Error, signal, and the placement of Ctenophora sister to all other animals

Nathan V. Whelan[a,1], Kevin M. Kocot[b], Leonid L. Moroz[c], and Kenneth M. Halanych[a]

[a]Molette Biology Laboratory for Environmental and Climate Change Studies, Department of Biological Sciences, Auburn University, Auburn, AL 36849; [b]School of Biological Sciences, University of Queensland, St. Lucia, QLD 4101, Australia; and [c]Department of Neuroscience and McKnight Brain Institute, University of Florida, Gainesville, FL 32610

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved March 24, 2015 (received for review February 18, 2015)

Elucidating relationships among early animal lineages has been difficult, and recent phylogenomic analyses place Ctenophora sister to all other extant animals, contrary to the traditional view of Porifera as the earliest-branching animal lineage. To date, phylogenetic support for either ctenophores or sponges as sister to other animals has been limited and inconsistent among studies. Lack of agreement among phylogenomic analyses using different data and methods obscures how complex traits, such as epithelia, neurons, and muscles evolved. A consensus view of animal evolution will not be accepted until datasets and methods converge on a single hypothesis of early metazoan relationships and putative sources of systematic error (e.g., long-branch attraction, compositional bias, poor model choice) are assessed. Here, we investigate possible causes of systematic error by expanding taxon sampling with eight novel transcriptomes, strictly enforcing orthology inference criteria, and progressively examining potential causes of systematic error while using both maximum-likelihood with robust data partitioning and Bayesian inference with a site-heterogeneous model. We identified ribosomal protein genes as possessing a conflicting signal compared with other genes, which caused some past studies to infer ctenophores and cnidarians as sister. Importantly, biases resulting from elevated compositional heterogeneity or elevated substitution rates are ruled out. Placement of ctenophores as sister to all other animals, and sponge monophyly, are strongly supported under multiple analyses, herein.

phylogenomics | Metazoa | Ctenophora | Porifera | Cnidaria

Resolving relationships among extant lineages at the base of the metazoan tree is integral to understanding evolution of complex animal traits, including nervous systems and gastrulation. Historically, sponges and placozoans, both of which have relatively simple body plans and lack neurons, have been considered to diverge from other animals earlier than ctenophores, cnidarians, and bilaterians (1). Phylogenomic studies have resulted in controversial hypotheses placing either Placozoa (Fig. 1*A*) (2), ctenophores (ctenophore-sister hypothesis) (Fig. 1*B*) (3–7), or a clade of ctenophores and sponges (Fig. 1*C*) (6) as sister to all remaining animals. Others (7–10) have claimed nontraditional findings resulted from systematic error and argued for traditional placement of sponges as sister to all remaining animals (Eumetazoa, or Porifera-sister hypothesis) (Fig. 1*D*) and a sister relationship between ctenophores and cnidarians (Coelenterata) (Fig. 1*D*). Limited statistical support for various hypotheses and conflict among, and even within studies, has undermined confidence in our understanding of early animal evolution. Basal metazoan relationships must be resolved with greater consistency before a consensus viewpoint is widely accepted.

Long-branch attraction (LBA) (11), which occurs when two divergent lineages are artificially inferred as related because of substitutional saturation (11), is perhaps the most often evoked explanation for controversial or spurious phylogenetic results (7–10, 12, 13). Additional sources of systematic error include poor taxon or character sampling (10, 14, 15), large amounts of missing data (16), and model misspecification (16, 17). Such errors have been implicated as influencing the position of ctenophores in metazoan phylogeny studies (5–8). For example, Ryan et al. (6) recovered a sister relationship between sponges and ctenophores in analyses where taxa with high amounts of missing data were excluded, and support for this was highest in Bayesian inference with the CAT (17) substitution model. However, ctenophores were recovered as sister to all other extant animals in maximum-likelihood analyses with greater taxon sampling (Bayesian inference never converged for datasets with more than 19 taxa). The CAT model is a site-heterogeneous model that may handle LBA artifacts better than site-homogeneous substitution models like GTR (18). Therefore, LBA plausibly influenced the phylogenetic position of ctenophores in analyses of Ryan et al. (6) that recovered ctenophores-sister. In Moroz et al. (5), strong nodal support for ctenophores sister to all other extant animals disappeared when both the strictest orthology criteria were enforced and ctenophore taxon sampling increased. Thus, further consideration of systematic error influencing phylogeny reconstruction at the base of the animal tree is desirable.

The ctenophore-sister hypothesis has challenged our understanding of early metazoan evolution, but given conflicting results (2–10), this and other hypotheses must be carefully scrutinized. Ideally, if robust datasets are assembled and causes of systematic error are accounted for, different datasets and analytical methods will converge on a single phylogenetic hypothesis (19). However, practical barriers exist in assembling robust datasets free of systematic error. For example, phylogenomic datasets are prone to missing data given the incomplete nature of transcriptome and even genome sequences, and orthology determination among

Significance

Traditional interpretation of animal phylogeny suggests traits, such as mesoderm, muscles, and neurons, evolved only once given the assumed placement of sponges as sister to all other animals. In contrast, placement of ctenophores as the first branching animal lineage raises the possibility of multiple origins of many complex traits considered important for animal diversification and success. We consider sources of potential error and increase taxon sampling to find a single, statistically robust placement of ctenophores as our most distant animal relatives, contrary to the traditional understanding of animal phylogeny. Furthermore, ribosomal protein genes are identified as creating conflict in signal that caused some past studies to recover a sister relationship between ctenophores and cnidarians.

Author contributions: N.V.W., K.M.K., L.L.M., and K.M.H. designed research; N.V.W. and K.M.K. performed research; N.V.W. and K.M.K. analyzed data; and N.V.W., K.M.K., L.L.M., and K.M.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the NCBI Sequence Read Archive, www.ncbi.nlm.nih.gov/sra (accession no. PRJNA278284). Transcriptome assemblies, phylogenetic datasets, and an annotation file were deposited to figshare, figshare.com (doi: 10.6084/m9.figshare.1334306).

[1]To whom correspondence should be addressed. Email: nwhelan@auburn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1503453112/-/DCSupplemental.

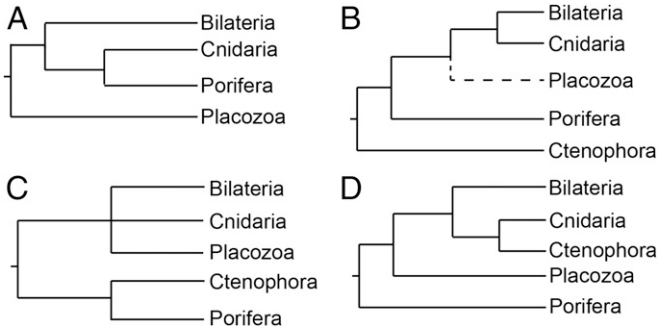

Fig. 1. Phylogenetic hypotheses from previous molecular studies. (*A*) Placozoa-sister hypothesis (2). (*B*) Ctenophora-sister hypothesis (3–7). Placozoa was not included in some studies that found support for this topology. (*C*) Ctenophora + Porifera-sister hypothesis (6). (*D*) Traditional Porifera-sister hypothesis (7–10).

distantly related species can be difficult (20, 21). Computational limitations of complex phylogenetic methods can also prevent using what may be the best theoretical phylogenetic method. Nevertheless, both data quality and appropriate methods should be emphasized if deep relationships of any organismal group are to be robustly resolved.

Here, we have assembled a more comprehensive phylogenomic dataset of metazoan lineages that branched early in animal evolution than previous studies to alleviate taxon sampling concerns. We have sequenced transcriptomes of eight additional species and used other deeply sequenced publicly available transcriptomes (including some not used in past studies). Additionally, we use a number of data-filtering steps to explore the sensitivity of these results to potential sources of error. This process includes strict orthology determination, removal of taxa and genes that may cause LBA, and removal of heterogeneous genes that may cause model misspecification. Regardless of how data were filtered, all maximum-likelihood analyses with model partitioning and all Bayesian inference analyses using a site-heterogeneous model recover ctenophores as sister to all other animals with strong support. We identify overreliance on ribosomal protein genes in some datasets (7, 9) as the source of incongruence among previous phylogenomic studies. We also find strong support for sponge monophyly in contrast to previous reports (7, 22–24).

Results

Datasets and Accounting for Biases. Orthology filtering of transcriptome and genome data from 76 species resulted in 251 orthologous groups (OGs) and 81,008 aligned amino acid sites (Tables S1 and S2). TreSpEx (25) further identified 83 "certain" paralogs (i.e., sequences TreSpEx classified as high-confidence paralogs) from 10 OGs. TreSpEx also identified 2,684 "uncertain" paralogs (i.e., sequences TreSpEx classified as possibly, but not definitively, paralogous) from 104 OGs. Datasets with certain and both certain and uncertain paralogs pruned were starting points for progressively filtering other causes of systematic error. Overall, 25 hierarchical datasets that had progressively fewer characters, but controlled for more potential causes of systematic errors, were assembled (Fig. 2 and Table S2) (all datasets have been deposited on figshare, doi 10.6084/m9.figshare.1334306). The percentage of gene occupancy and missing data ranged from 70–82% and 35–44%, respectively (Table S2). Other than progressive data filtering, differences between datasets analyzed here and those used in previous studies of basal metazoan relationships (3–10) are increased character sampling and less missing data compared with some studies and increased nonbilaterian taxon sampling. In contrast to Nosenko et al. (7) and Philippe et al. (9), both of which relied heavily on ribosomal proteins (i.e., 52% and 71%, respectively), our dataset did not contain a large representation of any one gene class (e.g., only 8 of 250 were ribosomal protein genes) (*SI Methods*).

Ctenophores Sister to Other Extant Animals and Monophyletic Sponges. All maximum-likelihood analyses with outgroups resulted in topologies with strong support [≥ 97% bootstrap support (BS)] for ctenophores sister to all other extant animal lineages (datasets 1–21 in Figs. 2 and 3 and Figs. S1–S4). Importantly, Phylobayes (26) analyses using the CAT-GTR+Γ model also resulted in phylogenies with Ctenophora sister to all other animals with 100% posterior probability (PP), and 100% PP for sponge monophyly (datasets 6 and 16 in Figs. 2 and 3 and Fig. S5 *B* and *C*). Inferred relationships among major sponge lineages (i.e., Demospongiae + Hexactinellida sister to Calcarea + Homoscleromorpha) were consistent with morphology (27, 28) and most other molecular analyses that have recovered monophyletic sponges (datasets 1–27 in Figs. 2 and 3 and Figs. S1–S5) (4, 5, 8, 9, 28–30). Alternative hypotheses of basal animal relationships (Fig. 1) were rejected by every phylogenetic analysis as measured by the approximately unbiased (AU) (31) test ($P \leq 0.001$) (Table 1). Overall, our results overwhelmingly reject alternative hypotheses to ctenophores sister to all other extant animals (Table 1).

Ribosomal protein genes can have conflicting signal with most other genes (32). The datasets of Philippe et al. (9) and Nosenko et al. (7), which recovered cnidarians and ctenophores sister, had high proportions of ribosomal protein genes (67 of 128 and 87 of 122 genes, respectively). Nosenko et al. (7) analyzed a dataset without ribosomal protein genes and recovered ctenophores sister to all other animals, but a similar analysis has not been done for the original Philippe et al. (9) dataset. If certain topologies are recovered only with the use of a high proportion of one group of genes (e.g., ribosomal protein genes), this may indicate a phylogenetic signal that conflicts with the true evolutionary history. As such, we analyzed the Philippe et al. (9) dataset with ribosomal proteins removed (67 of 128) using maximum-likelihood and Bayesian inference, and neither reconstruction placed ctenophores and cnidarians sister as in the original study (Fig. S5 *D* and *E*).

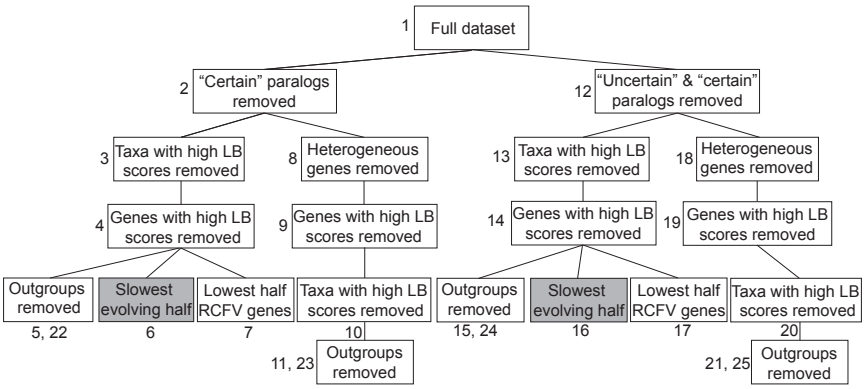

Fig. 2. Hierarchy of datasets with progressive data filtering. Numbers associated with each datasets are references throughout the text. Datasets 5, 11, 15, 21 have Choanoflagellates as an outgroup, and datasets 22, 23, 24, 25 have all outgroups removed. RAxML analyses were used for each dataset, and shaded datasets were also analyzed with PhyloBayes. Data matrix statistics (e.g., number of sites, percentage of missing data, and so forth) for each dataset can be found in Table S2.

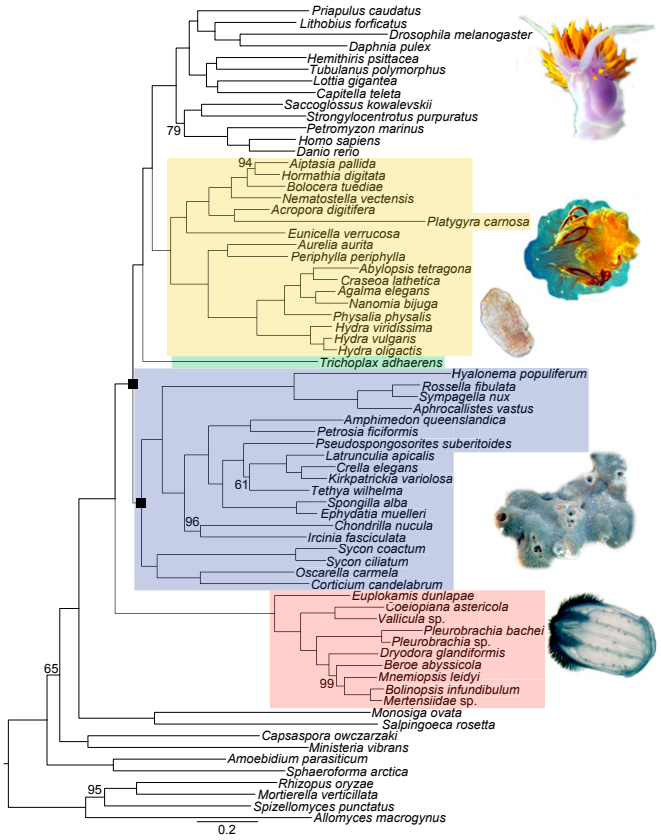

Fig. 3. Reconstructed maximum-likelihood topology of metazoan relationships inferred with dataset 10. Maximum likelihood and Bayesian topologies inferred with other datasets (Fig. 2) have identical basal branching patterns (Figs. S1–S5). Nodes are supported with 100% bootstrap support unless otherwise noted. Support, as inferred from each dataset (Fig. 2), for nodes covered by black boxes are in Table 1.

The maximum-likelihood analysis recovered strong support for ctenophores as sister to all other metazoan lineages (BS = 93) (Fig. S5*D*). However, Bayesian inference (Fig. S5*E*) recovered sponges as sister to all other metazoans, but support for this and other deep nodes were low (PP ≤ 90).

Systematic Biases and Their Effect on Phylogenetic Inference. Long-branch (LB) scores (28), a measurement for identifying taxa and OGs that could cause LBA, were calculated for each species and OG with TreSpEx (25). In total, we identified six "long-branched" taxa, all nonmetazoans (Fig. S6*A* and Table S2), and 28 OGs with high LB scores compared with other OGs (Fig. S6 *B* and *C*). We found complete congruence in relationships among basal metazoan phyla in trees inferred with (datasets 1, 2, 8, 12, and 18 in Fig. 2) and without (datasets 3–7, 9–11, 13–17, 19–21, and 22–25 in Fig. 2) taxa and genes that had high LB scores, and nodal support for critical nodes showed little variation among analyses (Fig. 3 and Figs. S1–S5). Removing OGs with high amino acid compositional heterogeneity (datasets 7–11, 17–21, 23, and 25 in Fig. 2) also had no effect on branching order (Fig. 3 and Figs. S2 *A–E*, S3 *E* and *F*, S4 *A–E*, and S5*A*). Topologies inferred with only the slowest evolving half of OGs assembled here (datasets 6 and 16 in Fig. 2) (i.e., least saturated and least prone to homoplasy; see Fig. S7 for saturation plots) recovered high support for ctenophores sister to all other animals and sponge monophyly with both maximum-likelihood (BS = 100) (Fig. 3 and Figs. S1*F* and S3*D*) and Bayesian inference using the CAT-GTR model (PP = 1) (Fig. 3 and Fig. S5 *B* and *C*). Importantly, our datasets of the slowest evolving half of OGs were of a broad range of protein classes (*SI Methods*; figshare), rather than consisting of a majority of ribosomal proteins (7, 9).

Inaccurate orthology assignment can also introduce systematic error into phylogenomic analyses. Although relationships among basal lineages were unaffected, removal of paralogs as identified by TreSpEx appeared to have the greatest effect on support for some critical nodes. For example, most topologies with both certain and uncertain paralogs removed had strong support for sponge monophyly (i.e., ≥ 95% BS) (datasets 12–14 and 18–20 in Figs. 2 and 3 and Figs. S2*F*, S3 *A*, *B*, and *F*, and S4 *A* and *B*), but four analyses with only certain paralogs removed recovered low support (< 90% BS) for sponge monophyly (datasets 5, 7, 9, and 10 in Figs. 2 and 3 and Figs. S1*E* and S2 *A*, *D*, and *E*).

Because outgroup sampling has the potential to influence rooting of the animal tree, we explored outgroup sampling as well. When all outgroups except two choanoflagellates were removed (datasets 5, 11, 15, and 21 in Fig. 2), inferred nonbilaterian relationships were identical as in analyses we performed with full outgroup sampling (datasets 5, 11, 15, and 21 in Figs. 2 and 3 and Figs. S1*E*, S2*E*, S3*C*, and S4*C*), but support for sponge monophyly decreased. In these analyses the leaf-stability indices for homoscleromorph and calcareous sponges were less than 0.94, but in all other analyses they were greater than 0.97 (Fig. S5 *F* and *G*). Regardless, when choanoflagellates were the only outgroup, ctenophores were still recovered as the deepest split within the animal tree with 100% BS support. Analyses with all outgroup taxa removed (datasets 22–25 in Fig. 2) recovered identical relationships among major metazoan lineages as other analyses (Figs. S4 *D–F* and S5*A*). However, we observed low support for relationships among ctenophores, sponges, and placozoans in these analyses. This resulted from the long placozoan branch being attracted to ctenophores in the absence of outgroup taxa as indicated by bootstrap tree topologies and leaf-stability index for *Trichoplax* of less than 0.92, whereas leaf-stability indices were greater than 0.99 in all other analyses (Fig. S5 *F* and *G*).

Discussion

Placement of Ctenophores Sister to all Remaining Animals Is Not Sensitive to Systematic Errors. Every analysis conducted herein strongly supported the ctenophore-sister hypothesis (Fig. 3 and Table 1). A major hurdle to wide acceptance of ctenophores as sister to other animals has been that different analyses have yielded conflicting hypotheses of early animal phylogeny (2–9). Sensitivity to the selected model of molecular evolution has been especially problematic (2–9). In contrast, both maximum-likelihood analyses using data partitioning and Bayesian analyses using the CAT-GTR model of our datasets resulted in identical branching patterns among ctenophores, sponges, placozoans, cnidarians, and bilaterians. Past critiques of studies that found ctenophores to be sister to all other animals have emphasized the CAT model as the most appropriate model for deep phylogenomics because it is an infinite mixture model that accounts for site-heterogeneity (7, 8, 29). Notably, when the CAT-GTR model was used here (datasets 6 and 16 in Fig. 2), we recovered ctenophores-sister to all other metazoans (Fig. 3 and Fig. S5 *B* and *C*).

The argument for LBA (7–10) or saturated datasets (7, 8) as the reason past studies found ctenophores to be sister to all other animals seems to have been overstated. The recovered position of ctenophores was identical in analyses with (datasets 1, 2, 8, 12, and 18 in Fig. 2 and Figs. S1 *A* and *B*, S2 *B* and *F*, and S3*F*) and without (datasets 3–7, 9–11, 13–17, and 19–25 in Fig. 2, and Figs. S1 *C–F*, S2 *A* and *C–E*, S3 *A–E*, S4, and S5 *A–C*) taxa and genes with high LB scores, and analyses with the slowest evolving genes (datasets 6 and 16 in Fig. 2 and Fig. S7) also recovered ctenophores sister to all other animals (Fig. 3 and Figs. S1*F*, S3*D*, and S5 *B* and *C*). Furthermore, despite the long internal branch leading to the ctenophore clade, the position of this lineage did not change in any analysis including those when outgroups were removed (datasets 5, 11, 15, 21, and 22–25 in Fig. 2 and Figs. S1*E*, S2*E*, S3*C*, and S4 *C–F*). If this branch was being artificially attracted toward outgroups, then employment of different outgroup schemes would be expected

Table 1. AU test *P* values for alternative hypotheses of animal relationships and support for ctenophora-sister and sponge monophyly

Dataset (no.)	Porifera-sister	Coelenterata	Placozoa-sister	Porifera + Ctenophora	Ctenophora-sister	Porifera monophyly
Full dataset (1)	2.E-04	5.E-06	1.E-76	2.E-05	100 BS	95 BS
"Certain" paralogs removed (2)	2.E-04	1.E-05	3.E-03	5.E-79	100 BS	90 BS
Taxa with high LB scores removed (3)	6.E-06	2.E-06	1.E-62	2.E-07	100 BS	94 BS
Genes with high LB scores removed (4)	1.E-02	2.E-04	2.E-32	3.E-47	100 BS	94 BS
Choanoflagellate-only outgroup (5)	4.E-45	1.E-04	1.E-05	6.E-104	100 BS	68 BS
All outgroups removed (22)	N/A	3.E-55	N/A	N/A	N/A	84 BS
Slowest evolving half of genes (6)	2.E-78	2.E-11	2.E-04	1.E-05	100 BS/100 PP	95 BS/100 PP
Genes with lowest half of RCFV values (7)	3.E-06	2.E-57	9.E-47	2.E-03	100 BS	53 BS
Heterogeneous genes removed (8)	7.E-52	2.E-49	9.E-05	3.E-05	100 BS	90 BS
Genes with high LB scores removed (9)	9.E-41	9.E-05	2.E-04	2.E-66	99 BS	90 BS
Taxa with high LB scores removed (10)	7.E-22	5.E-61	9.E-06	7.E-06	100 BS	88 BS
Choanoflagellate-only outgroup (11)	1.E-03	3.E-36	1.E-02	6.E-104	92 BS	82 BS
All outgroups removed (23)	N/A	2.E-04	N/A	N/A	N/A	92 BS
"Certain" and "uncertain" paralogs (12)	6.E-39	1.E-05	4.E-44	4.E-05	100 BS	99 BS
Taxa with high LB scores removed (13)	6.E-05	8.E-06	6.E-30	8.E-09	100 BS	99 BS
Genes with high LB scores removed (14)	8.E-105	9.E-05	1.E-45	5.E-102	99 BS	97 BS
Choanoflagellate-only outgroup (15)	5.E-59	5.E-39	7.E-05	1.E-72	100 BS	66 BS
All outgroups removed (24)	N/A	2.E-07	N/A	N/A	N/A	92 BS
Slowest evolving half of genes (16)	1.E-04	1.E-52	1.E-11	1.E-07	100 BS/100 PP	94 BS/100 PP
Genes with lowest half of RCFV values (17)	5.E-09	1.E-68	5.E-07	3.E-72	100 BS	93 BS
Heterogeneous genes removed (18)	7.E-07	4.E-10	6.E-77	8.E-09	99 BS	100 BS
Genes with high LB scores removed (19)	2.E-46	1.E-08	2.E-04	1.E-57	100 BS	96 BS
Taxa with high LB scores removed (20)	1.E-03	3.E-92	1.E-66	3.E-113	100 BS	100 BS
Choanoflagellate-only outgroup (21)	2.E-61	1.E-31	2.E-04	4.E-40	100 BS	77 BS
All outgroups removed (25)	N/A	9.E-05	N/A	N/A	N/A	96 BS
Philippe et al. (7) Maximum likelihood	0.001	0.010	0.008	0.075	93 BS	42 BS
Philippe et al. (7) Bayesian inference	—	—	—	—	N/A	99 PP

Dataset numbers are as in Fig. 2.

to result in different ctenophore placement. Maximum-likelihood and Bayesian inference using the CAT-GTR model of the least saturated datasets (datasets 6 and 16 in Fig. 2 and Fig. S7) recovered identical basal relationships as our other analyses (Fig. 3 and Figs. S1*F*, S3*D*, and S5 *B* and *C*), also indicating homoplasy and model choice did not bias results. Given the consistency among our analyses that were designed to have different levels of potential biases, we conclude that the ctenophore-sister hypothesis is robust to systematic errors.

Rather than focusing on long branches, fast evolving genes, or model misspecification as influencing the position of ctenophores, the individual genes underlying datasets that resulted in a sister relationship between ctenophores and cnidarians (7, 9) should be the focus of identifying problems with phylogenetic reconstruction. A benefit of phylogenomic datasets is that multiple gene classes and many parts of the genome are analyzed. As such, phylogenomic datasets should not rely too heavily on a single gene class. Past molecular studies that found support for Coelenterata and the Porifera-sister (7, 9) hypotheses appear to have been strongly affected by a disproportionate reliance (i.e., > 50%) on ribosomal protein genes. Nosenko et al. (7) and Philipe et al. (9) found support for ctenophores sister to cnidarians, but Nosenko et al. (7) recovered ctenophores sister to all other animals when ribosomal proteins were excluded. Ribosomal protein

datasets from these studies are less saturated than datasets assembled here based on linear regression of patristic distance versus uncorrected genetic distance (Fig. S7) (7–9, 25). This lower mutational saturation has been the primary rational for emphasizing ribosomal genes when inferring deep animal relationship (7). However, standard measurements of sequence saturation (7, 8, 25) average across the length of the sequence. Thus, a sequence with a few variable, highly saturated sites may appear less saturated than a sequence with numerous variable sites but less saturation per site. Furthermore, extremely low mutation rates can indicate selection and result in too little phylogenetic information, both of which could lead to the inference of incorrect relationships (33). Our maximum-likelihood analysis of Philipe et al.'s (9) dataset with ribosomal genes removed recovered support for ctenophores as sister to all other animals (Fig. S5*D*). Basal relationships were poorly resolved in the Bayesian analysis, which may be a result of too few characters, but ctenophores and cnidarians were not recovered as sister (Fig. S5*E*). Notably, a study focused only on myxozoan cnidarians (34) used the same matrix as Philippe et al. (9), but added two highly divergent cnidarians and recovered ctenophores sister to all other animals. Ribosomal protein genes have previously been identified as a potential source of phylogenetic error (32), and the above indicates that ctenophores sister to cnidarians as in Philippe et al. (9) was caused by either limited cnidarian taxon sampling, misleading signal in ribosomal genes, or both. More work is needed to assess saturation, possible convergent evolution, and selective pressures of ribosomal proteins in ctenophores, sponges, placozoans, and cnidarians. However, differences in topologies when ribosomal proteins are included or excluded strongly imply a misleading signal in ribosomal protein genes. Put simply, it appears highly improbable that all genes other than ribosomal protein genes could be recovering an incorrect phylogeny.

Sponges Are Monophyletic. Sponge monophyly, although less controversial than the phylogenetic positions of ctenophores, cnidarians, and placozoans remains an important question as several studies have supported sponge paraphyly (7, 22–24). In regards to inferring the characteristics of the metazoan ancestor, sponge paraphyly, coupled with sponges being at the base of the metazoan phylogeny, is an attractive hypothesis that implies the metazoan ancestor was sponge-like. However, sponge monophyly was recovered in all of our analyses and best supported when the strictest orthology criteria were applied with TreSpEx (Fig. 3 and Table 1), which also removes sequences resulting from sample contamination (e.g., endosymbionts).This observation suggests that spurious paralogs or sequence contamination may have been a source of error when sponges were found paraphyletic, but datasets that have recovered sponge paraphyly were also much smaller than those analyzed here (e.g., refs. 7, 22–24). Sponge monophyly and the well-supported ctenophore-sister hypothesis complicates inferring the ancestral condition of metazoan and other major metazoan groups (e.g., Placozoa + Cnidaria + Bilateria) because many sponge characteristics are likely apomorphic traits. Our robust support of sponge monophyly agrees with morphology (35) and most other large molecular datasets (5, 6, 8–10).

Conclusions

For more than a century, sponges were traditionally considered sister to all other extant metazoans because unlike ctenophores, cnidarians, and bilaterians, they lack true tissues and body symmetry (36, 37). Sponges also possess choanocyte cells that are similar in morphology to choanoflagellates, the sister group to metazoans (36). However, Mah et al. (38) found that homology between choanoflagellates and sponge choanocytes is not as definitive as previously assumed. Similarly, some authors have argued that Placozoans are sister to all other animals because they lack neural and muscular systems and also share similarities in mitochondrial genome size with choanoflagellaes (2, 39). A common theme of these two hypotheses is the placement of morphologically simple animals near the base of the animal tree, but complexity is not a good proxy for metazoan evolution (40). Challenges to long-held viewpoints of morphological complexity and assumed improbability of convergent evolution must not be dismissed simply because they seem unlikely at face value, especially considering a growing body of evidence that supports convergent evolution of many animal traits including neurons (5, 6, 41–43). Furthermore, the Porifera-sister hypothesis lacks critical evaluation and homology of many characters in these taxa still need thorough analysis (44). Overall, findings presented here robustly support ctenophores as sister to all remaining animals.

Methods

Taxon and Character Sampling. Taxon sampling included previously available data and eight new transcriptomes from two choanoflagellates, three glass sponges (Hexactinellida), two demosponges, and a deep-sea cnidarian (Scyphozoa) (Table S1). Briefly, RNA was extracted, reverse-transcribed, and amplified using the SMART kit (Clontech), and sequenced on an Illumina HiSeq (*SI Methods*). Raw or assembled transcriptome data for 68 additional species were retrieved from public databases (Table S1).

Raw Illumina transcriptome data were digitally normalized using normalize-by-median.py (45) with a k-mer size of 20, a desired coverage of 30, and four hash tables with a lower bound of 2.5×10^9. Normalized Illumina reads were assembled using default parameters in Trinity v20131110 (46). Raw 454 transcriptome data were assembled with Newbler (47).

Orthology Determination and Data Filtering. Putative orthologs were determined for each species using HaMStR v13.2 (48) using the model organism core ortholog set. OGs, determined by HaMStR, were further processed using a custom pipeline that filtered OGs with too much missing data (i.e., OGs with less than 37 species), aligned sequences, and filtered potential paralogs (https://github.com/kmkocot) (*SI Methods*).

TreSpEx (25), which requires individual gene trees for each OG, was used to identify putative paralogs and exogenous contamination missed by our initial orthology inference approach. Gene trees were inferred with RAxML v.8.0.2 (49) with 100 rapid bootstrap replicates followed by a full maximum-likelihood inference; each tree was inferred with the LG+Γ model, which was by far the most common best-fitting model when the complete dataset was partitioned (see below). Paralogs in the initial dataset were detected with TreSpEx using the automated BLAST method and the prepackaged *Capitella teleta* and *Helobdella robusta* BLAST databases. This method identified two classes of paralogs: "certain" or sequences that are high-confidence paralogs, and "uncertain" or sequences that are potential paralogs. From the initial, 251 gene dataset, we created one dataset by removing certain paralogs and another dataset with both certain and uncertain paralogs pruned (Fig. 2); after pruning, OGs with fewer than 37 taxa were removed. LB scores (25) were calculated for each taxon and OG with TreSpEx. Following Struck (25), these values were plotted in R (Fig. S6) (50), and outliers were identified as taxa or genes that could cause LBA artifacts. After removal of taxa and genes, each OG was ranked by evolutionary rate with a custom python script (https://github.com/nathanwhelan) following Telford et al. (51). Datasets with only the slowest half of remaining genes were then generated to assess if fast evolving homoplasious genes were biasing inferences (Fig. 2).

We used BaCoCa (52) and two metrics (χ^2-test of heterogeneity and relative composition frequency variability; RCFV) (53) to identify genes with amino acid compositional heterogeneity. Some datasets were further filtered by removing non-choanoflagellate outgroups and all outgroup taxa to determine if outgroup choice affects inferred relationships. Saturation of each filtered dataset with full outgroup sampling was explored with TreSpEx and plotted in R (Fig. S7) to provide a further metric to compare datasets.

Phylogenetics. In addition to removing compositionally heterogeneous genes from some datasets, two approaches were used to handle site-heterogeneity: (*i*) partitioning schemes for each dataset and associated protein substitution models were determined using the relaxed clustering method in PartitonFinder (54) with 20% clustering and the corrected Akaike information criterion; (*ii*) a site-heterogeneous mixture model, CAT-GTR+Γ, was used in PhyloBayes (26). Maximum-likelihood topologies were inferred with RAxML using partitions as indicated by ParitionFinder, associated best-fit substitution models, and the gamma parameter to model rate-heterogeneity. Nodal support was measured with 100 fast bootstrap replicates. Phylobayes analyses were run with two chains until the maxdiff statistic between chains was below 0.3 as measured by bpcomp (26). Convergence was also assessed with tracecomp (26) to ensure each parameter had a maximum discrepancy between chains of less than 0.3 and an effective sample size of at least 50. Computational demands and convergence

issues prevented us from using the CAT-GTR+Γ model for most datasets. Therefore, Bayesian phylogenies are only reported for the two analyses of the slowest evolving half of OGs. Leaf-stability indices (55) for each taxon were measured in PhyUtility (56) to identify potentially unstable taxa in each dataset.

Maximum-likelihood and Bayesian inference trees were also inferred from the Philippe et al. (9) dataset with ribosomal proteins removed (i.e., 67 of 128 genes) to determine if a single gene class-biased phylogenetic inference. Ribosomal proteins were filtered from the original dataset following data matrix annotations (9) and the matrix was split into individual genes for model testing using a custom R script (https://github.com/nathanwhelan).

The AU test (31) was used to determine if a priori hypotheses of basal metazoan relationships could be rejected (Fig. 1). Topological constraints were enforced in RAxML and the most likely tree given this constraint was inferred with the same partitioning scheme and models used for unconstrained phylogenetic inference. Per site log-likelihoods for trees were calculated in RAxML and AU tests were performed in Consel (57).

ACKNOWLEDGMENTS. We thank members of the Molette Biology Laboratory for Environmental and Climate Change Studies at Auburn University for help with bioinformatics and data collection, especially Damien Waits. This work was made possible in part by a grant of high-performance computing resources and technical support from the Alabama Supercomputer Authority and was supported by the US National Aeronautics and Space Administration (Grant NASA-NNX13AJ31G) and in part by National Science Foundation (Grant 1146575). This is Molette Biology Laboratory Contribution 36 and Auburn University Marine Biology Program Contribution 128.

1. Dohrmann M, Wörheide G (2013) Novel scenarios of early animal evolution—Is it time to rewrite textbooks? *Integr Comp Biol* 53(3):503–511.
2. Dellaporta SL, et al. (2006) Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. *Proc Natl Acad Sci USA* 103(23): 8751–8756.
3. Dunn CW, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452(7188):745–749.
4. Hejnol A, et al. (2009) Assessing the root of bilaterian animals with scalable phylogenomic models. *Proc Biol Sci* 276(1802):4261–4270.
5. Moroz LL, et al. (2014) The ctenophore genome and the evolutionary origins of neural systems. *Nature* 510(7503):109–114.
6. Ryan JF, et al.; NISC Comparative Sequencing Program (2013) The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* 342(6164):1242592.
7. Nosenko T, et al. (2013) Deep metazoan phylogeny: When different genes tell different stories. *Mol Phylogenet Evol* 67(1):223–233.
8. Philippe H, et al. (2011) Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol* 9(3):e1000602.
9. Philippe H, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19(8):706–712.
10. Pick KS, et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol* 27(9):1983–1987.
11. Felsenstein J (1978) Cases in which parsimony and compatability methods will be positively misleading. *Syst Zool* 27(4):401–410.
12. Boussau B, et al. (2014) Strepsiptera, phylogenomics and the long branch attraction problem. *PLoS ONE* 9(10):e107709.
13. Straub SCK, et al. (2014) Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Mol Phylogenet Evol* 80:169–185.
14. Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol* 46(3):239–257.
15. Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: The beginning of incongruence? *Trends Genet* 22(4):225–231.
16. Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol* 30(1):197–214.
17. Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* 7(Suppl 1):S4.
18. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci* 17:57–86.
19. Philippe H, Delsuc F, Brinkmann H, Lartillot N (2005) Phylogenomics. *Annu Rev Ecol Evol Syst* 36:541–562.
20. Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLOS Comput Biol* 5(1):e1000262.
21. Gabaldón T (2008) Large-scale assignment of orthology: Back to phylogenetics? *Genome Biol* 9(10):235.
22. Borchiellini C, et al. (2001) Sponge paraphyly and the origin of Metazoa. *J Evol Biol* 14(1):171–179.
23. Sperling EA, Pisani D, Peterson KJ (2007) Poriferan paraphyly and its implications for Precambrian paleobiology. *Geol Soc Lond Spec Publ* 286:355–368.
24. Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Mol Biol Evol* 26(10):2261–2274.
25. Struck TH (2014) TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol Bioinform Online* 10:51–67.
26. Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* 62(4):611–615.
27. van Soest RWM (1984) Deficient *Merlia normani* Kirkpatrick, 1908, from the Curacao reefs, with a discussion on the phylogenetic interpretation of sclerosponges. *Contrib Zool* 54(2):211–219.
28. Gazave E, et al. (2012) No longer Demospongiae: Homoscleromorpha formal nomination as a fourth class of Porifera. *Hydrobiologia* 687(1):3–10.
29. Dohrmann M, Janussen D, Reitner J, Collins AG, Wörheide G (2008) Phylogeny and evolution of glass sponges (Porifera, Hexactinellida). *Syst Biol* 57(3):388–405.
30. Voigt O, Adamski M, Sluzek K, Adamska M (2014) Calcareous sponge genomes reveal complex evolution of α-carbonic anhydrases and two key biomineralization enzymes. *BMC Evol Biol* 14(1):230.
31. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51(3):492–508.
32. Bleidorn C, et al. (2009) On the phylogenetic position of Myzostomida: Can 77 genes get it wrong? *BMC Evol Biol* 9(1):150.
33. Edwards SV (2009) Natural selection and phylogenetic analysis. *Proc Natl Acad Sci USA* 106(22):8799–8800.
34. Nesnidal MP, Helmkampf M, Bruchhaus I, El-Matbouli M, Hausdorf B (2013) Agent of whirling disease meets orphan worm: phylogenomic analyses firmly place Myxozoa in Cnidaria. *PLoS ONE* 8(1):e54576.
35. Ax P (1996) *Multicellular Animals: A New Approach to the Phylogenetic Order in Nature* (Springer, Berlin).
36. Nielsen C (2008) Six major steps in animal evolution: Are we derived sponge larvae? *Evol Dev* 10(2):241–257.
37. Srivastava M, et al. (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* 466(7307):720–726.
38. Mah JL, Christensen-Dalsgaard KK, Leys SP (2014) Choanoflagellate and choanocyte collar-flagellar systems and the assumption of homology. *Evol Dev* 16(1):25–37.
39. Osigus H-J, Eitel M, Bernt M, Donath A, Schierwater B (2013) Mitogenomics at the base of Metazoa. *Mol Phylogenet Evol* 69(2):339–351.
40. Halanych KM (1999) Metazoan phylogeny and the shifting comparative framework. *Recent Developments in Comparative Endocrinology and Neurobiology*, eds Roubos EW, Wendelaar-Bonga SE, Vaudry H, De Loof A (Shaker, Maastrict, The Netherlands), pp 3–7.
41. Dunn CW, Giribet G, Edgecombe GD, Hejnol A (2014) Animal phylogeny and its evolutionary implications. *Annu Rev Ecol Evol Syst* 45:371–395.
42. Liebeskind BJ, Hillis DM, Zakon HH (2015) Convergence of ion channel genome content in early animal evolution. *Proc Natl Acad Sci USA* 112(8):E846–E851.
43. Moroz LL (2015) Convergent evolution of neural systems in ctenophores. *J Exp Biol* 218(Pt 4):598–611.
44. Halanych KM (2015) The ctenophore lineage is older than sponges? That cannot be right! Or can it? *J Exp Biol* 218(Pt 4):592–597.
45. Brown T, Howe C, Zhang A, Pyrkosz Q, Brom AB (2012) A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv:1203.4802.
46. Haas BJ, et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8):1494–1512.
47. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380.
48. Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* 9:157.
49. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
50. R Core Development Team (2014) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).
51. Telford MJ, et al. (2013) Phylogenomic analysis of echinoderm class relationships supports Asterozoa. *Proc Biol Sci* 281(1786):20140479.
52. Kück P, Struck TH (2014) BaCoCa—A heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Mol Phylogenet Evol* 70:94–98.
53. Zhong M, et al. (2011) Detecting the symplesiomorphy trap: A multigene phylogenetic analysis of terebelliform annelids. *BMC Evol Biol* 11(1):369.
54. Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A (2014) Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol Biol* 14:82.
55. Thorley JL, Wilkinson M (1999) Testing the phylogenetic stability of early tetrapods. *J Theor Biol* 200(3):343–344.
56. Smith SA, Dunn CW (2008) Phyutility: A phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24(5):715–716.
57. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246–1247.

Supporting Information

Whelan et al. 10.1073/pnas.1503453112

SI Methods

Transcriptome Sequencing. Frozen aliquots of *Acanthoeca spectabilis* (ATCC PRA-103) and *Salpingoeca pyxidium* (ATCC 50929) cultures were purchased from ATCC and a 10-μL aliquot of each culture was used for RNA extraction with the RNAqueous Micro RNA extraction kit (Life Technologies). Notably, *Acanthoeca spectabilis* was grown in a xenic culture with the bacteria *Klebsiella pneumoniae* (ATCC 700831) and *Enterobacter aerogenes* (ATCC 13048) and *Salpingoeca pyxidium* was grown in a xenic culture with *Klebsiella pneumoniae* (ATCC 700831). Tissue cuttings were taken from the apical portion of the sponges *Latrunculia apicalis*, *Kirpatrickia variolosa*, *Hyalonema populiferum*, and *Rossella fibulata*. One "orb" of the unusual hexactinellid *Sympagella nux* was taken for RNA extractions. A tissue clip was taken from the margin of the bell of the deep-sea cnidarian *Periphylla periphylla*. RNA was extracted from all sponges and *P. periphylla* with TRIzol (invitrogen). The Qiagen RNeasy kit and on-column DNase digestion was used to purify TRIzol extracted RNA. cDNA libraries of all eight taxa were constructed with the SMART cDNA library construction kit (Clontech Laboratories). cDNA library construction followed manufacturer's instructions except that the provided 3′ oligo was replaced with the Cap-Trsa-CV oligo following Meyer et al. (1). The Advantage 2 PCR system (Clontech) was used to amplify full-length cDNA using a minimum number of PCR cycles (~17–21). Amplified cDNA was shipped to Hudson Alpha Institute for Biotechnology in Huntsville, AL for Illumina library preparation and 2 × 100-base pair (*A. spectabilis*, *S. pyxidium*, *K. variolosa*, *L. apicalis*) or 2 × 150-base pair (*H. populiferum*, *R. fibulata*, *S. nux*, *P. periphylla*) paired-end sequencing on an Illumina HiSEq. 2500 using ~one-eighth lane per taxon.

Initial Orthology Inference Procedure. Before any filtering steps, all sequences were backed up and new line characters were removed from interleaved sequences with nentferner.pl (https://github.com/mptrsen/HaMStRad). All sequences shorter than 50 bp were then removed. Any gene with fewer than 37 taxa (~47%) was also removed from the dataset to minimize missing data. HaMStR sometimes pulls two identical sequences from a single taxon for any given gene so the script uniqHaplo.pl (raven.iab.alaska.edu/~ntakebay/teaching/programming/perl-scripts/uniqHaplo.pl) was used to remove duplicate sequences. As an initial data-filtering step before sequence alignment, if an ambiguous amino acid (i.e., an X) was present in the first 20 or last 20 amino acids then the 5′ or 3′ end, respectively, was trimmed to this X as it may indicate sequencing errors.

Sequences were aligned using MAFFT (2) with the "auto" and "localpair" parameters and 1,000 maximum iterations. MAFFT outputs sequences in an interleaved format so nentferner.pl was again used to remove newline characters from sequences. Unalignable regions were then removed with aliscore (3) and alicut (4). Alignment columns with only gaps were subsequently removed, and any gene with an alignment less than 50-bp long after trimming was removed. For each gene, a custom javascript (https://github.com/kmkocot) then removed any sequence that did not overlap other sequences by at least 20 amino acids to minimize the number of end gaps in alignments that would likely not add to information content in phylogenetic inference. Finally, any gene that had fewer than 37 taxa after alignment and trimming was removed to further minimize missing data.

The last step removed putative paralogs with PhyloTreePrunner (5), which is a tree-based method. Gene trees for each OG were inferred using FastTreeMP (6) with the "slow" and "gamma" parameters. Gene trees and corresponding OGs were input into PhyloTreePrunner and inferred paralogs were removed from each OG. Alignments of each OG were concatenated using FASconCAT v1.0 (7). An automated wrapper script for this procedure is available from https://github.com/kmkocot.

OG Annotation. To ensure no one gene class dominated our matrices, OGs were annotated with PFAM domains (8) using the Trinotate pipeline (trinotate.github.io). An annotation spreadsheet was placed on figshare (doi: 10.6084/m9.figshare.1334306)

1. Meyer E, et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* 10:219.
2. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.
3. Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Syst Biol* 58(1):21–34.
4. Kück P (2009) *ALICUT: A Perlscript Which Cuts ALISCORE Identified RSS* (Department of Bioinformatics, Zoologisches Forschungsmuseum, Bonn, Germany).
5. Kocot KM, Citarella MR, Moroz LL, Halanych KM (2013) PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol Bioinform Online* 9(3940):429–435.
6. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5(3):e9490.
7. Kück P, Meusemann K (2010) FASconCAT: Convenient handling of data matrices. *Mol Phylogenet Evol* 56(3):1115–1118.
8. Finn RD, et al. (2014) Pfam: the protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230.

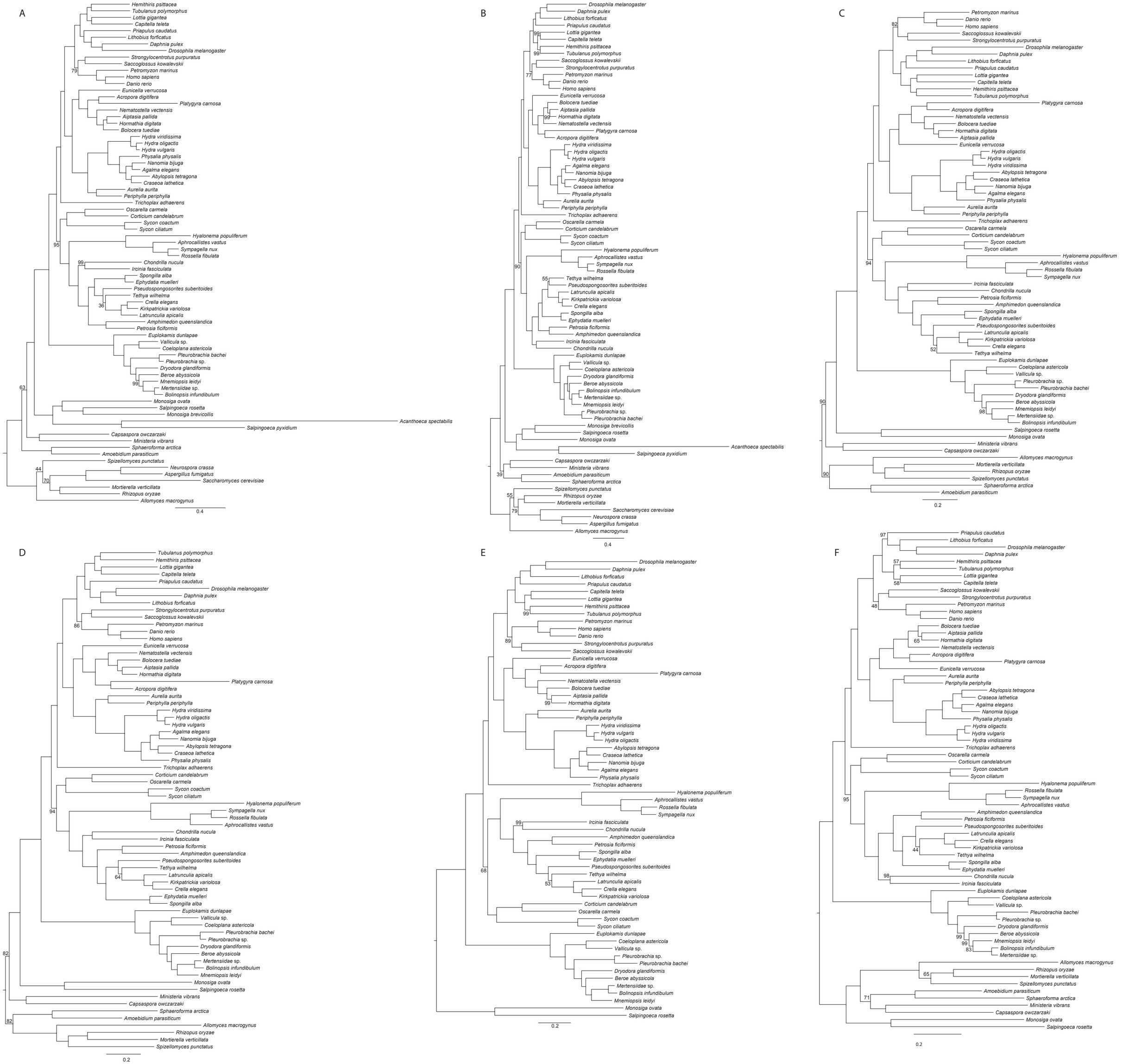

Fig. S1. Maximum-likelihood topologies. (*A*) Full dataset (Fig. 2, dataset 1). (*B*) Dataset with certain paralogs removed (Fig. 2, dataset 2). (*C*) Dataset with certain paralogs and taxa with high LB scores removed (Fig. 2, dataset 3). (*D*) Dataset with certain paralogs and taxa and gene with high LB scores removed (Fig. 2, dataset 4). (*E*) Dataset with certain paralogs, taxa and genes with high LB scores, and all outgroups except choanoflagellates removed (Fig. 2, dataset 5). (*F*) Dataset with slowest evolving half of genes after certain paralogs and taxa and genes with high LB scores were removed (Fig. 2, dataset 6). Bootstrap values for each node are 100% unless otherwise noted.

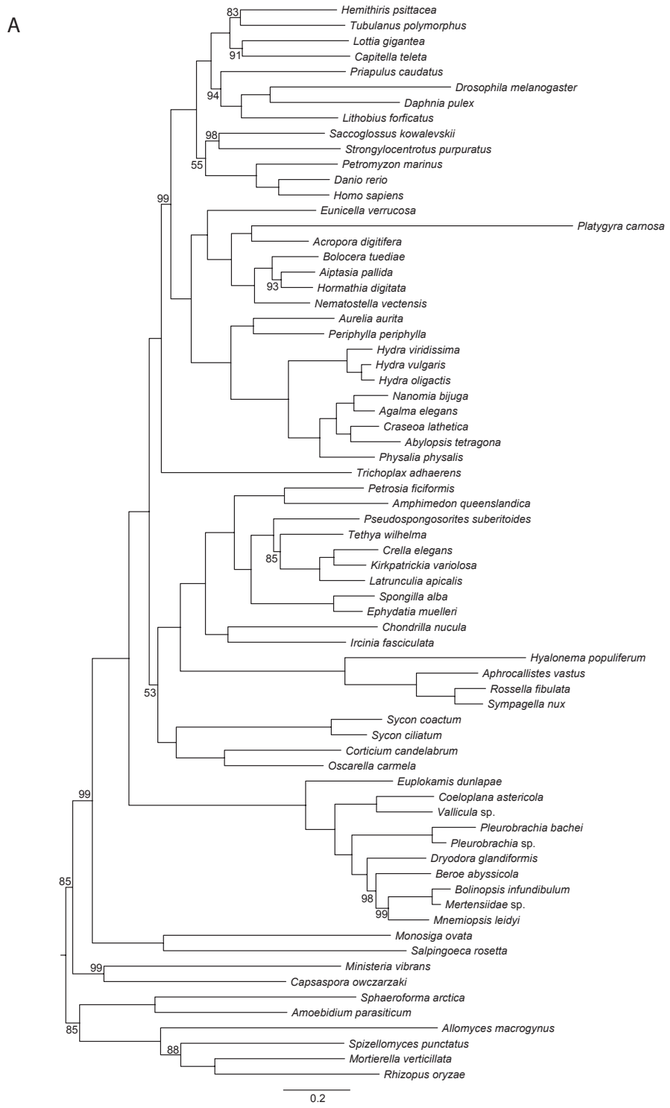

Fig. S2. Maximum-likelihood topologies. (*A*) Dataset with only genes with the lowest 50% of RCFV values after certain paralogs and taxa and genes with high LB scores were removed (Fig. 2, dataset 7). (*B*) Dataset with certain paralogs and heterogeneous genes removed (Fig. 2, dataset 8). (*C*) Dataset with certain paralogs, heterogeneous genes, and genes with high LB scores removed (Fig. 2, dataset 9). (*D*) Dataset with certain paralogs, heterogeneous genes, and taxa and genes with high LB scores removed (Fig. 2, dataset 10). (*E*) Dataset with certain paralogs, heterogeneous genes, taxa and genes with high LB scores, and all outgroups except choanoflagellates removed (Fig. 2, dataset 11). (*F*) Dataset with certain and uncertain paralogs removed (Fig. 2, dataset 12). Bootstrap values for each node are 100% unless otherwise noted.

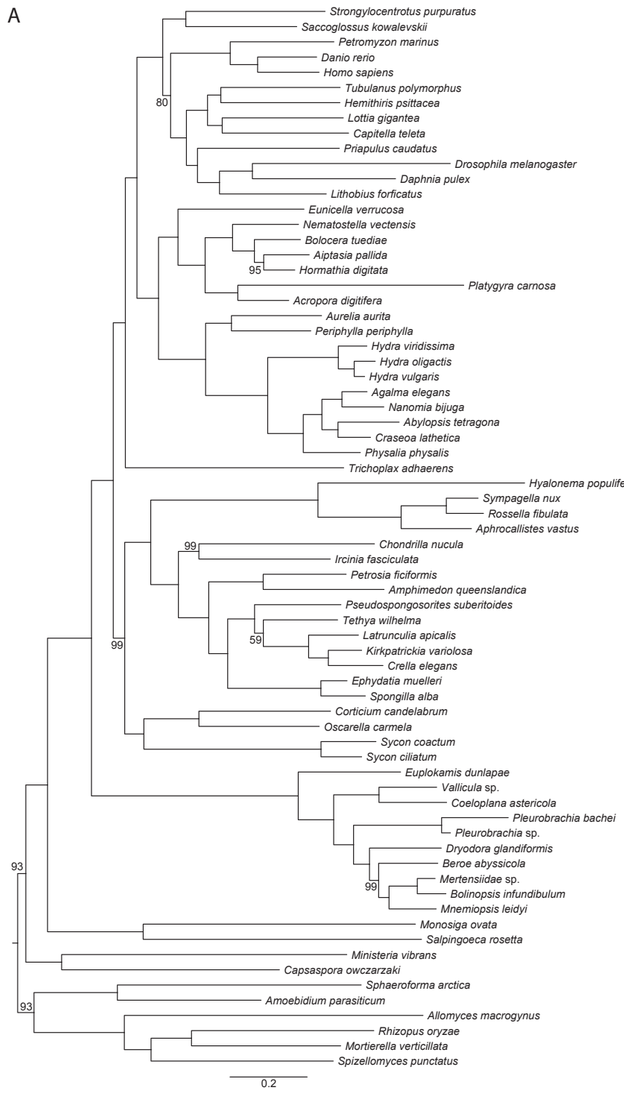

Fig. S3. Maximum-likelihood topologies. (*A*) Dataset with certain and uncertain paralogs and taxa with high LB scores removed (Fig. 2, dataset 13). (*B*) Dataset with certain and uncertain paralogs and taxa and gene with high LB scores removed (Fig. 2, dataset 14). (*C*) Dataset with certain and uncertain paralogs, taxa and genes with high LB scores, and all outgroups except choanoflagellates removed (Fig. 2, dataset 15). (*D*) Dataset with slowest evolving half of genes after certain and uncertain paralogs and taxa and genes with high LB scores were removed (Fig. 2, dataset 16). (*E*) Dataset with only genes with the lowest 50% of RCFV values after certain and uncertain paralogs and taxa and genes with high LB scores were removed (Fig. 2, dataset 17). (*F*) Dataset with certain and uncertain paralogs and heterogeneous genes removed (Fig. 2, dataset 18). Bootstrap values for each node are 100% unless otherwise noted.

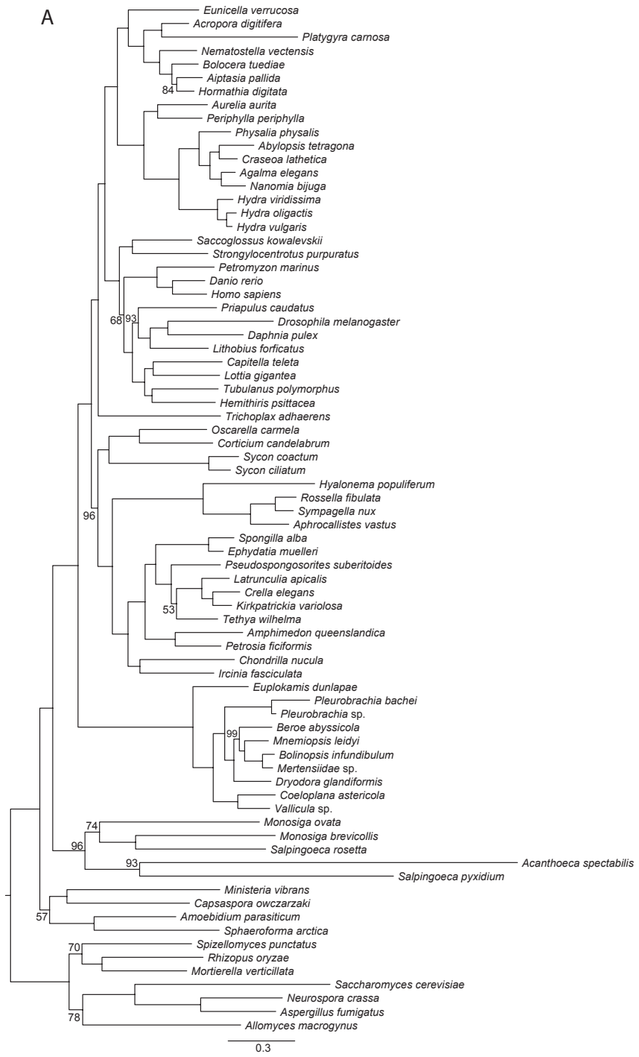

Fig. S4. Maximum-likelihood topologies. (*A*) Dataset with certain and uncertain paralogs, heterogeneous genes, and genes with high LB scores removed (Fig. 2, dataset 19). (*B*) Dataset with certain and uncertain paralogs, heterogeneous genes, and taxa and genes with high LB scores removed (Fig. 2, dataset 20). (*C*) Dataset with certain and uncertain" paralogs, heterogeneous genes, taxa and genes with high LB scores, and all outgroups except choanoflagellates removed (Fig. 2, dataset 21). (*D*) Dataset with certain paralogs, taxa and genes with high LB scores, and all outgroups removed (Fig. 2, dataset 22). (*E*) Dataset with certain paralogs, heterogeneous genes, taxa and genes with high LB scores, and all outgroups removed (Fig. 2, dataset 23). (*F*) Dataset with certain and uncertain paralogs, taxa and genes with high LB scores, and all outgroups removed (Fig. 2, dataset 24). Bootstrap values for each node are 100% unless otherwise noted.

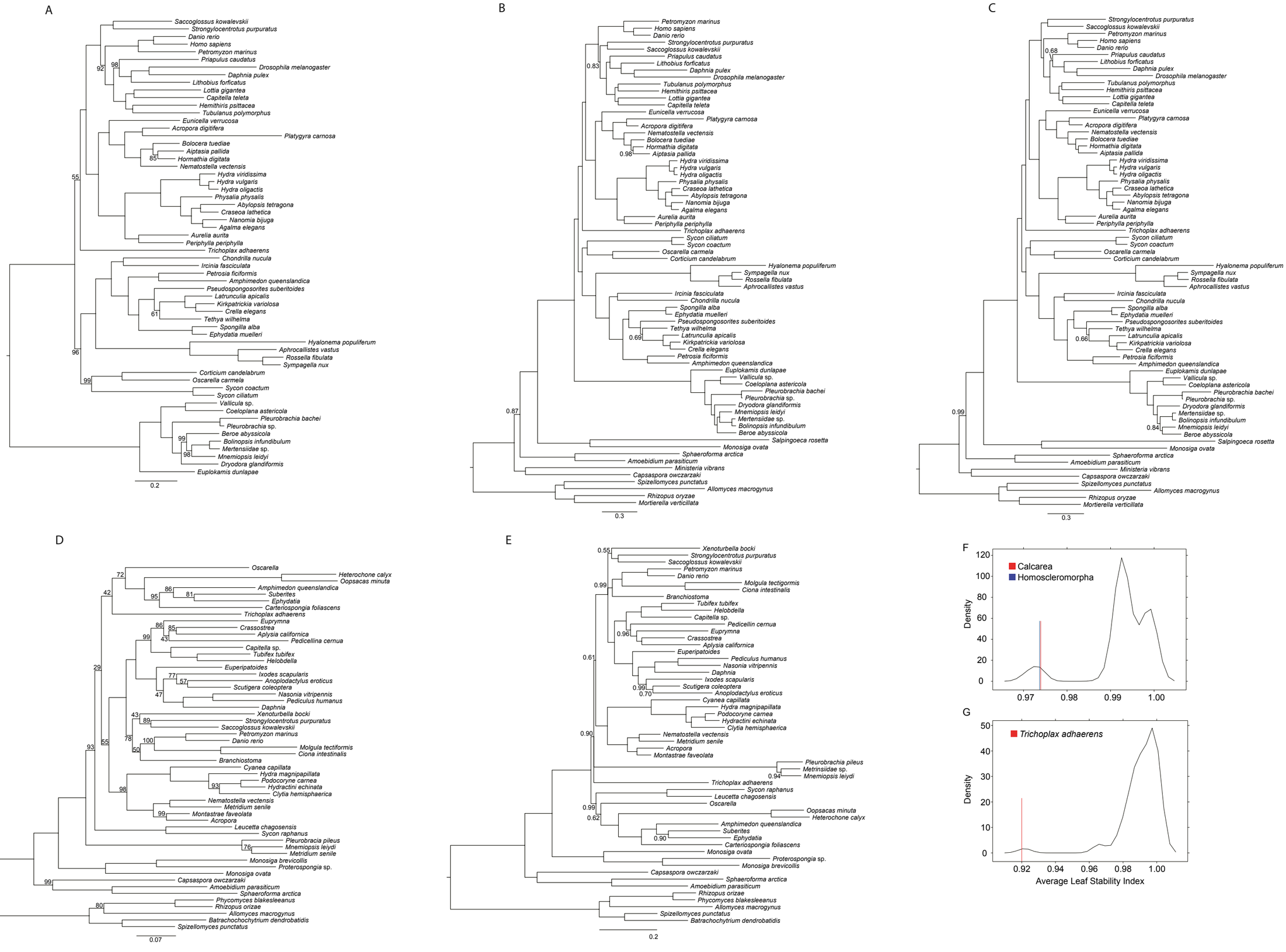

Fig. S5. (*A*) Maximum-likelihood topology of dataset with certain and uncertain paralogs, heterogeneous genes, taxa and genes with high LB scores, and all outgroups removed (Fig. 2, dataset 25). (*B*) Bayesian inference topology of dataset of slowest evolving half of genes after certain paralogs and taxa and genes with high LB scores were removed (Fig. 2, dataset 6). (*C*) Bayesian inference topology of dataset of slowest evolving half of genes after certain and uncertain paralogs and taxa and genes with high LB scores were removed (Fig. 2, dataset 16). (*D*) Maximum-likelihood phylogeny of Philippe et al. (1) dataset with ribosomal protein genes removed. Chimeric taxa are labeled only by their genus name as in the original publication. (*E*) Bayesian inference topology of Philippe et al. (1) dataset with ribosomal protein genes removed. Chimeric taxa are labeled only by their genus name as in the original publication. Bootstrap values or posterior probabilities for each node are 100% unless otherwise noted. (*F*) Density plot of average leaf stability indices for each taxon across all datasets. (*G*) Density plot of average leaf stability indices for each taxon in datasets without outgroups.

1. Philippe H, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19(8):706–712.

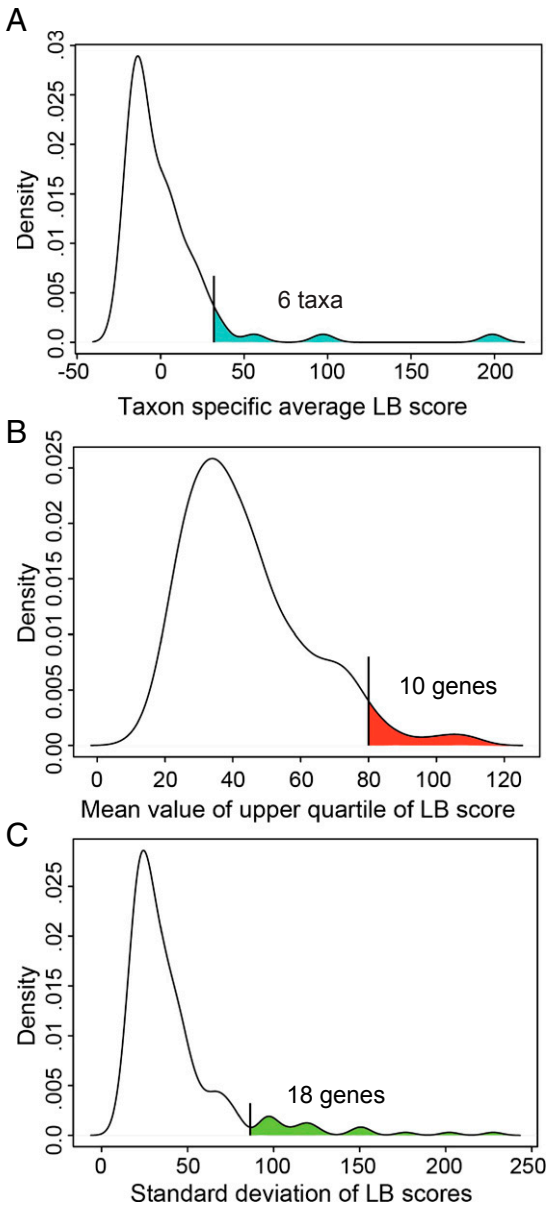

Fig. S6. Density plots of (*A*) average LB scores for each taxon, (*B*) average upper quartile LB score for each OG, and (*C*) SD of LB score for each OG. Shaded areas include taxa or genes that were considered to have "high" LB scores and were trimmed from respective datasets (see Fig. 2).

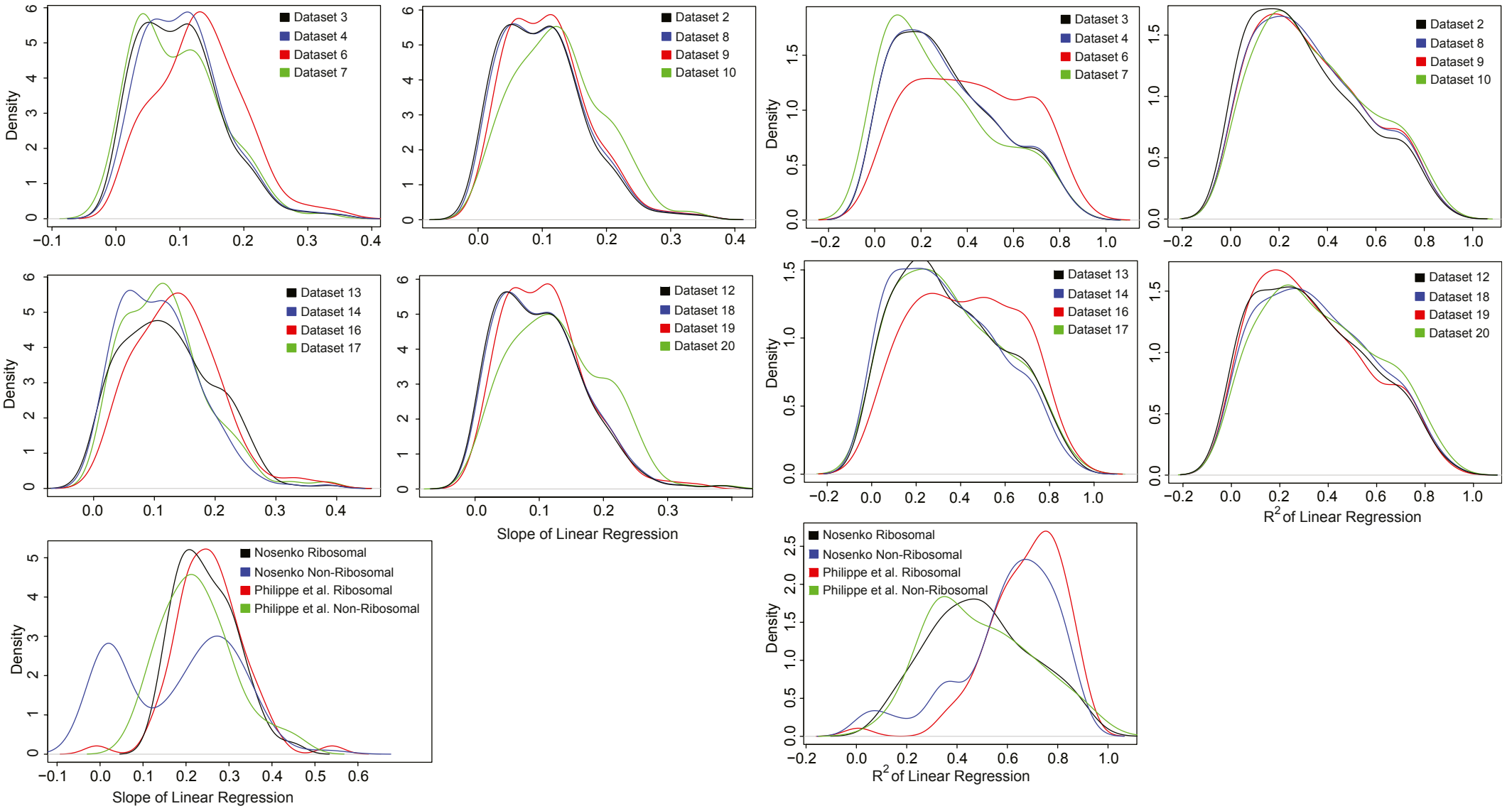

Fig. S7. Density plots of gene specific slope and R^2 of linear regression between patristic and uncorrected pairwise distances for each dataset. Datasets with higher values for each metric are less saturated. Dataset numbers are referenced as in Fig. 2.

Table S1. Raw data for each species used in orthology searches

Species	Sequence type	No. of reads, transcripts or predicted proteins	Source or collection locality	NCBI or other accession
Fungi				
Allomyces macrogynus	Whole genome	19,446 proteins	Broad Institute	www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html
Aspergillus fumigatus†	Whole genome	11,485 proteins	InParanoid database	
Mortierella verticillata	Whole genome	12,569 proteins	Broad	www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html
Neurospora crassa†	Whole genome	9,822 proteins	InParanoid database	
Rhizopus oryzae	Whole genome	16,971 proteins	InParanoid database	
*Saccharomyces cerevisiae**	Whole genome	6,590 proteins	InParanoid database	
Spizellomyces punctatus	Whole genome	9,424 proteins	Broad Institute	www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html
Ichthyosporea				
Amoebidium parasiticum	Illumina	105,237 transcripts	Broad Institute	www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html
Capsaspora owczarzaki	Whole genome	101,23 proteins	Broad Institute	www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html
Ministeria vibrans	Illumina	19,953,296	NCBI SRA	SRR343051
Sphaeroforma arctica	Whole genome	18,730 proteins	Broad Institute	www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html
Choanoflagellata				
Acanthoeca spectabilis*	Illumina	41,408,758 reads	ATCC	SRR1915695
*Monosiga brevicollis**	Whole genome	10336 proteins	Joint Genome Institute	
Monosiga ovata	Sanger	29,495 sequences	dbEST	
Salpingoeca pyxidium*	Illumina	34,177,888 reads	ATCC	SRR1915694
Salpingoeca rosetta	Whole genome	11,731 proteins	Broad Institute	www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html
Placozoa				
Trichoplax adhaerans	Whole genome	11,179 proteins	Joint Genome Institute	
Cnidaria				
Acropora digitifera	Whole genome	33,366 proteins	OIST Marine genomics Unit	
Abylopsis tetragona	Illumina	43150352 reads	NCBI SRA	SRR871525
Bolocera tuediae	454	546,903 reads	NCBI SRA	SRR504347
Agalma elegans	Illumina	107,996,364 reads	NCBI SRA	SRR871526
Nanomia bijuga	Illumina	105,066,886 reads	NCBI SRA	SRR871527
Aiptasia pallida	Illumina	536616844 reads	NCBI SRA	SRR696721; SRR696732; SRR696745
Aurelia aurita (strain Roscoff)	454	1,099,012 reads	NCBI SRA	SRR040475; SRR040476; SRR040477; SRR040478; SRR040479
Eunicella verrucosa	Illumina	70,071,835 reads	NCBI SRA	SRR1324944; SRR1324945
Hydra oligactis	454	1,017,333 reads	NCBI SRA	SRR040466; SRR040467; SRR040468; SRR040469
Hydra viridissima	454	1,048,810 reads	NCBI SRA	SRR040470; SRR040471; SRR040472; SRR040473
Hormathia digitata	454	546,846 reads	NCBI SRA	SRR504348
Hydra vulgaris	Sanger	167,982 sequences	dbEST	
Nematostella vectensis	Whole genome	27,273 proteins	Joint Genome Institute	
Periphylla periphylla	Illumina	44, 443,566	45°55.33N 125°05.47 W	SRR1915828
Physalia physalis	Illumina	72,963,546 reads	NCBI SRA	SRR871528
Platygyra carnosus	Illumina	72,963,546 reads	NCBI SRA	SRR402974; SRR402975
Craseoa lathetica	Illumina	76,466,398 reads	NCBI SRA	SRR871529
Porifera				
Amphimedon queenslandica	Sanger	63,542 sequences	dbEST	
Aphrocallistes vastus	Illumina	preassembled	Evolution & Research Archive	
Chondrilla nucula	Illumina	preassembled	Dataverse Network	(1)
Corticium candelabrum	Illumina	preassembled	Dataverse Network	(1)
Crella elegans	Illumina	preassembled	Dryad	doi: 10.5061/dryad.50dc6/3
Ephydatia muelleri	Illumina	156,515,380 reads	NCBI SRA	SRR1041944
Hyalonema populiferum	Illumina	62,031,394 reads	43°55.376N 127°24.65W	SRR1916923
Ircinia fasciculata	Illumina	preassembled	Dataverse Network	(1)
Kirkpatrickia variolosa	Illumina	59,234,424 reads	77°54.228S 170°57.915E	SRR1916957

Table S1. Cont.

Species	Sequence type	No. of reads, transcripts or predicted proteins	Source or collection locality	NCBI or other accession
Latrunculia apicalis	Illumina	49,432,908 reads	82°54.228S 175°57.915 E	SRR1915755
Oscarella carmela	Illumina	Preassembled	www.compagen.org	
Petrosia ficiformis	Illumina	preassembled	Dataverse Network	(1)
Pseudospongosorites suberitoides	Illumina	preassembled	Dataverse Network	(1)
Rossella fibulata	Illumina	79,607,548 reads	76°14.716S 174°30.247E	SRR1915835
Spongilla alba	Illumina	preassembled	Dataverse Network	(1)
Sycon coactum	Illumina	preassembled	Dataverse Network	(1)
Sympagella nux	Illumina	30,065,060 reads	28°10.26N 89°56.80W	SRR1916581
Tethya wilhelma	454	442,949 reads	NCBI SRA	ERR216193
Ctenophora				
Beroe abyssicola	Illumina	45,444,644 reads	NCBI SRA	SRR777787
Bolinopsis infundibulum	Illumina	43,377,170 reads	NCBI SRA	SRR786491
Coeloplana astericola	Illumina	41,685,220 reads	NCBI SRA	SRR786490
Dryodora glandiformis	Illumina	41,269,166 reads	NCBI SRA	SRR777788
Euplokamis dunlapae	Illumina	68,302,698 reads	NCBI SRA	SRR777663
Mertensiidae sp.	Illumina	47,454,246 reads	NCBI SRA	SRR786492
Mnemiopsis leidyi	Illumina	49,782,900 reads	NCBI SRA	SRR789900
Pleurobrachia bachei	Whole genome	80,151 proteins	neurobase.rc.ufl.edu/pleurobrachia	
Pleurobrachia sp.	Illumina	50626422 reads	NCBI SRA	SRR789901
Vallicula sp.	Illumina	49,090,372 reads	NCBI SRA	SRR786489
Bilateria				
Homo sapiens	Whole genome	N/A	HaMStR core orthologs	
Danio rerio	Whole genome	54,390 proteins	EnSEMBL	
Petromyzon_marinus	Whole genome	12,444 proteins	EnSEMBL	
Saccoglossus kowalevskii	Sanger	202,190 sequences	dbEST	
Strongylocentrotus purpuratus	Whole genome	28,474 proteins	InParanoid	
Capitella teleta	Whole genome	32,415 proteins	Joint Genome Institute	
Hemithiris psittacea	Illumina	60,730,022 reads	NCBI SRA	SRR1611556
Lottia gigantea	Whole genome	22,899 proteins	Joint Genome Institute	
Tubulanus polymorphus	Illumina	39,262,732 reads	NCBI SRA	SRR1611583
Daphnia pulex	Whole genome	30,137 proteins	Joint Genome Institute	
Drosophila melanogaster	Whole genome	N/A	HaMStR core orthologs	
Priapulus caudatus	Illumina	57,331,982 reads	NCBI SRA	SRR1611567
Lithobius forficatus	Illumina	57,098,550 reads	NCBI SRA	SRR1159752

Taxa in bold were sequenced in this study.
*Taxa identified as possibly causing LBA based on LB scores.

1. Riesgo A, Farrar N, Windsor PJ, Giribet G, Leys SP (2014) The analysis of eight transcriptomes from all Poriferan classes reveals surprising genetic complexity in sponges. *Mol Biol Evol* 31(5):1102–1120.

Table S2. Data statistics for each phylogenetic dataset

Dataset (no.)	Taxa	Genes	Sites	Gene occupancy, %	Missing data (including gaps), %
Full dataset (1)	76	251	81,008	75	40
Certain paralogs removed (2)	76	250	80,735	75	40
Taxa with high LB scores removed (3)	70	250	80,735	75	41
Genes with high LB scores removed (4)	70	231	73,323	75	41
All outgroups except Choanoflagellates removed (5)	62	231	73,323	73	44
All outgroups removed (6)	60	231	73,323	73	44
Slowest evolving half of genes (7)	70	115	33,403	80	35
Genes with lowest half of RCFV values (8)	70	115	41,933	81	43
Heterogeneous genes removed (9)	76	228	66,571	76	38
Genes with high LB scores removed (10)	76	210	59,733	75	38
Taxa with high LB scores removed (11)	70	210	59,733	82	38
All outgroups except Choanoflagellates removed (12)	62	210	59,733	74	40
All outgroups removed (13)	60	210	59,733	75	40
Certain and uncertain paralogs removed (14)	76	209	60,768	72	41
Taxa with high LB scores removed (15)	70	209	60,768	78	42
Genes with high LB scores removed (16)	70	191	54,193	78	42
All outgroups except Choanoflagellates removed (17)	62	191	54,193	70	44
All outgroups removed (18)	60	191	54,193	71	44
Slowest evolving half of genes (19)	70	89	23,680	78	35
Genes with lowest half of RCFV values (20)	70	96	33,069	80	40
Heterogeneous genes removed (21)	76	195	52,543	73	39
Genes with high LB scores removed (22)	76	178	46,542	73	39
Taxa with high LB scores removed (23)	70	178	46,542	79	34
All outgroups except Choanoflagellates removed (24)	62	178	46,542	71	41
All outgroups removed (25)	60	178	46,542	72	41

Datasets nested in others have the same data filtered as corresponding upper-level datasets (see Fig. 2).