nature

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

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

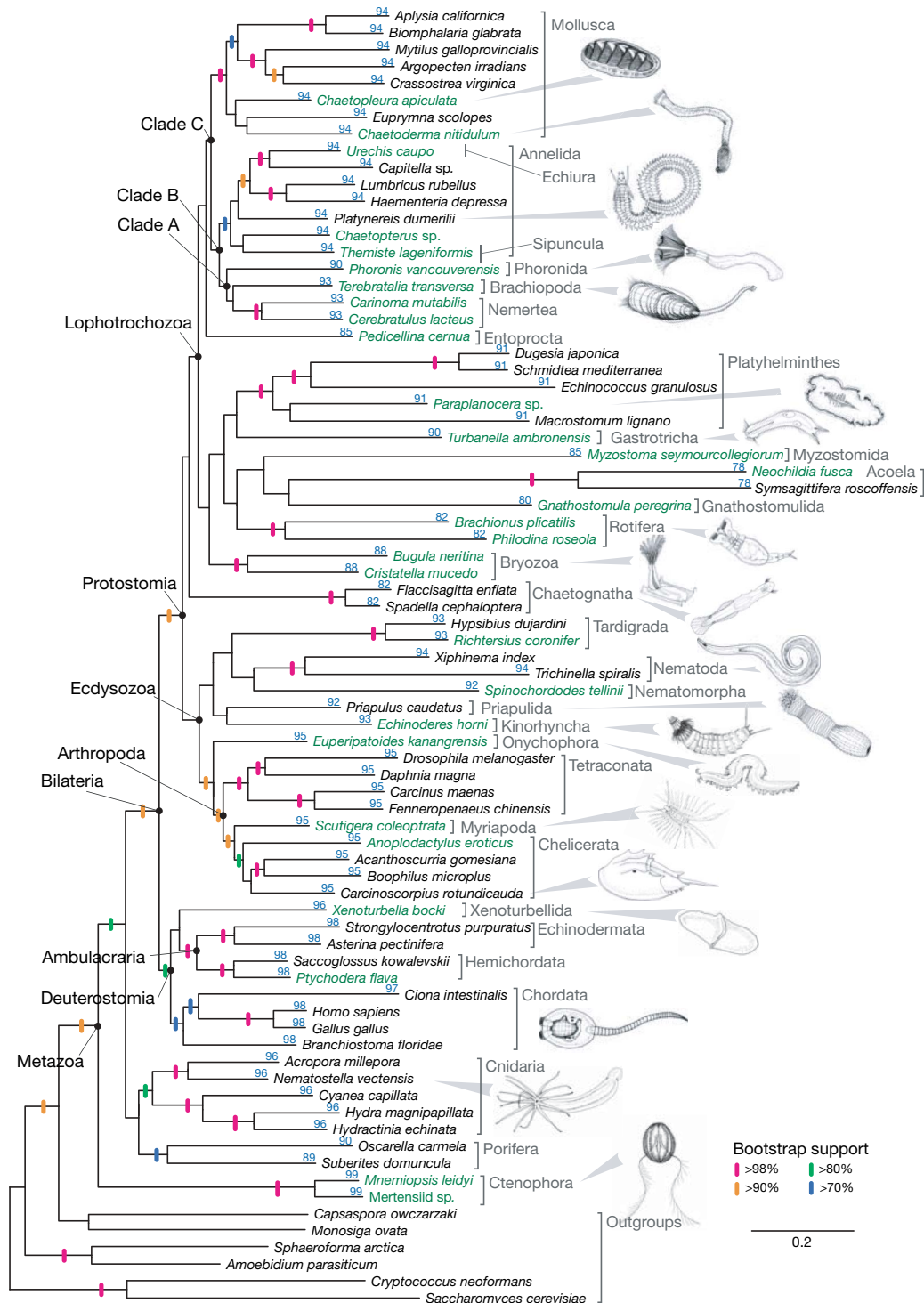**Figure 1 | Phylogram of the 77-taxon RaxML maximum likelihood analyses conducted under the WAG model.** The figured topology and branch lengths are for the sampled tree with the highest likelihood (1,000 searches, log likelihood = −796,399.2). Support values are derived from 1,000 bootstrap replicates. 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
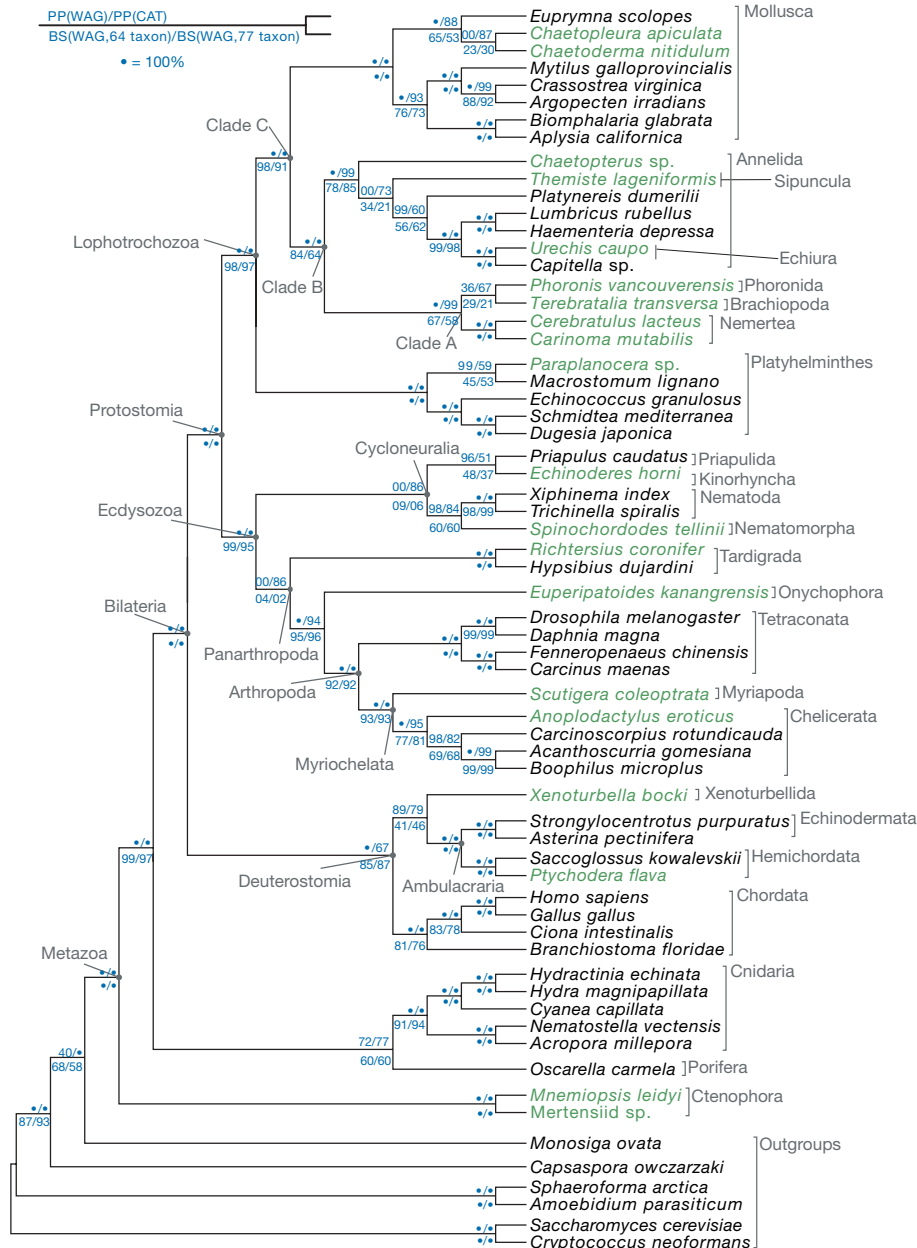

Figure 2 | Cladogram of the 64-taxon PhyloBayes bayesian analyses conducted under the CAT model. Posterior probabilities (PP) estimated under the CAT (15 PhyloBayes runs of 6,000 generations each; 1,200 generation burn-in) and WAG+I+Γ (8 MrBayes runs of two-million generations each; 125,000 generation 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.

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 (2007).
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).

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).