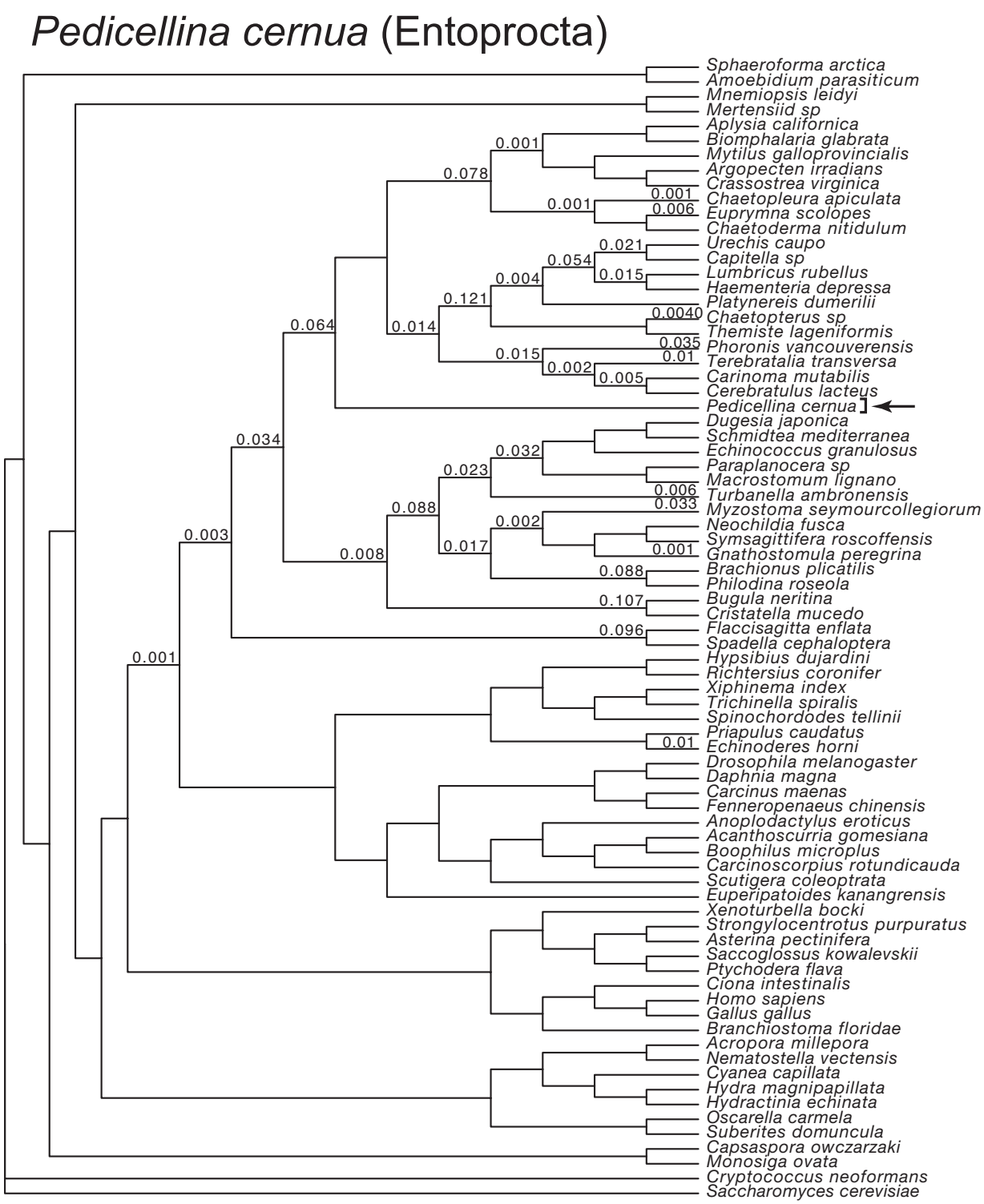
Supplementary Figure 8d



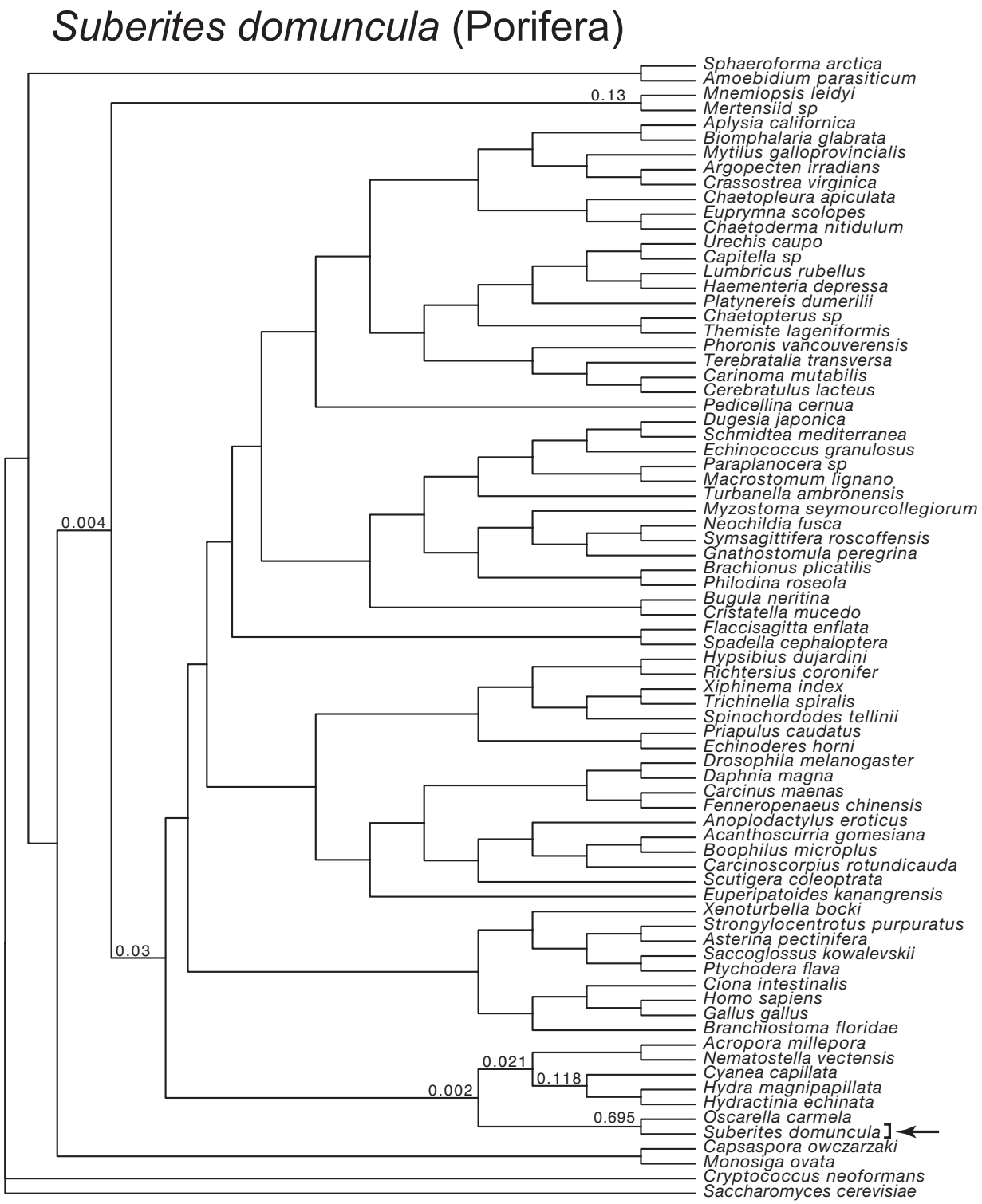
Supplementary Figure 8e



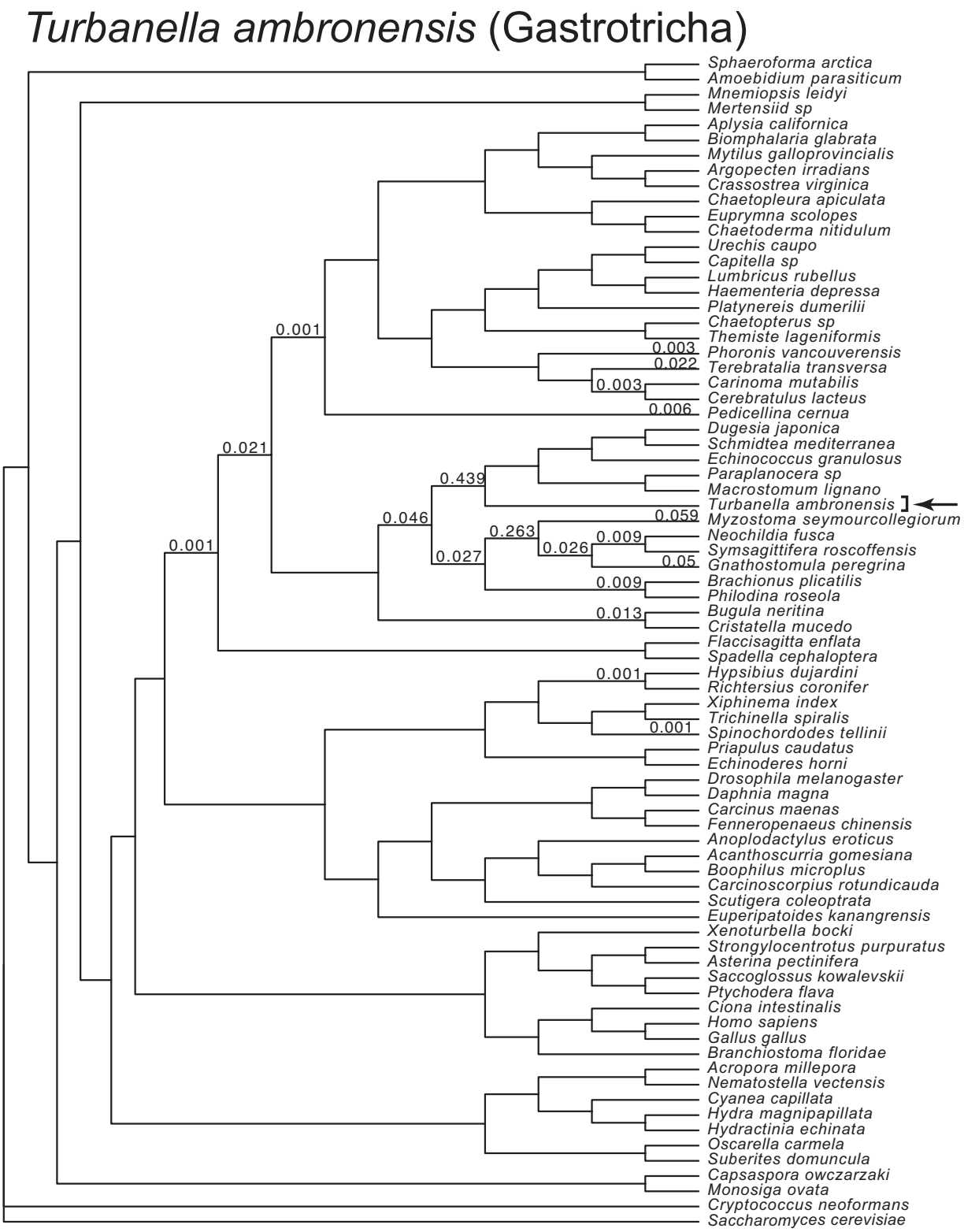
Supplementary Figure 8f



Supplementary Figure 8g

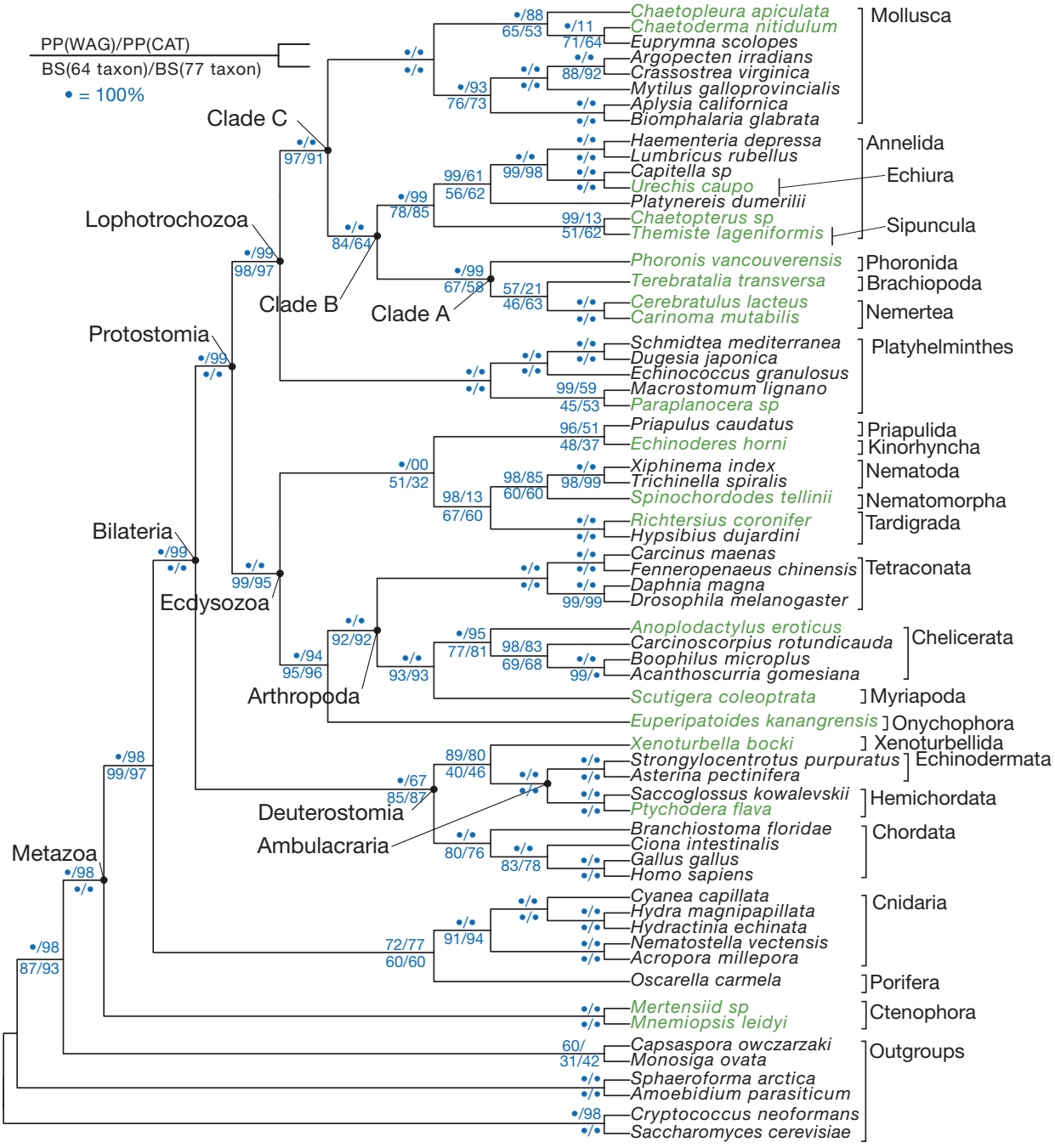


Supplementary Figure 8h

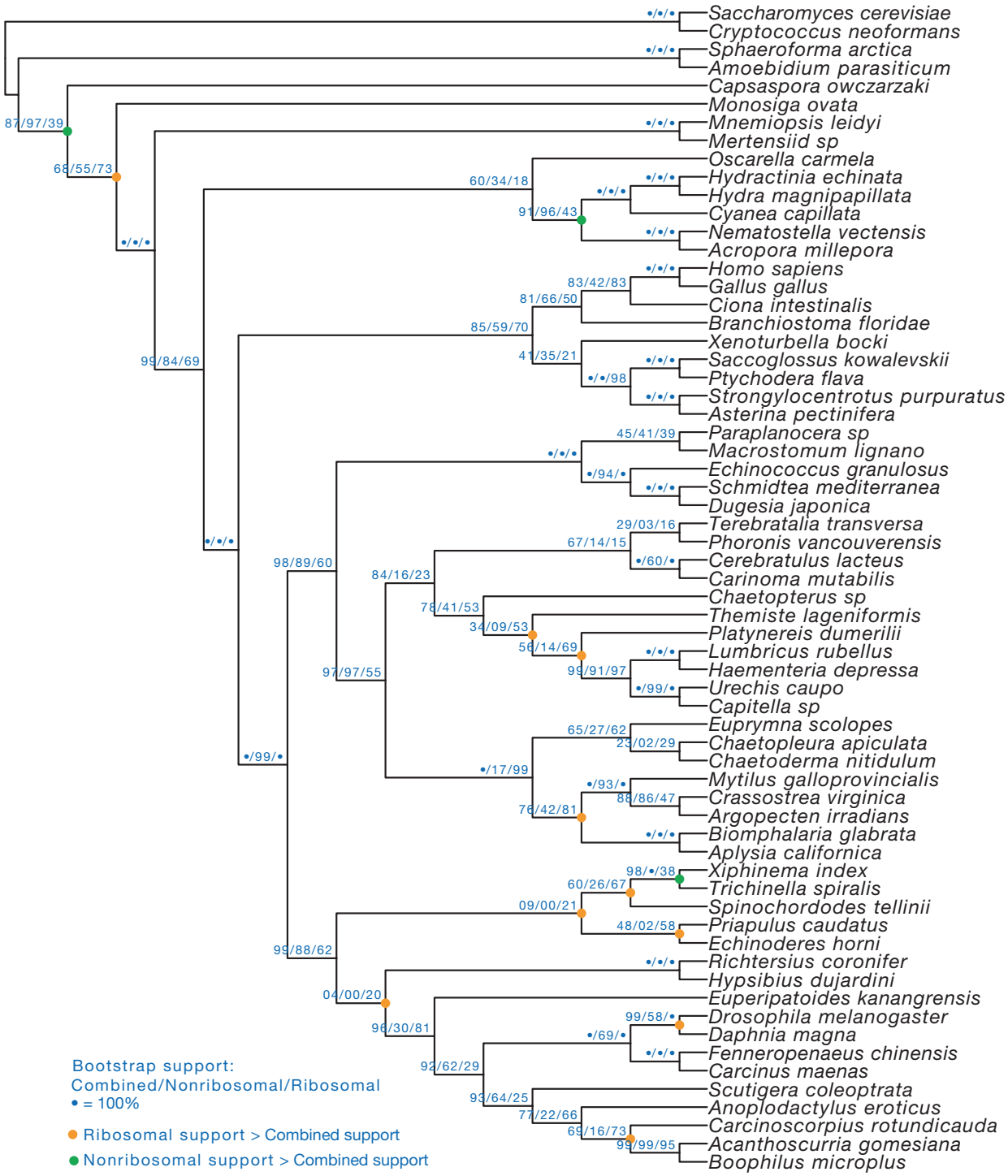


Supplementary Figure 8i

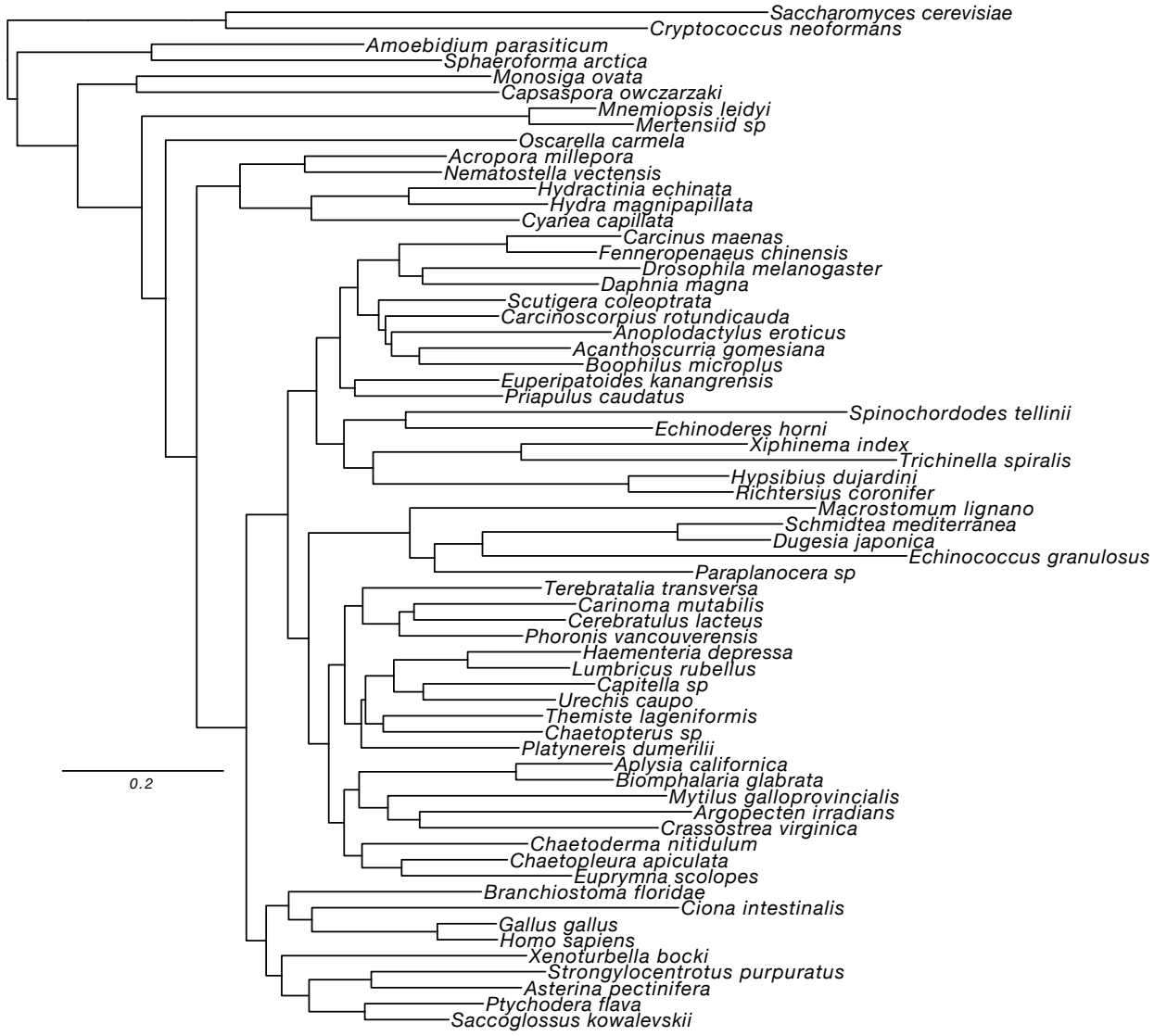
Supplementary Figure 8 | Alternative positions of unstable taxa in 77-taxon ML bootstrap analyses. The topology is the best-known tree found in the 1000 ML searches. The number along each branch indicates the fraction of trees in which the focal taxon attaches along that branch. Unlike bipartition support (as in consensus trees), terminal branches can also have values (which indicate that the focal taxon attaches along the terminal branch). The fraction for the most likely position is indicated along the branch subtending the node that gives rise to the stem of the focal taxon. Values of 0 are omitted for clarity.



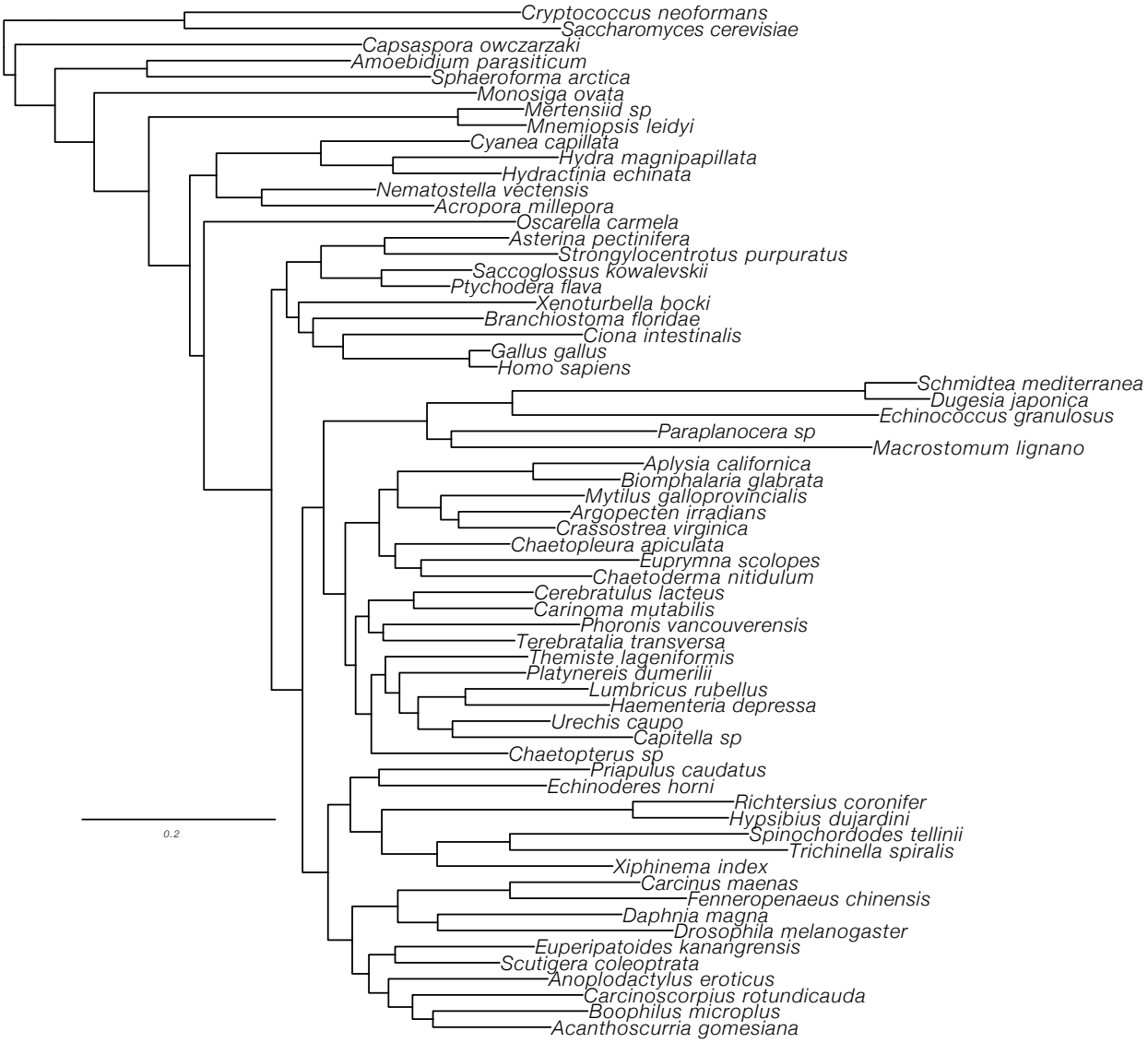
Supplementary Figure 9 | Cladogram of 64-taxon analyses. The figured topology is for the best known tree found in 1000 searches (WAG+Mixed rates model, log likelihood= -699741.6), with support values calculated from the same bootstrap and posterior treesets as Fig. 2.



Supplementary Figure 10a



Supplementary Figure 10b



Supplementary Figure 10c

Supplementary Figure 10 | Independent analyses of ribosomal and non-ribosomal proteins (calculated for the 64 stable taxa). a- ML bootstrap support from ribosomal and non-ribosomal proteins (1000 bootstrap replicates, RaxML, WAG+Mixed rates model), mapped onto the cladogram from Fig. 2. Bootstrap support for the combined matrix is reproduced from Fig. 2 for convenience. b,c- Phylograms of the sampled trees with the highest likelihood (500 replicate searches, RaxML, WAG+Mixed rates model) for the 110 non-ribosomal proteins (log likelihood = -491006.9) and 40 ribosomal (log likelihood = -207699.5), respectively.

Supplementary Notes

Additional references cited in Supplementary Information

41 F. Marletaz, E. Martin, Y. Perez et al., *Curr Biol* **16** (15), R577 (2006).
42 E.T. Munoz, L.D. Bogarad, and M.W. Deem, *BMC Genomics* **5**, 30 (2004).
43 Y. Yang, S. Cun, X. Xie et al., *FEBS Lett* **538** (1-3), 183 (2003).
44 D. Q. Matus, R. R. Copley, C. W. Dunn et al., *Curr Biol* **16** (15), R575 (2006).