www.sciencemag.org/content/342/6164/1242592/suppl/DC1

Supplementary Materials for

The Genome of the Ctenophore *Mnemiopsis leidyi* and Its Implications for Cell Type Evolution

Joseph F. Ryan, Kevin Pang, Christine E. Schnitzler, Anh-Dao Nguyen, R. Travis Moreland, David K. Simmons, Bernard J. Koch, Warren R. Francis, Paul Havlak, NISC Comparative Sequencing Program, Stephen A. Smith, Nicholas H. Putnam, Steven H. D. Haddock, CaseyW. Dunn, Tyra G. Wolfsberg, James C.Mullikin, Mark Q. Martindale, Andreas D. Baxevanis

Corresponding author. E-mail: andy@mail.nih.gov

Published 13 December 2013, *Science* **342**, 1242592 (2013)
DOI: 10.1126/science.1242592

The PDF file includes:

Materials and Methods
Figs. S1 to S10
Tables S1 to S31
References

Ryan et. al.: The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution

Supplementary Materials:
Materials and Methods
Tables S1 to S31
Figure S1 to S10
References

Material and Methods

Source material and sequencing

We used cteno-dippers (i.e., a beaker secured to a broom handle) to carefully collect adult *Mnemiopsis leidyi* from the Vineyard Sound near Woods Hole, Massachusetts, USA. We isolated genomic DNA from the embryos of two self-fertilized adults. DNA from one embryo pool was used to construct a library for sequencing using a Roche 454 Genome Sequencer FLX machine (Roche Applied Science, Indianapolis, IN). One picotiter plate was run in the 100-cycle mode using FLX chemistry yielding an average read length of 236 bases. Eight more plates were run in the 200-cycle mode using Titanium chemistry with an average read length of 334 bases. We generated 7,334,972 raw reads from these nine runs, which yielded 2.5 Gb of sequence. The initial libraries were made with 'GS FLX Titanium Rapid Library Preparation Kit' and subsequent libraries were made with the same kit along with the 'GS FLX Titanium Library Paired End Adaptors Kit'.

Genome assembly

Using the Phusion assembler (*54*), we assembled this data into 24,884 contigs, constituting 150,340,428 bases of sequence. The N50 of our contig assembly is 11,936 bases.

DNA from the other embryo pool was used to create two mate pair libraries for Illumina GA-iiX sequencing, one with a 3-kb insert and the other with a 4-kb insert. Paired reads were 51 bases each. After removing duplicate read-pairs, 4.2 million and 2.6 million pairs remained for the 3 and 4-kb libraries, respectively. We mapped these mate-pair reads to the assembly using Illumina's short read aligner ELAND and integrated these mappings into Phusion's scaffolding process. The final assembly consists of 5,100 scaffolds resulting in 160-fold physical coverage and an N50 of 187 kb. See Table S5 for scaffold size frequencies and Table S6 for scaffold gap frequencies.

Estimation of variation

To estimate the level of genetic variation between animals, we sequenced a lane of genomic material from the embryos of another self-fertilized animal using an Illumina Mi-Seq and aligned these reads to our assembly. We identified 589,252 positions with single nucleotide variations (SNVs) across 118,613,222 bases that were covered well enough to call SNVs. Therefore, we conservatively estimate the occurrence of SNVs between these two animals to occur once every 200 bases. This result suggests that there is ample variation between animals for population studies.

Evaluation of completeness and correctness of genome assembly

We estimated the completeness of our genome assembly by aligning publicly available (GenBank) *M. leidyi* ESTs (15,752) to the assembly using BLAT (*55*) (version 34x12) with default parameters. We used baa.pl (*56*) (version 0.20) with default settings to generate the following statistics: (1) 99.4% of the transcripts were mapped with BLAT; (2) 98.2% of the positions in the mapped transcripts were aligned; and (3) 95.2% of the transcripts mapped to a single scaffold.

We also generated 79 Mb of paired-end reads and 83 Mb of single-end reads of RNA-seq data from mixed-stage *M. leidyi* embryos (15-30h post-fertilization) through Illumina GA-II sequencing. We assembled these two runs of RNA-seq reads using Trinity (*57*) (version r2012-10-05) into 32,630 and 27,315 transcripts respectively. As was done with the public ESTs, we used BLAT to align these transcripts to our scaffolds. We used baa.pl (*56*) with default settings to generate the following statistics for the RNA-seq transcripts: (1) 99.2% of the transcripts were mapped with BLAT; (2) 98.1% of the positions in the mapped transcripts were aligned; and (3) 92.7% of the transcripts mapped to a single scaffold.

Large repeats and segmental duplications can be particularly problematic for assemblies based on next-generation sequencing data (*58*). To be sure that we have not collapsed large amounts of identical repeats, we generated a genome-size estimation based on the occurrence of kmers in our reads. Using a kmer of size 17 and excluding low-count kmers (<= 4x) we estimate the genome size to be 200,738,694 base pairs. This is 45MB larger than our genome assembly. We suspect that most of this difference in size is due to SNVs in our sequencing data due to the variation introduced by our sample material (embryos from a self-fertilized hermaphrodite). To test this possibility, we simulated 454 data using the Art next generation sequencing read simulator (*59*), with one haplotype coming from our assembly and another haplotype from a version of the assembly that was permuted once every 200 bases and coverage being 12x. When we estimate the genome size of the simulated data with two haplotypes, we see over-estimates of size in the range 50MB, which is the difference that we see between our assembly and the estimated genome size. Therefore, we suspect that our 156MB genome assembly closely resembles the true genome size and that hidden genome complexity based on collapsed repeats is minimal.

Detecting Novel repeats

Repeats were identified in the *M. leidyi* genome using a large-scale sequence substring-matching program called VMatch (*60*). Repeats were selected according to their length (length >= 50bp), percent identity (100%), and orientation (both direct and reverse-complemented). In a post-processing step, we required five matches to qualify as a repeat. 10.07% of the *M. leidyi* genome was identified as being comprised of repeats (Table S3).

The VMatch algorithm was also used to evaluate the genomic repeat levels for several other species of interest, including *A. queenslandica, M. brevicollis, N. vectensis, T. adhaerens*, and *S. rosetta* (Table S3). *H. magnipapillata* repeats were not evaluated using VMatch due to the excessive number of transposable elements in the *H. magnipapillata* genome (~57%). Total genomic repeats range from 2.10% (for *T. adhaerens*) to 24.69% (for *N. vectensis*) of each individual genome.

Detecting Known repeats

RepeatMasker (*61*), is a tool that identifies known repetitive DNA elements including low-complexity sequences and interspersed repeats. We used RepeatMasker (with crossmatch) to screen the *M. leidyi* genome for known repeats (Table S4).

We determined that 2.49% of the *M. leidyi* genome is comprised of previously classified repeats (Table S4). We did not scan the *H. sapiens, D. melanogaster*, or *N. vectensis* genomes since these genomes were used to generate the Repbase database (*62*) used by RepeatMasker.

RNA sequencing and assembly

We generated 162 MB of RNA-seq data from mixed stage embryos. We mapped this RNA-seq data to our genome assembly using the TopHat (*63*) read-mapping software (length = 51; inner-distance = 295) (version_1.2.0). We assembled these tophat mappings into 49,850 transcript fragments using Cufflinks (*64*).

Overview of gene prediction pipeline

Our gene prediction pipeline consisted of several rounds of automated and manual steps. We loaded the Cufflink fragments, along with 15,775 publicly available EST sequences and 161 publicly available cDNA sequences into PASA (*65*) (Program to Assemble Spliced Alignments, version_08_22_2010). Prior to submitting the *M. leidyi* genome for gene prediction analysis, we masked the genome with our repeat library, except for regions where RNA-seq mappings overlapped. The RNA-seq overlap consisted of 53,244 regions and 7,387,140 base pairs of sequence. We generated protein-coding gene models using various gene prediction programs, including FGENESH (*66*); AUGUSTUS (*67*) (version_2.3.1); HMMgene (*68*) (version_1.1); and GenomeScan (*69*) (version_0.1). We submitted all predicted gene models to EvidenceModeler (*70*) (EVM, version_r03062010), which also considers EST and RNA-seq evidence (through PASA) as well as sequence similarity (through BLASTP of predicted proteomes). The total number of gene predictions for each program is reported in Table S22.

Gene prediction with FGENESH

The FGENESH pipeline, consisting of three main steps, was implemented using partially repeat-masked *M. leidyi* genomic sequence (5100 scaffolds). First, known mRNAs, namely *M. leidyi* RNA-seq data (49,850 Cufflinks transcripts) and publicly available mRNAs (161) and ESTs (15,752), were mapped to the *M. leidyi* genome. Next, genes were predicted using ProtMap, based on homology to known proteins from the following organisms: *H. sapiens, D. melanogaster, A. thaliana, C. intestinalis, B. floridae, C. teleta, D. purpureum, L. gigantea, M. brevicollis, N. vectensis,* and *T. adhaerens*. Finally, FGENESH performed an *ab initio* prediction of genes in the remaining regions having neither mapped mRNAs nor genes predicted based on protein homology. FGENESH predicted a total of 16,367 genes in the *M. leidyi* genome (Table S22).

Gene prediction with AUGUSTUS

AUGUSTUS uses species-specific parameters that can be trained on sets of annotated genes (e.g., the Markov chain transition probabilities of coding and non-coding regions). We trained AUGUSTUS with the *A. queenslandica* gene set. We then ran AUGUSTUS gene modeling by incorporating known gene structure evidence (“hints”) from extrinsic sources, including *M. leidyi* RNA-seq reads (49,850 Cufflinks transcripts) and publicly available mRNAs (161) and ESTs (15,752). Hints were used to search against the *M. leidyi* genome using BLAT (*55*), and the results were combined into a single hints file. AUGUSTUS was executed using this gene structure evidence and partially repeat-masked *M. leidyi* genomic sequence (5100 scaffolds). AUGUSTUS generated a total number of 29,359 *M. leidyi* gene predictions (Table S22).

Gene prediction with HMMgene

HMMgene, a program based on a hidden Markov model of gene structure, was also used to generate *M. leidyi* gene models. HMMgene was run against the partially repeat-masked (VMatch repeats not overlapping RNA-seq reads) *M. leidyi* genomic sequence (5100 scaffolds). HMMgene predicted a total of 13,948 *M. leidyi* gene models (Table S22).

Gene prediction with GenomeScan

We used the GenomeScan gene prediction program to generate gene models using the partially repeat-masked (VMatch repeats not overlapping RNA-seq reads) *M. leidyi* genomic sequence (5100 scaffolds). For the GenomeScan run we supplied sequence similarity information based on BLASTX homology of known proteins from the following organisms: *H. sapiens, D. melanogaster, A. thaliana, C. intestinalis, B. floridae, C. teleta, D. purpureum, L. gigantea, M. brevicollis, N. vectensis,* and *T. adhaerens*. BLASTX results (using the parameters -G 9, -E 2, and -e 0.05) were converted to a “genoa” file in order to submit them to GenomeScan. GenomeScan predicted a total of 6,443 *M. leidyi* gene models (Table S22).

Gene prediction with PASA

We implemented the PASA annotation pipeline. We first used Genomic Alignment and Mapping Program (*71*) (GMAP) version_9_28_2007) to align the known ESTs, mRNAs and RNA-seq reads to the partially-masked *M. leidyi* genome and transcript assemblies were generated. We then trained PASA using the program options for longest ORFs and extraction of terminal exons. The output of PASA was a GFF-formatted set of 153,004 “validated” transcripts, which we used as transcript evidence for consensus gene modeling using EVM.

Choosing between gene predictions with EVM

We used the EvidenceModeler software (EVM) to combine *ab initio* gene predictions and protein and transcript alignments, from all other gene modeling programs outlined above, into weighted consensus gene structures. EVM requires an evidence weights file as input where each piece of evidence is manually assigned a weight (1 to 10) based on data confidence levels. To determine initial weights, we compared the gene predictions from each program (through manual inspection) to 92 experimentally verified transcripts. We determined these 92 transcripts using RACE-PCR (Clontech SMART™ RACE kit) from cDNA of mixed developmental stages of several individuals. Our initial comparisons of gene models from each program to our RACE sequences showed that FGENESH and Augustus greatly outperformed the HMMgene and GenomeScan programs. This is perhaps not surprising, since our runs of HMMgene and GenomeScan did not take into consideration transcript information. Based on these initial evaluations, we set the initial weighting of FGENESH and Augustus higher than HMMgene and GenomeScan.

We performed several rounds of weighting adjustments based on subsequent manual inspections after each run (Table S23). We compared EVM predictions to the 92 full-length RACE transcripts by visually evaluating their gene structures using their designated transcript tracks in the JBrowse genome browser. Specifically, we inspected the gene structure characteristics (e.g., exons, introns, splice junctions, and other genomic features) with regard to quantity, length, and location and categorized any discrepancies as joins, splits, partials, or missed predictions (Table S24).

In our initial EVM run, we generated 14,537 predicted genes, 85,446 exons, and 2,037 scaffolds. These were determined using four gene prediction programs (FGENESH, AUGUSTUS, GenomeScan, and HMMgene), the PASA validated transcripts, and RNA-seq (Cufflinks transcripts) data. Manual evaluation of EVM consensus gene models was then conducted using JBrowse as described above. (Table S25). In the course of our manual inspection of gene prediction results, we noticed that many of the mispredictions corresponded with incorrect GenomeScan and HMMgene predictions, despite the output from these two programs being heavily down-weighted. Removing GenomeScan and HMMgene gene models from the next round of EVM runs led to improved results.

The second EVM run with only FGENESH and AUGUSTUS predictions produced 14,835 predicted genes, 86,712 exons, and 1,998 scaffolds. When we compared these results to our RACE sequences we found that these results were an improvement over the previous run (Table S25).

Since FGENESH consistently outperformed AUGUSTUS we tried removing the AUGUSTUS predictions from consideration and ran EVM with FGENESH only. These new set of results resulted in a 12% increase in the total number of gene predictions (16,845 predicted genes, 89,564 exons, and 1,915 scaffolds) and resulted in an overall improvement in terms of accuracy towards predicting the RACE sequences (Table S25). We were not able to improve upon this run.

To make sure our RACE sequences were accurately represented in our gene models, we processed one final run that included our test set (the RACE sequences) in PASA and reran EVM with FGENESH as the only set of predictions. This run produced 16,545 gene predictions, 91,482 exons, and 1,748 scaffolds. In the few cases where RACE sequences were wrongly predicted or missed completely, we manually replaced or added the correct sequence. We later added a few additional manual additions and changes based on manual refinements made during subsequent studies (*72-74*). Statistics for the final gene set are in Table S9. Statistics for the genome, in terms of coding vs. non-coding sequences, are in Table S10.

Nested intronic genes

Genes that occur in the introns of other genes are called nested intronic genes. Table S2 shows that *M. leidyi* has a high number of nested intronic genes compared to other genomes where these genes have been characterized. We see some examples of interesting functional relationships between nested and "host" genes. For example, a gene likely involved in the cell cycle, ML41156a (BLAST hit to human cyclin B2) is embedded within the first intron of a gene likely involved in DNA Repair, ML41157a (BLAST hit to human DNA-repair protein RAD52). Similarly, we found a gene that codes for a likely RNA-binding protein, ML431110a (BLAST hit to human RAD52) embedded within the intron of a likely RNA-editing enzyme, uridylyltransferase (BLAST hit to human ZCCHC11). A more in-depth analysis of these genes may lead to some interesting regulatory interactions between nested genes and their host genes.

We suspect that some of the predicted nested intronic genes are incorrect predictions. We manually examined many of these nested gene models in some depth and found examples of (1) seemingly bona fide nested genes, (2) likely spurious predictions based on rare isoforms of the "host" gene, and (3) many cases that are not clear.

An example of bona fide nested genes includes ML000128a and ML000127a, which are situated within the fourth and seventh introns of ML000126a respectively (Fig. S5). ML000126a has an E2F_TDP domain that spans the fourth and fifth intron, indicating that it truly encompasses the ML00128a gene. ML000128a is a G-protein coupled receptor with an Ldl_recept_a domain. According to Pfam, the Ldl_recept_a domain does not co-occur with E2F_TDP domain, suggesting that the 2-exon ML00128a gene is not part of an ML00126a isoform. ML00127a is translated on the opposite strand from ML00126a, and it has two clearly defined Glyco_hydro domains, suggesting that it is a bona fide gene.

ML000317a is a single-exon prediction within the 16th intron of the ML000316a gene model. These two predictions are in the same orientation (Fig. S6). We find evidence, based on paired end reads from independent transcriptomic sequencing (Dr. William Browne, University of Miami, personal communication), suggesting that the single-exon gene model ML000317a is actually an exon within an isoform of the "host" ML000316a. This makes ML000317a likely a misprediction and not a bona fide nested gene. These findings are noted on the gene wiki pages of both the ML000316a (http://research.nhgri.nih.gov/mnemiopsis/wiki/index.php/ML000316a) and ML000317a (http://research.nhgri.nih.gov/mnemiopsis/wiki/index.php/ML000317a), and will be corrected in the next version (ML2.3) of the gene models. We encourage users to make annotations on the gene wiki pages, so that they can be incorporated into future versions of the *M. leidyi* gene models.

As additional *M. leidyi* RNA-seq data and sequence data from other ctenophores become available, we will be able to improve upon our existing gene models. We expect the number of nested genes to decrease, but to remain high compared to other animals. A more in depth look at these nested genes, will likely lead to interesting discoveries.

Testing for potential non-metazoan contaminants in ML2.2 gene models

To address concerns about possible contamination in the *M. leidyi* gene models, we computed an alien index for each ML2.2 gene model, as described in Gladyshev et al., 2008 (*75*). An alien index is computed as the log-transformed difference between the best BLAST e-value to a metazoan hit in the NCBI non-redundant (NR) database, and the best BLAST e-value to a non-metazoan hit in NR. Of the 16,548 M. leidyi gene models, we found 136 models (<1%) with alien indices above the threshold of 45 (i.e., possible non-metazoan contaminants or horizontally transferred sequences). Ninety-six of these had corresponding transcript data suggesting through independent sequencing evidence that these were not the results of contamination. We included notes in the corresponding wiki pages of the remaining 40 models at the *Mnemiopsis* Genome Portal warning that these sequences could possibly be non-metazoan contaminants.

Ortholog assignments

We generated sets of genes with putative orthology using sequence similarity based on BLAST (*76*) and relative position in a predetermined phylogenetic tree. We used as our input tree Table S26. We assigned bit scores to hits between each pair of genes by summing those for initial BLASTP high-scoring segments found on the same pair of genes, in consistent order, and overlapping less than five percent (with bit scores penalized proportional to the amount of overlap). We determined orthologous sets of genes at each tree node in two steps. First, if a set or gene from one child of the node was in a mutual best hit-relation with a set or gene from the other child, they were combined into a new set. Second, we considered in descending order all hits within this node's subtree and between the subtree, and all outgroup genes. A hit to an outgroup gene blocked any further merging of a gene or set (until another tree node was visited), while a hit between two sets or genes within the subtree, neither previously blocked, resulted in these being merged into a new set. This orthology computation was based on that described in Putnam and coworkers, 2007 (*77*) with further refinement of the blocking rules.

The use of a predetermined phylogenetic tree in the clustering algorithm makes these clusters unsuitable for definitive phylogenetic analyses. As such, we did not use the resulting clusters for gene selection in the phylogenetic analyses.

Lineage-specific genes

We used the clusters of orthologous genes and identified 7,171 clusters containing only *M. leidyi* genes. These 7,171 clusters contained 9,288 genes. An RNA transcript is associated with 7,798 (84%) of these 9,928 genes. We calculated the number of single-species clusters for each species in our analysis and found that *M. leidyi* is eleventh out of 23 species in terms of numbers of lineage-specific clusters (Fig. S7).

Gene duplications

We looked in the clusters of orthologous genes generated in for lineage-specific duplications. By definition, a cluster with more than one gene from a particular cluster contains N-1 duplications from that lineage, where N is the number of genes in the cluster from that species. We looked at clusters of genes that must have been in the metazoan ancestor by only considering where at least one or both of the non-metazoan eukaryotes *C. owczarzaki* and *M. brevicolis* are present and at least one gene from a metazoan is present. *M. leidyi* has fewer duplicates than all 21 animals in the study other than *Schistosoma mansoni* (Fig. S8)

Conserved synteny

Conservation of long-range gene linkage has been observed among diverse metazoan genomes (*77-80*). We compared the gene composition of the 16 largest *M. leidyi* scaffolds, where each contains genes representing from 20 to 41 ortholog groups to the reconstructed linkage groups of the bilaterian ancestor (*80*) and found no scaffolds showing significant conservation of synteny (relative to the null hypothesis of a random distribution of genes across scaffolds). While a future genome assembly with longer scaffolds would allow a more sensitive search for conservation, the current scaffolds are long enough to detect conservation of the level observed in other studies. For comparison, between chordates and the mollusk *Lottia gigantea*: 38% of scaffolds in the same size range (20-41 ortholog groups) showed significant conservation with at least one ancestral bilaterian linkage group (*81*). Our analysis takes into account the relatively small number of gene families that *M. leidyi* shares with bilaterians, in that the metric we used to compare scaffold sizes between genomes was the number of distinct ortholog groups. If a large percentage of the *M. leidyi* genes that did not cluster with other animal genes represent true orthologs that we were unable to detect, correct identification of these difficult orthologs would increase our statistical power to detect a lower level of synteny conservation. However, if there was extensive conserved synteny, we would expect to detect this given the data and methods employed.

OrthoMCL clusters

Our ortholog assignments required a starting phylogenetic tree and therefore were not suitable for phylogenetic inference. We therefore generated a set of clusters using OrthoMCL (*82*) to use in phylogenetic analysis using default parameters. We used the same set of species that were used to construct the ortholog clusters above, plus added 6-frame translations of a de novo assembly of our *M. leidyi* RNA-seq (35,203 transcripts assembled with Trinity (*57*)) to identify genes present in the transcriptome but missing from the ML2.2 protein set. Of the 57,620 *M. leidyi* Trinity transcript translations incorporated into the clusters, 57,161 were incorporated into clusters that also included ML2.2 proteins or that consisted only of other *M. leidyi* Trinity transcript translations. There were 459 transcript translations were incorporated into 132 clusters that did not include an ML2.2 protein. These transcripts represent gene-structure improvements to existing predictions or genes that were missed completely in ML2.2 and will ultimately be incorporated into the next set of gene predictions (ML2.3).

Gene losses

Using the OrthoMCL clusters we calculated putative losses by identifying absences from clusters where a gene from at least one of the non-metazoan eukaryotes *C. owczarzaki* and *M. brevicolis* and one metazoan are present. We calculated this value for each animal in our analysis. *M. leidyi* has 2,129 putative losses, comparable with *D. melanogaster* with 2,123. The losses in *M. leidyi* are above average but fall within one half a standard deviation of the median of 1,875 (Fig. S9).

Pfam domain content

We used the HMMER suite v3.0b (*83*) and Pfam v24.0 database (*84*) to identify protein domains in the *M. leidyi* genome. We also searched the predicted proteomes of several other animal and related non-animals. We have compiled a list of domains that appear to have been lost in *M. leidyi* (Tables S27 and S28). Each proteome was scanned using the HMMER program hmmscan on default settings. For each domain model in Pfam, if the domain was detected in at least one peptide sequence with a HMMER independent expect-value equal to or below .01, the domain was considered to be present in that proteome. We also detected an additional 94 domains in the *M. leidyi* genome (with E-Values better than 0.001) that did not occur in any of our gene predictions; we count these as present in Table S27 and they are not included in Table S28). These regions may be pseudogenes or coding genes that were not annotated.

Construction of the Genome Set amino acid matrix

We started with the 242-gene data matrix (104,840 columns, 12 animals) from the *A. queenslandica* genome paper (*80*). In this study, the authors generated mutual best hits from 17 proteomes to genes in the *M. brevicolis* genome. Hits with E-values ≥ 0.001 were discarded. To avoid paralogs, hits were only kept if the score of the second best hit in both directions was less than half of the score of the best hit. These hits are considered filtered mutual best hits. We used this same criteria to produce filtered mutual best hits from the predicted proteomes of the following species: *Mnemiopsis leidyi*, *Salpingoeca rosetta*, *Capsaspora owczarzaki*, *Sphaeroforma arctica*, *Spizellomyces punctatus*, and *Saccharomyces cerevisiae*. For each of the 242 genes in the original data matrix, a file with the full-length amino acid sequence from each species with a filtered mutual best hit was created. These sequences were aligned using MAFFT (*85*) with the following command: `mafft --ep 0 --genafpair --maxiterate 1000 FILE.fasta > FILE.mafft; einsi FILE.mafft > FILE.einsi`. Next we trimmed each alignment with Gblocks (*86*) using the following command (where X=0.65 * number of seqs in FILE.einsi): `Gblocks FILE.einsi -b2=X -b3=10 -b4=5 -b5=a`. These 242 individual alignments were then concatenated to create the Genome Set.

In subsequent analyses we varied the non-animal outgroups as follows. The complete matrix (Opisthokonta) consists of 13 animals plus *M. brevicolis, S. rosetta, C. owczarzaki*, *S. punctatus*, *S. arctica, and S. cerevisiae*. The Holozoa matrix consists of all species in the Opisthokonta matrix except for *S. punctatus*, *S arctica, and S. cerevisiae.* The Choanimalia matrix consists of all species in the Holozoa matrix except for *C. owczarzaki*. The animalia matrix includes only the 13 animals.

Construction of the EST Set amino acid matrix

To identify homologs between many taxa, we BLASTed each sequence from each taxa-specific dataset (transcriptome or predicted proteome) against OrthoDB (*87*). We categorized hits by the ortholog hit in OrthoDB and created fasta files for each ortholog. We generated alignments using MAFFT and trimmed these alignments with Gblocks. Individual phylogenies were created for each of these alignments and orthologs were identified using the method described in Hejnol et al. (*88*). We analyzed final orthologs for completeness and created a concatenated matrix with approximately 50% occupancy. The final dataset includes 88,384 sites and 406 gene regions from the following 70 taxa: *Oscarella carmela, Asterina pectinifera, Strongylocentrotus purpuratus, Capitella telata, Branchiostoma floridae, Helobdella robusta, Lottia gigantea, Saccoglossus kowalevskii, Drosophila melanogaster, Anoplodactylus eroticus, Daphnia pulex, Petromyzon marinus, Ciona intestinalis, Acropora palmata, Acropora millepora, Nematostella vectensis, Anemonia viridis, Capsaspora owczarzaki ATCC 30864, Saccharomyces cerevisiae, Cryptococcus neoformans, Monosiga brevicollis, Phycomyces blakesleeanus, Batrachochytrium dendrobatidis, Spizellomyces punctatus, Salpingoeca rosetta, Amphimedon queenslandica, Lubomirskia baicalensis, Mnemiopsis, Hydra magnipapillata, Trichoplax adhaerens, Cyanea capillata, Crassostrea virginica,* *Schmidtea mediterranea, Cerebratulus lacteus, Ephydatia muelleri, Pleurobrachia pileus, Suberites domuncula, Hydractinia echinata, Clytia hemisphaerica, Metridium senile, Porites astreoides, Montastraea faveolata, Halocynthia roretzi, Euprymna scolopes, Boophilus microplus, Isodiametra pulchra, Symsagittifera roscoffensis, Monosiga ovata, Oscarella lobularis, Gallus gallus, Meara stichopi, Terebratalia transversa, Euperipatoides kanangrensis, Xiphinema index, Oopsacas minuta, Convolutriloba longifissura, Mertensiid, Paraplanocera oligoglena, Nemertoderma westbladi, Echinoderes horni, Ptychodera flava, Sphaeroforma arctica, Rhizopus orizae, Podocoryna carnea, Amoebidium parasiticum, Sycon raphanus, Xenoturbella bocki, Carteriospongia foliascens, Aiptasia pallida,* and *Leucetta chagosensis.*

In subsequent analyses we varied the non-animal outgroups as follows. The complete matrix (Opisthokonta) consists of 58 animals plus *M. brevicolis*, *S. rosetta*, *M. ovata*, *C. owczarzaki*, *S. arctica, A. parasiticum*, *S. cerevisiae*, *S. punctatus*, *R. orizae*, *C. neoformans, B. dendrobatidis, and C. neoformans*. The Holozoa matrix consists of all species in the Opisthokonta matrix, except for *S. cerevisiae, S. punctatus, R. orizae, C. neoformans, B. dendrobatidis,* and *C. neoformans*. The Choanimalia matrix consists of all species in the Holozoa matrix except for *C. owczarzaki*, *S. arctica,* and *A. parasiticum*. The Animalia matrix includes only the 58 animals.

Note: *E. muelleri* was inadvertently included twice in all matrices. The two *E. muelleri* entries went together 100% of all tree runs with a near-zero length node. We pruned one of the *E. muelleri* branches from all trees. Numbers above were adjusted accordingly.

Testing for potential non-metazoan contaminants in the Genome or EST matrices

The only gene model in our Genome or EST matrices with an alien index above the suggested threshold of 45 was ML02315a (alien index=51.35), which was present in both sets. This gene has both RNA-seq data and public EST data that aligned to the model and is therefore likely not a contaminant.

Maximum-likelihood phylogenetic analyses of amino acid matrices

We ran a maximum-likelihood analysis on both the genome matrices and the EST matrices using RaxML (version 7.2.8). For each matrix, we generated a best tree with the following command: `raxmlHPC -m PROTGAMMAGTR -s MATRIX.phy -n NAME -q MODELFILE`. Next, 100 bootstraps were generated with the following command: `raxmlHPC -b RANDOMSEED -N 100 -m PROTGAMMAGTR -s MATRIX.phy -n NAME.bs -q MODELFILE`. All trees are available in Figure S1.

Bayesian analyses of amino acid matrices

We ran two instances of Phylobayes (version 3.2e) on each matrix with the following commands: `pb -d MATRIX.nex NAME.01; pb -d MATRIX.nex NAME.02`
For the Genome Set, we ran: `bpcomp -x BURN-IN NAME.01 NAME.02`

For the EST Set, we ran: `readpb -x BURN-IN NAME` for each individual run. We set the burn-in to be approximately 1/3 of the total length of the shortest chain. Table S29 gives statistics including the burn-in used for each run. Bayesian trees resulting from the genome matrices are available as Figure S1. Despite an average runtime of 205 days per run, none of the runs on the EST Set converged. Consequently, we report trees from all eight runs. An unexpected result in Bayesian analyses of the Genome Set is that *T. adhaerens* forms a clade with *M. leidyi* and *A. queenslandica*, which is positioned sister to the rest of animals. We are not aware of this result ever being reported in the literature.

Calculating support values for each scenario

For each phylogenetic scenario in Figure 2a, we constructed constraint trees and used Paup* (*89*) to determine the percentage of bootstrap trees (in the case of maximum-likelihood analysis) or post-burn-in trees (in the case of Bayesian analysis) that fit a particular scenario (i.e., constraint tree). These values are reported in Table 1 of the main paper.

Phylogenetic analysis of individual genes

It is well documented that phylogenies performed on a number individual genes will produce phylogenetic trees that are incongruent with a tree produced from the concatenation of these same set of genes (*90*). Nevertheless, it is interesting to see how well individual genes support competing hypotheses. We have therefore, performed analyses on each individual gene from both the Genome Set and the EST Set. In our analyses of the 406 genes from the EST set (with all outgroups included), we found that, of the 361 which included at least one sponge, one ctenophore, and one non-metazoan outgroup, 19 of these produced a topology congruent with ctenophores as the sister group to the rest of animals (Fig. 2D) and nine were congruent with sponges as the sister group to the rest of animals (Fig. 2C). In our analyses of the 242 genes from the Genome set (with all outgroups included), we found that of the 196 genes, which included at least one sponge, one ctenophore and one non-metazoan outgroup, that 18 of these produced a topology congruent with ctenophores as the sister group to the rest of animals (Fig. 2D) and 15 were congruent with sponges as the sister group to the rest of animals (Fig. 2C).

For this analysis we generated a parsimony tree for each gene matrix using RAxML (version 7.7.8) and then used this parsimony tree as a starting tree under RAxML-Light (version 1.0.9) with the PROTGAMMAAUTO specified as the model. Under this model setting RAxML conducts an ML estimate of all available pre-defined AA models (excluding GTR) every time the model parameters are optimized during the tree search.

Branch length comparison

For every tree in our phylogenetic analyses of the Genome and EST Sets, except those without a non-metazoan outgroup, we calculated the distance from the root in terms of branch length for each taxon. These values are included in Tables S30 and S31.

To determine branch lengths, we opened each tree in FigTree v.1.3.1 (*91*) and rooted on branch separating the metazoan clade from the non-metazoan clade. We exported this rooted tree and opened the Nexus formatted tree in TreeStat v. 1.2 (*92*). We used TreeStat to calculate the root-tip-lengths for each taxa. In all trees, the lengths of the *M. leidyi* branches are similar to those of *Drosophila melanogaster*.

Maximum likelihood analysis of gene content

We assembled a presence/absence matrix of the OrthoMCL clusters and analyzed these data with RAxML version 7.2.6 (SSE3 version) under the GTR gamma model of rate heterogeneity as was done in previous studies (*93, 94*). The substitution probability matrix depends on the branch lengths as well as the instantaneous rate matrix which, in turn, depends on both the equilibrium state frequencies and parameters for rate changes between states. Though only a single rate parameter is used in the RAxML binary model, the equilibrium frequencies can, to a certain extent, account for differences in the rates of gain and loss. Though the rate parameter is constant for gains and losses, the rate matrix is therefore not necessarily symmetric. The equilibrium frequencies were estimated from the data and do deviate from the observed frequencies as expected for asymmetric rates

Using input data and the NCBI taxonomy tree to initially infer a binary and taxonomy-constrained tree topology as a starting tree we ran: (raxmlHPC-PTHREADS-SSE3 -T 4 -s min2.I1.5.phy -m BINCAT -g known_relationships.tre -p 12345 -n S1). This returns a list of files ended with '.S1', in which the file 'RAxML_bestTree.S1' is the binary constraint starting tree. We used this resulting tree (via '-t' option) to infer the integer weights for features (via '-f u' option), up-weighting the congruent ones while down-weighting those that were incongruent: (raxmlHPC-PTHREADS-SSE3 -T 4 -f u -m BINGAMMA -s min2.I1.5.phy -t RAxML_bestTree.S1 -# 100 -p 12345 -n S2). This returns a list of files ending with '.S2', in which the file 'RAxML_weights.S2' is the weight vector. Using the weight vector to infer the binary and taxonomy-constrained best tree as a final tree topology we ran: (raxmlHPC-PTHREADS-SSE3 -T 4 -m BINGAMMA -s min2.I1.5.phy -t RAxML_bestTree.S1 -a RAxML_weights.S2 -p 12345 -n S3). This will return a list of files ending with '.S3', in which the file 'RAxML_bestTree.S3' is the final tree. To see if the result is the same without the starting tree we ran: (raxmlHPC-PTHREADS-SSE3 -T 4 -m BINGAMMA -s min2.I1.5.phy -a RAxML_weights.S2 -p 12345 -n S4) We also ran a rapid bootstrap analysis: (raxmlHPC-PTHREADS-SSE3 -T 4 -m BINGAMMA -s min2.I1.5.phy -a RAxML_weights.S2 -f a -# 100 -x 234544 -p 12345 -n S5). To see the effect of weighting we ran: (raxmlHPC-PTHREADS-SSE3 -T 4 -m BINGAMMA -s min2.I1.5.phy -f a -# 100 -x 234544 -p 12345 -n S6). All analyses produced the topology seen in Figure 4.

Maximum parsimony analysis of gene content

We ran a maximum parsimony analyses using Paup*, using the same character weighting as the ML analyses, and recovered 100% bootstrap support for Ctenophora as the sister group to all animals. This was the case whether characters were coded as ordered or Dollo.

Bayesian analysis of gene content

We also performed Bayesian analyses of the unweighted gene presence/absence matrix using Mr. Bayes. The data type was set to restriction (i.e., binary) and rates model was set to invgamma. Two runs, each with one million generations, were completed. The average standard deviation of split frequencies was 0.000000 by 80000 generations, indicating quick convergence. These analyses recover Parahoxozoa and place *M. leidyi* as sister to all other sampled animals, both with posterior probabilities of 100%. Some relationships within Parahoxozoa were inconsistent with well-supported relationships (e.g., Cnidaria was not recovered as monophyletic).

Hypothesis testing of gene content data using CONSEL

To see if this result is significantly better than the *a priori* alternative hypotheses presented in Figure 2, we built constraint trees corresponding to each hypothesis and generated trees with branch lengths for each starting tree (`raxmlHPC-SSE3 -m MULTIGAMMA -K GTR -n fig2a -s aln.phy -a weights -g fig2a.tre`). We then generated per site log likelihoods for these trees (`raxmlHPC-SSE3 -f g -m MULTIGAMMA -K GTR -n 6trees -s aln.phy -a weights -z 6trees.tre`). Next, we used the per site log likelihoods as input to CONSEL version 1.20 (*95*) to generate p-values for each of the alternative hypotheses (`seqmt --puzzle RAxML_perSiteLLs.6trees; makermt RAxML_perSiteLLs; consel RAxML_perSiteLLs; catpv RAxML_perSiteLLs`). Under the Approximately Unbiased (AU) test, Kishino-Hasegawa (KH) test, and the Bootstrap Probability (BP) test, the relationship of *M. leidyi* as the sister group to all other animals (Ct,) is significantly (0.05) better than all of the other hypotheses presented in Figure 2 (Table S12). Under the Shimodaira-Hasegawa (SH) test the relationship of *M. leidyi* as sister to all other animals except *A. queenslandica* (Po,) cannot be rejected; however it is well-known that the SH test is very conservative and includes more trees in the confidence set than is appropriate (*96*).

Hypothesis testing of gene content data using SOWH

We performed a Swofford–Olsen–Waddell–Hillis (SOWH) test on the gene content data. Under the SOWH test, the relationship of *M. leidyi* as the sister group to all other animals (Ct,) is significantly ($P < 0.05$) better than all of the other hypotheses presented in Figure 2. We followed the procedure as described by Goldman and coworkers, 2000 (*97*). Briefly, we performed the following: (1) performed a maximum likelihood analysis on the original alignment using the weights file generated as described in S9.1 with and without a constraint tree representing the hypothesis being tested, (2) generated 100 data sets with seq-gen using the alpha and branch lengths obtained from the ML tree produced with the constraint (NOTE: since seq-gen does not produce binary matrices, the frequency of zeroes and ones were assigned to A and T respectively, while the frequency of C and G were set to 0. Since ones and zeroes are required by RAxML for a two-state matrix, occurrences of A and T were substituted with zeroes and ones in the seq-gen output), (3) performed maximum likelihood analyses on each of these 100 simulated datasets with and without the constraint tree being tested, and (4) compared the distribution of the differences in likelihoods with the difference in likelihood between the best tree and the tree generated with the constraint on the real data set with the pnorm function in R.

Phylogenetic analysis of Near Intron Pairs

Lehmann and coworkers assembled and analyzed a large set of near intron pairs (i.e., mutually exclusive introns situated within close proximity) from 48 species using maximum parsimony (*98*). The *M. leidyi* genome assembly was used for this original analysis, but gene models were not publicly available at the time. In their original study *M. leidyi* was placed sister to insects with very low support using unweighted maximum parsimony. Jörg Lehmann supplied us with a version of the near intron pair matrix that used the ML2.2 gene models to increase the number of orthologs considered from 1,847 to 2,955 for the 4,405-gene dataset (Jörg Lehmann, University of Leipzig, personal communication). We conducted a weighted likelihood analysis on this new matrix with RAxML (raxmlHPC-PTHREADS-SSE3 version 7.7.8) and constrained known relationships.

We constructed a tree of the taxa in the matrix used in OrthoMCL with known bilaterian relationships, leaving the non-bilaterian groups in a polytomy. Using this tree as input to the RAxML option (-g) we generated a weight matrix in RAxML (raxmlHPC-PTHREADS-SSE3 -T 4 -f u -g known_relationships.tre -m MULTIGAMMA -n 01.int.weights -s concatenatedAll28.REAL_NIPS.phy) to up-weight congruent and down-weight incongruent columns. We used these weights to infer the most likely tree (raxmlHPC-PTHREADS-SSE3 -T 4 -m MULTIGAMMA -K GTR -n 03.restrained_w_known -s concatenatedAll28.REAL_NIPS.phy -a RAxML_weights.01.int.weights -g known_relationships.tre). Bootstraps were generated using (raxmlHPC-PTHREADS-SSE3 -T 4 -m MULTIGAMMA -K GTR -n 03.restrained_w_known.boot -s concatenatedAll28.REAL_NIPS.phy -a RAxML_weights.01.int.weights -b 4321 -N autoMRE -g known_relationships.tre) and then applied with (raxmlHPC -f b -z RAxML_bootstrap.03.restrained_w_known.boot -t RAxML_bestTree.03.restrained_w_known -s concatenatedAll28.REAL_NIPS.phy -m MULTIGAMMA -n 03.restrained_w_known_boots_applied).

This analyses resulted in a topology with *M. leidyi* as the sister group to the rest of animals with low support (Fig. S2).

Presence and absence of *M. leidyi* neural components

The presence and absence of neural components in *M. leidyi* (Tables S15-17) was determined using the following methods: A neural synaptic gene set of protein sequences of *Homo sapiens* was collected from the NCBI database. This set of neural genes was used as reciprocal best BLASTP and TBLASTN queries against the *M. leidyi* predicted proteome and genome respectively. The domain architecture of candidate protein sequences was predicted using the SMART domain prediction database (*99*) utilizing the outlier homologues and PFAM domains settings. The domain architecture of each protein was then compared to the human proteins, noting when diagnostic domains were absent.

The presence or absence of neural components in *N. vectensis* (*Nv*), *H. magnipapillata* (*Hm*), *T. adhaerens* (*Ta*), *A. queenslandica* (*Aq*), *M. leidyi* (*Ml*), and *M. brevicolis* (*Mb*) were verified these candidates using the same approach with queries from the datasets of two other studies (*100, 101*).

Missing synaptic scaffolding components are also absent from the transcriptomes of seven other ctenophore species

We searched for neuroligin and glutamate receptors in the transcriptomes of seven other ctenophore species. We found no neuroligin in any of the transcriptomes. We found candidate glutamate receptors in these other species, but our phylogenetic analysis (Fig. S3) shows that these sequences are related equally to AMPA, NMDA, kainate-type, and delta2-like glutamate receptors, suggesting that the ctenophore sequences descended from an ancestral sequence that went on to differentiate into these four classes of glutamate receptors. The top five sequences for neuroligin and glutamate receptors queries (i.e., those with TBLASTN hits to proteins of interest with E-values less than 0.001) have been submitted to GenBank and accession numbers are included in Table S20.

Phylogenetic distribution of neural components

Behavioral sensitivity to several neurotransmitters (acetylcholine, serotonin, epinephrine, and norepinephrine) has been shown in ctenophores (*102*), and immunoreactivity to the neuropeptide FMRFamide, and preprohormone vasopressin has also been reported (*103*). The neurotransmitter synthesis genes acetylcholinesterase and glutamate decarboxylase are present in the *M. leidyi* genome; however, the monoamine enzymes DOPA decarboxylase and dopamine-beta hydroxylase (Table S17) required for the production of the chatelcholamines epinephrine, norepinephrine, and dopamine, are absent, and immunological investigations have failed to detect the presence of serotonin in the ctenophore *Bolinopsis infundibulum* (*104*). Metabotropic glutamate receptors (mGluR and GABAb) are present in the *M. leidyi* genome. Seven-transmembrane receptors with high homology to serotonin, dopamine receptors, and a large number of G-protein coupled receptors are also present, suggesting that these receptors target alternative neuropetides in ctenophores. Neurotransmitter transport genes found in all other metazoans such as SNAP, syntaxin, synaptotagmin, and synaptobrevin are present, but the gene synapsin is missing (also absent in placozoans and sponges). Gap junction genes (innexins), found in cnidarians and bilaterians, are also present within the *M. leidyi* genome (12 predicted paralogs), but are absent in placozoans and sponges.

Presence and absence of M. leidyi mesoderm components

We determined the presence and absence of mesoderm components in *M. leidyi* (Tables S18-19) using the same methods used to determine neural components above. The following references detail the role of these genes in mesoderm: twist and snail (*105*), tinman, bagpipe, and ladybird (*106, 107*), nk 2.1, Pax3, and Myf5 (*108*), Eomesodermin (*109*), myoD and myogenin (*110*).

Trancriptome data from other ctenophore species

Many genes involved in mesodermal cell types that are present in cnidarians and bilaterians are absent from *M. leidyi*. Since these genes are also missing from non-metazoans and sponges, it is parsimonious to assume that these absences are primary (i.e., these genes arose after ctenophores and sponges split from Parahoxozoa) rather than secondary losses. To reduce the possibility that these were secondarily lost in the lineage leading to *M. leidyi*, we searched for these missing genes in deeply sequenced transcriptomic data from seven ctenophore species: *Bathyctena chuni*, *Beroe forskalii*, *Charistephane fugiens*, *Euplokamis dunlapae*, *Hormiphora californensis*, *Lampea lactea*, and *Thalassocalyce inconstans*. The top five sequences for each query (i.e., those with TBLASTN hits to proteins of interest with E-values less than 0.001) have been submitted to GenBank and accession numbers are included in Table S20.

RNA was isolated from adult ctenophores that were collected in either the Gulf of California or Monterey Bay. Total RNA was extracted using RNeasy kit (Qiagen). Preparation of RNA-seq libraries was done using Illumina TruSeq kit for paired end reads. Total RNA was sequenced at the University of Utah. Sequencing was done using the Illumina HiSeq2000 platform on a paired-end protocol with 100 cycles. Sequence data was assembled as described in Francis and coworkers, 2013 (*111*). The following number of unique transcripts were produced: *B. chuni*=399,516, *B. forskalii*=194,616, *C. fugiens*=285,056, *E. dunlapae*=276,866, *H. californensis*=528,793, *L. lactea*=198,516, and *T. inconstans*=194,483. The high numbers reflect minor assembly differences at single loci.

General pipeline for analyzing additional ctenophore transcripts

One or more human RefSeq proteins were used as a TBLASTN query against the seven ctenophore transcriptomes and a Trinity assembly of the *M. leidyi* RNA-seq data. All hits with E-values ≤ 0.001 were considered candidates. These candidate transcripts were used as BLASTX queries against the Human RefSeq database (edited to include only a single isoform per Entrez Gene ID). Any human sequence that among the top 10 hits (with E-values ≤ 0.001) for any of the ctenophore transcripts was marked for inclusion in our downstream phylogenetic analyses. All candidate ctenophore transcripts were translated using the r2012-08-15 version of TransDecoder (http://transdecoder.sourceforge.net/) with the following command: transcripts_to_best_scoring_ORFs.pl -t candidates.fa --search_pfam Pfam-A.hmm. All translations of ctenophore candidates and marked human proteins were aligned using MAFFT v6.864b (*85*) with the following command: mafft –auto FASTAFILE. The resulting alignment was trimmed with Gblocks version 0.91b (*86*) using the following command: Gblocks -b2=Z -b3=10 -b4=5 -b5=a, where Z = 0.65 x the number of sequences. Sequences where occupancy was less than 50% were removed. We used RAxML version 7.2.8 (*112*) to generate a maximum likelihood tree and 10 bootstraps with the following command: raxmlHPC -f a -x RANDOMSEED -m PROTGAMMALG -p RANDOMSEED -N 10 -n NAME -s ALIGNMENT.phy.

If there is weak evidence uniting any ctenophore candidate with the human gene being tested, the pipeline is rerun on the subset of best candidates. These reruns often lead to longer alignments and better resolution in the resulting trees.

A small percentage of ctenophore hits were not translated by transdecode. Of the 263 total hits that were in the top five for each query, 12 were missed by transdecode. We performed manual analysis on these 12.

Search for Neuroligin (synaptic scaffolding) in other ctenophore transcriptomes

Human NLGN1 (accession= NP_055747) was used as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. We generated an alignment of 92 unique ctenophore sequences and 20 human proteins (NCBI Gene ids = AADAC, AADACL2, AADACL3, AADACL4, ACHE, BCHE, CEL, CES1, CES2, CES3, CES4A, CES5A, LIPE, NLGN1, NLGN2, NLGN3, NLGN4X, RPL3, RPL3L, and TG) and a maximum likelihood tree using the general pipeline described in S13.3. The human NLGN formed a polytomy with a small subset of the ctenophore sequences as well as LIPE and the AADAC proteins (zero bootstrap support). The majority of ctenophore sequences formed a large clade that was most closely related to the CES proteins.

We next removed this large clade of ctenophores and reran the pipeline from S13.3. In this tree, a very long branch *L. lactea* sequence is sister to NLGN3 with a bootstrap value of 20; the highest bootstrap supporting this sequence and the NLGN sequences is 50. An *E. dunlapae* sequence is sister to all the NLGN proteins plus the *L. lactea* sequence with a support value of 10.

We next reran the pipeline, this time including only the human TG, NLGN and RPL3 proteins, as well as the *L. lactea* and *E. dunlapae* sequences. With the subtraction of most of the ctenophore sequence, the number of columns in the alignment expanded from 59 to 532. Very few of the *L. lactea* sequence aligned – upon examination the 199 amino acids of this sequence included seven stop codons. This sequence was removed from this analysis based on it having less than 50% occupancy. The resulting tree had the *E. dunlapae* sequence as sister to the RPL proteins (or sister to TG + NLGN proteins) with 80% bootstrap support. We were unable to find additional candidates using *Nematostella vectensis* sequences as queries.

None of these analyses show any reasonable support for a ctenophore neuroligin and the final analysis shows fairly strong support that the best candidate is not a neuroligin.

Search for Ionotropic glutamate receptors (synaptic scaffolding) in other ctenophore transcriptomes

We used human GRIA2 (accession= NP_000817), GRIN1 (accession= NP_000823), GRIK2 (accession= NP_001159719), and GRID2 (accession= NP_001501) as TBLASTN queries against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. GRIA2 is an AMPA-type glutamate receptor, GRIN1 is an NMDA-type glutamate receptor, GRIK2 is a kainate-type glutamate receptor, and GRID2 is a delta2-like glutamate receptor.

We generated an alignment starting with 136 ctenophore sequences (including Trinity isoforms) and 28 human proteins (NCBI Gene ids = CASR, GRIA1, GRIA2, GRIA3, GRID1, GRID2, GRIK1, GRIK2, GRIK3, GRIK4, GRIK5, GRIN1, GRIN2A, GRIN2B, GRIN2C, GRIN2D, GRIN3A, GRIN3B, GRM1, GRM2, GRM3, GRM4, GRM5, GRM6, GRM7, GRM8, and TAS1R3). Our pipeline produced an alignment of only four columns. We next aligned each ctenophore sequence separately against the human set (using the pipeline alignment procedure) and kept only those in which the resulting alignment included less than 50% gaps in the ctenophore sequence. This resulted in 34 ctenophore sequences (including two *M. leidyi* sequences and one published *Pleurobrachia* sequence with NCBI accession ADV31314). The resulting tree led to a clade of human ionotropic glutamate receptor (iGluR) sequences and a clade of ctenophore sequences that were bisected by the non-iGluR human outgroup sequences. Given the length of the outgroup branch and the fact that the top BLAST hits of all the ctenophore sequences were the iGluR human sequences, we surmise that the attachment of the outgroup branch is likely unreliable. When we run this tree without outgroups the ctenophore sequences form a monophyletic group with 100% bootstrap support. We were unable to find additional candidates using *Capitella telata* sequences as queries.

These analyses support the ctenophore iGluR candidates being descendants of an ancestral sequence that would later give rise to AMPA, NMDA, kainate-type, and GRID glutamate receptors after bilaterians split from ctenophores. As such, ctenophores do not have bona fide AMPA, NMDA, kainate-type, and GRID iGluRs (Fig. S3).

Search for Twist (myogenic genes) in other ctenophore transcriptomes

We used human TWIST1 (accession=NP_000465) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. We generated an alignment of 30 ctenophore and 20 human proteins (NCBI Gene ids = TWIST1, TWIST2, HAND2, HAND1, TCF15, SCXB, SCXA, TCF21, MSC, ATOH7, NEUROD1, NHLH2, NHLH1, TAL1, TAL2, and LYL1) and a maximum likelihood tree using the general pipeline described above. The original analysis produced a clade that included the human Twist genes, a *M. leidyi* sequence, and two *E. dunlapae* sequences with a support value of 60%. In this same tree there was a large clade of ctenophore sequences that grouped sister to all the human genes. To see if we could get better support for the Twist clade, we removed all ctenophore sequences except for the three that formed a clade with human Twist and reran the analysis. An unrooted view of this tree results in a clade of human LYL1 and TAL proteins sister to the ctenophore sequences with a support value of 40%.

We ran a third analysis this time running the standard pipeline but running BLAST without SEG filtering. This yielded more ctenophore sequences and a larger alignment: 80 columns from 97 sequences as opposed to 66 columns from 46 sequences in the first analysis. Again we recover the majority of ctenophore sequences in a cluster separate from the human sequences. The only exceptions are (1) a very long branch of *B. forskalii* sequences that is sister to a clade that includes human FERD3L and human PTF1A, and (2) a longish *E. dunlapae* branch that is sister to human ID4 sequence. In this tree, the human TWIST1 and TWIST2 sequences form a clade with 18 of the other 26 human proteins. We were unable to find additional candidates using *Capitella telata* sequences as queries.

From these analyses, it seems unlikely that a true Twist gene exists in ctenophores.

Search for Snail (myogenic genes) in other ctenophore transcriptomes

We used human SNAI1 (accession=NP_005976) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. We generated an alignment of 1,431 ctenophore and 547 human proteins. The alignment that resulted from MAFFT and Gblocks consisted of only seven columns and any tree resulting from this alignment would not be informative, consequently we must rely on reciprocal best BLAST. We identified one *T. inconstans* sequence and two *E. dunlapae* sequences, which are reciprocal best BLAST hits with human Snail genes. This result suggests that Snail was present in the last common ctenophore ancestor, but was lost in the lineage leading to *M. leidyi*. We were unable to find additional candidates using *Capitella telata* sequences as queries.

Search for Lbx/ladybird, NK4/tinman, NK3/bagpipe, and NK2/vnd (myogenic genes) in other ctenophore transcriptomes

We fetched the human ANTP class homeodomains from HomeoDB (*113*) and used these sequences as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. After translating transcripts that were hit, we filtered hits based on the presence of the following: 1) complete homeodomains with “WFQN,” which is present in all ANTP homeodomains in HomeoDB, 2) absence of LIM, Homez, Pou, Homeobox_KN, and/or PBC domains (these are not associated with ANTP homeodomains). This list produced 822 candidate ANTP homeodomains, of which 124 were unique. We aligned these ctenophore homeoboxes to the ANTP alignment included in the additional file 1 of Ryan and coworkers, 2010 (*114*). We then ran RAxML with 100 fast bootstraps. We found no ctenophore sequences in the Lbx/ladybird, NK4/tinman, NK3/bagpipe, or NK2/vnd clades. Furthermore, we found no evidence of ctenophore Hox or ParaHox genes in this analysis.

Search for Myf5, Mrf4, Myogenin, and MyoD (myogenic genes) in other ctenophore transcriptomes

We used human Myf5 (accession=NP_005584), Mrf4 (accession=NP_002460), Myogenin (accession=NP_002470), and MyoD (accession=NP_002469) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. We ran TBLASTN both with and without SEG filtering and found no hits with E-values below 0.001. These human genes are identifiable by TBLASTN to protostome transcriptomes at much lower E-values. We were unable to find additional candidates using *Capitella telata* sequences as queries.

Search for Noggin (myogenic genes) in other ctenophore transcriptomes

We used human NOG (accession=NP_005441) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. The search produced no hits with E-values < 0.2. We ran the searches with and without SEG filtering. This human gene is identifiable by TBLASTN to protostome transcriptomes at much lower E-values. We were unable to find additional candidates using *Capitella telata* sequences as queries.

Search for Eomesodermin (myogenic genes) in other ctenophore transcriptomes

We used human Eomes (accession=NP_005433) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. We generated an alignment of 126 ctenophore (58 unique) and 21 human proteins. The alignment that resulted from MAFFT and Gblocks consisted of 177 columns. The RAxML analysis produced a clade with the human proteins Tbr1 and Eomes, (bootstrap=90%). Sister to this clade was human TBX21 (bootstrap=30%). These results suggest that that ctenophore sequences are not direct orthologs to human Eomes. We were unable to find additional candidates using *Danio rerio* sequences as queries.

Search for GATA (myogenic genes) in other ctenophore transcriptomes

We used human GATA1 (accession=NP_002040) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. We generated an alignment of 60 candidate transcripts, of which 33 were unique. There were 31 columns in the alignment. The resulting tree produced a clade that included ctenophore sequences from six species of ctenophores along with the human GATA genes and the human TRPS1 and ZGLP1 genes (bootstrap=60). Like GATA proteins in other animals, these ctenophore sequences contain two GATA zinc finger domains (PFAM accession=PF00320). From this analysis we can say that GATA is present in *B. chuni*, *B. forskalii, E. dunlapae, H. californensis,* and *T. inconstans,* and was lost in *M. leidyi*.

Search for Troponin (myogenic genes) in other ctenophore transcriptomes

We used human TNNI1 (accession=NP_003272) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. The search produced no hits with E-Value < 0.001. We ran the searches with and without SEG filtering. This human gene is identifiable by TBLASTN to protostome transcriptomes at much lower E-values.

Search for FGF (myogenic genes) in other ctenophore transcriptomes

We used human FGF1 (accession=NP_001138364) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. The search produced no hits with E-values < 0.001. We ran the searches with and without SEG filtering. This human gene is identifiable by TBLASTN to protostome transcriptomes at much lower E-values.

Search for Nodal (myogenic genes) in other ctenophore transcriptomes

We used human Nodal (accession=NP_060525) as a TBLASTN query against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. We generated an alignment with 145 ctenophore sequences (50 unique) and 31 human sequences. The RAxML analysis produced a clade that included human Nodal with seven ctenophore sequences from four species. Support for this clade was 10%. Included in this clade is a *M. leidyi* sequence (accession=AEP16385) that was analyzed as part of a genome analysis of *M. leidyi* TGF-beta and was named Tgf2 (*115*). In this detailed study, it was determined that no *M. leidyi* sequence (including Tgf2) was orthologous to Nodal.

To test whether the low support of the clade of Nodal and ctenophore sequences was due to long ctenophore branches, we reran this analysis, pruning all ctenophore sequences besides those that appeared in the Nodal clade. In this analysis only two of the ctenophore sequences formed a clade with human Nodal. The bootstrap support for this clade was 20%. The other ctenophore sequences formed poorly supported clades with other human sequences. To test whether the low support of this Nodal clade may have been due to the longer human branches, we ran another maximimum-likelihood analysis on a smaller subset of human sequences. In this tree, the two ctenophore NODAL proteins formed a sister clade to a clade of Nodal and GDF5 with 20% bootstrap support. From this analysis and consistent with our previous work (*115*), we conclude that there is not convincing evidence that a Nodal ortholog is present in either *M. leidyi* or the other seven ctenophore transcriptomes that we have investigated.

Search for Shh/hh (myogenic genes) in other ctenophore transcriptomes

We used human Shh (accession= NP_000184) as a TBLASTN query against against the seven ctenophore transcriptomes and the Trinity assembly of the *M. leidyi* RNA-seq data. We recovered two hits to partial *B. chuni* transcripts. One coded 266 amino acids and the other coded 340 amino acids. The latter was incomplete at the C-terminal end. Like the full length *M. leidyi* candidate (ML073718a) and unlike the hedgehog proteins from cnidarians and bilaterians, a “Hog” domain (Pfam accession=PF01079) was present in these sequences, but there no “Hedge” domain (Pfam accession=PF01085) was present (Fig. S10). Turning off SEG filtering for the same TBLASTN search, we identified an additional two *B. chuni* candidates. The two were 970 and 981 amino acids long. Both contained “Hog” domains, but no “Hedge” domains. These two also contained MMR_HSR1 and MACPF domains. As such, we see no proper hedgehog gene in any of the ctenophore transcriptomes.

Genome browser

As part of this project, we have launched the *Mnemiopsis leidyi* genome Project Portal. At this site, users can perform BLAST searches against the genome sequence, proteome, and transcriptome. An interface allows users to retrieve individual sequences or scaffolds, and assemble custom data sets. We have implemented JBrowse, a JavaScript-based genome browser (*116*), for viewing the *M. leidyi* genome assembly, gene predictions, RNA-seq data, and public EST and mRNA sequences. Gene pages are editable wiki pages and users are encouraged to annotate sequences of interest. The *Mnemiopsis* web site is freely accessible at http://research.nhgri.nih.gov/mnemiopsis/

Genomic scaffolds are named using a six-character convention (e.g. MLXXXX); the ML designates the species (*Mnemiopsis leidyi*), and the individual scaffolds are numbered from 0001 to 5100 (e.g., ML4323). Gene identifiers (e.g., ML103316a) are prefixed with the scaffold on which the gene is located (in this example, “ML1033”), followed a non-padded integer that is unique in combination with the scaffold identifier (in this case “16”), followed by a lower-case letter that corresponds to the genes isoform (in this case “a”). A gene identifier is typically (but not always) ordered by its most 5’ position on the scaffold. Newly added genes are assigned the next integer, regardless of its position on the scaffold.

Supplementary Table S1: Summary of phylogenetic placements of ctenophores

Below is a table of results from a series of phylogenetic analyses that include ctenophores. The result is left blank if the reference refers only to the relationship of ctenophores rather than the branching pattern of the five earliest branching lineages. Abreviations are as follows: Porifera(Po), Ctenophora(Ct), Placozoa(Tr), Cnidaria(Cn) and Bilateria(Bi). Multiple "Po" entries indicate that a study included data from multiple poriferans and recovered a paraphyletic Porifera. Parahoxozoa indicates a monophyletic group consisting of Cnidaria, Bilateria and Placozoa (if included). Taxa in square brackets indicates a sister relationship to all other animals based on a rooted tree. Modeled after Table 1 of Wallberg and coworkers, 2004 (*117*) and Ryan and Baxevanis 2007 (*118*).

Authors	**Year**	**Data/Meth.**	**Result**	**Hypothesis**
Lang (*119*)	1884	Morph-MP		Ct->Platyhelminthes
Hyman (*120*)	1940	Morph	(Po,((Cn,Ct),Bi))	Coelenterata
Hadzi (*121*)	1953	Morph		Ct->Protostomia
Brusca & Brusca (*122*)	1990	Morph		Coelenterata
Ehlers (*123*)	1993	Morph		(Ct,Bi)
Ruppert & Barnes (*124*)	1994	Morph		Coelenterata
Nielsen (*125*)	1995	Morph	(Po,(Tr,(Cn,(Bi,Ct,Bi))))	Ct->Deuterostomia
Nielsen (*126*)	1996	Morph	(Po,(Tr,(Cn,(Ct,Bi))))	(Ct,Bi)
Ax (*127*)	1996	Morph		(Ct,Bi)
Margulis & Schwartz (*128*)	1998	Morph		Coelenterata
Nielsen (*129*)	2001	Morph		(Ct,Bi)
Wainright et al. (*130*)	1993	Ribo-ML	(Po,(Ct,((Tr,Cn),Bi)))	Parahoxozoa [Po]
Katayama et al. (*131*)	1995	Ribo-DI	((((Po,Ct),Tr),Cn),Bi)	Diploblastica
""		Ribo-MP	((((Po,Ct),Tr),Cn),Bi)	Diploblastica
""		Ribo-ML	((Po,(Ct,Tr)),(Cn,Bi))	

Hanelt et al. (*132*)	1996	Ribo-DI	((((Po,Ct),Tr),Cn),Bi)	Diploblastica [Bi]
Van de Peer & Wachter (*133*)	1997	Ribo-DI	(((Po,(Po,Ct)),(Tr,Cn)),Bi)	Parahoxozoa (Po,Ct) [Bi]
Abouheif et al. (*134*)	1998	Ribo-MP	(Po,(Ct,(Tr,(Cn,Bi))))	Parahoxozoa [Po]
Collins (*135*)	1998	Ribo-MP	(Po,(Po,Ct,(Tr,(Cn,Bi)))	Parahoxozoa [Po]
""		Ribo-ML	(Po,((Po,Ct),(Tr,(Cn,Bi))))	Parahoxozoa (Po,Ct) [Po]
Halanych (*136*)	1998	Ribo-MP	(Po,(Tr,((Cn,(Ct,Cn)),Bi)));	Coelenterata [Po]
""		Ribo-MP	(Po,(Tr,(Ct,Cn,Bi))))	
Lipscomb et al. (*137*)	1998	Ribo-MP	(Po,Po,Ct,((Tr,Cn),Bi));	Parahoxozoa [Po]
Winnepenninckx et al. (*138*)	1998	Ribo-DI	((Po,(Po,Ct)),(Cn,(Tr,Bi)))	Parahoxozoa [Po]
Zrzavý et al. (*139*)	1998	Ribo-MP	((Po,(Po,Ct)),(Tr,(Cn,(Cn, Bi))))	Parahoxozoa [Po]
Kim et al. (*140*)	1999	Ribo-DI	(Po,(Po,(Ct,(Tr,(Cn,Bi)))))	Parahoxozoa [Po]
""		Ribo-ML	(Po,(Po,(Ct,(Tr,(Cn,Bi)))))	Parahoxozoa [Po]
Giribet et al. (*141*)	1999	Ribo-MP	(Po,(Ct,(Cn,(Tr,Bi))))	Parahoxozoa [Po]
Siddall & Whiting (*142*)	1999	Ribo-MP	((Po,Ct),(Cn,(Tr,Bi)))	Parahoxozoa (Po,Ct)
Medina et al. (*143*)	2001	Ribo-ML	(Po,(Po,(Ct,(Cn,Bi))))	Parahoxozoa [Po]
""		Ribo-ML	((Cn,Ct),(Po,(Po,Bi))	Coelenterata
""		Ribo-ML	(Po,(Ct,(Cn,Bi))	Parahoxozoa [Po]
""		Ribo-MP	(Po,Po,(Ct,(Cn,Bi)))	Parahoxozoa [Po]
Peterson & Eernisse (*144*)	2001	Ribo-MP	(Po,(Po,Ct,(Tr,(Cn,Bi))))	Parahoxozoa [Po]
Podar et al. (*145*)	2001	Ribo-ML	(Po,(Ct,(Tr,(Cn,Bi))))	Parahoxozoa [Po]

Collins et al. (*146*)	2002	Ribo-MP	(Po,(Po,(Ct,(Cn,(Tr,Bi)))))	Parahoxozoa [Po]
Jondelius et al. (*135*)	2002	Ribo-ML	(Po,(Po,(Ct,(Cn,(Tr,Bi)))))	Parahoxozoa [Po]
Martinelli & Spring (*147*)	2003	Ribo-ML	(Po,(((Ct,Tr),Cn),Bi))	
Zrzavý & Hypša (*148*)	2003	Ribo-MP	(Po,(Ct,((Tr,Cn),Bi)))	Parahoxozoa [Po]
""		Ribo-MP	(Po,(Ct,(Tr,(Cn,Bi))))	Parahoxozoa [Po]
""		Ribo-MP	(Po,((Ct,Tr),(Cn,Bi)))	
Wallberg et al. (*117*)	2004	Ribo-MP	(Po,Po,(Ct,(Cn,(Tr,Bi))))	Parahoxozoa [Po]
Dunn et al. (*149*)	2008	EST-ML	(Ct,((Po,Cn),Bi))	[Ct]
Philippe et al. (*150*)	2009	EST-Ba	(Po,(Tr,((Cn,Ct),Bi)))	Coelenterata [Po]
Hejnol et al. (*88*)	2009	EST-ML	(Ct,(Po,(Tr,(Po,(Cn,Bi)))))	[Ct]
Schierwater et al. (*151*)	2009	MT,EST, Morph	(Bi,(Tr,(Po,(Cn,Ct))))	[Bi]
Pick et al. (*152*)	2010	EST-Ba	(Po,(Ct,(Cn,(Pl,Bi))))	Parahoxozoa [Po]
Srivastava et al. (*80*)	2010	Geno-ML	(Ct,(Po,(Pl,(Cn,Bi))))	Parahoxozoa [Ct]
Mallat et al. (*153*)	2012	rRNA-ML	((Po,Ct),(Cn,Tr),Bi)	Parahoxozoa (Po,Ct)

Supplementary Table S2: Nested intronic genes in *Mnemiopsis leidyi* and other animal genomes

Numbers for bilaterian species are from Kumar, 2009 and references therein (*154*). Genome assembly versions from which numbers were generated are in parentheses after gene number. “% Nested” indicates the percent of genes in the genome that nested genes comprise.

Species	**Number of Nested Intronic Genes**	**Total Number of Genes**	**% Nested**
M. leidyi	1,323	16,554	~8.0
D. melanogaster	792	14,601 (r5.1)	~5.4

C. elegans	429	20,061 (WSI176)	~2.1
C. briggsae	233	19,500	~1.2
H. sapiens	158	28,755 (r36.2)	~0.5

Supplementary Table S3: Genomic repeat content as detected by VMatch

Organism	**Total bases**	**Masked bases**	**% Masked**
M. leidyi	155,865,547	15,693,038	10.07
A. queenslandica	147,463,102	15,964,649	10.83
M. brevicollis	41,633,360	5,135,212	12.33
N. vectensis	356,613,585	88,039,756	24.69
T. adhaerens	105,632,827	2,220,249	2.10
S. rosetta	55,440,309	4,644,212	8.38

Supplementary Table S4: Known genomic repeats (RepeatMasker)
Ml=Mnemiopsis leidyi, Aq=Amphimedon queenslandica, Mb=Monosiga brevicolis, Ta=Trichoplax adhaerens, Sr=Salpingoeca rosetta

	Ml	***Aq***	***Mb***	***Ta***	***Sr***
Sequences	5100	3579	218	1415	154
Total length (bp)	155,865,547	147,463,102	41,633,360	105,632,827	55,440,309
GC level (%)	38.86%	35.87%	54.89%	32.74%	56.01%
bases masked (bp)	3,874,208	5,678,584	1,258,976	1,541,813	9,862,436
bases masked (%)	2.49%	3.85%	3.02%	1.46%	17.79%
Retroelements	0.49%	0.93%	0.23%	0.06%	1.68%
SINEs	0.01%	0.00%	0.01%	0.00%	0.00%
Penelope	0.01%	0.05%	0.03%	0.00%	0.02%
LINEs	0.38%	0.25%	0.09%	0.02%	0.60%
CRE/SLACS	0.00%	0.00%	0.00%	0.00%	0.00%
L2/CR1/Rex	0.31%	0.18%	0.00%	0.01%	0.11%
R1/LOA/Jockey	0.01%	0.01%	0.03%	0.00%	0.25%
R2/R4/NeSL	0.00%	0.00%	0.00%	0.00%	0.00%
RTE/Bov-B	0.02%	0.00%	0.00%	0.00%	0.00%
L1/CIN4	0.01%	0.01%	0.01%	0.00%	0.01%
LTR elements	0.11%	0.67%	0.13%	0.04%	1.08%
BEL/Pao	0.01%	0.06%	0.01%	0.00%	0.01%
Ty1/Copia	0.00%	0.09%	0.04%	0.00%	0.04%
Gypsy/DIRS1	0.05%	0.48%	0.03%	0.03%	0.30%

Retroviral	0.02%	0.02%	0.03%	0.01%	0.08%
DNA transposons	0.26%	0.41%	0.17%	0.04%	3.88%
hobo-Activator	0.05%	0.12%	0.08%	0.01%	0.35%
Tc1-IS630-Pogo	0.00%	0.06%	0.01%	0.00%	0.09%
En-Spm	0.02%	0.03%	0.02%	0.01%	0.55%
MuDR-IS905	0.00%	0.00%	0.00%	0.00%	0.00%
PiggyBac	0.00%	0.01%	0.00%	0.00%	0.01%
Tourist/Harbinger	0.00%	0.04%	0.00%	0.00%	0.00%
Other (Mirage, P-element, Transib)	0.01%	0.01%	0.00%	0.00%	0.00%
Rolling-circles	0.00%	0.00%	0.00%	0.00%	0.00%
Unclassified	0.02%	0.04%	0.00%	0.01%	0.01%
Total interspersed repeats	0.77%	1.38%	0.40%	0.11%	5.57%
Small RNA	0.03%	0.01%	0.04%	0.03%	0.10%
Satellites	0.05%	0.01%	0.01%	0.01%	0.04%
Simple repeats	1.16%	1.05%	1.97%	0.68%	11.58%
Low complexity	0.49%	1.40%	0.61%	0.64%	0.51%

Supplementary Table S5: Scaffold size frequencies

Scaffold Size	Frequency	Total base pairs
> 2 kbp	2342	152300948
> 5 kbp	1517	149739537
> 10 kbp	1253	147912957
> 20 kbp	1076	145345415
> 50 kbp	794	135886036
> 100 kbp	509	115294090
> 120 kbp	429	106527368
> 200 kbp	215	72812389
> 500 kbp	27	16400572
> 1 Mbp	1	1222598

Supplementary Table S6: Scaffold gap frequencies

Gap Size	Frequency	Total gaps

>= 10 bp	1882	5527158
>= 100 bp	1508	5520825
>= 1 kbp	1274	5378496
>= 5 kbp	350	3040555
>= 10 kbp	90	1252099

Supplementary Table S7: GC content (% of genome)

Genome-wide GC content was calculated for *M. leidyi* and several other species. The GC content of the *M. leidyi* genome in the same range as *A. queenslandica*, *N. vectensis*, and *H. sapiens*. *Hydra magnipapillata (Hm)*, *Trichoplax adhaerens* (*Ta*), *Amphimedon queenslandica* (*Aq*), *Mnemiopsis leidyi* (*Ml*), *Nematostella vectensis* (*Nv*), *Homo sapiens* (*Hs*), *Monosiga brevicolis* (*Mb*), *Salpingoeca rosetta* (*Sr*)

	Hm	***Ta***	***Aq***	***Ml***	***Nv***	***Hs***	***Dm***	***Mb***	***Sr***
GC	28	32	36	38	40	40	42	54	56

Supplementary Table S8: Dinucleotide odds ratios

The quantity [XpY]/[X][Y] was calculated over all dinucleotides for *M. leidyi* and other selected species. The mean value of each dinucleotide ratio for these species was also calculated. The dinucleotide odds ratios for *M. leidyi* show little notable divergence from the calculated mean values of each dinucleotide of the other sampled genomes. *Mnemiopsis leidyi* (*Ml*), *Amphimedon queenslandica* (*Aq*), *Hydra magnipapillata (Hm)*, *Monosiga brevicolis* (*Mb*), *Nematostella vectensis* (*Nv*), *Trichoplax adhaerens* (*Ta*), *Salpingoeca rosetta* (*Sr*), *Drosophila melanogaster* (*Dm*), *Homo sapiens* (*Hs*)

Dinucleotide	***Ml***	***Aq***	***Hm***	***Mb***	***Nv***	***Ta***	***Sr***	***Dm***	***Hs***	μ
AA	1.04	0.98	1.16	1.13	1.11	0.95	0.83	1.19	1.11	1.06
AC	1.02	1.04	0.99	0.97	1.00	0.92	1.30	0.82	0.83	0.99
AG	1.02	1.04	0.99	0.97	1.00	1.10	0.81	0.82	1.17	0.99
AT	0.83	0.98	0.93	0.95	0.89	1.04	0.83	0.95	0.89	0.92
CA	1.19	1.22	1.19	1.29	1.17	1.10	1.79	1.15	1.17	1.25
CC	1.11	0.93	1.02	0.96	1.25	0.78	0.77	1.13	1.25	1.02
CG	0.83	0.31	0.51	0.82	0.75	0.78	0.89	0.91	0.25	0.67
CT	1.02	1.04	0.99	0.97	1.00	1.10	0.81	0.82	1.17	0.99
GA	1.02	1.04	0.79	0.97	1.00	0.92	0.81	0.99	1.00	0.95
GC	1.11	0.93	1.02	1.23	1.00	1.17	1.15	1.36	1.00	1.11

GG	1.11	0.93	1.02	0.96	1.25	0.78	0.77	1.13	1.25	1.02
GT	1.02	1.04	0.99	0.97	1.00	0.92	1.30	0.82	0.83	0.99
TA	0.83	0.88	0.85	0.38	0.78	0.95	0.41	0.71	0.78	0.73
TC	1.02	1.04	0.79	0.97	1.00	0.92	0.81	0.99	1.00	0.95
TG	1.19	1.22	1.19	1.29	1.17	1.10	1.79	1.15	1.17	1.25
TT	1.04	0.98	1.16	1.13	1.11	0.95	0.83	1.19	1.11	1.06

Supplementary Table S9: Summary statistics for final gene models

	Mean (bp)	**Mode (bp)**
Intron Length	898	297
Length of Intergenic Regions	2463	45492
Exon Length	314	132
Length of Predicted Transcript	5799	315
Protein Length	463	132
Scaffold Length	30562	1012

Supplementary Table S10: Total coding percentage of *Mnemiopsis leidyi* genome

	Percent
Coding	14.76
Non-coding	85.24

Supplementary Table S11: Non-metazoan outgroups used in the analyses of the Genome and EST sets

Name	**Genome Set**	**EST Set**

Opisthokonta	*Sphaeroforma arctica* *Saccharomyces cerevisiae* *Spizellomyces punctatus* *Capsaspora owczarzaki* *Monosiga brevicollis* *Salpingoeca rosetta*	*Spizellomyces punctatus* *Batrachochytrium dendrobatidis* *Cryptococcus neoformans* *Saccharomyces cerevisiae* *Phycomyces blakesleeanus* *Rhizopus orizae* *Amoebidium parasiticum* *Sphaeroforma arctica* *Capsaspora owczarzaki* *Monosiga ovata* *Monosiga brevicollis* *Salpingoeca rosetta*
Holozoa	*Capsaspora owczarzaki* *Monosiga brevicollis* *Salpingoeca rosetta*	*Capsaspora owczarzaki* *Monosiga ovata* *Monosiga brevicollis* *Salpingoeca rosetta*
Choanimalia	*Monosiga brevicollis* *Salpingoeca rosetta*	*Monosiga ovata* *Monosiga brevicollis* *Salpingoeca rosetta*
Animalia	None	None

Table S12: Estimated branch lengths for each Genome Set tree

For each tree, we calculated the branch length from root to tip for each taxa. Table is sorted based on average branch lengths across all analyses (longest to shortest). Species are in the first column and the latin names are abbreviated to first letter of the first part and first three of the second. Opist.=Opisthokonta (all opisthokontan outgroups included), Holo=Holozoa (only holozoan outgroups included), Cho=Choanimalia (only choanoflagellate outgroups included), ML=maximum-likelihood, Bayes=Bayesian, Avg=Average.

	Opist. ML	**Holo. ML**	**Cho. ML**	**Opist. Bayes**	**Holo. Bayes**	**Cho. Bayes**
Cele	1.04	1.06	1.23	2.03	2.30	2.76
Ppac	1.04	1.06	1.23	2.02	2.29	2.76
Mlei	0.67	0.69	0.73	1.26	1.37	1.75
Dmel	0.68	0.69	0.85	1.06	1.14	1.59
Hrob	0.63	0.64	0.80	0.90	0.97	1.41
Aque	0.52	0.53	0.66	0.88	0.94	1.31
Tadh	0.51	0.53	0.68	0.70	0.75	1.16

Ctel	0.48	0.49	0.65	0.65	0.69	1.14
Lgig	0.46	0.47	0.62	0.62	0.66	1.11
Spur	0.45	0.46	0.62	0.58	0.62	1.06
Hsap	0.43	0.44	0.60	0.57	0.62	1.06
Bflo	0.40	0.41	0.56	0.50	0.54	0.98
Nvec	0.37	0.38	0.54	0.45	0.48	0.92
Median	0.51	0.53	0.66	0.7	0.75	1.16
STDEV	0.22	0.23	0.23	0.53	0.62	0.62

Table S13: Estimated branch lengths for each EST Set tree

For each tree, we calculated the branch length from root to tip for each taxon. Since each of the two independent Bayesian analyses did not converge for any of the analyses of the EST Sets, we report both trees in the main text. However, for the branch lengths analyses we calculate a single tree using bpcomp for each analysis. Table is sorted based on average branch lengths across all analyses (longest to shortest). Species are in the first column and the Latin names are abbreviated to first letter of the first part and first three (or four) of the second. Opist.=Opisthokonta (all opisthokontan outgroups included), Holo=Holozoa (only holozoan outgroups included), Cho=Choanimalia (only choanoflagellate outgroups included), ML=maximum-likelihood, Bayes=Bayesian, Avg=Average.

	Opist. ML	**Holo. ML**	**Cho. ML**	**Opist. Bayes**	**Holo. Bayes**	**Cho. Bayes**
Smed	0.93	0.96	1.03	1.47	1.83	2.11
Clon	0.82	0.84	0.91	1.37	1.67	1.92
Ipul	0.82	0.85	0.92	1.35	1.66	1.93
Sros	0.80	0.82	0.89	1.27	1.54	1.45
Msti	0.75	0.75	0.82	1.17	1.37	1.62
Omin	0.69	0.71	0.77	1.14	1.34	1.60
Xind	0.71	0.72	0.79	0.99	1.16	1.41
Poli	0.71	0.72	0.79	0.98	1.14	1.40
Ppil	0.57	0.59	0.63	1.00	1.28	1.52

Nwes	0.65	0.66	0.73	0.95	1.12	1.36
Dmel	0.68	0.70	0.76	0.92	1.08	1.33
Mlei	0.56	0.58	0.61	0.96	1.22	1.47
Msp	0.56	0.58	0.61	0.94	1.21	1.45
Hrob	0.62	0.64	0.70	0.82	0.95	1.20
Hror	0.63	0.64	0.71	0.82	0.96	1.19
Cint	0.62	0.64	0.71	0.81	0.95	1.18
Dpul	0.58	0.59	0.66	0.77	0.90	1.14
Ehor	0.58	0.59	0.66	0.74	0.86	1.10
Srap	0.53	0.54	0.60	0.78	0.89	1.12
Chem	0.55	0.56	0.63	0.76	0.87	1.11
Bmic	0.56	0.57	0.64	0.73	0.85	1.09
Aque	0.51	0.52	0.58	0.76	0.88	1.11
Lbai	0.52	0.53	0.59	0.75	0.84	1.07
Cfol	0.52	0.53	0.59	0.73	0.83	1.06
Hmag	0.54	0.55	0.62	0.70	0.81	1.05
Esco	0.55	0.56	0.63	0.69	0.81	1.05
Pmar	0.54	0.55	0.61	0.66	0.77	1.00
Emue	0.49	0.50	0.55	0.71	0.80	1.04
Hech	0.52	0.53	0.60	0.67	0.77	1.01
Cvir	0.53	0.53	0.60	0.65	0.76	1.01
Lcha	0.49	0.50	0.56	0.70	0.79	1.02
Aero	0.51	0.52	0.59	0.65	0.77	1.01
Sdom	0.47	0.48	0.54	0.68	0.76	1.00
Apall	0.53	0.55	0.62	0.64	0.72	0.96
Ocar	0.48	0.49	0.55	0.68	0.76	0.99

Lgig	0.51	0.52	0.59	0.64	0.75	0.99
Ctel	0.51	0.52	0.59	0.64	0.75	0.99
Pcar	0.51	0.52	0.59	0.64	0.74	0.98
Xboc	0.48	0.49	0.55	0.65	0.75	1.00
Tadh	0.46	0.47	0.54	0.65	0.77	1.01
Spur	0.50	0.51	0.58	0.60	0.71	0.96
Ttra	0.48	0.49	0.56	0.60	0.70	0.95
Olob	0.46	0.46	0.53	0.64	0.72	0.95
Clac	0.48	0.49	0.56	0. 60	0.70	0.94
Ccap	0.48	0.49	0.55	0.60	0.69	0.93
Apec	0.48	0.49	0.56	0.58	0.68	0.92
Ekan	0.46	0.47	0.54	0.59	0.69	0.93
Ggal	0.46	0.47	0.53	0.57	0.66	0.89
Bflo	0.45	0.46	0.53	0.55	0.64	0.87
Avir	0.44	0.44	0.51	0.53	0.59	0.83
Skow	0.43	0.44	0.50	0.50	0.60	0.84
Apalm	0.43	0.44	0.51	0.52	0.59	0.83
Msen	0.43	0.44	0.51	0.52	0.59	0.83
Mfav	0.43	0.44	0.51	0.52	0.59	0.83
Amil	0.42	0.42	0.49	0.51	0.57	0.82
Pfla	0.41	0.42	0.48	0.49	0.58	0.82
Past	0.41	0.42	0.49	0.48	0.57	0.80
Nvec	0.40	0.41	0.48	0.49	0.55	0.79
Median	0.515	0.525	0.59	0.68	0.77	1.01
STDEV	0.12	0.12	0.12	0.23	0.30	0.29

Supplementary Table S14: Hypotheses comparisons of likelihood analysis of gene content

Values in blue indicate values larger than the significance level, 0.05, and indicate inclusion within the confidence set. See Figure 2 for graphical representation of hypotheses. Hypotheses were determined *a priori*. AU = the p-value from the approximately unbiased test, NP = bootstrap probability of the selection, BP = same as np but calculated directly from the replicates with $r_k = 1$, PP = Bayesian posterior probability calculated by the BIC approximation, KH = the p-value from the Kishino-Hasegawa test, SH = the p-value from the Shimodaira-Hasegawa test, WKH = the p-value from the weighted Kishino-Hasegawa test, WSH = the p-value from the weighted Shimodaira-Hasegawa test. SOWH = the p-value from the Swofford–Olsen–Waddell–Hillis test. Asterisk indicates that the SH and WSH tests tend to include more trees in the confidence set than is necessary (*96*).

Hypothesis **(figure)**	**AU**	**NP**	**BP**	**PP**	**KH**	**SH**	**WKH**	**WSH**	**SOWH**
(Ct,) Fig. 2d	0.973	0.973	0.973	1	0.972	0.991	0.972	0.998	1
(Po,) Fig. 2c	0.027	0.027	0.027	2 x 10^{-14}	0.028	0.32*	0.028	0.086*	0
(Bi,) Fig. 2f	1 x 10^{-7}	3 x 10^{-07}	0	2 x 10^{-220}	0	0	0	0	0
(Cn,Ct) Fig. 2a	3 x 10^{-41}	3 x 10^{-15}	0	8 x 10^{-129}	0	0	0	0	0
(Ct,Bi) Fig. 2b	8 x 10^{-45}	5 x 10^{-16}	0	1 x 10^{-129}	0	0	0	0	0
(Tr,) Fig. 2e	4 x 10^{-53}	2 x 10^{-17}	0	1 x 10^{-93}	0	0	0	0	0

Supplementary Table S15: Presence and absence of notch pathway components

A check mark in a green box indicates the presence of the gene. A check mark in a yellow box indicates uncertainty because of missing diagnostic domains. A grey box with a minus sign (-) indicates the absence of the gene. See Table S14 for *M. leidyi* gene identifiers. *Hs=Homo sapiens*, *Nv=Nematostella vectensis*, *Hm=Hydra magnipapillata*, *Ta=Trichoplax adhaerens*, *Aq=Amphimedon queenslandica*, *Ml=Mnemiopsis leidyi*, *Mb=Monosiga brevicolis*, Choano=Choanoflagellata.

	Bilateria	Cnidaria		Placozoa	Porifera	Ctenophora	Choano
	Hs	***Nv***	***Hm***	***Ta***	***Aq***	***Ml***	***Mb***
Notch	√	√	√	√	√	√	-
Delta	√	√	√	√	√	√	-
O-fut	√	√	√	√	√	-	√
Fringe	√	√	-	-	√	-	-
furin	√	√	√	-	-	√	-
Tace=Adam17	√	√	√	√	√	√	√
Kuzbanian=Adam10	√	√	√	√	√	√	√
Presenillin	√	√	√	√	√	√	√
Nicastrin	√	√	√	√	√	√	√
APH1	√	√	√	√	√	√	√
Pen2	√	√	√	√	√	√	√
Su(H)	√	√	√	√	√	√	√
Mastermind (Co-A)	√	√	-	-	-	-	-
SMRT (Co-R)	√	-	-	-	-	-	-
Numb	√	√	√	√	-	√	-
Hes/Hey	√	√	√	√	√	√	-
Strawberry notch	√	√	√	√	√	-	√
Neuralized	√	√	-	√	√	-	-
Mindbomb	√	√	√	-	√	√	-
Deltex	√	√	-	√	√	√	√
Nedd4/sudx	√	√	√	√	√	√	√
notchless	√	√	√	√	√	√	√
Total	22	16-21	8-17	14-17	17-18	12-16	10-12

Supplementary Table S16: *M. leidyi* identifiers corresponding to genes in Supplementary Table S15

A yellow box indicates uncertainty because of missing diagnostic domains.

Notch	ML128617a
Delta	ML21438a
furin	ML07022a
Tace=Adam17	ML17408a
Kuzbanian=Adam10	ML03054a
Presenillin	ML01594a
Nicastrin	ML102219a
APH1	ML305514a
Pen2	ML0708a, MLRB070835a
Su(H)	ML141212a
Numb	ML00718a

Hes/Hey	ML065313a
Mindbomb	ML26791a
Deltex	ML030223a
Nedd4/sudx	ML044111a
notchless	ML45849a

Supplementary Table S17: Presence and absence of post-synaptic genes
A check mark indicates the presence of the gene and a grey box with a minus sign (-) indicates the absence of the gene. See Table S16 for *M. leidyi* gene identifiers. *Hs=Homo sapiens*, *Nv=Nematostella vectensis*, *Hm=Hydra magnipapillata*, *Ta=Trichoplax adhaerens*, *Aq=Amphimedon queenslandica*, *Ml=Mnemiopsis leidyi*, *Mb=Monosiga brevicolis*, Choano=Choanoflagellata

	Bilateria	**Cnidaria**		**Placozoa**	**Porifera**	**Ctenophora**	**Choano**
	Hs	***Nv***	***Hm***	***Ta***	***Aq***	***Ml***	***Mb***
Alpha Catenin	√	√	-	√	√	√	-
AMPA iGluR	√	√	√	-	-	-	-
Beta Catenin	√	√	√	√	√	√	√
CamKII	√	√	√	√	√	√	√
CASK	√	√	√	√	-	-	-
Citron	√	√	√	√	√	√	√
Classical Cadherin	√	√	√	√	√	√	-
Cortactin	√	√	√	√	√	√	√
CRIPT	√	√	√	√	√	√	√
p120/δ Catenin	√	√	√	√	√	√	-
DLG	√	√	√	√	√	√	√
Ephrin Receptor	√	√	√	√	√	√	√
ErbB Receptor	√	-	-	-	√	-	-
Erbin	√	√	√	-	-	-	-
GKAP	√	√	√	√	√	√	-
GRIP	√	-	-	-	√	-	-
Homer	√	√	√	√	√	√	√
IP3R	√	√	√	√	√	√	√
K+ channel shaker	√	√	√	√	√	√	√
LIN-7	√	√	√	√	√	√	-
MAGI	√	√	√	√	√	√	-
mGluR	√	√	√	√	√	√	-
Neuroligin	√	√	-	-	-	-	-
NMDA iGluR	√	√	√	-	-	-	-
NOS	√	√	√	√	√	√	-

PICK1	√	√	√	√	√	√	-
PKC	√	√	√	√	√	√	√
PMCA	√	√	√	√	√	√	√
Shank	√	√	√	-	√	√	√
SPAR	√	√	√	√	√	√	-
stargazin	√	-	-	-	-	-	-
SynGAP	√	√	√	-	√	√	-
Total	32	29	27	23	26	24	13

Supplementary Table S18: *M. leidyi* identifiers corresponding to genes in Supplementary Table S17

Alpha Catenin	ML257617a
Beta Catenin	ML073715a
CamKII	ML070269a
Citron	ML154113a
Classical Cadherin	ML00359a
Cortactin	ML045237a
CRIPT	ML04051a
Delta Catenin	ML002622a
DLG	ML01744a
Ephrin Receptor	ML35913a
GKAP	ML03326a
Homer	ML06361a
IP3R	ML00401a
K+ channel shaker	ML18152a
LIN-7	ML05296a
MAGI	ML02503a
mGluR	ML17995a
NOS	ML074215a
PICK1	ML19124a
PKC	ML13931a
PMCA	ML11054a
Shank	ML01578a
SPAR	ML11651a
SynGAP	ML038810a

Supplementary Table S19: Presence and absence of Dopamine / Norepinephrine / Epinephrine Biosynthetic Pathway components

A check mark indicates the presence of the gene and a grey box with a minus sign (-) indicates the absence of the gene. The identifier for the Qdpr gene in *M. leidyi* is

ML08064. Qdpr = quinoid dihydropteridine reductase, Th = tyrosine hydroxylase, Slc18A2 = Homo sapiens solute carrier family 18 member 2, Ddc = dopa decarboxylase, Dbh = dopamine beta-hydroxylase, Pnmt = phenylethanolamine N-methyltransferase, *Hs=Homo sapiens*, *Nv=Nematostella vectensis*, *Hm=Hydra magnipapillata*, *Ta=Trichoplax adhaerens*, *Aq=Amphimedon queenslandica*, *Ml=Mnemiopsis leidyi*, *Mb=Monosiga brevicolis*, *Cowc=Capsaspora owczarzaki*, Choano=Choanoflagellata

	Bilateria	**Cnidaria**		**Placozoa**	**Porifera**	**Ctenophora**	**Choano**
	Hs	***Nv***	***Hm***	***Ta***	***Aq***	***Ml***	***Mb***
Qdpr	√	√	√	√	√	√	(in *Cowc*)
Th	√	-	-	-	-	-	-
Slc18A2	√	-	-	-	-	-	-
Ddc	√	√	√	√	-	-	-
Dbh	√	√	-	√	√	-	-
Pnmt	√	-	-	-	-	-	-

Supplementary Table S20: Presence and absence of mesoderm components in model genomes

A check mark indicates the presence of the gene and a grey box with a minus sign (-) indicates its absence. See Table S19 for *M. leidyi* gene identifiers. *Hs=Homo sapiens*, *Nv=Nematostella vectensis*, *Hm=Hydra magnipapillata*, *Ta=Trichoplax adhaerens*, *Aq=Amphimedon queenslandica*, *Ml=Mnemiopsis leidyi*, *Mb=Monosiga brevicolis*, Choano=Choanoflagellata

	Bilateria	**Cnidaria**		**Placozoa**	**Porifera**	**Ctenophora**	**Choano**
	Hs	***Nv***	***Hm***	***Ta***	***Aq***	***Ml***	***Mb***
Twist	√	√	-	-	-	-	-
MyoD family myogenin/Myf5	√	√	-	-	-	-	-
Brachyury	√	√	√	√	(a)	√	(b)
Snail	√	√	√	√	-	(c)	-
Eomesodermin/TBR2	√	-	-	-	-	-	-
gli/glis	√	√	√	-	√	√	-
Tinman/Nkx-2.6	√	√	√	√	√	-	-
nk2.1/NKX2-1	√	√	√	√	√	-	-
Bagpipe/NKX3-2	√	√	√	-	-	-	-
Ladybird	√	√	-	-	-	-	-
Pax3	√	√	-	-	-	-	-
Gsc	√	√	√	√	-	-	-
Forkhead/HNF3 FoxA/group 1	√	√	√(d)	√	-	-	-
Mef2	√	√	√	√	√	√	(b)
GATA	√	√	√	√	√	(e)	(f)

Muscle LIM/CSRP3	√	√	√	-	√	√	√
Troponin T	√	-	-	-	-	-	-
Troponin I	√	-	-	-	-	-	-
Troponin C	√	-	-	-	-	-	-

(a) Brachyury in *A. queenslandica* but in 3 other sponges(*155*, *156*)
(b) Brachyury and Muscle LIM are present in *Capsaspora owczarzaki*
(c) Snail is present in the ctenophores *Euplokamis dunlapae* and *Thalassocalyce inconstans*
(d) Forkhead is a pseudogene in *Hydra magnipapillata* but it is present in *Hydra vulgaris (157)*
(e) GATA is found in several other ctenophores
(f) Genes with single Znf GATA domains are present in *C. owczarzaki*

Supplementary Table S21: *M. leidyi* identifiers corresponding to genes in Supplementary Table S20

Brachyury	ML174736a
gli/glis	ML145833a
Mef2	ML07781a
Muscle LIM/CSRP3	ML02959a

Supplementary Table S22: Accession numbers of sequences from other ctenophore transcriptomes that were top TBLASTN hits to myogenic and synaptic scaffolding genes of interest. The row headers are the human NCBI Gene IDs used as the TBLASTN queries. In all cases except for ACCESSIONs (in red), these genes were determined not to be orthologous to the gene of interest.

	B. chuni	*B. forskalii*	*C. fugiens*	*E. dunlapae*	*H. californensis*	*L. lactea*	*T. inconstans*
GRIA2, GRIN1, GRIK2, GRID2	KF317296 KF317294 KF317293 KF317292 KF317295 KF317321 KF317322 KF317300	KF317345 KF317346 KF317347 KF317348 KF317344 KF317358	KF317388 KF317389 KF317387 KF317390 KF317391	KF317429 KF317428 KF317426 KF317427 KF317425 KF317440	KF317468 KF317466 KF317465 KF317467 KF317469 KF317476 KF317475	KF317491 KF317490 KF317493 KF317497 KF317492 KF317494 KF317496	KF317522 KF317524 KF317521 KF317523 KF317520 KF317531
TWIST1	KF317289 KF317290 KF317287 KF317302 KF317301	KF317332 KF317328 KF317330 KF317331 KF317329	KF317374 KF317375 KF317373 KF317372 KF317371	KF317408 KF317407 KF317411 KF317410 KF317409	KF317452 KF317454 KF317453		KF317505 KF317506 KF317507 KF317504 KF317503
SNAI1	KF317279 KF317277 KF317276 KF317278 KF317280	KF317323 KF317325 KF317324 KF317326 KF317327	KF317369 KF317366 KF317367 KF317370 KF317368	KF317402 KF317403 KF317404 KF317405 KF317406	KF317449 KF317447 KF317450 KF317448	KF317484 KF317482 KF317483 KF317485 KF317481	KF317498 KF317501 KF317500 KF317499 KF317502

NLGN1	KF317281 KF317284 KF317282 KF317283 KF317285	KF317340 KF317338 KF317339 KF317342 KF317341	KF317383 KF317386 KF317382 KF317384 KF317385	KF317418 KF317417 KF317419 KF317420 KF317421	KF317462 KF317463 KF317460 KF317459 KF317461	KF317489 KF317488	KF317517 KF317514 KF317513 KF317515 KF317516
MYF5, MYF6, MYOG, MYOD1	KF317297 KF317286 KF317288 KF317316	KF317349 KF317350 KF317351 KF317353 KF317352 KF317360 KF317359		KF317422 KF317423	KF317464		KF317534 KF317535
NOG	KF317291			KF317424			
EOMES	KF317309 KF317307 KF317310 KF317306 KF317308	KF317337 KF317336 KF317334 KF317335 KF317333	KF317376	KF317412 KF317413 KF317416 KF317415 KF317414	KF317458 KF317457 KF317456 KF317455 KF317451	KF317486	KF317508 KF317511 KF317512 KF317509 KF317510
GATA1	KF317319 KF317318 KF317320 KF317317	KF317365 KF317361 KF317362 KF317364 KF317363	KF317399 KF317397 KF317400 KF317401 KF317398	KF317446 KF317443 KF317444 KF317442 KF317445	KF317479 KF317478 KF317477 KF317480	KF317487	KF317536 KF317537 KF317525 KF317526 KF317527
TNNC1, TNNT1, TNNI1	KF317533 KF317532			KF317441			
NODAL	KF317313 KF317312 KF317314 KF317311 KF317315	KF317343 KF317354 KF317356 KF317355 KF317357	KF317392 KF317395 KF317396 KF317393 KF317378 KF317394	KF317435 KF317437 KF317436 KF317439 KF317438	KF317472 KF317470 KF317471 KF317473 KF317474	KF317495	KF317519 KF317518 KF317529 KF317530
FGF	KF317305 KF317303 KF317304		KF317377 KF317380 KF317381 KF317379				
SHH	KF317299 KF317298			KF317431 KF317434 KF317433 KF317432 KF317430			KF317528

Table S23: ANTP class homeodomains from other ctenophore transcriptomes

```
B.chuni.2 PRKDRTIFSKIQLFRLERRFHDQKYLSSYEKALIAHSLQLTQQQVQTWFQNRRAKAKREE
B.chuni.3 EKRVRTIFSISQLFRLERRFNAQKYLTASERARLAYSLQLTETQVKIWFQNRRAKWKREM
B.chuni.5 KKKTRTTFSRAQVFGLERKFATQKYLSTHDRILLAAALQMTECQVKIWFQNRRTKWRREK
B.chuni.6 RRGPRTTIKSKQLEILKDAFNSTPKPTRHIREQLATKTGLNMRVIQVWFQNRRSKERRMK
B.chuni.7 QRRNRTQFSTYQLDQLEAEFEKSHYPDVLTREELANGLELTEARVQVWFQNRRAKWRKKQ
B.chuni.8 ERKARTVFTNDQLQLLENRFKSQPRLTSLEREELAEQIQLSATQIRVWFQNKRAKMKRDK
B.chuni.11 PRKTQTVFSTHQLSNLERRYNSNNYITTEERKCIADSLHMSVPQVKNWFQNKRNKERKLG
B.chuni.12 QRRNRITFTAAQLDGLEKSFQETHYPDVFGREEIAYSLQLTEQRVQVWFQNRRAKWRKRE
B.chuni.16 KRKPRTCFSVGQMLVLENTFQQTKYVSITDRGKLAGSLGLTDSQIKVWFQNRRSKWRKTM
B.chuni.18 KWRNRHTFTATQLEELEQVFNSTHYPDIFTREELAMKHNLTESRVQVWFQNRRAKHRKNE
B.chuni.20 SRKPRTVFTWEQLKQMEETFKQKKYLCTEDRMILAQKLDLTDEQIKVWFQNRRQKWKKGG
B.chuni.22 QRRNRTNYSAIQLNELEIVFSKSKYPDIFTREELALRLGVPEARIQVWFQNRRAKWRKQG
B.chuni.32 IKRPRTTITAKQLETLKSAYENSPKPARHVREQLSTETGLDMRVVQVWFQNRRAKEKRMK
B.chuni.36 GTNDRTIYTPVQAVRLEELFTERPHITKEQRDTLSGELTIHPDRIKVWFQNRRAKQRRED
B.chuni.39 PKYSRKNFSPEQIEGLEAAFSQHRFVKKDLRKQLAKQLSLSERSISYWFQNKRARSKPPI
B.chuni.40 KRNERHGIDDHQATTLKDWFQRYTYLTIENRKIVSIETGLPEKTVMYWFQNQRRKIKRQQ
B.chuni.41 PRRPRTIFSASQLLELEERFRYQKYLSTAERSCLAFTLGLSEEQVKVWFQNRRSKWKKGE
B.chuni.49 QRRNRTTFSSVQLHELERAFQQSHYPDVFTREELAMRLDLTEARVQVWFQNRRAKWRKRE
B.chuni.60 KRRKRTTISSNSKEILEQYYNTNPLPSTDEIGNLSNNLALDKRVIRIWFQNRRAIGKRLS
B.chuni.64 PRKPRTHFTDSQIEDLEKVFEDKKYLSATERQIIANDLNLQEEQVKNWFQNRRSRWRKDC
B.chuni.68 SRRCRTKIEPDMLDLLETKYQESHFISPFERKSLSETLGITERAVIYWFQNRRRKDIKNM
B.chuni.83 PRKPRTIFTAAQLLELEQKFRHQKYLSTSERSCLAFMLGLSEEQIKVWFQNRRSKWKKGG
B.chuni.111 TMRHRTRFTSSQLEKLEDAFSDTQYPDLSSRESISRDIGLSESCVQVWFQNRRARWRKSL
C.fugiens.1 KKKTRTTFSRAQVFGLERKFATQKYLSTHDRILLAAALHMTECQVKIWFQNRRTKWRREK
C.fugiens.4 QRRNRTNYSAAQLNELEIVFQRTRYPDIFTREELSLRLGIPEARIQVWFQNRRAKWRKRT
C.fugiens.8 PRKPRTHFTNGQVEELEKIFEDKKYLDAKEREVVAVDLNLAEEQVKNWFQNRRSKWRKDI
C.fugiens.10 PRKTQTVFSSFQVQHLEQKYNSANYISTEERQKIASRLQMSVPQVKNWFQNKRNKERKMG
C.fugiens.11 QRRNRTNYSAIQLNELEMVFGKSKYPDIFTREELALRLGVPEARIQVWFQNRRAKWRKQS
C.fugiens.12 RRKARTVFSDDQLQGLERKFKMQKYLSVPERMELAGMLSLSETQVKTWFQNRRMKWKKQG
C.fugiens.13 KRRRRTVFTERQLQGLEEAYSKSQYLDRESRLELCKKLSLSLHTVVYWFQNKRAISRRRG
C.fugiens.16 KRRRRTEYNEYQVAYLEMAFTENHYPSIAMREELADHIQIPEARIQIWFQNRRNKFRNHG
C.fugiens.20 KWRNRHTFTQAQLDELEKVFSKTHYPDIFTREDLAMKHNLTESRVQVWFQNRRAKHRKSE
C.fugiens.22 IKRPRTTITAKQLETLKNAYENSPKPARHVREQLSTETGLDMRVVQVWFQNRRAKEKRMK
C.fugiens.30 KRKPRTSFTNLQLFELERKFYHKKYLASSERKKLAQLLNLTDIQVKTWFQNRRNKHKRSK
C.fugiens.33 QRRNRTQFTTYQLDQLEAEFDKSHYPDVLTREELANGLDLTEARVQVWFQNRRAKWRKKQ
C.fugiens.38 CSKRTKFSNEQLRVLEHYYHQVNKYVVGANKTALCHVTGLELNTILMWFQNKRAREKKQR
C.fugiens.52 RCRPRTSFTNFQLAHLEVAFSQTHYADIHTREELAKRLQLHESRIQVWFQNRRAKFRKAG
C.fugiens.55 PKYTRKNFSPEQIEGLEAAFQEHRFVKKDLRRELARKLNLSERSISYWFQNKRARSKPPI
C.fugiens.79 PRKPRTIFSAAQLNELEERFKYQKYLSTSERSCLAYSLGLTEEQIKVWFQNRRSKWKKGD
C.fugiens.80 RRKARTVFTDDQLKGLETQFGTQKYLSVPERMELAVSLRLSETQVKTWFQNRRMKWKKQV
C.fugiens.81 EKRVRTIFSISQLFRLERRFNAQKYLSASERARLAYSLQLTETQVKIWFQNRRAKWKREM
C.fugiens.89 RRKARTVFSDEQLTGLEDKFRVQKYLSVPERVELAVSLDLSETQVKTWFQNRRMKWKKGQ
C.fugiens.99 ERKARTVFTNEQLQLLENRFKAQPRLTSLEREELAQQIKLSATQIRVWFQNKRAKVKRDR
C.fugiens.123 RRKARTVFSDVQLEGLEKKFRSQKYLSVPERLDVATGLGLSETQVKTWFQNRRMKWKKQV
C.fugiens.128 GKKPRTIFSRDQVQKLEEAYLSKRYLTRRERKDLATAAQLSHTQVKIWFQNRRAKAKLKD
C.fugiens.164 RKRTRTTFSSAQVYELEKKFQRSQYLSAVDRLNLAAALSMSDVQVKRWFQNRRCKERHRA
C.fugiens.178 QRRNRITFTAAQLEGLEKSFQETHYPDVYGREEIAYSLQLTEQRVQVWFQNRRAKWRKRQ
L.lactea.2 QRRNRTNYSAVQLNELEIVFSKSKYPDIFTREELALRLGVPEARIQVWFQNRRAKWRKQG
L.lactea.3 SRKNRTSFSDKQLSVLEGVFRYKMYVSVTDRTMLSSRLELSDMQIKTWFQNRRTKWKKDN
L.lactea.8 PRKPRTHFTCSQVGDLEKVFEDKKYLSASERQTIAIDLNLQEEQVKNWFQNRRSRWRKDC
L.lactea.9 QRRNRTNYSAAQLNELELVFQRTRYPDIFTREELSLRLGIPEARIQVWFQNRRAKWRKRA
L.lactea.30 RRKARTVFSDIQLEGLERKFRNQKYLSVPERLDIATGLGLSETQVKTWFQNRRMKWKKQV
L.lactea.44 RRGPRTTIKSKQLEILKDAFNKTPKPTRHIREQLATKTGLNMRVIQVWFQNRRSKERRMK
L.lactea.46 HVNSRTMFTVHQVAAMEERYLKSSTILKPEREKFGMGLGLSETAIRTWFQNRRARQKRQE
T.inconstans.4 PRKPRTHFTDTQIEDLEKVFEDKKYLSANERQIIANDLSLHEEQVKNWFQNRRSRWRKDC
T.inconstans.8 KRRKRTTISSNSKEILEQYYQTNPLPSTDEIGNLSNNLSLDKRVIRIWFQNRRAIGKRLS
T.inconstans.14 PRKPRTIFSAAQLNELEERFKYQKYLSTTERSCLAYSLGLTEEQIKVWFQNRRSKWKKGD
T.inconstans.21 ERKPRTHFTDTQIEDLEKVFEDKKYLSANERQIIANDLSLHEEQVKNWFQNRRSRWRKDC
T.inconstans.23 PRKTQTVFTSHQLNNLEAKYRRNNYITTDERVSIAETLHMSVPQVKNWFQNKRNKERKMG
T.inconstans.37 QRRNRTQFTTCQLDQLEAEFDRSHYPDVLTREELAKCLGLTEARVQVWFQNRRAKWRKKQ
T.inconstans.51 YAGGARKQRRNRTNELEIVFSKSKYPDIFTREELALRLGVPEARIQVWFQNRRAKWRKQG
T.inconstans.71 QRRNRITFTAAQLEGLEKSFQETHYPDVFGREEIAYSLQLTEQRVQVWFQNRRAKWRKRQ
T.inconstans.84 ERKARTVFTNDQLQLLENRFRSQPRLTSLEREELAEQIQLSATQIRVWFQNKRAKMKRDR
T.inconstans.99 KWRNRHTFTQAQIDELEQVFATTHYPDIFTREELANKHKLTEARVQVWFQNRRAKHRKNE
B.forskalii.5 ERKARTVFTNDQLQLLEDRFKSQPRLTSLEREELAEQIQLSATQIRVWFQNKRAKMKRDK
B.forskalii.7 PRKPRTHFTDSLIEDLEKVFEDKKYLSANERIIIANDLGLQEEQVKNWFQNRRSRWRKDC
B.forskalii.8 RHRTRFNDDQVAAMERYYNHVSRYARPDNGLPELIRETGLTHDTIMLWFQNKRARDKRKG
B.forskalii.12 AKYARRNFSATQIRGLEAAFREQRFIKKEVREQLAKDLGLSERSISYWFQNKRARSKGPI
B.forskalii.21 PRKPRTHFTDSQIEDLEKVFEDKKYLSANERIIIANDLGLQEEQVKNWFQNRRSRWRKDC
B.forskalii.25 VKRPRTVLSSVQRKVFKEAFDRTPRPCRKEREKLSSQTGLSVRVVQVWFQNQRAKVKKLA
B.forskalii.27 TKRFRSSFSTEQLRTMEDTFRHRPYLSTAQVEELAGNLQLSSRQVKIWFQNRRTKLKKQV
B.forskalii.31 QRRNRTQFSTYQLDQLEAEFVKSHYPDVLTREELANGLDLTEARVQVWFQNRRAKWRKKE
B.forskalii.33 QRRNRITFTAAQLEGLEKSFQETHYPDVFGREEIAYSLQLTEQRVQVWFQNRRAKWRKRE
B.forskalii.40 SRRFRTTFTSCQLQALEGAFRQTHYPDMYMREELAMRIDLTEARVQVWFQNRRAKWRKRE
B.forskalii.41 KRRRRTVFTERQLQGLEEAYSRSQYLDRESRLELCKKLSLSLHTVVYWFQNKRAISRRRG
B.forskalii.46 SKKPRTIFSREQVQKLEDAYQTKRYLTRRERKELAADAEISHTQVKIWFQNRRAKAKLKD
B.forskalii.50 IKRPRTTITAKQLETLKTAYENSPKPARHVREQLSTETGLDMRVVQVWFQNRRAKEKRMK
B.forskalii.56 PRKTQTVFTRQQLSELEAEYNVNNYISTEDRNEIARRLNMTVPQVKNWFQNKRNKERKMG
B.forskalii.59 PRKPRTIFSATQLNELEERFKYQKYLSTTERSCLAYSLGLTEEQIKVWFQNRRSKWKKGD
B.forskalii.68 KRRKRTTISSNSKEILEQYYQTNPLPSTEEIGNLSNNLSLDKRVIRIWFQNRRAIGKRLS
B.forskalii.81 EKRNRTKFTWEQLDRLEAEYQREQFAVADRKDALARELGVPARTISLWFQNRRAKQRREQ
B.forskalii.84 KKKTRTTFSRAQVFGLERKFATQKYLSTHDRILLAAALQMTECQVKIWFQNRRAKAKLKD
B.forskalii.86 PRKDRTIFSKIQLFRLERRFHAQKYLSSYEKALIAHSLQLTQQQVQTWFQNRRAKAKREE
E.dunlapae.1 RRKARTVFSDDQLNGLENKFKDQKYLSVPERVELAVSLELSETQVKTWFQNRRMKWKKGQ
E.dunlapae.2 QRRNRTQFTTYQLDQLEGEFDKSHYPDVLTREELAHSLELTEARVQVWFQNRRAKWRKKQ
E.dunlapae.5 PRKDRTIFTKIQLFRLERRFHDQKYLSSYEKALIAHSLHLTQQQVQTWFQNRRAKAKREE
E.dunlapae.6 RRKARTVFTDDQLQGLESQFGAQKYLSVPERMELAVSLRLSETQVKTWFQNRRMKWKKQV
```

```
E.dunlapae.7 RRKARTVFSDVQLEGLERKFRSQKSLSVPERMDIASGLGLSETQVKTWFQNRRMKWKKQI
E.dunlapae.9 ERKARTVFTNAQLQLLEDRFKAQPRLTSLEREELAEQMKLSATQIRVWFQNKRAKMKRDK
E.dunlapae.10 KWRNRHTFTAAQLEELEHVFNTTHYPDIFTREELAMKHNLTEARVQVWFQNRRAKFRKTL
E.dunlapae.11 QRRNRTNYSAAQLNELEMVFQRTRYPDIFTREELSLRLGIPEARIQVWFQNRRAKWRKRA
E.dunlapae.14 RRKARTVFSDVQLEGLERKFRSQKYLSVPERMDIASGLGLSETQVKTWFQNRRMKWKKQI
E.dunlapae.16 TTRVRTVLNDRQLRILRTCYNNNPRPDALMKEQMTKLTGLSARVIRVWFQNKRCKDKKKA
E.dunlapae.17 PRKPRTIFTASQLLELEQKFRYQKYLSTSERSCLAFGLGLSEEQIKVWFQNRRSKWKKGG
E.dunlapae.25 PRKPRTIFSASQLNELEERFKYQKYLSTTERSCLAYSLGLTEEQIKVWFQNRRSKWKKGD
E.dunlapae.27 GECDRHTFTAAQLEELEHVFNTTHYPDIFTREELAMKHNLTEARVQVWFQNRRAKFRKTE
E.dunlapae.30 VMRHRTRFNNEQLFVLENSFNDSQYPDLAARENIAGLVGLSENCVQVWFQNRRARWRKIV
E.dunlapae.33 KRRKRTTISSNSKEILEQYYQTNPLPSTEEIGNLSNNLALDKRVIRIWFQNRRAIGKRLS
E.dunlapae.36 SRKPRTVFTWEQLKQMEETFKLKKYLCTEDRMLLAQKLDLTDEQIKVWFQNRRQKWKKGG
E.dunlapae.38 GKKPRTIFTREQVQRLEEAYTQKRYLTRRERKELAKEADISHTQVKIWFQNRRAKAKLKD
E.dunlapae.41 PRKPRTHFTISQIDDLEKVFVDKKYLSVTERSTLAYNLGLQEEQVKNWFQNRRSRWRKDC
E.dunlapae.48 TVRHRTRFTVSQLDQLEQVFNKTHYPDLSLREHLAMRTSLTEACVQIWFQNRRARWRKAV
E.dunlapae.50 RHRTRFNDDQIIAMEQYYNNVSHYARPDNGLPDLIQSTGLSHDTIMLWFQNKRARDKRKV
E.dunlapae.54 MKRPRTTITAKQLETLKTAYEKSPKPARHVREQLSTETGLDMRVVQVWFQNRRAKEKRLK
E.dunlapae.55 QRRNRTNYSAAQLNELEIIFQKSRYPDIFTREEMSLRLGIPEARIQVWFQNRRAKWRKQV
E.dunlapae.57 EEGTLANGKKPRTIFTREQVQRLEEAYTQKRYLTRRERKELHTQVKIWFQNRRAKAKLKD
E.dunlapae.59 RRKARTVFSDDQLQGLERKFKMQKYLSVPERMELANMLALSETQVKTWFQNRRMKWKKQG
E.dunlapae.71 KWRNRHTFTAAQLEELEHVFNTTHYPDIFTREELAMKHNLTEARVQVWFQNRRAKFRKTE
E.dunlapae.79 QRRNRTTFSSIQLHELERAFQQSHYPDVFTREELAMRLDLTEARVQVWFQNRRAKWRKRE
E.dunlapae.88 QRRNRTNYSAAQLNELEMVFQRTRYPDIFTREELSLRLGIPEARIQVWFQNCHVGVMSPC
E.dunlapae.91 QRRNRTNYSAAQLNELEMVFQRTRYPDIFTREELSLRLGIPEARIQVWFQNRRATRRRPQ
E.dunlapae.109 TKISRRNFSVNQIEGLEAVYLKHRFVKRDMRRELAKRLNLSERSVSYWFQNKRARSKDPV
E.dunlapae.113 IKRYRSSFSTDQLRRMEDTFRHRPYLSTSQVEELAGSLKLSNRQVKIWFQNRRTKLKKQV
E.dunlapae.115 QRRNRTNYSAAQLNELEMVFQRTRYPDIFTREELSLRLGIPEARIQVWFQNRRAKWRKRT
E.dunlapae.116 RKRTRTTFSSAQVYELEKKFQRSQYLSAVDRLNLASALNMSDVQVKRWFQNRRCKERHRA
E.dunlapae.123 SRRFRTKIEPHMLSHLEEMYQRKHFISCEERKELADLLGMTERAVVYWFQNRRRKDMRNY
E.dunlapae.142 RRGPRTTIKSKQLEILKDAFSSTPKPTRHIREQLAAKTGLNMRVIQVWFQNRRSKERRMK
E.dunlapae.144 VKRPRTVLSSAQRRVFKEAFDRSPRPCRKEREKLSSQTGLSVRVVQVWFQNQRAKVKKLA
E.dunlapae.170 WRRRRTVFTQEELSVLESVYSQNKFLNPDLKAEILSKVNVPGNVIVMWFQNRRAKDRSAG
E.dunlapae.185 QRRNRTNYSAAQLNELEMVFQRTRYPDIFTREELSLRLGIPEARIQVWFQNVMSPCSNEK
E.dunlapae.204 SNEQRHTFTAAQLEELEHVFNTTHYPDIFTREELAMKHNLTEARVQVWFQNRRAKFRKTE
E.dunlapae.219 PRRPRTIFSASQLLELEERFRYQKYLSTAERSCLAFTLGLSEEQVKVWFQNRRSKERRMK
E.dunlapae.225 GRPLRRRFTDDQLQGLESQFGAQKYLSVPERMELAVSLRLSETQVKTWFQNRRMKWKKQG
```

Table S24: Total genes predicted by various gene-modeling programs

	FGENESH	AUGUSTUS	HMMGene	GenomeScan
Gene models	16,367	29,359	13,948	6,443

Table S25: EVidenceModeler (EVM) weight assignments

	1st Run	2nd Run	3rd Run	Final
Transcripts	10	10	10	10
RNA-seq	5	5	5	5
FGENESH	3	5	5	5
AUGUSTUS	2	2	--	--
GenomeScan	2	--	--	--
HMMgene	1	--	--	--
RACE	--	--	--	10

Table S26: Manual comparison of EVM gene predictions with various settings vs. genes in which the structure has been determined with RACE PCR

Gene Name	Scaffold	1st Run	2nd Run	3rd run	Final
Innexin1	ML2599	√	√	√	√
Un119	ML3221	√	√	√	√

Mago	ML2679	√	√	√	√
Nova1	ML2207	√	√	√	√
Nova2	ML1054	√	√	√	√
Syn	ML3422	√	√	√	√
Cornichon	ML4816	√	Miss	Join	√
Tropomyosin	ML0772	√	√	√	√
Tubby	ML0823	√	√	√	√
Kv1/4	ML0780	Miss	Miss	Miss	√
Titinlike	ML1985	Miss	√	√	√
Neb	ML2528	Join	Join	Join	√
Eph	ML3591	√	√	√	√
Kv3	ML0252	Join	Join	Join	Join
Chibby	ML0412	√	√	√	√
DIXD	ML0866	Partial	Partial	Miss	√
PaxA	ML0693	Split	Split	Split	√
PaxB	ML0788	Join	Join	Join	√
SixB	ML0649	Partial	Partial	Partial	√
SixC	ML2235	Join	Join	√	√
SixD	ML2340	Miss	Miss	Miss	√
SixE	ML0120	Join	Join	√	Join
SixA	ML0146	Miss	Miss	Miss	√
Six	ML0179	√	√	√	√
Lhx3	ML0681	Join	Join	Join	√
Evx	ML4500	√	Split	Split	√
Ml4195	ML0547	√	√	√	√
M4121	ML0144	√	√	√	√
Msxlike	ML0001	Miss	Miss	Miss	√
MN3	ML0500	Join	Join	Join	√
DL2	ML1029	Miss	√	√	√
Mleng	ML2705	√	√	√	√
MINK1	ML0633	√	√	√	√
Pbx	ML0476	√	√	√	√
Lim/Isl	ML0530	Join	Join	√	√
Lhx1	ML1325	Join	Join	Join	Join
Smad4	ML0219	Split	Join	Join	√
Smad1/5a	ML0930	Miss	Miss	Miss	√
Smad1/5b	ML0120	Join	Join	√	√
Smad6	ML1970	√	√	√	√
Smad2/3	ML0117	Join	Join	√	√
TgfR	ML0859	√	√	√	Join
TgfR	ML0821	√	√	√	√
TgfR	ML1311	Join	Join	√	√
TgfR	ML0465	√	√	√	√
SMURF	ML2068	√	√	√	√
MlBmp	ML2188	Join	Join	√	√

MlInh	ML1022	Join	Join	√	√
MlMst	ML2002	Partial	Partial	Partial	√
MlMst2	ML2002	√	√	√	√
MITG2	ML3487	Miss	Miss	Miss	Miss
MITGF	ML0482	Join	Join	√	√
MITG3	ML1932	Partial	Join	Join	Partial
MITG4	ML3588	Miss	Miss	Miss	√
MITG5	ML3689	Miss	Miss	Join	√
Piwi4	ML0091	√	√	√	√
HINT1	ML0737	√	√	√	√
HINT2	ML3209	√	Miss	Join	Join
Opsin	ML1305	√	Join	Join	√
Piwi2	ML1497	Miss	Miss	Miss	√
Piwi	ML0742	Join	Join	Join	√
Pygopus	ML2207	√	√	√	√
Tcf	ML1122	√	√	√	√
Dishevelled	ML0053	Partial	Partial	Partial	√
Beta-catenin	ML0737	√	√	√	√
Sec. frizzled	ML2235	Join	Join	Join	Join
FrizzledB	ML0346	Partial	Partial	Partial	Join
FrizzledA	ML0032	Join	Join	√	√
Wnt9	ML1010	√	√	√	√
WntX	ML0105	Miss	√	√	√
Wnt6	ML0602	Partial	Partial	Partial	√
WntA	ML0752	Join	Join	√	√
COE	ML0447	√	√	√	√
Bar	ML0315	√	√	√	√
NKL1	ML0585	√	√	√	√
Tlx-like	ML0143	√	√	√	√
Prd1	ML0211	√	√	√	√
Prd2	ML1011	√	√	√	√
Prd3	ML0015	Join	Join	Partial	Partial
Glis	ML1458	√	√	√	√
Nanos1	ML1302	Split	Split	Split	√
Nanos2	ML2208	√	√	Split	√
Vasa	ML0471	√	√	Miss	√
Zn finger	ML0317	Join	Join	Partial	√
Opsinlike	ML1790	Partial	Partial	Partial	√
Mef2	ML1232	√	√	√	√
Syntaxin	ML3422	√	√	√	√
Rag1	ML0821	Join	Join	Join	√
Paraxis	ML0324	Join	Join	Join	Join
Atonal	ML1043	Miss	Miss	√	√
Musashi	ML0617	√	√	√	√
Opsin	ML1204	Partial	Partial	Partial	√

Table S27: EVidenceModeler (EVM) consensus gene predictions

	Genes	Exons	Scaffolds
1st Run	14537	85446	2037
2nd Run	14835	86712	1998
3rd Run	16845	89564	1915
Final	16545	91482	1748

Table S28: Species tree used for gene clustering algorithm (Newick)

(((((((((((*Homo_sapiens*,*Gallus_gallus*),*Xenopus_tropicalis*),*Danio_rerio*),*Ciona_intestinalis*),*Branchiostoma_floridae*),*Strongylocentrotus_purpuratus*),(((*Lottia_gigantea*,(*Capitella_teleta*,*Helobdella_robusta*)),*Schistosoma_mansoni*),((*Pristionchus_pacificus*,*Caenorhabditis_elegans*),((*Drosophila_melanogaster*,*Daphnia_pulex*),*Ixodes_scapularis*)))),(*Nematostella_vectensis*,*Hydra_magnipapillata*))),*Trichoplax_adhaerens*),*Amphimedon_queenslandica*),*Mnemiopsis_leidyi*);

Table S29: Pfam domains - patterns of presence-absence

Columns 2-6 indicate the presence absence pattern that defines column 1. The “TOTAL” column indicates how many Pfam domains fit this pattern. Co=*Capsaspora owczarzaki*, *Mb=Monosiga brevicolis*, *Sr=Salpingoeca rosetta*, *Ml=Mnemiopsis leidyi*, *Aq=Amphimedon queenslandica*, *Ta=Trichoplax adhaerens*, *Nv=Nematostella vectensis*, *Bf=Branchiostoma floridae*, Ce=Caenorhabditis *elegans*, *Ci=Ciona intestinalis*, *Ct=Capitella teleta*, *Dm=Drosophila melanogaster*, *Dr=Danio rerio*, *Gg=Gallus gallus*, *Hs=Homo sapiens*, *Lg=Lottia gigantea*

	Non-metazoan	Ctenophora	Porifera	Placozoa	Cnidaria	Bilateria	TOTAL
	Co,Mb,Sr	***Ml***	***Aq***	***Ta***	***Nv***	***Bf, Ce, Ci, Ct, Dm, Dr, Gg, Hs, Lg***	
Lost in Ml	√	X	√	√	√	√	190
Lost in Aq	√	√	X	√	√	√	122
Bilaterian	X	X	X	X	X	√	1786
Parahoxozoa	X	X	X	√	√	√	50
Coelenterata	X	√	X	X	√	√	56
All animals but Porifera	X	√	X	√	√	√	58
All animals but Ctenophora	X	X	√	√	√	√	62

Table S30: Domains absent from *M. leidyi* but present in *A. queenslandica, T.*

***adhaerens*, *N. vectensis*, at least one bilaterian and at least one non-metazoan**

Pfam name	Pfam ID	Pfam description
3-HAO	PF06052.5	3-hydroxyanthranilic acid dioxygenase
4HBT	PF03061.15	Thioesterase superfamily
AICARFT_IMPCHas	PF01808.11	AICARFT/IMPCHase bienzyme
AIRC	PF00731.13	AIR carboxylase
AOX	PF01786.10	Alternative oxidase
Adeno_IVa2	PF02456.8	Adenovirus IVa2 protein
Ald_Xan_dh_C	PF01315.15	Aldehyde oxidase and xanthine dehydrogenase, a/b hammerhead domain
Ald_Xan_dh_C2	PF02738.11	Molybdopterin-binding domain of aldehyde dehydrogenase
Alpha_L_fucos	PF01120.10	Alpha-L-fucosidase
Arginosuc_synth	PF00764.12	Arginosuccinate synthase
BAAT_C	PF08840.4	BAAT / Acyl-CoA thioester hydrolase C terminal
BRE	PF06113.5	Brain and reproductive organ-expressed protein (BRE)
BSMAP	PF12280.1	Brain specific membrane anchored protein
Bac_rhamnosid	PF05592.4	Bacterial alpha-L-rhamnosidase
Baculo_LEF5_C	PF11792.1	Baculoviridae late expression factor 5 C-terminal domain
Branch	PF02485.14	Core-2/I-Branching enzyme
BtpA	PF03437.8	BtpA family
CDKN3	PF05706.5	Cyclin-dependent kinase inhibitor 3 (CDKN3)
CHGN	PF05679.9	Chondroitin N-acetylgalactosaminyltransferase
COLFI	PF01410.11	Fibrillar collagen C-terminal domain
CO_deh_flav_C	PF03450.10	CO dehydrogenase flavoprotein C-terminal domain
CR6_interact	PF10147.2	Growth arrest and DNA-damage-inducible proteins-interacting protein 1
CTP_transf_3	PF02348.12	Cytidylyltransferase
CYTH	PF01928.14	CYTH domain
Catalase	PF00199.12	Catalase
Catalase-rel	PF06628.5	Catalase-related immune-responsive
Caveolin	PF01146.10	Caveolin
Chitin_bind_3	PF03067.8	Chitin binding domain
Churchill	PF06573.4	Churchill protein
Cob_adeno_trans	PF01923.11	Cobalamin adenosyltransferase
CorA	PF01544.11	CorA-like Mg2+ transporter protein
DAHP_synth_1	PF00793.13	DAHP synthetase I family
DNA_alkylation	PF08713.4	DNA alkylation repair enzyme
DNA_binding_1	PF01035.13	6-O-methylguanine DNA methyltransferase, DNA binding domain
DREV	PF05219.5	DREV methyltransferase

Pfam name	Pfam ID	Pfam description
DUF1075	PF06388.4	Protein of unknown function (DUF1075)
DUF1103	PF06513.4	Repeat of unknown function (DUF1103)
DUF1183	PF06682.5	Protein of unknown function (DUF1183)
DUF1228	PF06779.7	Protein of unknown function (DUF1228)
DUF1242	PF06842.5	Protein of unknown function (DUF1242)
DUF1258	PF06869.5	Protein of unknown function (DUF1258)
DUF143	PF02410.8	Domain of unknown function DUF143
DUF1493	PF07377.5	Protein of unknown function (DUF1493)
DUF1604	PF07713.6	Protein of unknown function (DUF1604)
DUF1647	PF07801.4	Protein of unknown function (DUF1647)
DUF1713	PF08213.4	Mitochondrial domain of unknown function (DUF1713)
DUF1736	PF08409.4	Domain of unknown function (DUF1736)
DUF1855	PF08910.3	Protein of unknown function (DUF1855)
DUF1903	PF08991.3	Domain of unknown function (DUF1903)
DUF2054	PF10218.2	Uncharacterized conserved protein (DUF2054)
DUF2201	PF09967.2	Predicted metallopeptidase (DUF2201)
DUF2209	PF09974.2	Uncharacterized protein conserved in archaea (DUF2209)
DUF2215	PF10225.2	Uncharacterized conserved protein (DUF2215)
DUF2225	PF09986.2	Uncharacterized protein conserved in bacteria (DUF2225)
DUF2315	PF10231.2	Uncharacterised conserved protein (DUF2315)
DUF2348	PF09807.2	Uncharacterized conserved protein (DUF2348)
DUF2373	PF10180.2	Uncharacterised conserved protein (DUF2373)
DUF2721	PF11026.1	Protein of unknown function (DUF2721)
DUF2962	PF11176.1	Protein of unknown function (DUF2962)
DUF3074	PF11274.1	Protein of unknown function (DUF3074)
DUF3128	PF11326.1	Protein of unknown function (DUF3128)
DUF3184	PF11380.1	Protein of unknown function (DUF3184)
DUF3377	PF11857.1	Domain of unknown function (DUF3377)
DUF3639	PF12341.1	Protein of unknown function (DUF3639)
DUF3657	PF12394.1	Protein of unknown function (DUF3657)
DUF3752	PF12572.1	Protein of unknown function (DUF3752)
DUF3754	PF12576.1	Protein of unknown function (DUF3754)
DUF543	PF04418.5	Domain of unknown function (DUF543)
DUF563	PF04577.7	Protein of unknown function (DUF563)
DUF608	PF04685.6	Protein of unknown function, DUF608
DUF803	PF05653.7	Protein of unknown function (DUF803)
DUF818	PF05677.5	Chlamydia CHLPS protein (DUF818)
DUF872	PF05915.5	Eukaryotic protein of unknown function (DUF872)

Pfam name	Pfam ID	Pfam description
DUF971	PF06155.5	Protein of unknown function (DUF971)
EBP	PF05241.5	Emopamil binding protein
EF-1_beta_acid	PF10587.2	Eukaryotic elongation factor 1 beta central acidic region
EPSP_synthase	PF00275.13	EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)
ERG2_Sigma1R	PF04622.5	ERG2 and Sigma1 receptor like protein
ETC_C1_NDUFA4	PF04800.5	ETC complex I subunit conserved region
Ependymin	PF00811.11	Ependymin
Erv26	PF04148.6	Transmembrane adaptor Erv26
Exostosin	PF03016.8	Exostosin family
FAD_binding_5	PF00941.14	FAD binding domain in molybdopterin dehydrogenase
FA_FANCE	PF11510.1	Fanconi Anaemia group E protein FANCE
Fe_dep_repress	PF01325.12	Iron dependent repressor, N-terminal DNA binding domain
Fer2_2	PF01799.13	[2Fe-2S] binding domain
Ferritin	PF00210.17	Ferritin-like domain
Fibrinogen_C	PF00147.11	Fibrinogen beta and gamma chains, C-terminal globular domain
Flavodoxin_2	PF02525.10	Flavodoxin-like fold
FliL	PF03748.7	Flagellar basal body-associated protein FliL
Folate_rec	PF03024.7	Folate receptor family
Frataxin_Cyay	PF01491.9	Frataxin-like domain
FtsH_ext	PF06480.8	FtsH Extracellular
GBA2_N	PF12215.1	beta-Glucocerebrosidase 2 N terminal
Galactosyl_T_2	PF02709.7	Galactosyltransferase
Glyco_hydro_2_N	PF02837.11	Glycosyl hydrolases family 2, sugar binding domain
Glyco_hydro_30	PF02055.9	O-Glycosyl hydrolase family 30
Glyco_transf_10	PF00852.12	Glycosyltransferase family 10 (fucosyltransferase)
Glyco_transf_54	PF04666.6	N-Acetylglucosaminyltransferase-IV (GnT-IV) conserved region
Glyco_transf_64	PF09258.3	Glycosyl transferase family 64 domain
Glycophorin_A	PF01102.11	Glycophorin A
HemN_C	PF06969.9	HemN C-terminal region
HisKA	PF00512.18	His Kinase A (phosphoacceptor) domain
IF3_C	PF00707.15	Translation initiation factor IF-3, C-terminal domain
IF3_N	PF05198.9	Translation initiation factor IF-3, N-terminal domain
IIGP	PF05049.6	Interferon-inducible GTPase (IIGP)
IRK	PF01007.13	Inward rectifier potassium channel
Interfer-bind	PF09294.3	Interferon-alpha/beta receptor, fibronectin type III
IspD	PF01128.12	Uncharacterized protein family UPF0007
Kinetochor_Ybp2	PF08568.3	Central kinetochore-associated
LMWPc	PF01451.14	Low molecular weight phosphotyrosine protein phosphatase

Pfam name	Pfam ID	Pfam description
LuxC	PF05893.7	Acyl-CoA reductase (LuxC)
MMtag	PF10159.2	Kinase phosphorylation protein
MRP-S33	PF08293.4	Mitochondrial ribosomal subunit S27
MRP-S35	PF10246.2	Mitochondrial ribosomal protein MRP-S35
MaoC_dehydratas	PF01575.12	MaoC like domain
Menin	PF05053.6	Menin
Methyltransf_5	PF01795.12	MraW methylase family
MitoNEET_N	PF10660.2	Iron-containing outer mitochondrial membrane protein N-terminus
Myb_DNA-bind_2	PF08914.4	Rap1 Myb domain
NUC194	PF08163.5	NUC194 domain
Neur_chan_memb	PF02932.9	Neurotransmitter-gated ion-channel transmembrane region
Nnf1	PF03980.7	Nnf1
Nuf2	PF03800.7	Nuf2 family
O-FucT	PF10250.2	GDP-fucose protein O-fucosyltransferase
Ocular_alb	PF02101.8	Ocular albinism type 1 protein
PAC2	PF09754.2	PAC2 family
PAF-AH_p_II	PF03403.6	isoform II
PAPA-1	PF04795.5	PAPA-1-like conserved region
PD40	PF07676.5	WD40-like Beta Propeller Repeat
PEP-utilizers_C	PF02896.11	PEP-utilising enzyme, TIM barrel domain
PNPOx_C	PF10590.2	Pyridoxine 5'-phosphate oxidase C-terminal dimerisation region
PPR	PF01535.13	PPR repeat
Peptidase_C15	PF01470.10	Pyroglutamyl peptidase
Peptidase_M19	PF01244.14	Membrane dipeptidase (Peptidase family M19)
Peptidase_M49	PF03571.8	Peptidase family M49
Peptidase_S37	PF05576.4	PS-10 peptidase S37
Peptidase_S49	PF01343.11	Peptidase family S49
Peroxin-3	PF04882.5	Peroxin-3
Phlebovirus_NSM	PF07246.4	Phlebovirus nonstructural protein NS-M
PhzC-PhzF	PF02567.9	Phenazine biosynthesis-like protein
PigN	PF04987.7	Phosphatidylinositolglycan class N (PIG-N)
Pox_A32	PF04665.5	Poxvirus A32 protein
Pox_A_type_inc	PF04508.5	Viral A-type inclusion protein repeat
Pr_beta_C	PF12465.1	Proteasome beta subunits C terminal
Pyridox_oxidase	PF01243.13	Pyridoxamine 5'-phosphate oxidase
RRN7	PF11781.1	RNA polymerase I-specific transcription initiation factor Rrn7
Rib_hydrolayse	PF02267.10	ADP-ribosyl cyclase
Ribosomal_L17	PF01196.12	Ribosomal protein L17

Pfam name	Pfam ID	Pfam description
Ribosomal_L28	PF00830.12	Ribosomal L28 family
Ribosomal_L9_N	PF01281.12	Ribosomal protein L9, N-terminal domain
Ribosomal_S21	PF01165.13	Ribosomal protein S21
Ribosomal_S6	PF01250.10	Ribosomal protein S6
SAICAR_synt	PF01259.11	SAICAR synthetase
SE	PF08491.3	Squalene epoxidase
SKI	PF01202.15	Shikimate kinase
SRA1	PF07304.4	Steroid receptor RNA activator (SRA1)
SUV3_C	PF12513.1	Mitochondrial degradasome RNA helicase subunit C terminal
Scramblase	PF03803.8	Scramblase
Sec2p	PF06428.4	GDP/GTP exchange factor Sec2p
Sec39	PF08314.4	Secretory pathway protein Sec39
SecA_DEAD	PF07517.7	SecA DEAD-like domain
Seipin	PF06775.7	Putative adipose-regulatory protein (Seipin)
Selenoprotein_S	PF06936.4	Selenoprotein S (SelS)
Sigma70_ner	PF04546.6	Sigma-70, non-essential region
TMP-TENI	PF02581.10	Thiamine monophosphate synthase/TENI
TPPII	PF12580.1	Tripeptidyl peptidase II
TPX2	PF06886.4	Targeting protein for Xklp2 (TPX2)
TRP	PF06011.5	Transient receptor potential (TRP) ion channel
TYW3	PF02676.7	Methyltransferase TYW3
Tmemb_14	PF03647.6	Transmembrane proteins 14C
Tmemb_40	PF10160.2	Predicted membrane protein
Translin	PF01997.9	Translin family
TrmB	PF01978.12	Sugar-specific transcriptional regulator TrmB
Trp_halogenase	PF04820.7	Tryptophan halogenase
Trp_syntA	PF00290.13	Tryptophan synthase alpha chain
UCR_TM	PF02921.7	Ubiquinol cytochrome reductase transmembrane region
UDPG_MGDP_dh	PF00984.12	UDP-glucose/GDP-mannose dehydrogenase family, central domain
UDPG_MGDP_dh_C	PF03720.8	UDP-glucose/GDP-mannose dehydrogenase family, UDP binding domain
UPF0041	PF03650.6	Uncharacterised protein family (UPF0041)
UPF0054	PF02130.10	Uncharacterized protein family UPF0054
VPS11_C	PF12451.1	Vacuolar protein sorting protein 11 C terminal
Vitellogenin_N	PF01347.15	Lipoprotein amino terminal region
Vps55	PF04133.7	Vacuolar protein sorting 55
Wyosine_form	PF08608.5	Wyosine base formation
XLF	PF09302.4	XLF (XRCC4-like factor)
Xylo_C	PF12529.1	Xylosyltransferase C terminal

Pfam name	Pfam ID	Pfam description
dCMP_cyt_deam_2	PF08211.4	Cytidine and deoxycytidylate deaminase zinc-binding region
tRNA_Me_trans	PF03054.9	tRNA methyl transferase
zf-C4H2	PF10146.2	Zinc finger-containing protein

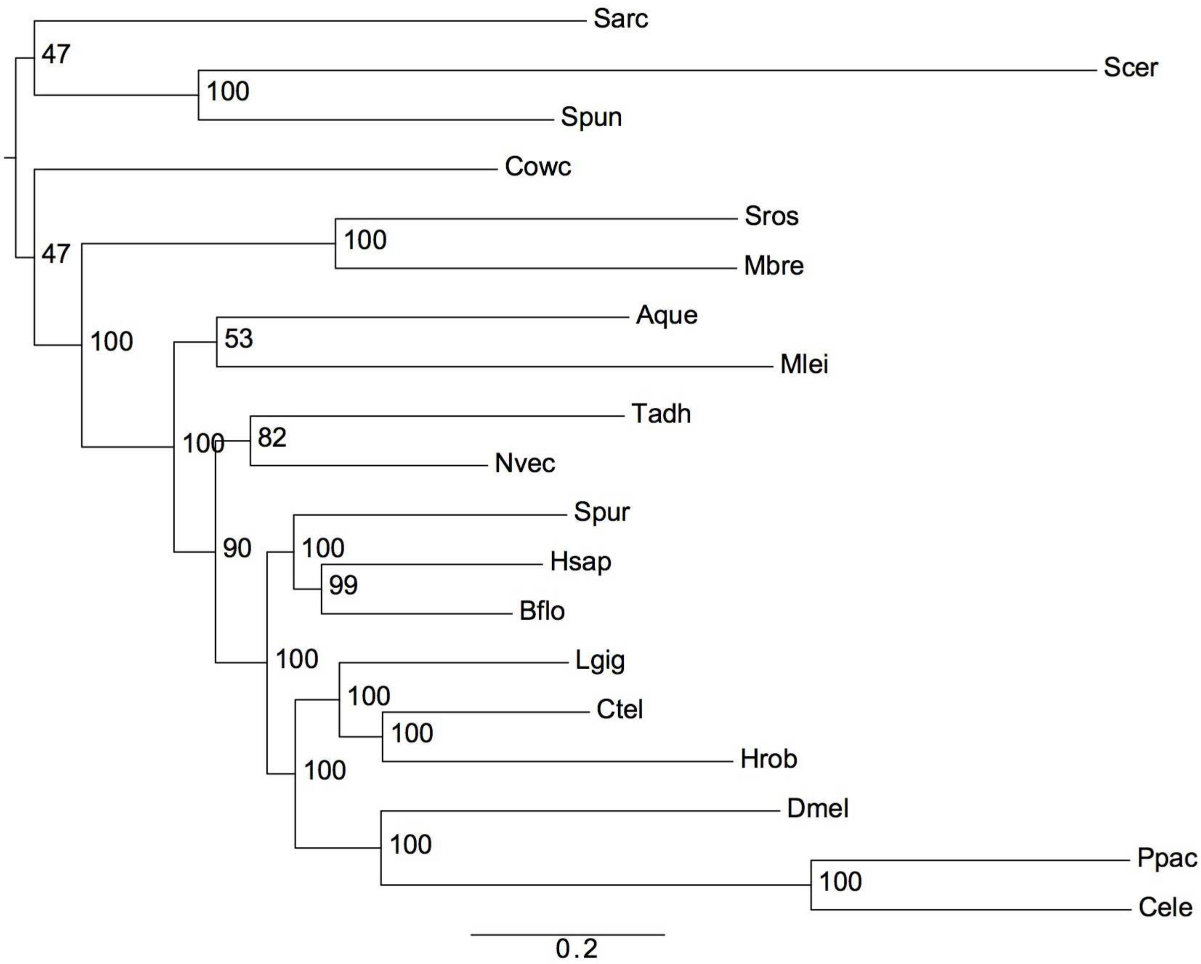
Table S31: PhyloBayes stats

We ran two instances of PhyloBayes for each data matrix (genome and EST) with varied outgroups (Opisthokonta, Holozoa, Choanimalia, Animalia). We used the burn-in below with bpcomp, which produced the reported maxdiff values. EST trees did not converge.

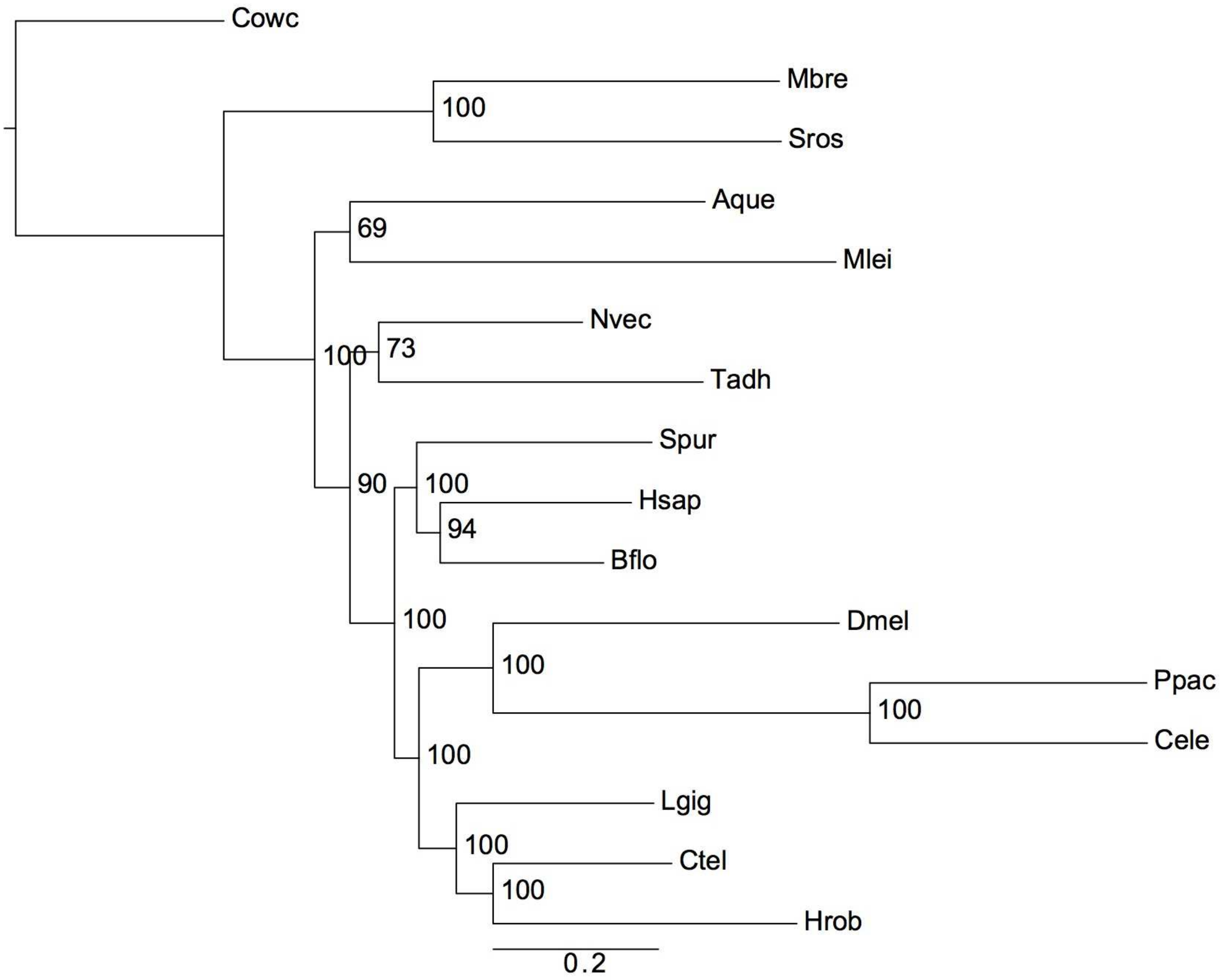
Dataset	Total cycles.01	Total cycles.02	Burn-in	Maxdiff
Genome.Opisthokonta	7964	8504	2500	0.05
Genome.Holozoa	12190	12895	4000	0
Genome.Choanimalia	8653	8643	2500	0.01
Genome.Animalia	14409	14451	4500	0
EST.Opisthokonta	10321	11814	3440	1
EST.Holozoa	12569	11124	3708	0.749153
EST.Choanimalia	15592	15747	5197	0.978199
EST.Animalia	9998	15801	3332	0.866308

Figure S1: Phylogenetic results of concatenated amino acid analyses

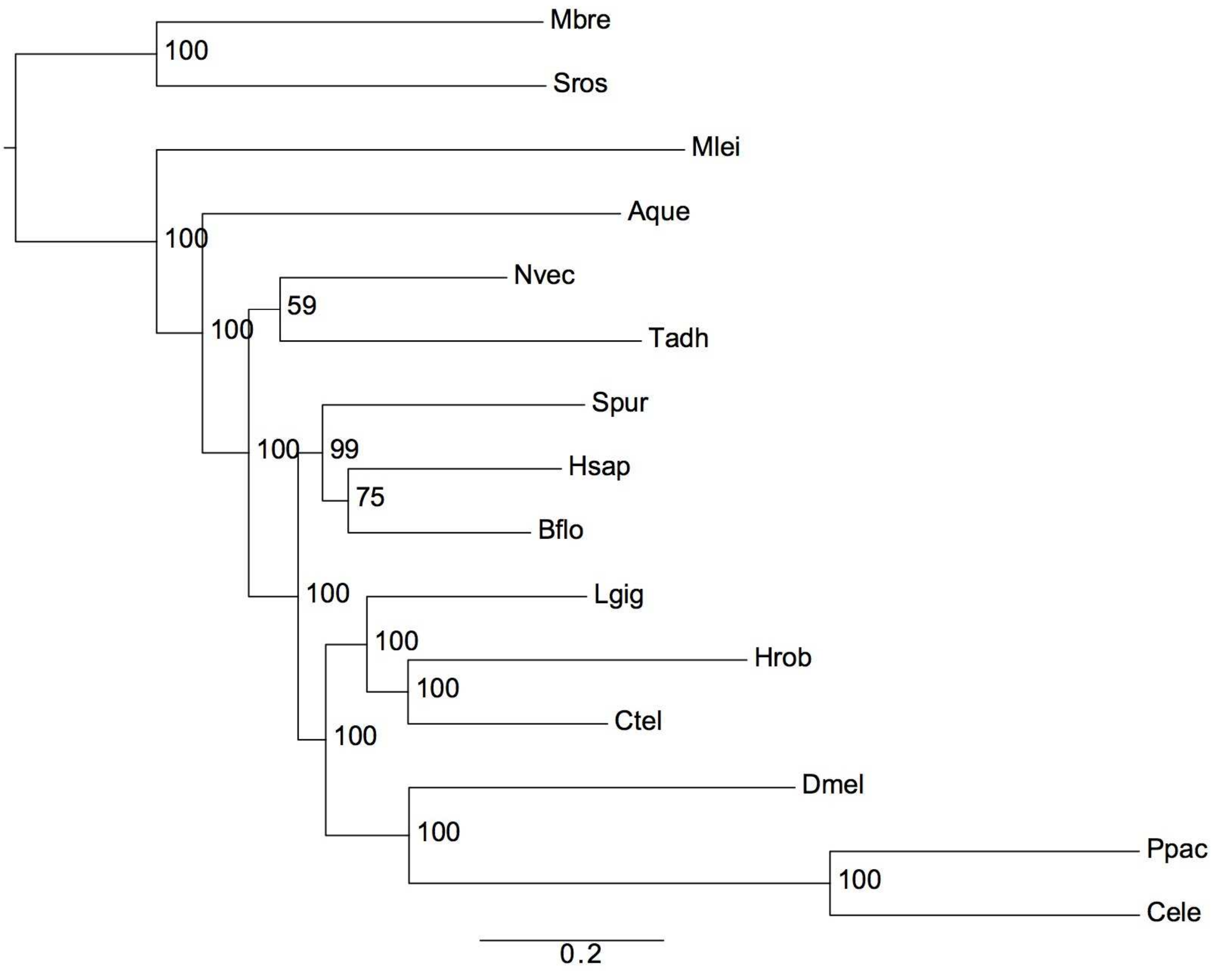
a) RaxML Genome.Opisthokonta dataset



b) RaxML Genome.Holozoa dataset

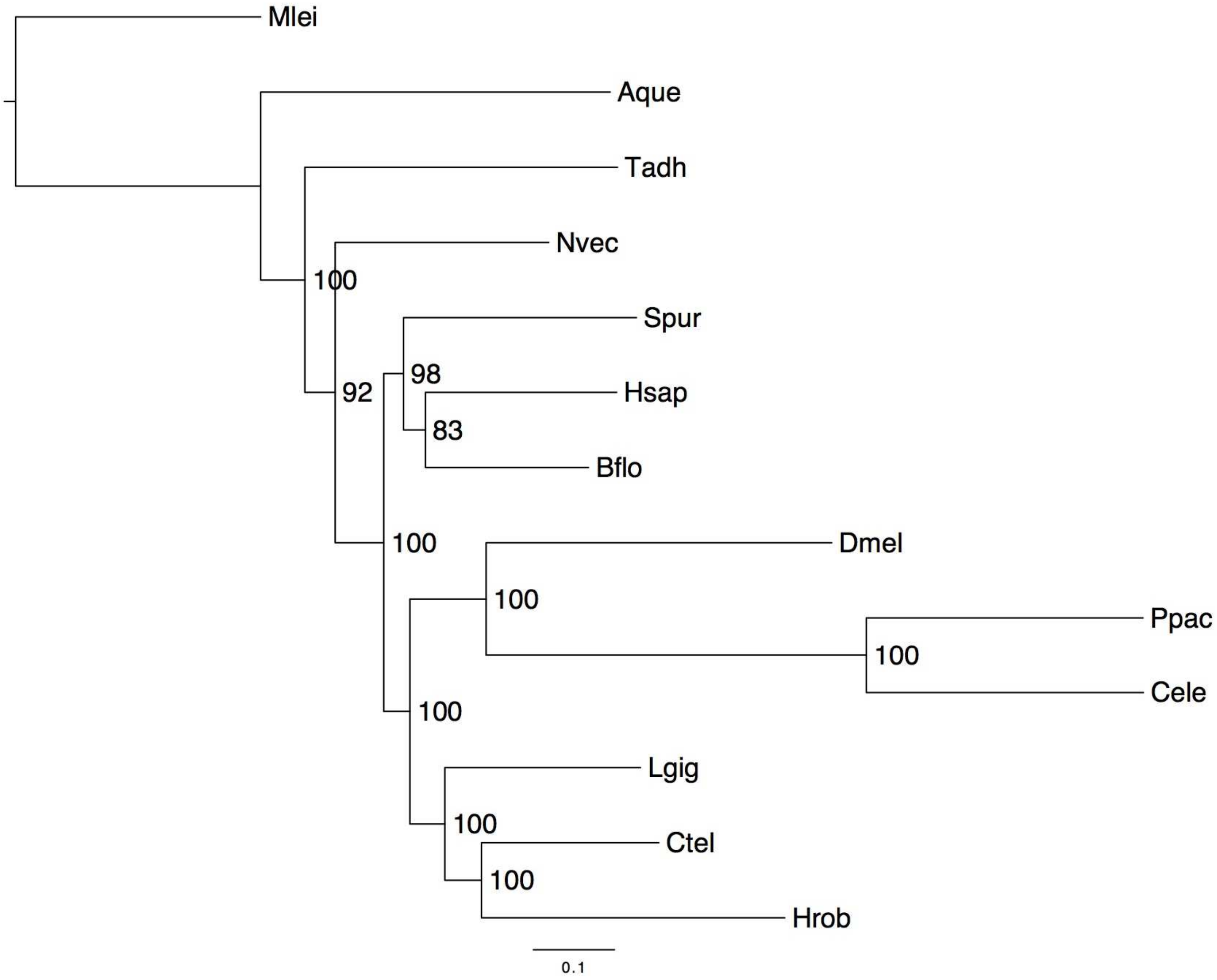


1255 **c) RaxML Genome.Choanimalia dataset**

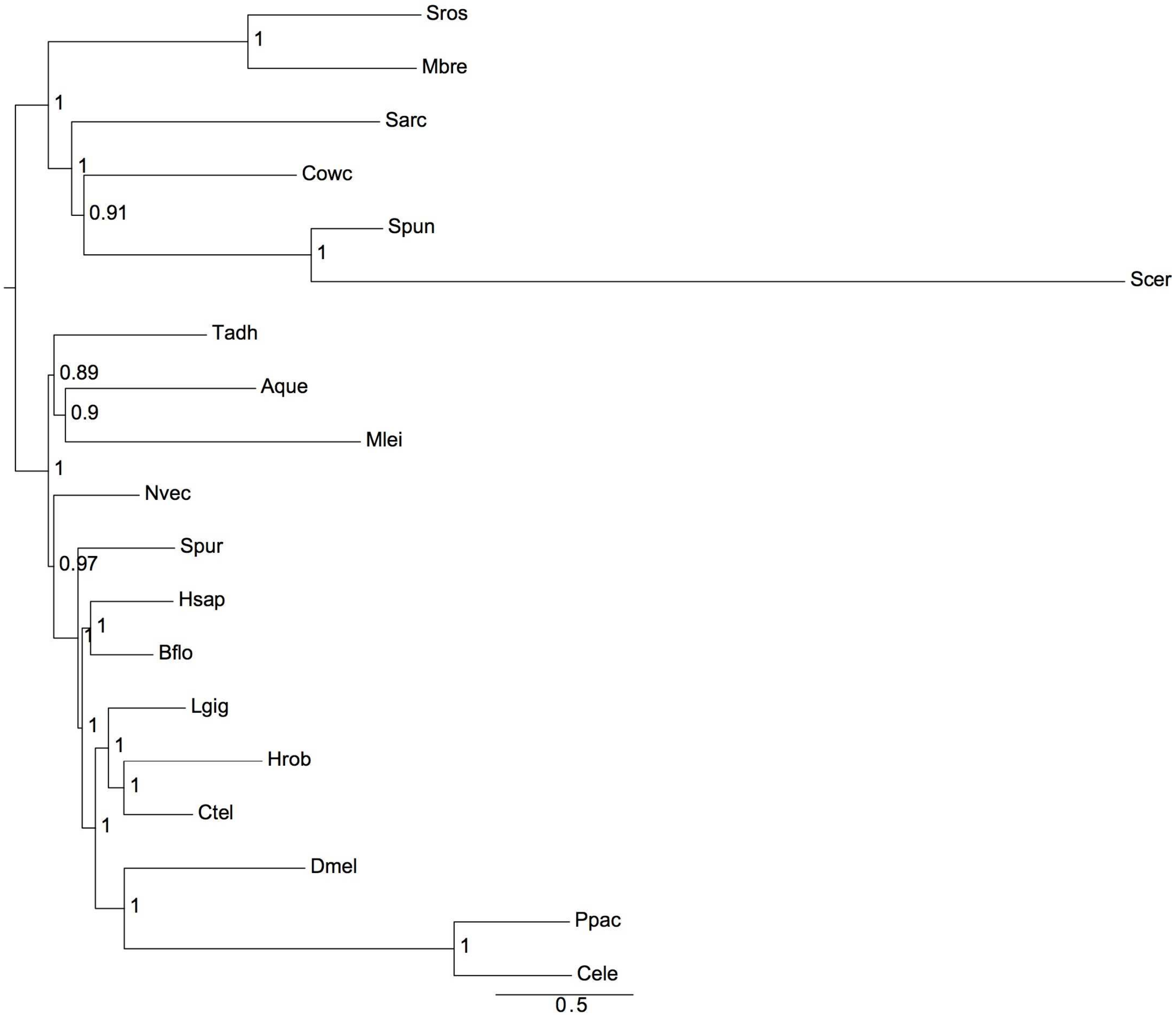


1256

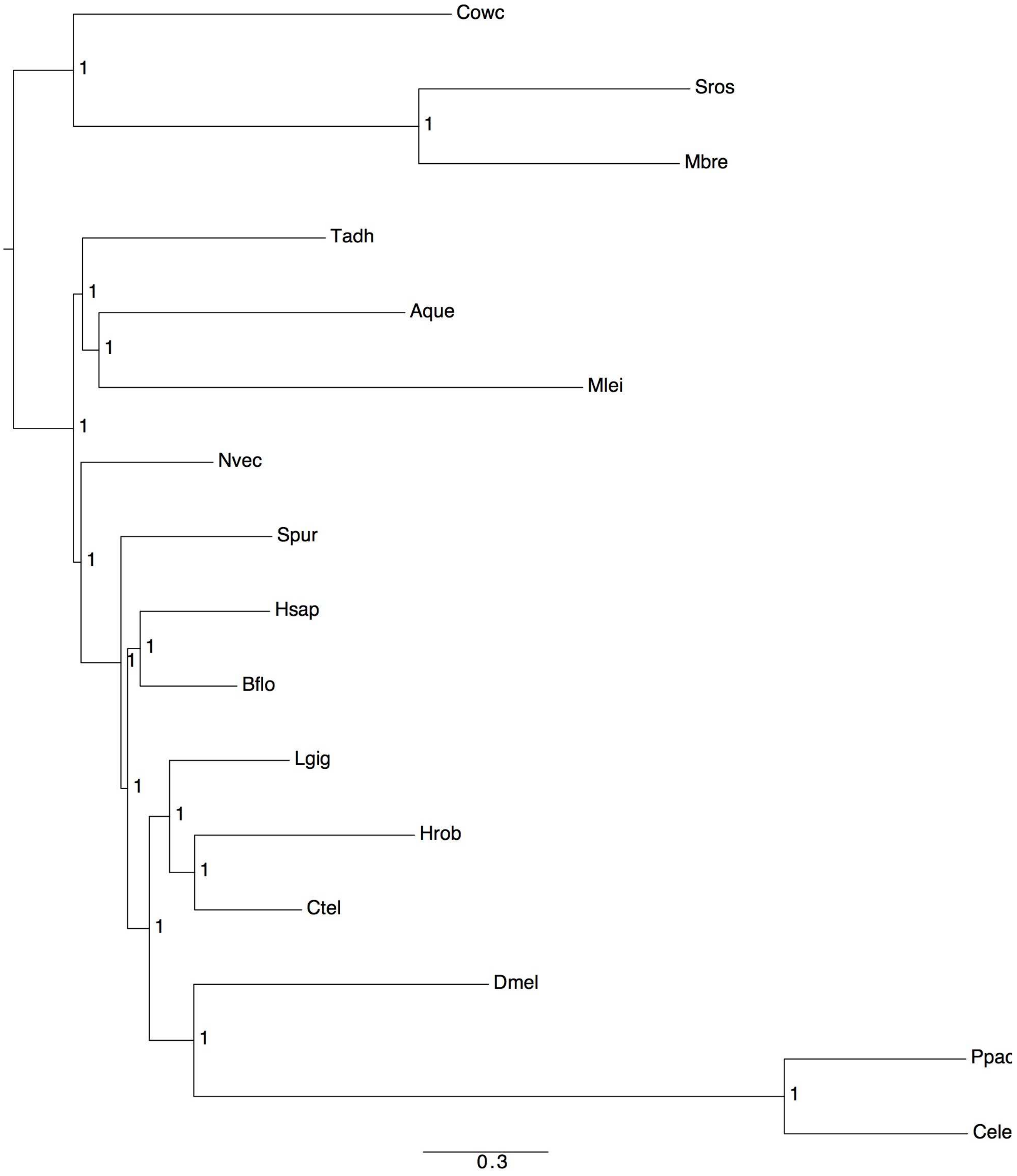
d) RaxML Genome.Animalia dataset



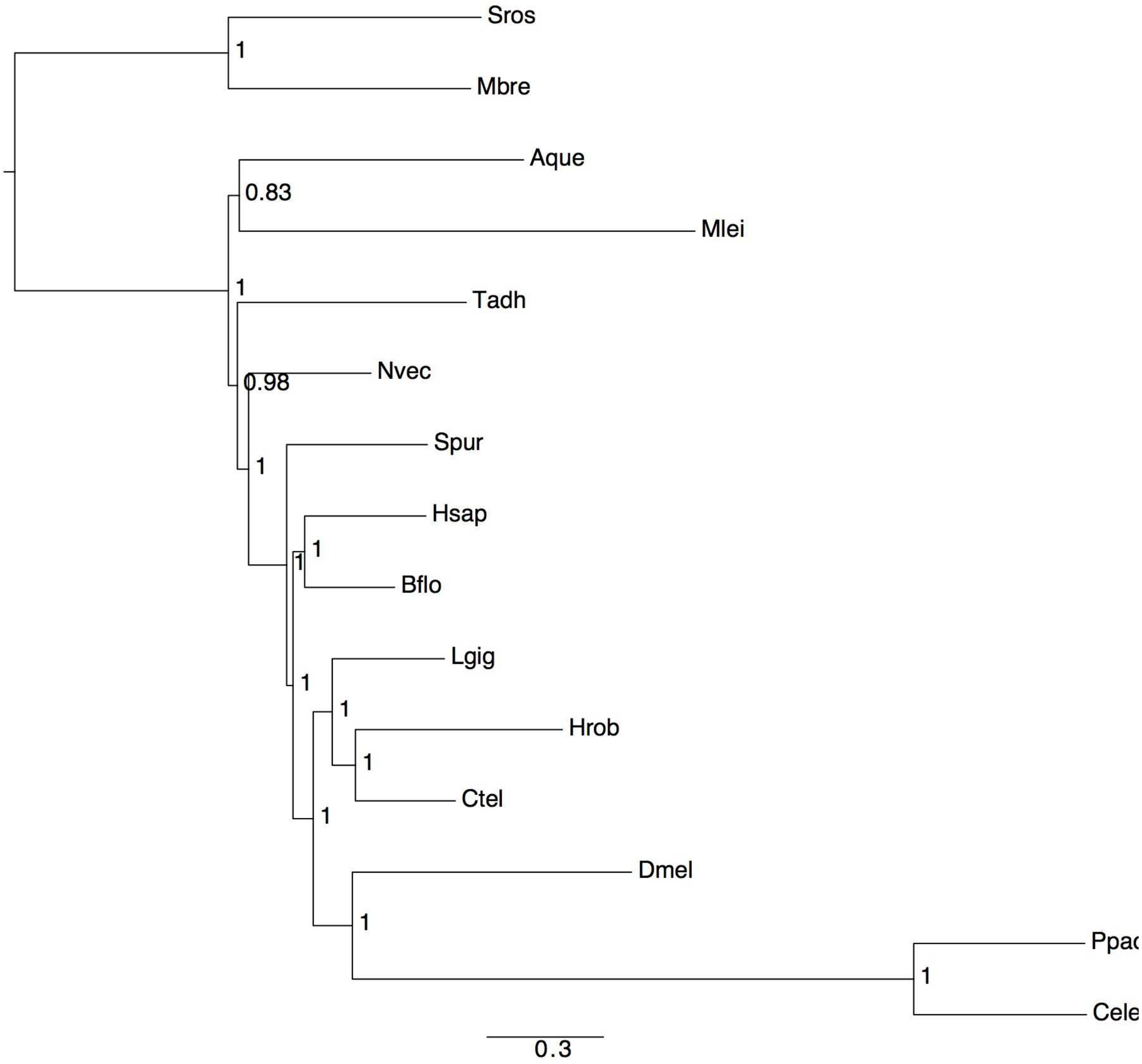
e) PhyloBayes Genome.Opisthokonta dataset



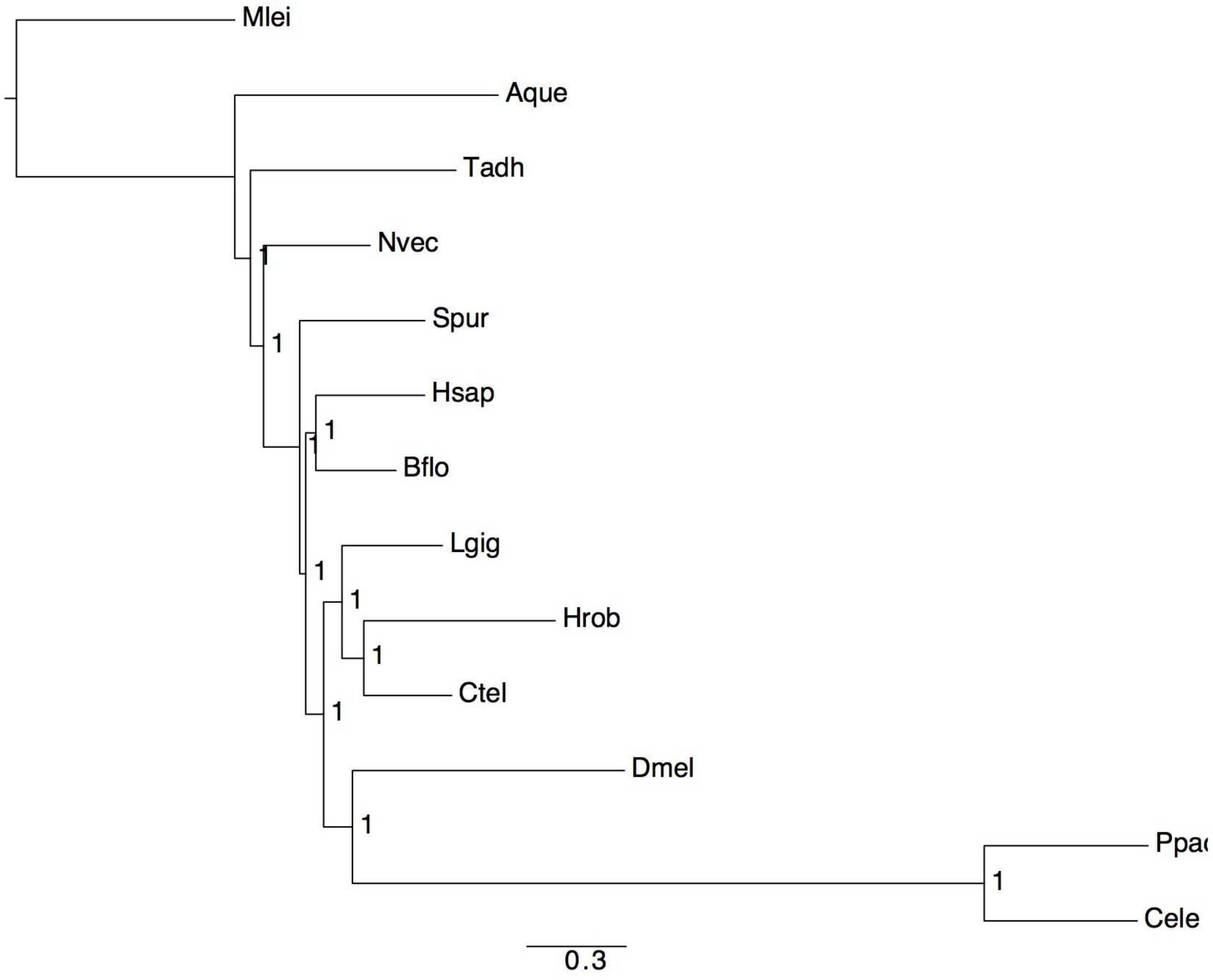
f) PhyloBayes Genome.Holozoa dataset



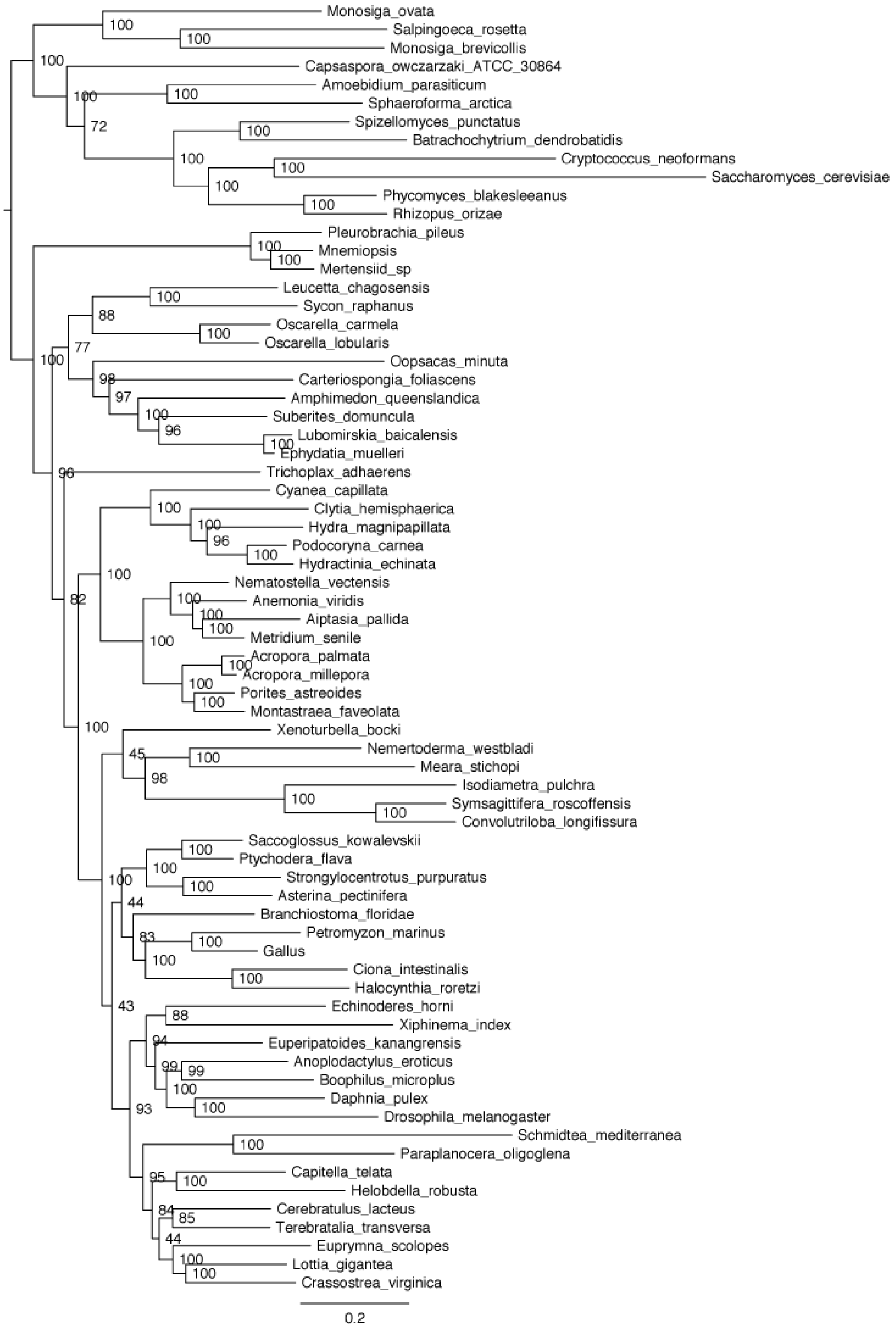
g) PhyloBayes Genome.Choanimalia dataset



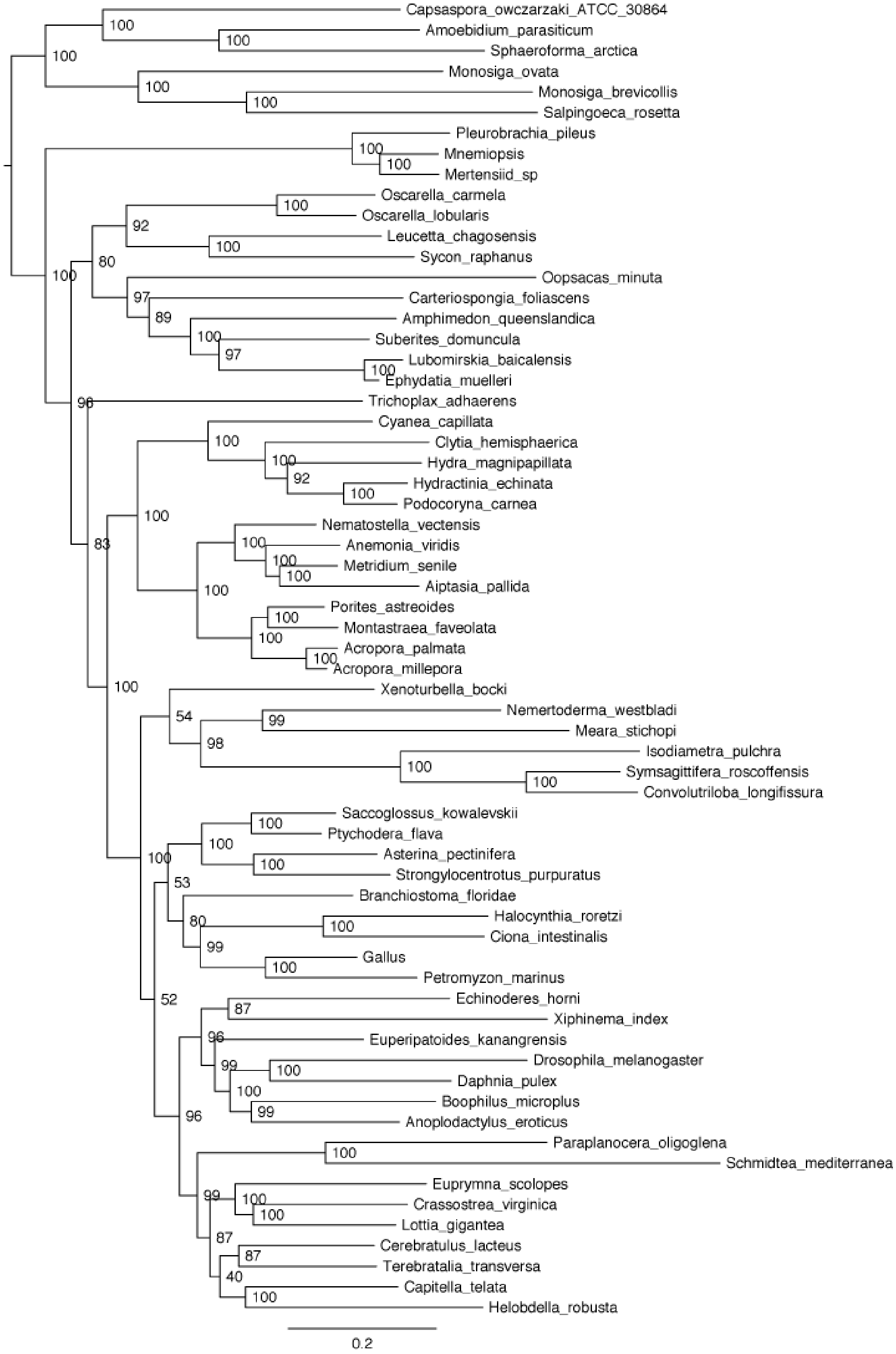
h) PhyloBayes Genome.Animalia dataset



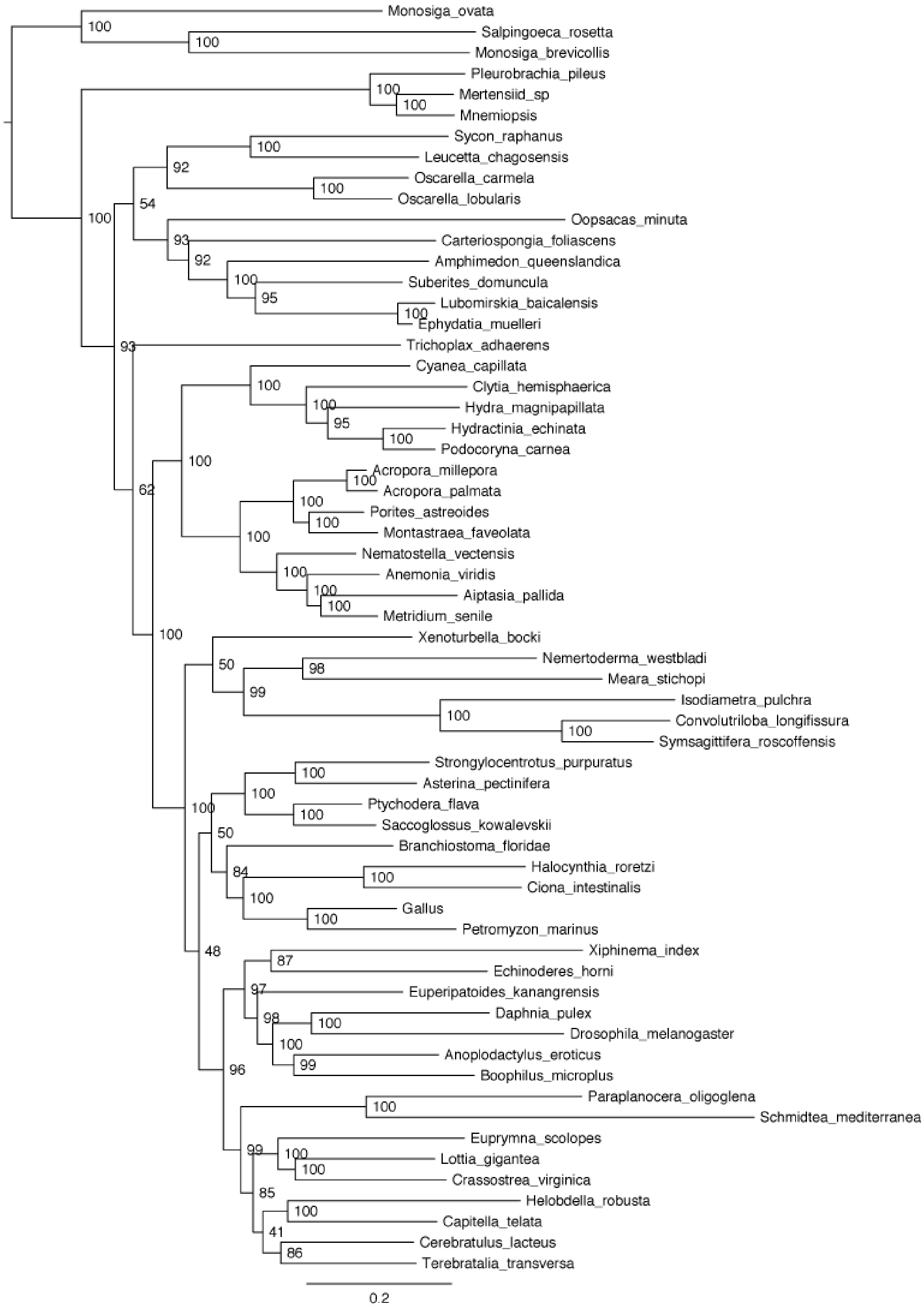
i) RaxML EST.Opisthokonta dataset



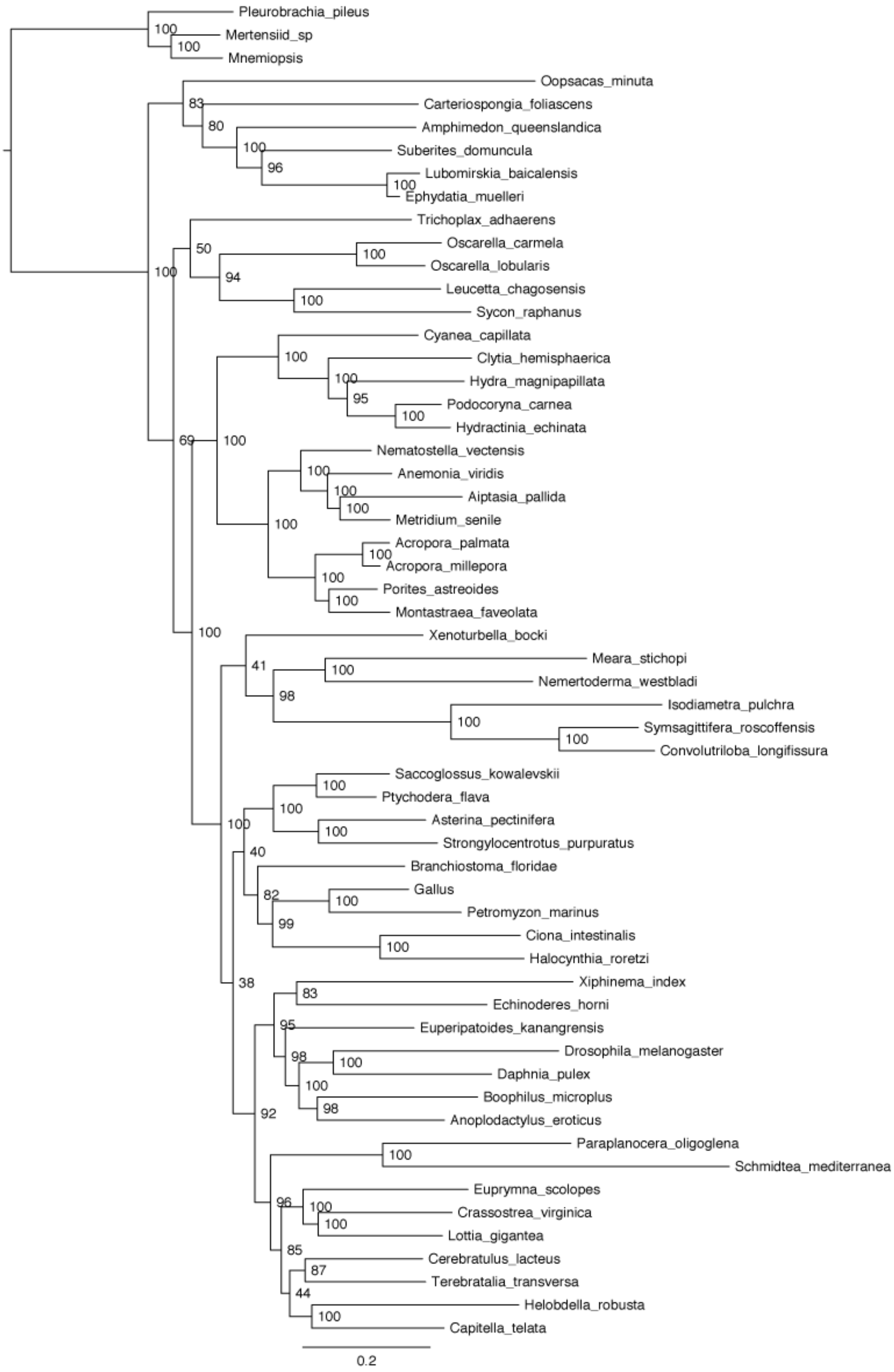
j) RaxML EST.Holozoa dataset



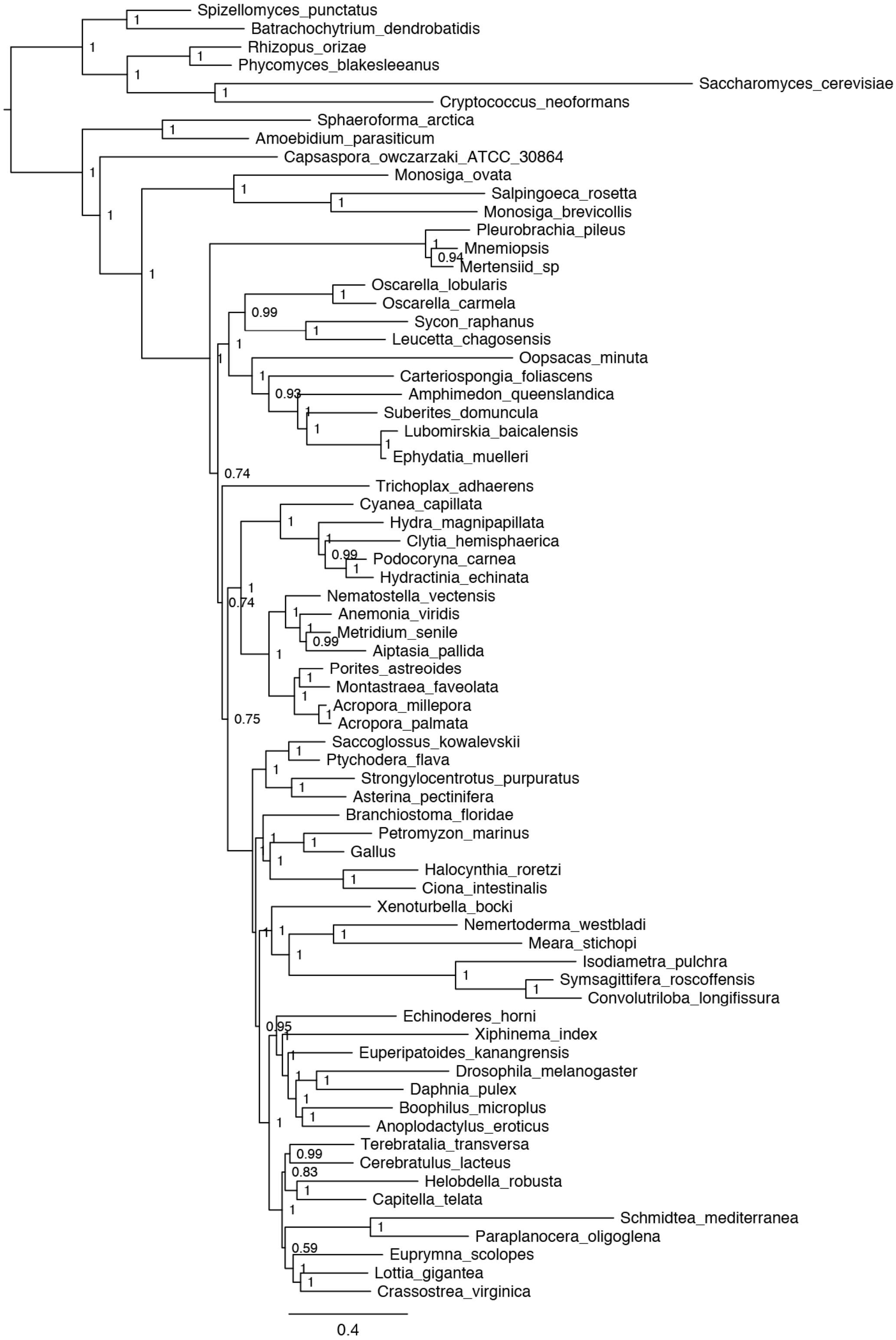
k) RaxML EST.Choanimalia dataset



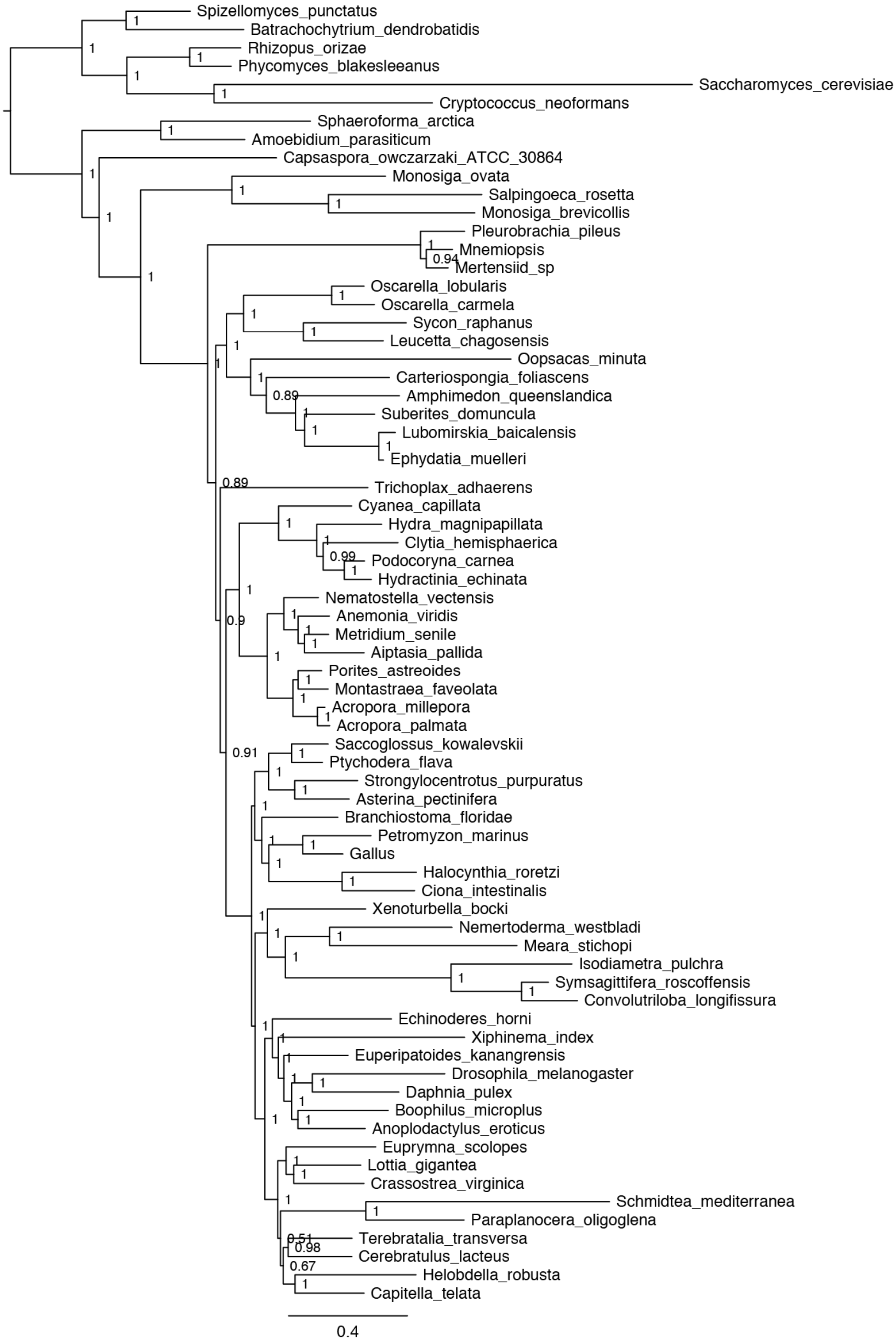
l) RaxML EST.Animalia dataset



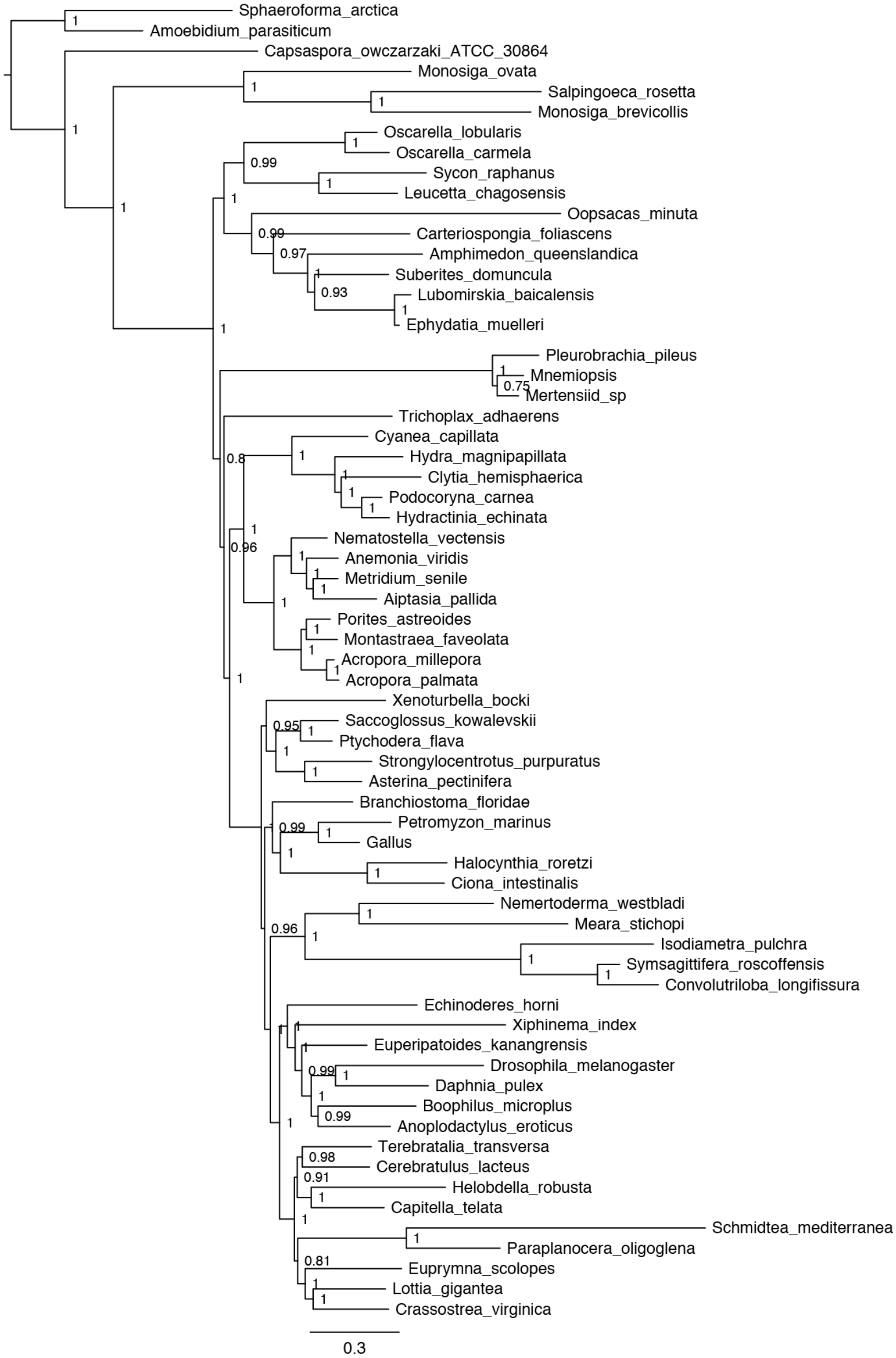
m) PhyloBayes EST.Opisthokonta dataset run 1



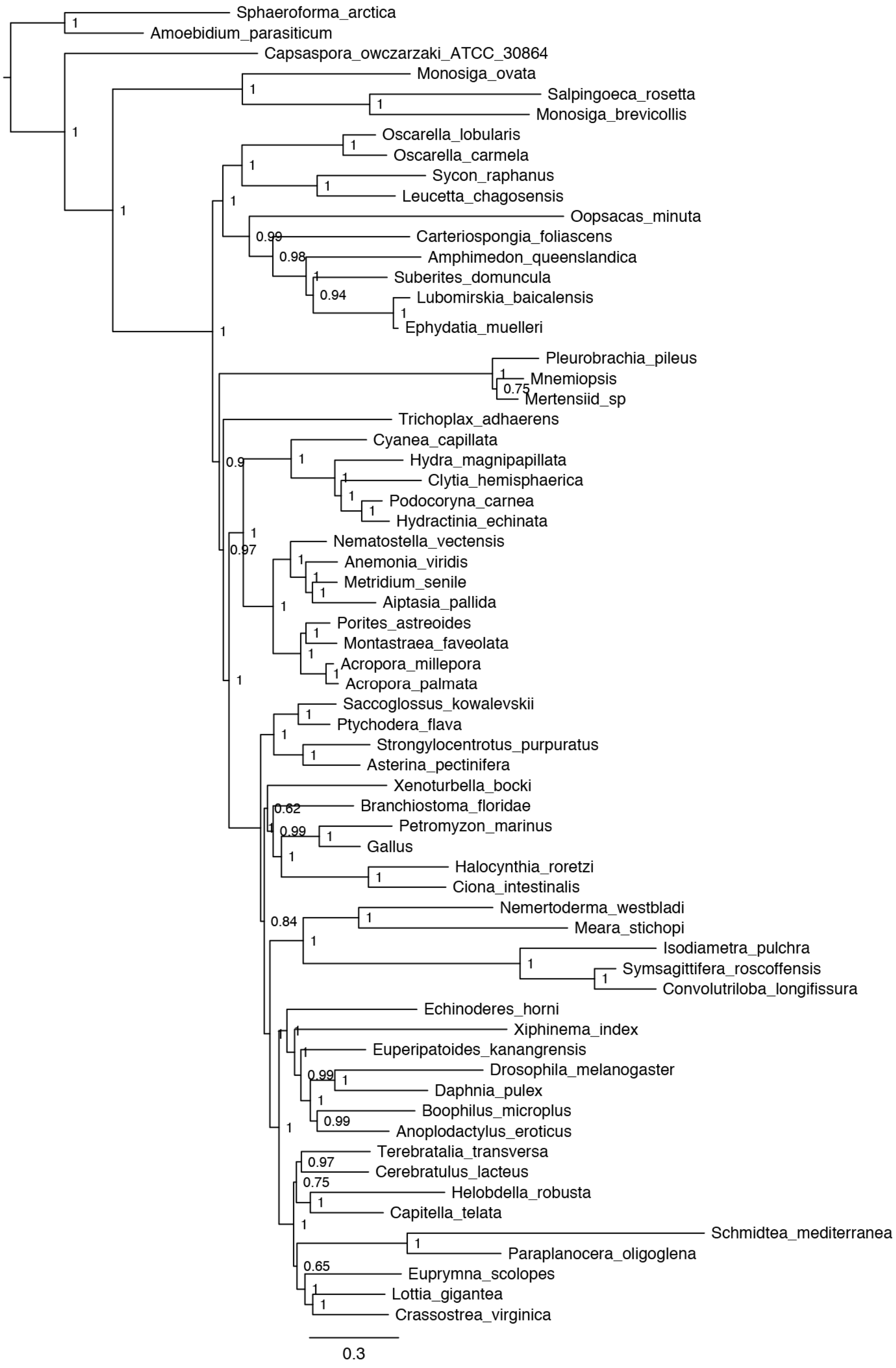
n) PhyloBayes EST.Opisthokonta dataset run 2



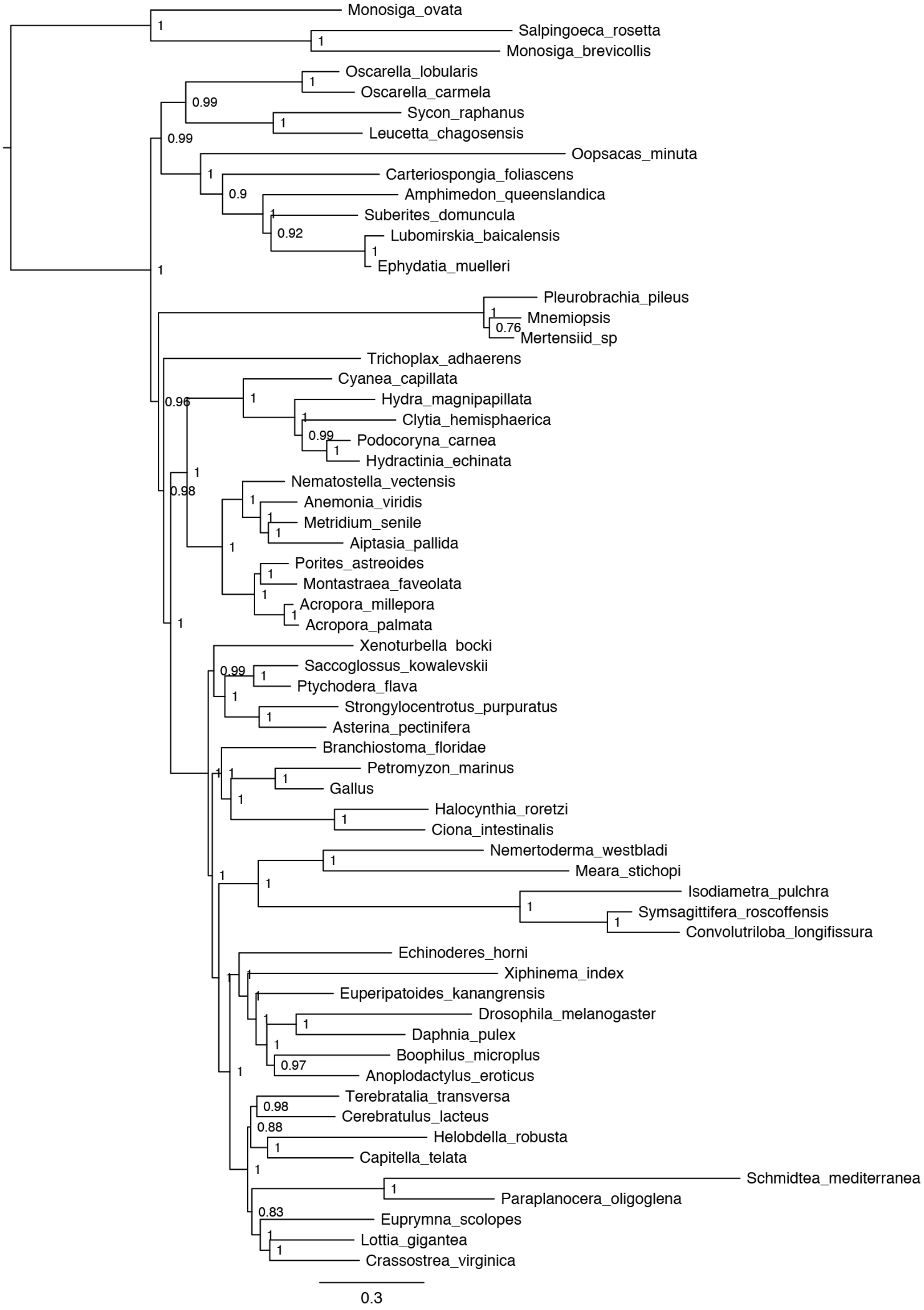
o) PhyloBayes EST.Holozoa dataset run 1



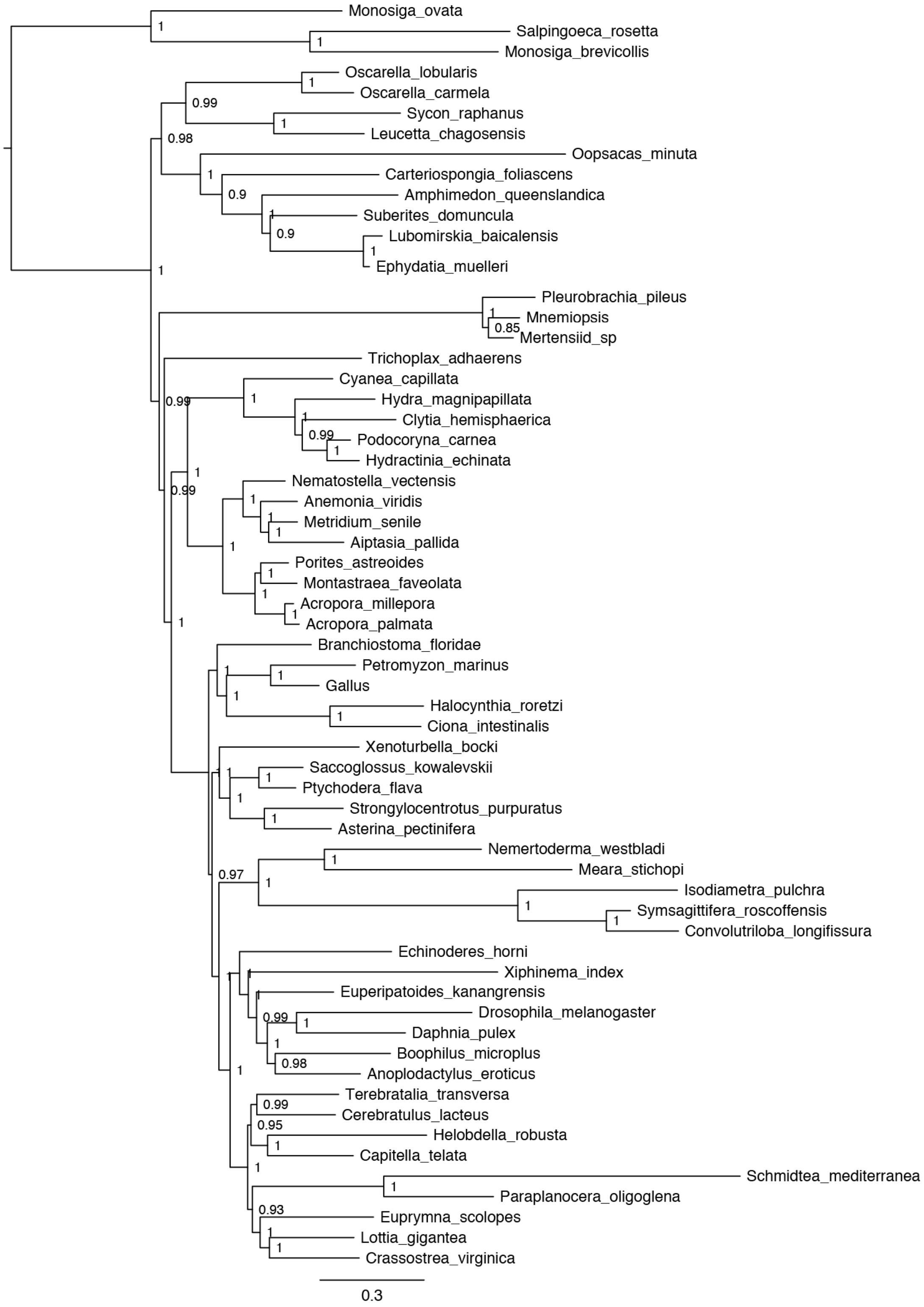
p) PhyloBayes EST.Holozoa dataset run 2



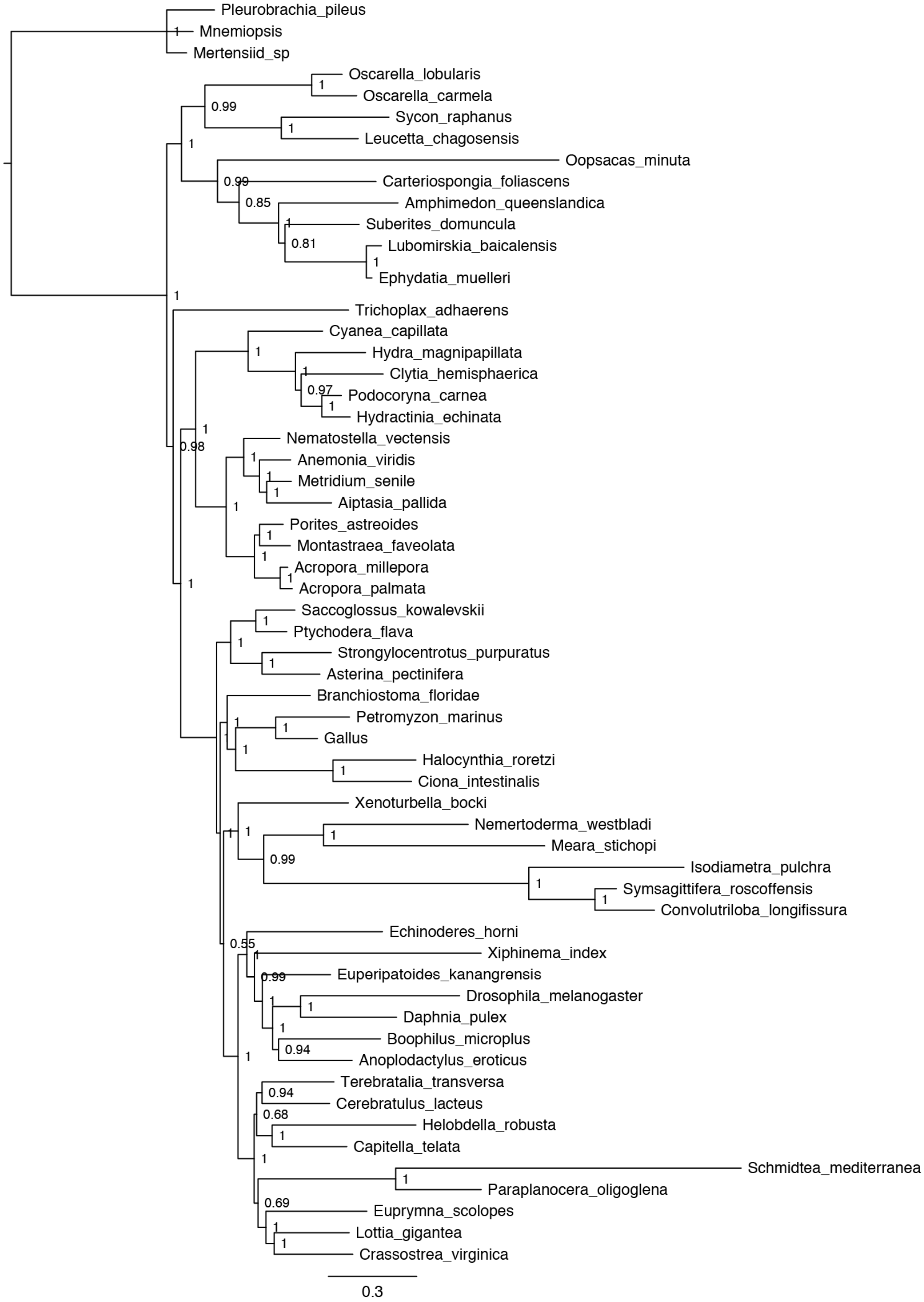
q) PhyloBayes EST.Choanimalia dataset run 1



r) PhyloBayes EST.Choanimalia dataset run 2



s) PhyloBayes EST.Animalia dataset run 1



t) PhyloBayes EST.Animalia dataset run 2

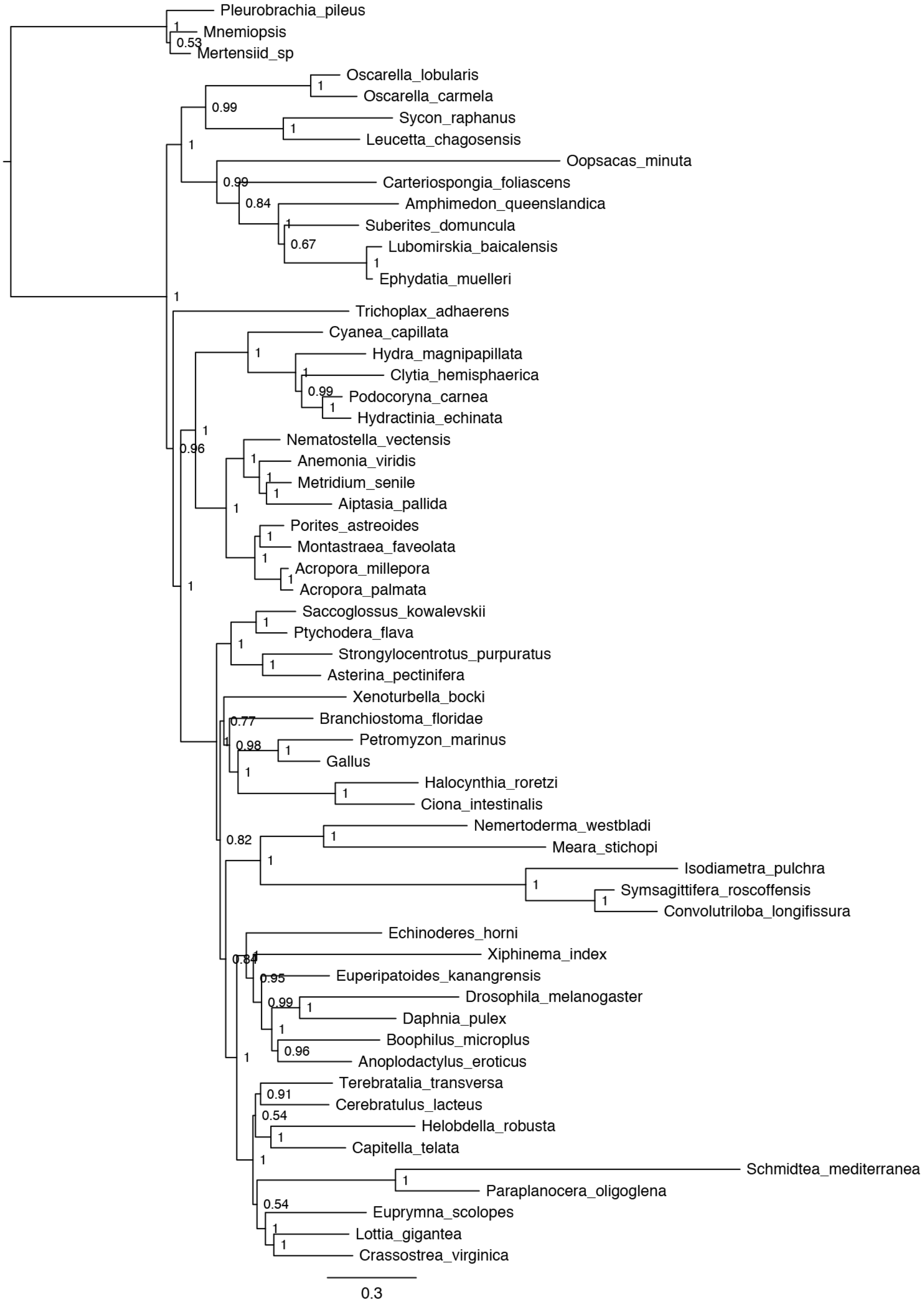


Figure S2: Maximum likelihood tree of an near intron pair matrix

A Hexagon at a node indicates the node was constrained in the analysis.

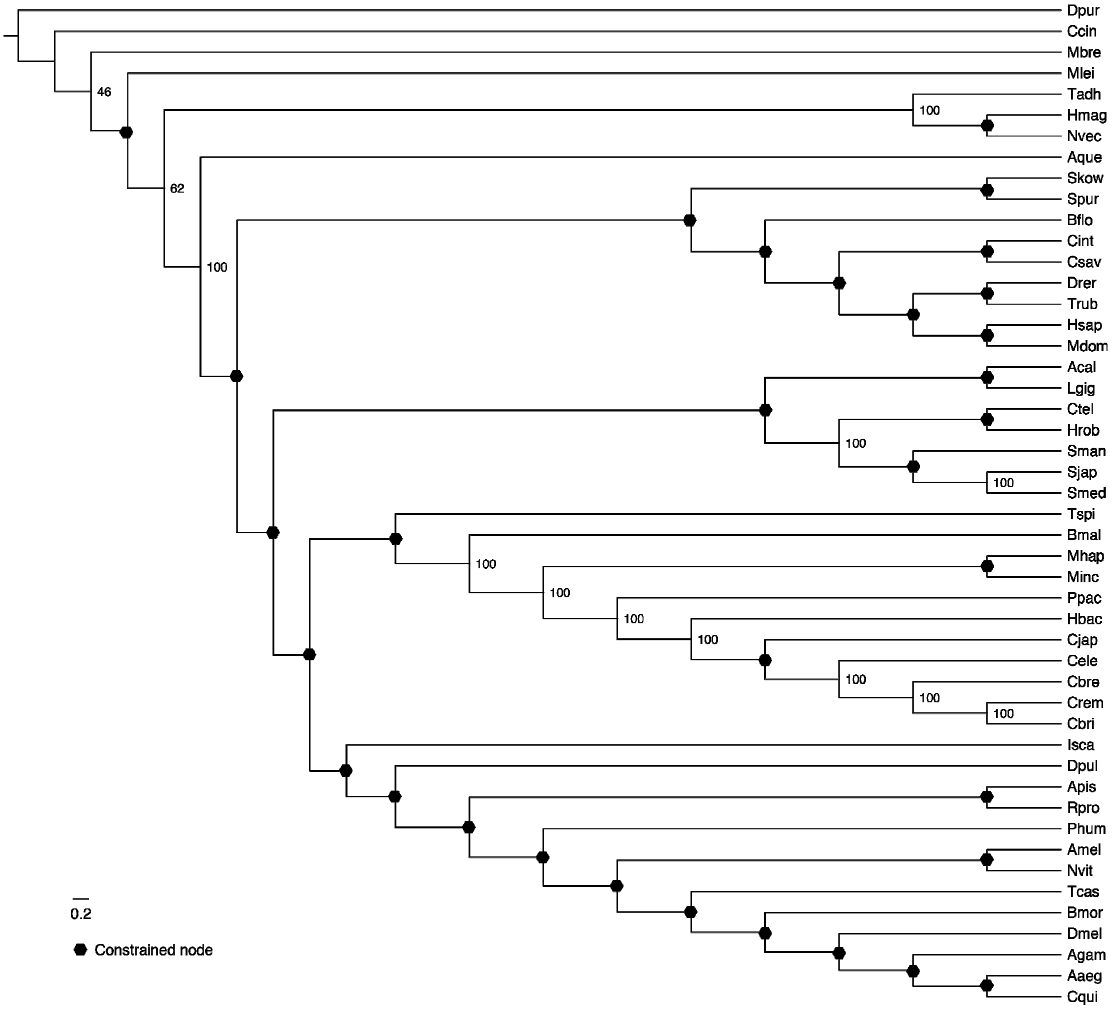
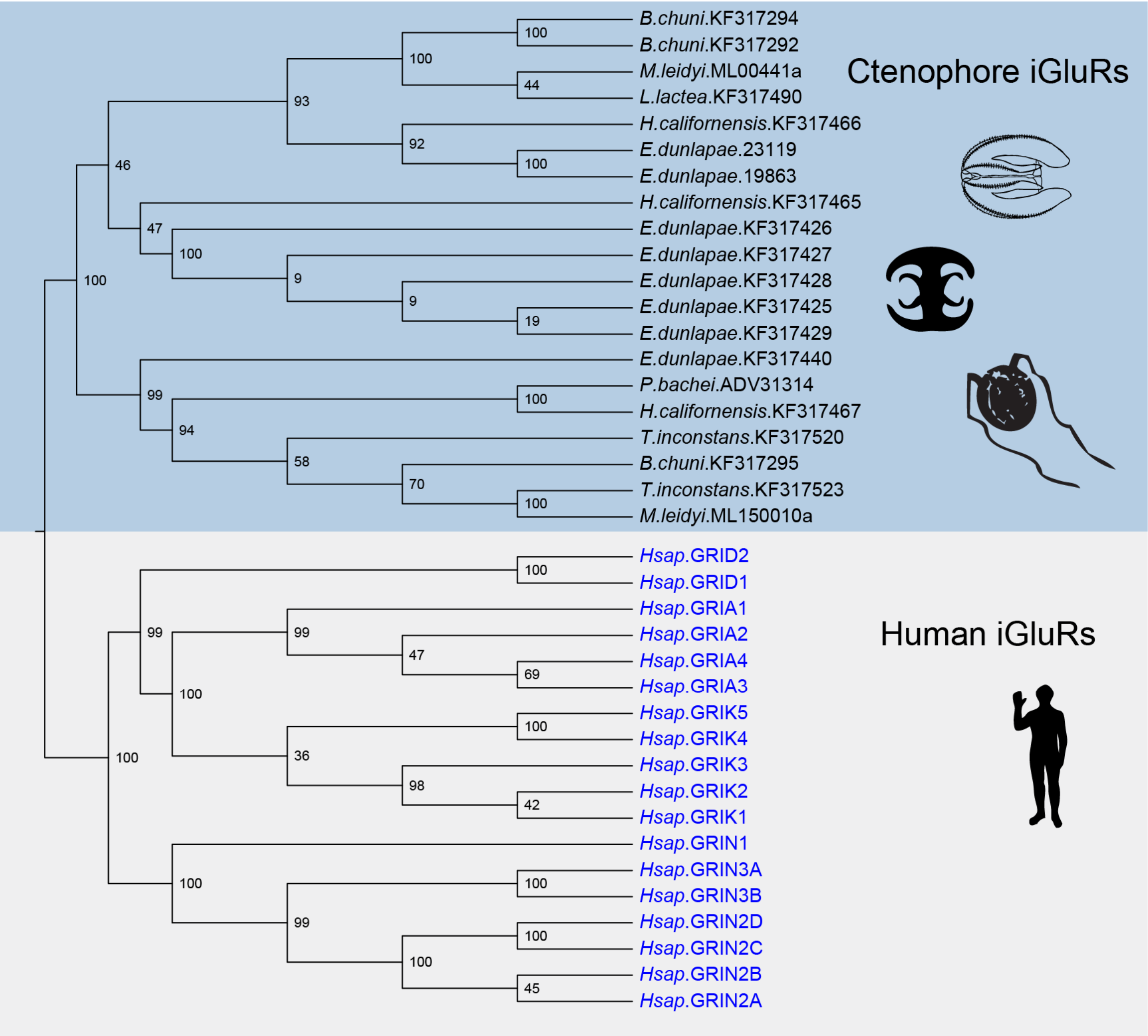
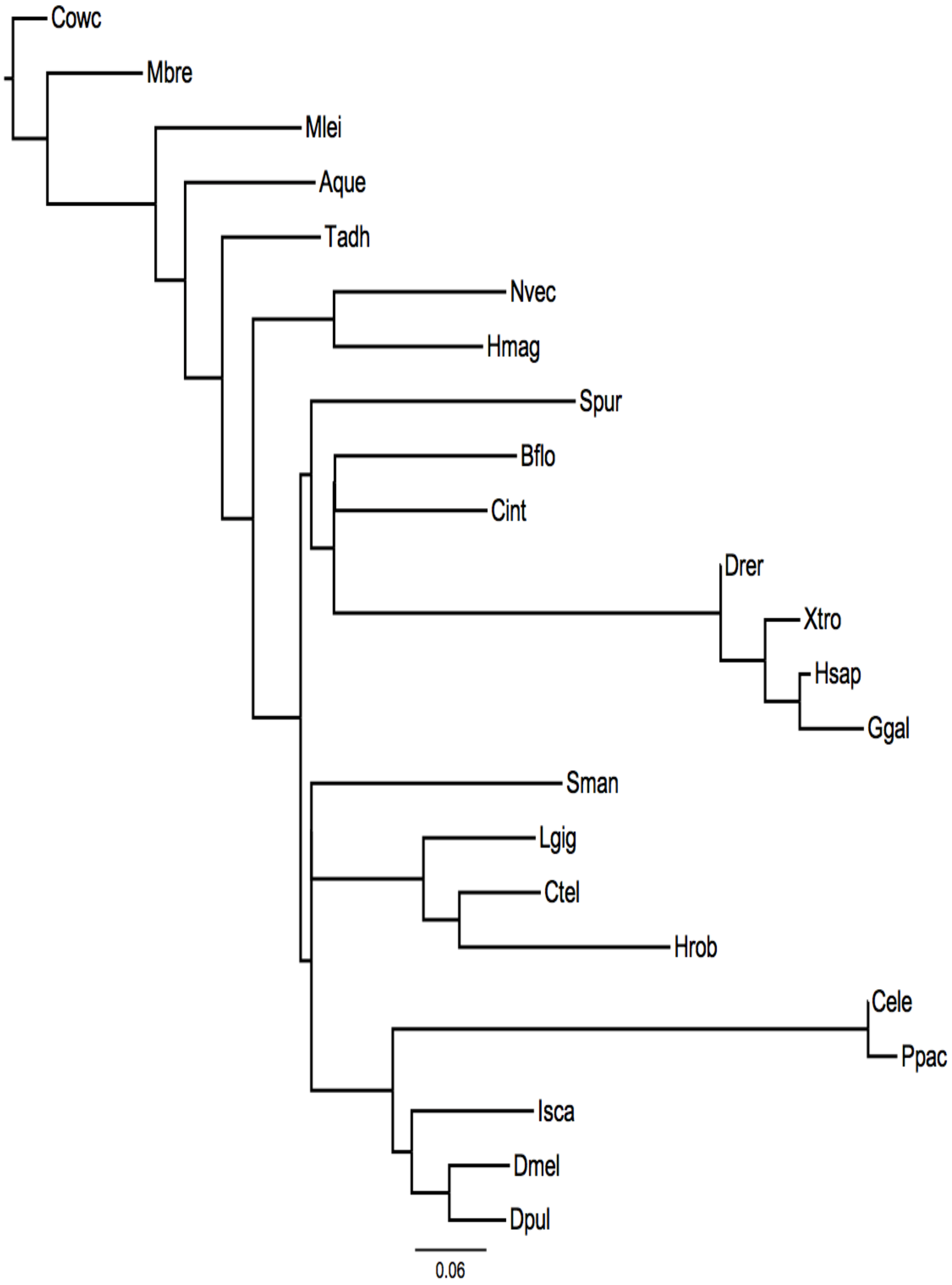


Figure S3: Ionotropic glutamate receptor phylogeny of human and ctenophore sequences from 8 ctenophores

The tree demonstrates that the ionotropic glutamate receptors (iGluR) of ctenophores are not direct orthologs to AMPA (GRIA), NMDA (GRIN), kainate-type (GRIK), or delta2-like (GRID) glutamate receptors. NOTE: The *Pleurobrachia bachei* sequence (ADV31314) is erroneously labeled as a kainate-type in GenBank.



1338 **Figure S4: Maximum-likelihood analysis of gene content data with known**
1339 **relationships constrained**
1340



1341

Figure S5: Example of bona fide nested genes
A screenshot from the *Mnemiopsis* Genome Browser shows ML000127a and ML000128a nested within introns of ML000126a.

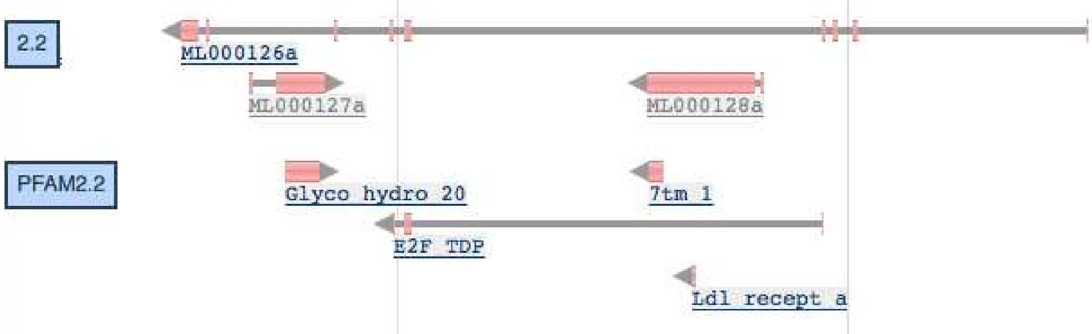


Figure S6: Example of likely spurious prediction of a nested gene

A screenshot from the *Mnemiopsis* Genome Browser shows ML000317a nested within an intron of ML000316a. Additional transcriptomic evidence that was not included in the initial gene prediction pipeline suggests that ML000317a is actually an exon incorporated in a rare isoform of ML000316a.

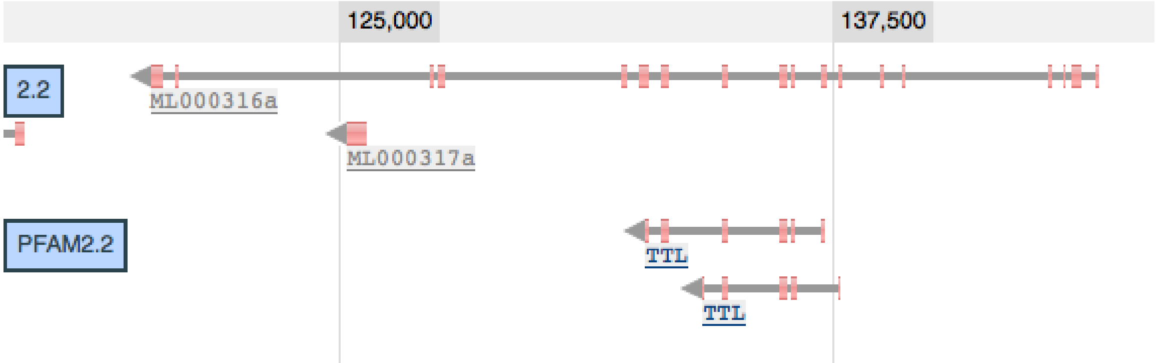


Figure S7: Comparison of lineage-specific genes

For each species, we count the occurrence of clusters that only include genes from that species (blue bars). We also count the number of genes within those clusters (red bars).

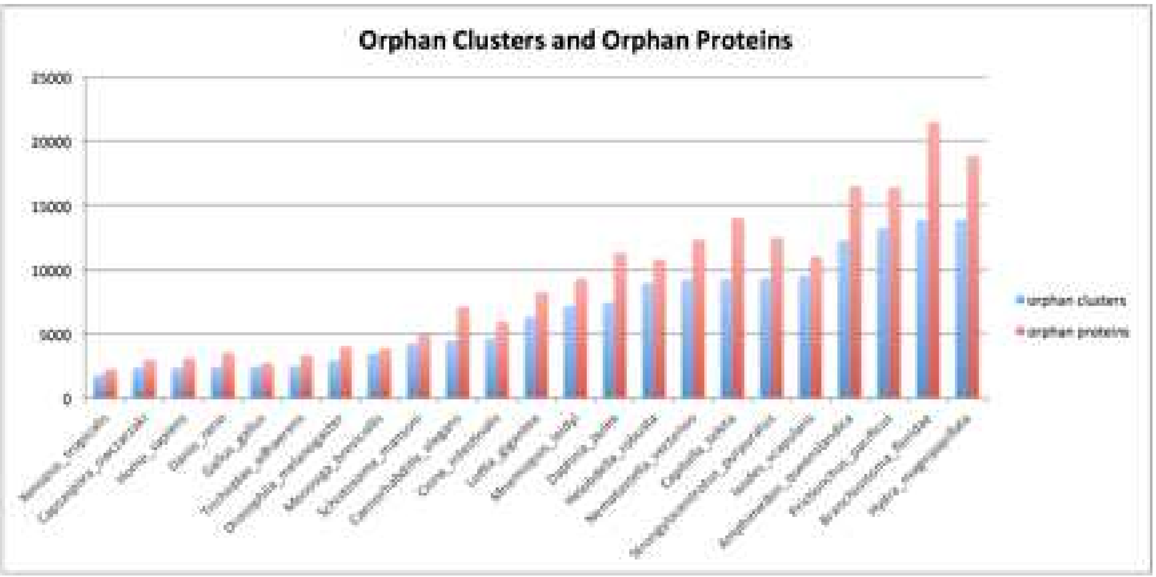


Figure S8: Comparison of gene duplications

For each cluster thought to represent a single gene present in filozoan and/or choanimalian ancestors where a species contains more than one gene, we count the number of genes in this cluster minus one and total this among all these clusters.

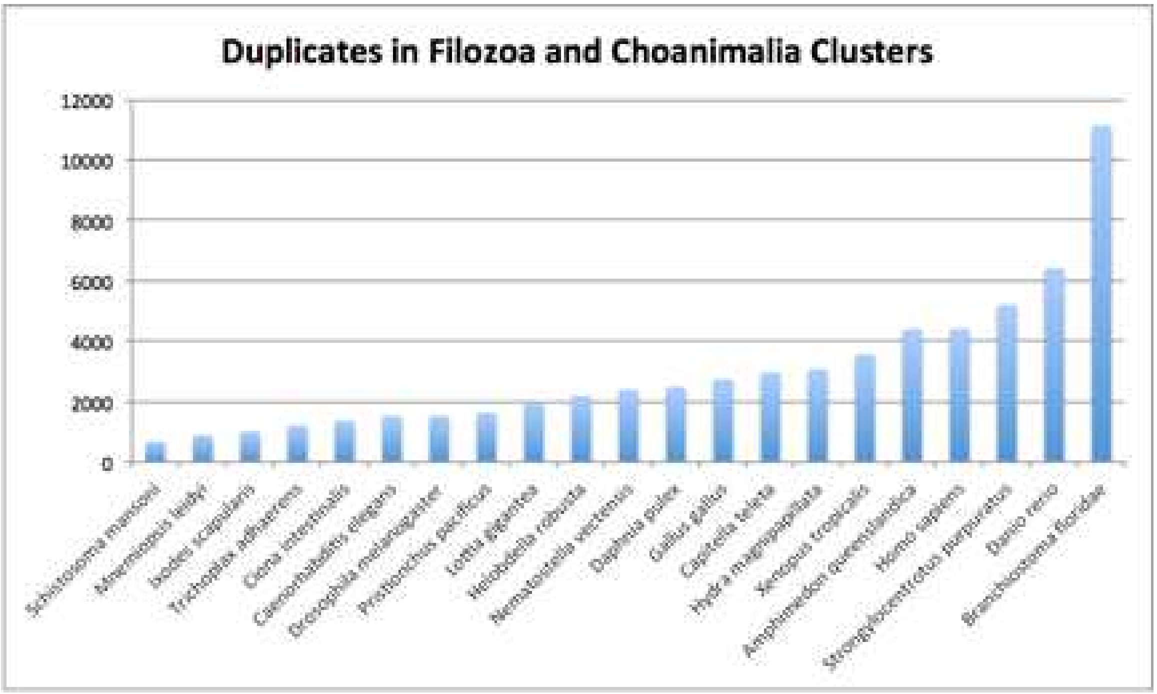


Figure S9: Comparison of gene losses

For each species, we count clusters in which a gene from this species is not present, but a gene from at least one metazoan and one non-metazoan are present.

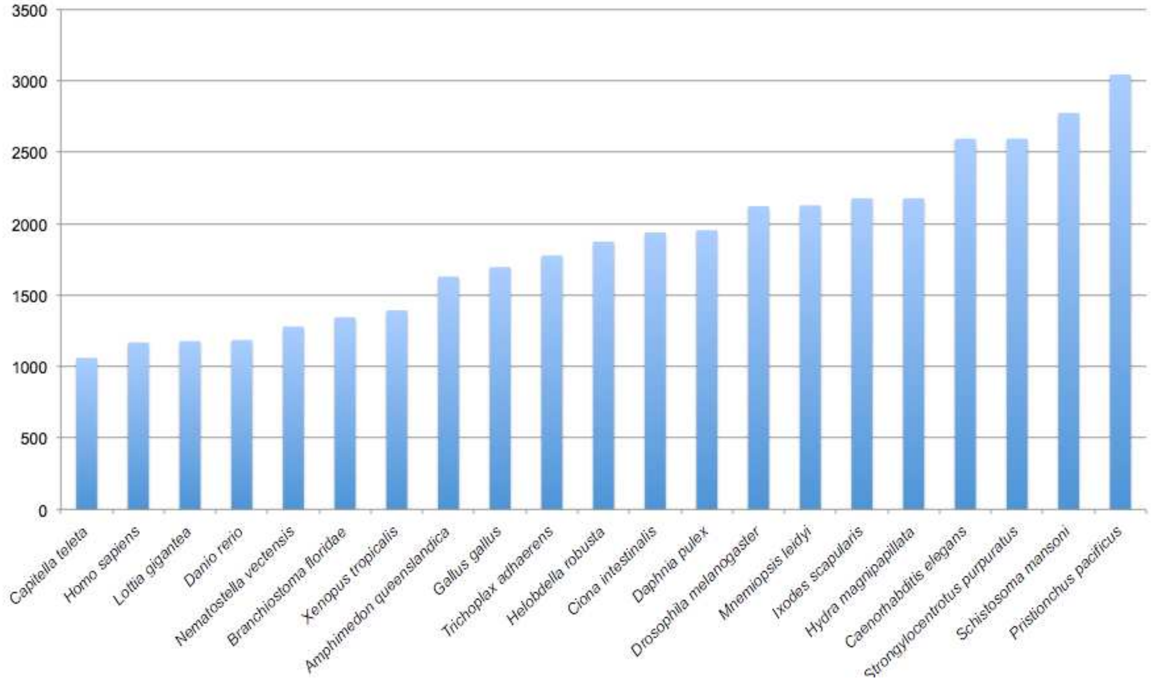
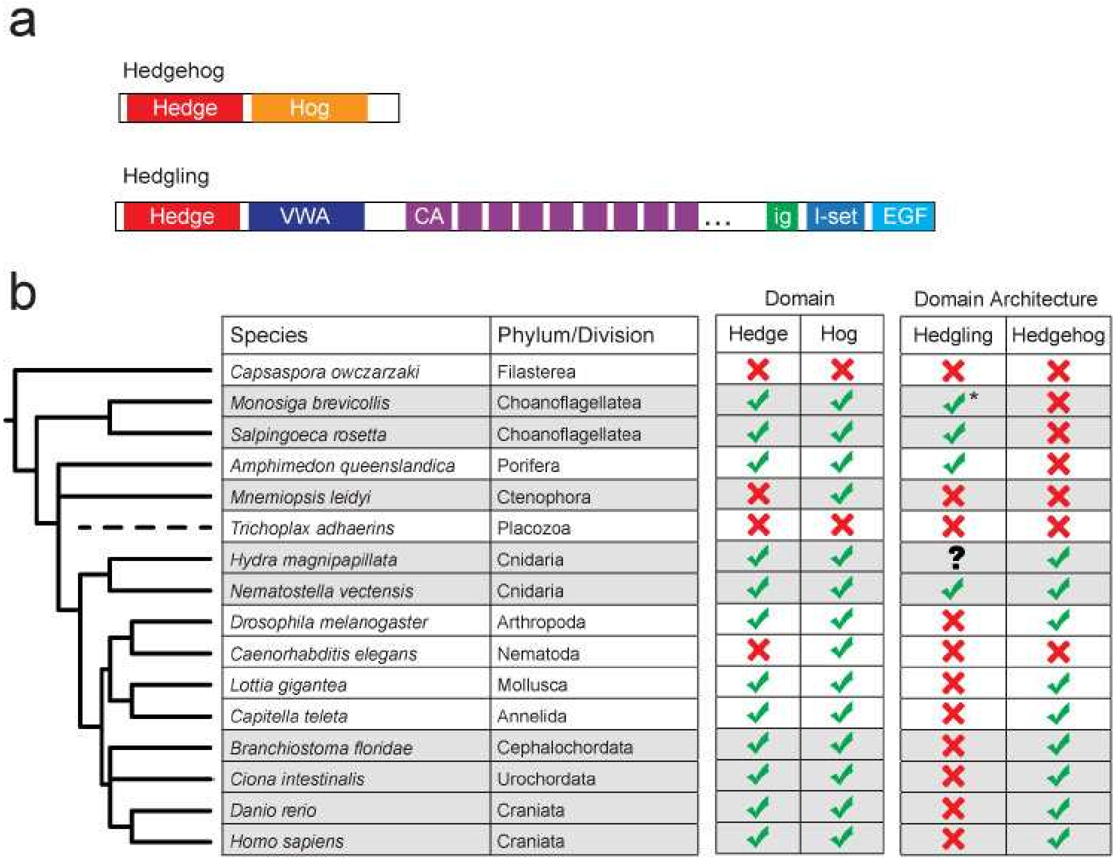


Figure S10: Hedgehog and Hedgling

a, Domain structure of Hedgehog and Hedgling gene products. **b,** Phylogenetic distribution of Hedge domains, Hog domains, Hedgling genes, and Hedgehog genes.



		Domain		Domain Architecture	
Species	Phylum/Division	Hedge	Hog	Hedgling	Hedgehog
Capsaspora owczarzaki	Filasterea	✗	✗	✗	✗
Monosiga brevicollis	Choanoflagellatea	✓	✓	✓*	✗
Salpingoeca rosetta	Choanoflagellatea	✓	✓	✓	✗
Amphimedon queenslandica	Porifera	✓	✓	✓	✗
Mnemiopsis leidyi	Ctenophora	✗	✓	✗	✗
Trichoplax adhaerens	Placozoa	✗	✗	✗	✗
Hydra magnipapillata	Cnidaria	✓	✓	?	✓
Nematostella vectensis	Cnidaria	✓	✓	✓	✓
Drosophila melanogaster	Arthropoda	✓	✓	✗	✓
Caenorhabditis elegans	Nematoda	✗	✓	✗	✗
Lottia gigantea	Mollusca	✓	✓	✗	✓
Capitella teleta	Annelida	✓	✓	✗	✓
Branchiostoma floridae	Cephalochordata	✓	✓	✗	✓
Ciona intestinalis	Urochordata	✓	✓	✗	✓
Danio rerio	Craniata	✓	✓	✗	✓
Homo sapiens	Craniata	✓	✓	✗	✓

References

54. J. C. Mullikin, Z. Ning, The phusion assembler. *Genome Res* **13**, 81-90 (2003).

55. W. J. Kent, BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

56. J. F. Ryan, Baa. pl: A tool to evaluate de novo genome assemblies with RNA transcripts. *arXiv preprint arXiv:1309.2087*, (2013).

57. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644-652 (2011).

58. C. Alkan, S. Sajjadian, E. E. Eichler, Limitations of next-generation genome sequence assembly. *Nature methods* **8**, 61-65 (2010).

59. W. Huang, L. Li, J. R. Myers, G. T. Marth, ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593-594 (2012).

60. S. Kurtz, The Vmatch large scale sequence analysis software. *Ref Type: Computer Program*, 4-12 (2003).

61. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-3.0. (2004).

62. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-467 (2005).

63. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

64. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).

65. B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr, L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654-5666 (2003).

66. A. A. Salamov, V. V. Solovyev, Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**, 516-522 (2000); published online EpubApr (

67. M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215-ii225 (2003).

68. A. Krogh, Two methods for improving performance of an HMM and their application for gene finding. *Center for Biological Sequence Analysis. Phone* **45**, 4525 (1997).

69. R. F. Yeh, L. P. Lim, C. B. Burge, Computational inference of homologous gene structures in the human genome. *Genome research* **11**, 803-816 (2001).

70. B. J. Haas, S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell, J. R. Wortman, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).

71. T. D. Wu, C. K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005).

72. B. J. Koch, J. F. Ryan, A. D. Baxevanis, The Diversification of the LIM Superclass at the Base of the Metazoa Increased Subcellular Complexity and Promoted Multicellular Specialization. *PLoS One* **7**, e33261 (2012).

73. C. E. Schnitzler, K. Pang, M. L. Powers, A. M. Reitzel, J. F. Ryan, D. Simmons, T. Tada, M. Park, J. Gupta, S. Y. Brooks, Genomic organization, evolution, and expression of photoprotein and opsin genes in Mnemiopsis leidyi: a new view of ctenophore photocytes. *BMC Biology* **10**, 107 (2012).

74. D. K. Simmons, K. Pang, M. Q. Martindale, Lim homeobox genes in the Ctenophore Mnemiopsis leidyi: the evolution of neural cell type specification. *Evodevo* **3**, 2 (2012).

75. E. A. Gladyshev, M. Meselson, I. R. Arkhipova, Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**, 1210-1213 (2008).

76. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

77. N. H. Putnam, M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, A. Terry, H. Shapiro, E. Lindquist, V. V. Kapitonov, J. Jurka, G. Genikhovich, I. V. Grigoriev, S. M. Lucas, R. E. Steele, J. R. Finnerty, U. Technau, M. Q. Martindale, D. S. Rokhsar, Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86-94 (2007).

78. N. H. Putnam, T. Butts, D. E. K. Ferrier, R. F. Furlong, U. Hellsten, T. Kawashima, M. Robinson-Rechavi, E. Shoguchi, A. T. J. K. Yu, The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071 (2008).

79. M. Srivastava, E. Begovic, J. Chapman, N. H. Putnam, U. Hellsten, T. Kawashima, A. Kuo, T. Mitros, A. Salamov, M. L. Carpenter, A. Y. Signorovitch, M. A. Moreno, K. Kamm, J. Grimwood, J. Schmutz, H. Shapiro, I. V. Grigoriev, L. W. Buss, B. Schierwater, S. L. Dellaporta, D. S. Rokhsar, The Trichoplax genome and the nature of placozoans. *Nature* **454**, 955-960 (2008).

80. M. Srivastava, O. Simakov, J. Chapman, B. Fahey, M. E. Gauthier, T. Mitros, G. S. Richards, C. Conaco, M. Dacre, U. Hellsten, C. Larroux, N. H. Putnam, M. Stanke, M. Adamska, A. Darling, S. M. Degnan, T. H. Oakley, D. C. Plachetzki, Y. Zhai, M. Adamski, A. Calcino, S. F. Cummins, D. M. Goodstein, C. Harris, D. J. Jackson, S. P. Leys, S. Shu, B. J. Woodcroft, M. Vervoort, K. S. Kosik, G. Manning, B. M. Degnan, D. S. Rokhsar, The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**, 720-726 (2010).

81. O. Simakov, F. Marletaz, S. J. Cho, E. Edsinger-Gonzales, P. Havlak, U. Hellsten, D. H. Kuo, T. Larsson, J. Lv, D. Arendt, R. Savage, K. Osoegawa, P. de Jong, J. Grimwood, J. Chapman, H. Shapiro, A. Aerts, R. P. Otillar, A. Y. Terry, J. L. Boore, I. V. Grigoriev, D. R. Lindberg, E. C. Seaver, D. A. Weisblat, N. Putnam, D. S. Rokhsar, Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526-531 (*2012*).

82. L. Li, C. J. Stoeckert, D. S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**, 2178-2189 (2003).

83. L. S. Johnson, S. R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).

84. M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, R. D. Finn, The Pfam protein families database. *Nucleic Acids Res* **40**, D290-301 (2012).

85. K. Katoh, G. Asimenos, H. Toh, Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* **537**, 39-64 (2009).

86. G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**, 564-577 (2007).

87. R. M. Waterhouse, E. M. Zdobnov, F. Tegenfeldt, J. Li, E. V. Kriventseva, OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res* **39**, D283-288 (2011).

88. A. Hejnol, M. Obst, A. Stamatakis, M. Ott, G. W. Rouse, G. D. Edgecombe, P. Martinez, J. Baguna, X. Bailly, U. Jondelius, M. Wiens, W. E. Muller, E. Seaver, W. C. Wheeler, M. Q. Martindale, G. Giribet, C. W. Dunn, Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci* **276**, 4261-4270 (2009).

89. D. L. Swofford, PAUP*: phylogenetic analysis using parsimony, version 4.0 b10. (2003).

90. A. Rokas, B. L. Williams, N. King, S. B. Carroll, Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798-804 (2003).

91. A. Rambaut, A. Drummond, FigTree v1. 3.1. *Program distributed by the author. Institute of Evolutionary Biology, University of Edinburgh. Edinburgh, United Kingdom*, (2009).

92. A. Rambaut, A. J. Drummond, TreeStat v1. 2: tree statistic calculation tool. *Program distributed by the author. Institute of Evolutionary Biology, University of Edinburgh. Edinburgh, United Kingdom*, (2008).

93. S. A. Berger, A. Stamatakis, R. Lucking, Morphology-based phylogenetic binning of the lichen genera Graphis and Allographa (Ascomycota: Graphidaceae) using molecular site weight calibration. *Taxon* **60**, 1450-1457 (2011).

94. H. Fang, M. E. Oates, R. B. Pethica, J. M. Greenwood, A. J. Sardar, O. J. L. Rackham, P. C. J. Donoghue, A. Stamatakis, D. A. de Lima Morais, J. Gough, A daily-updated tree of (sequenced) life as a reference for genome research. *Scientific Reports* **3**, (2013).

95. H. Shimodaira, M. Hasegawa, CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246-1247 (2001).

96. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51**, 492-508 (2002).

97. N. Goldman, J. P. Anderson, A. G. Rodrigo, Likelihood-based tests of topologies in phylogenetics. *Systematic Biology* **49**, 652-670 (2000).

98. J. Lehmann, P. F. Stadler, V. Krauss, Near intron pairs and the metazoan tree. *Molecular Phylogenetics and Evolution*, (2013).

99. I. Letunic, T. Doerks, P. Bork, SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* **40**, D302-305 (2012).

100. A. Alie, M. Manuel, The backbone of the post-synaptic density originated in a unicellular ancestor of choanoflagellates and metazoans. *BMC Evol Biol* **10**, 34 (2010).

101. O. Sakarya, K. A. Armstrong, M. Adamska, M. Adamski, I. F. Wang, B. Tidor, B. M. Degnan, T. H. Oakley, K. S. Kosik, A post-synaptic scaffold at the origin of the animal kingdom. *PLoS One* **2**, e506 (2007).

103. M. Jager, R. Chiori, A. Alie, C. Dayraud, E. Queinnec, M. Manuel, New insights on ctenophore neural anatomy: immunofluorescence study in Pleurobrachia pileus (Muller, 1776). *J Exp Zool B Mol Dev Evol* **316B**, 171-187 (2011).

104. A. Hay-Schmidt, The evolution of the serotonergic nervous system. *P Roy Soc Lond B Bio* **267**, 1071-1079 (2000).

105. M. Leptin, twist and snail as positive and negative regulators during Drosophila mesoderm development. *Genes Dev* **5**, 1568-1576 (1991).

106. N. Azpiazu, M. Frasch, tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of Drosophila. *Genes Dev* **7**, 1325-1340 (1993).

107. K. Jagla, M. Bellard, M. Frasch, A cluster of Drosophila homeobox genes involved in mesoderm differentiation programs. *Bioessays* **23**, 125-133 (2001).

108. T. Sato, D. Rocancourt, L. Marques, S. Thorsteinsdottir, M. Buckingham, A Pax3/Dmrt2/Myf5 regulatory cascade functions at the onset of myogenesis. *PLoS Genet* **6**, e1000897 (2010).

109. K. Ryan, N. Garrett, A. Mitchell, J. B. Gurdon, Eomesodermin, a key early gene in Xenopus mesoderm differentiation. *Cell* **87**, 989-1000 (1996).

110. P. J. Gianakopoulos, V. Mehta, A. Voronova, Y. Cao, Z. Yao, J. Coutu, X. Wang, M. S. Waddington, S. J. Tapscott, I. S. Skerjanc, MyoD directly up-regulates premyogenic mesoderm factors during induction of skeletal myogenesis in stem cells. *J Biol Chem* **286**, 2517-2525 (2011).

111. W. R. Francis, L. M. Christianson, R. Kiko, M. L. Powers, N. C. Shaner, S. H. D. Haddock, A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics* **14**, 167 (2013).

112. A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).

113. Y. f. Zhong, P. W. H. Holland, HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution & Development* **13**, 567-568 (2011).

114. J. F. Ryan, K. Pang, J. C. Mullikin, M. Q. Martindale, A. D. Baxevanis, The homeodomain complement of the ctenophore Mnemiopsis leidyi suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *Evodevo* **1**, 9 (2010).

115. K. Pang, J. F. Ryan, A. D. Baxevanis, M. Q. Martindale, Evolution of the TGF-beta signaling pathway and its potential role in the ctenophore, Mnemiopsis leidyi. *PLoS One* **6**, e24152 (2011).

116. M. E. Skinner, A. V. Uzilov, L. D. Stein, C. J. Mungall, I. H. Holmes, JBrowse: A next-generation genome browser. *Genome research* **19**, 1630-1638 (2009).

117. A. Wallberg, M. Thollesson, J. S. Farris, U. Jondelius, The phylogenetic position of the comb jellies (Ctenophora) and the importance of taxonomic sampling. *Cladistics* **20**, 558-578 (2004).

118. J. F. Ryan, A. D. Baxevanis, Hox, Wnt, and the evolution of the primary body axis: insights from the early-divergent phyla. *Biol Direct* **2**, 37 (2007).

119. A. Lang, Die Polycladen des Golfes von Neapel und der angrenzenden Meeresabschnitte. *Fauna u. Flora Neapel* **11**, 1-688 (1884).

120. L. H. Hyman, *The invertebrates*. (McGraw-Hill, New York,, ed. 1st, 1940), pp. v.

121. H. Hadzi, *Turbellarijska teorija knidarijev (Turbellarien-Theorie der Knidarien)*. Slovenian Academy of Sciences and Arts (Ljubljana, Slovenia, 1944), vol. 3, pp. 238.

122. R. C. Brusca, G. J. Brusca, Invertebrates. *Sinauer, Sunderland, Mass* **264**, (1990).

123. U. Ehlers, Ultrastructure of the spermatozoa of Halammohydra schulzei (Cnidaria, Hydrozoa): The significance of acrosomal structures for the systematization of the Eumetazoa. *Microfauna Marina* **8**, 115-130 (1993).

124. E. E. Ruppert, R. D. Barnes, R. S. Fox, *Invertebrate zoology*. (Saunders College Pub., 1994), vol. 6.

125. C. Nielsen, *Animal evolution: interrelationships of the animal phyla*. (Oxford University Press, Oxford, 1995).

126. C. Nielsen, Cladistic analyses of the animal kingdom. *Biological Journal of the Linnean Society* **57**, 385-410 (1996).

127. P. Ax, *Multicellular animals. A New Approach to the Phylogenetic Order in Nature, Volume I* (Springer, Berlin ; New York, 1996).

128. L. Margulis, K. V. Schwartz, *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth*. (Freeman New York, ed. 3, 1998).

129. C. Nielsen, *Animal evolution : interrelationships of the living phyla*. (Oxford University Press, Oxford, ed. 2nd, 2001), pp. X, 563 s.

130. P. O. Wainright, G. Hinkle, M. L. Sogin, S. K. Stickel, Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science* **260**, 340-342 (1993).

131. T. Katayama, H. Wada, H. Furuya, N. Satoh, M. Yamamoto, Phylogenetic position of the dicyemid mesozoa inferred from 18S rDNA sequences. *Biological Bulletin* **189**, 81-90 (1995).

132. B. Hanelt, D. Van Schyndel, C. M. Adema, L. A. Lewis, E. S. Loker, The phylogenetic position of Rhopalura ophiocomae (Orthonectida) based on 18S ribosomal DNA sequence analysis. *Molecular Biology and Evolution* **13**, 1187-1191 (1996).

133. Y. Van De Peer, R. De Wachter, Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. *Journal of Molecular Evolution* **45**, 619-630 (1997).

134. E. Abouheif, R. Zardoya, A. Meyer, Limitations of metazoan 18s rRNA sequence data: Implications for reconstructing a phylogeny of the animal kingdom and inferring the reality of the cambrian explosion. *Journal of Molecular Evolution* **47**, 394-405 (1998).

135. A. G. Collins, Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence. *Proc. Natl. Acad. Sci. USA* **95**, 15458-15463 (1998).

136. K. M. Halanych, Considerations for Reconstructing Metazoan History: Signal, Resolution, and Hypothesis Testing. *Integrative and Comparative Biology* **38**, 929-941 (1998).

137. D. L. Lipscomb, J. S. Farris, M. Ka?llersjo, A. Tehler, Support, ribosomal sequences and the phylogeny of the eukaryotes. *Cladistics* **14**, 303-338 (1998).

138. B. M. H. Winnepenninckx, Y. D. E. Van Peer, T. Backeljau, Metazoan Relationships on the Basis of 18S rRNA Sequences: A Few Years Later. *Integrative and Comparative Biology* **38**, 888-906 (1998).

139. J. Zrzavy, S. Mihulka, P. Kepka, A. Bezde?k, D. Tietz, Phylogeny of the Metazoa Based on Morphological and 18S Ribosomal DNA Evidence. *Cladistics* **14**, 249-285 (1998).

140. J. Kim, W. Kim, C. W. Cunningham, A new perspective on lower metazoan relationships from 18S rDNA sequences [2]. *Molecular Biology and Evolution* **16**, 423-427 (1999).

141. G. Giribet, W. C. Wheeler, The Position of Arthropods in the Animal Kingdom: Ecdysozoa, Islands, Trees, and the "Parsimony Ratchet". *Molecular Phylogenetics and Evolution* **13**, 619-623 (1999).

142. M. E. Siddall, M. F. Whiting, Long-branch abstractions. *Cladistics* **15**, 9-24 (1999).

143. M. Medina, A. G. Collins, J. D. Silberman, M. L. Sogin, Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. *Proc Natl Acad Sci U S A* **98**, 9707-9712 (2001).

144. K. J. Peterson, D. J. Eernisse, Animal phylogeny and the ancestry of bilaterians: Inferences from morphology and 18S rDNA gene sequences. *Evolution and Development* **3**, 170-205 (2001).

145. M. Podar, S. H. Haddock, M. L. Sogin, G. R. Harbison, A molecular phylogenetic framework for the phylum Ctenophora using 18S rRNA genes. *Mol Phylogenet Evol* **21**, 218-230 (2001).

146. A. G. Collins, Phylogeny of Medusozoa and the evolution of cnidarian life cycles. *Journal of Evolutionary Biology* **15**, 418-432 (2002).

147. C. Martinelli, J. Spring, Distinct expression patterns of the two T-box homologues Brachyury and Tbx2/3 in the placozoan Trichoplax adhaerens. *Dev Genes Evol* **213**, 492-499 (2003).

148. J. Zrzavy, V. Hypsa, Myxozoa, Polypodium, and the origin of the Bilateria: The phylogenetic position of "Endocnidozoa" in light of the rediscovery of Buddenbrockia [1]. *Cladistics* **19**, 164-169 (2003).

149. C. W. Dunn, A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, G. Giribet, Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745-749 (2008).

150. H. Philippe, R. Derelle, P. Lopez, K. Pick, C. Borchiellini, N. Boury-Esnault, J. Vacelet, E. Renard, E. Houliston, E. Queinnec, C. Da Silva, P. Wincker, H. Le Guyader, S. Leys, D. J. Jackson, F. Schreiber, D. Erpenbeck, B. Morgenstern, G. Worheide, M. Manuel, Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* **19**, 706-712 (2009).

151. B. Schierwater, M. Eitel, W. Jakob, H. J. Osigus, H. Hadrys, S. L. Dellaporta, S. O. Kolokotronis, R. Desalle, Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biol* **7**, e20 (2009).

152. K. S. Pick, H. Philippe, F. Schreiber, D. Erpenbeck, D. J. Jackson, P. Wrede, M. Wiens, A. Alie, B. Morgenstern, M. Manuel, G. Worheide, Improved phylogenomic taxon sampling noticeably affects non-bilaterian relationships. *Mol Biol Evol*, (2010).

153. J. Mallatt, C. W. Craig, M. J. Yoder, Nearly complete rRNA genes from 371 Animalia: updated structure-based alignment and detailed phylogenetic analysis. *Mol Phylogenet Evol* **64**, 603-617 (2012).

154. A. Kumar, An overview of nested genes in eukaryotic genomes. *Eukaryot Cell* **8**, 1321-1329 (2009).

155. T. Adell, V. A. Grebenjuk, M. Wiens, W. E. Muller, Isolation and characterization of two T-box genes from sponges, the phylogenetically oldest metazoan taxon. *Dev Genes Evol* **213**, 421-434 (2003).

156. M. Manuel, Y. Le Parco, C. Borchiellini, Comparative analysis of Brachyury T-domains, with the characterization of two new sponge sequences, from a hexactinellid and a calcisponge. *Gene* **340**, 291-301 (2004).

157. D. E. Martinez, M. L. Dirksen, P. M. Bode, M. Jamrich, R. E. Steele, H. R. Bode, Budhead, a fork head/HNF-3 homologue, is expressed during axis formation and head specification in hydra. *Dev Biol* **192**, 523-536 (1997).