Xenacoelomorpha is the sister group to Nephrozoa

Johanna Taylor Cannon¹, Bruno Cossermelli Vellutini², Julian Smith III³, Fredrik Ronquist¹, Ulf Jondelius¹, Andreas Hejnol²

¹*Naturhistoriska Riksmuseet, P.O. Box 50007, SE-104 05 Stockholm, Sweden*

²*Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, 5008 Bergen, Norway*

³*Winthrop University, 701 Oakland Avenue, Rock Hill, South Carolina 29733, USA*

The position of Xenacoelomorpha in the tree of life remains a major unresolved question in the study of deep animal relationships¹. Xenacoelomorpha, comprising Acoela, Nemertodermatida, and *Xenoturbella*, are bilaterally symmetrical marine worms that lack a number of features common to most other bilaterians, e.g. anus, excretory organs, and circulatory system. Two conflicting hypotheses are under debate: Xenacoelomorpha is sister group to all remaining Bilateria (=Nephrozoa, i.e. protostomes and deuterostomes)^{2,3} or is a clade inside Deuterostomia⁴. Thus, determining the phylogenetic position of this clade is pivotal for understanding the early evolution of bilaterian features, or alternatively as a case of drastic secondary loss of complexity. Here we show robust phylogenomic support for Xenacoelomorpha as the sister taxon of Nephrozoa. Our phylogenetic analyses based on eleven novel xenacoelomorph transcriptomes and using different models of evolution under maximum likelihood and Bayesian paradigms, strongly corroborate this result. Rigorous testing of 25 experimental datasets designed to exclude data partitions and taxa potentially prone to reconstruction biases indicates that long branch attraction,

saturation, and missing data are not influencing these results. The sister group relationship between Nephrozoa and Xenacoelomorpha supported by our phylogenomic analyses implies that the last common ancestor of bilaterians was likely a benthic, ciliated acoelomate worm with a single opening into an epithelial gut and that excretory organs, coelomic cavities and nerve cords evolved after xenacoelomorphs separated from the stem lineage of Nephrozoa.

Acoela have an essential role in hypotheses of bilaterian body plan evolution⁵. Acoels have been compared to cnidarian planula larvae because they possess characters such as sac-like gut, net-like nervous system, and lack excretory organs. However, they also share apomorphies with Bilateria such as bilateral symmetry and a mesodermal germ layer that gives rise to circular and longitudinal muscles. Classical systematics placed acoels in Platyhelminthes^{6,7}, or as a separate early bilaterian lineage⁸. When nucleotide sequence data became available, Acoela were placed as the sister group of Nephrozoa⁹. Nemertodermatida were originally classified within Acoela, but were soon recognized as a separate clade on morphological grounds¹⁰. Subsequently, nucleotide sequence data fueled a debate on whether nemertodermatids and acoels form a monophyletic group, the Acoelomorpha, or if nemertodermatids and acoels are independent early bilaterian lineages as suggested by several studies, for example^{11,12}. The enigmatic *Xenoturbella* was first placed together with Acoela and Nemertodermatida^{13,14} then a detailed ultrastructural appraisal supported its position as sister group of all other bilaterians¹⁵. The first molecular study suggested *Xenoturbella* to be closely related to molluscs¹⁶, whereas other analyses proposed a deuterostome affiliation^{17,18}. Recent analyses of molecular data reunited *Xenoturbella* with acoels and nemertodermatids²⁻⁴ to form a clade called Xenacoelomorpha (Figure 1a).

47
48 Current conflicting hypotheses suggest that Xenacoelomorpha are the sister group of
49 Deuterostomia⁴, are nested within Deuterostomia⁴, are the sister group of Nephrozoa^{2,3}, or are
50 polyphyletic, with *Xenoturbella* included within Deuterostomia and Acoelomorpha sister taxon
51 to remaining Bilateria¹⁹ (Figure 1b-e). The deuterostome affiliation derives support from three
52 lines of evidence⁴: an analysis of mitochondrial gene sequences, microRNA complements, and a
53 phylogenomic data set. Analyses of mitochondrial genes recovered *Xenoturbella* within
54 deuterostomes¹⁸. However, limited mitochondrial data (typically ~16kb total nucleotides, 13
55 protein-coding genes) is less efficient in recovering higher-level animal relationships than
56 phylogenomic approaches, especially in long-branching taxa¹. The one complete and few partial
57 mitochondrial genomes for acoelomorphs are highly divergent in terms of both gene order and
58 nucleotide sequence^{19,20}. Analyses of new complete mitochondrial genomes of *Xenoturbella* spp.
59 do not support any phylogenetic hypothesis for this taxon²¹. Philippe *et al.*⁴ propose that
60 microRNA data support Xenacoelomorpha within the deuterostomes; however, microRNA
61 distribution is better explained by a sister relationship between Xenacoelomorpha and Nephrozoa
62 both under parsimony^{4,22} and Bayesian inference²².
63
64 Phylogenomic analyses recovering xenacoelomorph taxa within Deuterostomia show branching
65 patterns that differ significantly between alternative models of evolution⁴. Conflicting results in
66 studies that used the same EST data for xenacoelomorphs^{2,4} suggest some degree of model
67 misspecification, missing data generating positively misleading signal, or long branch attraction
68 in either or both of these studies. Testing of hypotheses under alternative models of evolution,
69 dataset partitioning, and taxon selection schemes can identify possible weaknesses of a dataset.

Here, we use this approach to test the phylogenetic position of Acoela, Nemertodermatida, and *Xenoturbella*.

Novel Illumina RNAseq data were collected for six acoel species, four nemertodermatids, *Xenoturbella bocki*, and six additional diverse metazoans (Supplementary Data Table 1). Acoel and nemertodermatid species were selected to broadly represent diversity of these two clades, including two representatives of the earliest-branching clade of Acoela, Diopisthoporidae²³. With the exception of *Hofstenia miamia* in Srivastava *et al.*³, previous phylogenomic analyses of acoels have included only representatives of Convolutidae and Isodiametridae, which possess a number of highly derived morphological characters. Our datasets include 76 diverse metazoan taxa and 2 choanoflagellate outgroups (Supplementary Data Table 1). Our primary dataset consists of 212 orthologous groups (OGs), 44896 amino acid positions, and 31% missing data (Extended Data Table 1). Sequences were taken entirely from Illumina transcriptomes or predicted transcripts from genomic data. Gene occupancy per taxon ranged from 100% for *Homo sapiens* and *Drosophila melanogaster* to 8% for the nemertodermatid *Sterreria* sp., with median per-taxon gene occupancy of 90% and an average of 80% (Supplementary Data Table 2). Notably, gene coverage for key taxa is enhanced over previous phylogenomic analyses: *Xenoturbella bocki*, six acoels, and two nemertodermatids have >90% gene occupancy in our 212 OG dataset, whereas the best represented acoelomorph terminal in the Philippe *et al.*⁴ study had an occupancy of 63%.

Maximum likelihood analyses were conducted under the best-fitting model for each individual gene partition, or alternatively LG, or LG4X²⁴ over each independent partition. LG4X is

composed of four substitution matrixes designed to improve modeling of site heterogeneity²⁴. Bayesian analyses were conducted with the site heterogeneous CAT+GTR+ Γ model and GTR+ Γ . To further validate robustness of our results to variations in substitution model specification we performed Bayesian inference analyses under an independent substitution model using a back-translated nucleotide dataset derived from our amino acid alignment. To test whether any particular taxon was biasing our analyses due to artifacts such as Long Branch Attraction (LBA), we conducted a series of taxon pruning experiments. Additional datasets were analysed that minimized missing data, excluded taxa and individual genes identified to be potentially more subject to LBA artifacts, and genes or positions that were more saturated. Using our standard pipeline, for the best-sampled 56 taxa, we also generated a dataset with 336 OGs, 81451 amino acids, and 11% missing data. Lastly, using an independent pipeline for orthologous gene selection, we generated a set of 881 OGs. This larger dataset contains 77 OTUs, 337954 amino acid positions, and 63% matrix occupancy. In all, we generated 25 unique data matrixes to address the robustness of phylogenetic signal and sensitivity of our results to parameter changes (Extended Data Table 1).

Our analyses consistently supported monophyletic Xenacoelomorpha as sister group of Nephrozoa (Figures 2-4, Extended Data Figures 1-4, Extended Data Table 1). Within Xenacoelomorpha *Xenoturbella* is the sister taxon of Acoela + Nemertodermatida. Maximum likelihood analyses under all models (Figure 2), Bayesian analyses under the site-heterogeneous CAT+GTR+ Γ model (Figure 3), as well as analyses of back-translated nucleotides (Extended Data Figure 5) all recover this topology. We found no evidence of LBA influencing the position of Xenacoelomorpha or any other group in the tree. Differing outgroup schemes do not affect the

position of Xenacoelomorpha (Supplementary Figures 4-9); neither does exclusion of taxa or genes more subject to LBA (Supplementary Figures 14-17). Monophyletic Deuterostomia (excluding *Xenoturbella*), Ecdysozoa, and Spiralia are robustly recovered, with Ctenophora as the earliest branching metazoan in all maximum likelihood analyses, while Porifera holds this position in Bayesian analyses under the CAT+GTR+ Γ model (Figure 3). Taxon exclusion analyses where Acoelomorpha alone (Supplementary Data Figure 1) or *Xenoturbella* alone (Extended Data Figure 3) were included recovered these taxa as the first branch of Bilateria. Approximately unbiased (AU) tests strongly reject the alternative hypothesis constraining Xenacoelomorpha within Deuterostomia. Leaf stability indices (LSI) for all taxa in the primary 212 OG analysis were >97% (Supplementary Table 2), suggesting that improved matrix and taxon coverage in our analyses had a positive effect on overall taxon stability in comparison to Dunn *et al.*²⁵, where both included acoels had LSI of 78%. In our own calculations of leaf stability index from the dataset of Philippe *et al.*⁴, their 6 representative xenacoelomorph species have the 6 lowest leaf stabilities of all included taxa, ranging from 88%-79% (Supplementary Table 3).

To assess gene conflict, we conducted decomposition analyses using ASTRAL²⁶, which calculates the species tree that agrees with the largest number of quartets derived from each gene tree and their respective bootstrap replicates (Extended Data Figure 4). This analysis finds strong support for the position of Xenacoelomorpha (bootstrap 99%). Bleidorn *et al.*²⁷ and Whelan *et al.*²⁸ pointed to issues with incongruence in phylogenomic analyses of ribosomal protein genes versus other protein-coding genes. Notably, in our 212 OG set, only 5 ribosomal protein genes were retained after screening for paralogous groups. To investigate if this gene class may have

biased previous results, we generated an additional data matrix composed of 52 ribosomal protein genes that passed through our other filters for gene length and taxon presence. In maximum likelihood analyses of this dataset Xenacoelomorpha, Acoelomorpha, Nemertodermatida, Deuterostomia and Spiralia are all non-monophyletic (Supplementary Figure 21). Ribosomal protein genes are heavily represented in the xenacoelomorph data in previous studies, comprising >50% of the gene occupancy in most cases. Gene partition information was not made available for the study proposing a deuterostome position for Xenacoelomorpha⁴, so reanalysis of the data without ribosomal protein genes was not possible. We suggest insufficient data for key taxa and a reliance on ribosomal protein genes were biasing the results causing Xenacoelomorpha to group within Deuterostomia.

Within Xenacoelomorpha morphological complexity differs among the three groups, as should be expected in a clade of the same age as Nephrozoa. The simplest organization is evident in *Xenoturbella*, with a saclike gut opening to a simple mouth, basic epidermal nervous system, and no gonopores or secondary reproductive organs¹³. Nemertodermatida have an epithelial gut, but the mouth appears to be a transient structure¹⁰. Furthermore, the position and anatomy of the nervous system are variable, and there is a simple gonopore. The more than 400 nominal species of Acoela (compared to 18 nemertodermatids and 5 *Xenoturbella* species) exhibit considerable morphological variation: acoels have no intestinal lumen although a mouth opening and sometimes a pharynx is present²³. The nervous system is highly variable, there are one or two gonopores, and often accessory reproductive organs²³. The morphological evolution that occurred within Xenacoelomorpha provides an interesting parallel case to Nephrozoa.

The sister group relationship between Xenacoelomorpha and Nephrozoa allows us to infer the order in which bilaterian features were evolved^{12,29}. The bilaterian ancestor was likely a soft-bodied, small benthic worm^{5,23,29,30}. Mesoderm and body axis were established before the split between Xenacoelomorpha and Nephrozoa, whereas excretory organs evolved in the stem lineage of nephrozoans (Figure 4). Centralization of the nervous system appears to have evolved in parallel in the Xenacoelomorpha and Nephrozoa. Further investigations of the genomic architecture and biology of xenacoelomorphs will provide insights into molecular, developmental and cellular building blocks used for evolving complex animal body plans and organ systems.

REFERENCES

1. Dunn, C. W., Giribet, G., Edgecombe, G. D. & Hejnol, A. Animal phylogeny and its evolutionary implications. *Annu. Rev. Ecol. Evol. Syst.* **45**, 371–395 (2014).
2. Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. Biol. Sci.* **276**, 4261–4270 (2009).
3. Srivastava, M., Mazza-Curll, K. L., van Wolfswinkel, J. C. & Reddien, P. W. Whole-body acoel regeneration is controlled by Wnt and Bmp-Admp signaling. *Curr. Biol.* **24**, 1107–1113 (2014).

- 183 4. Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature*
184 **470**, 255–258 (2011).
185
- 186 5. Nielsen, C. *Animal Evolution: Interrelationships of the living Phyla*. (Oxford University
187 Press, 2012).
188
- 189 6. Ehlers, U. *Das phylogenetische System der Plathelminthes*. (G. Fischer, 1985).
190
- 191 7. Smith, J. P. S. I., Teyler, S. & Rieger, R. M. Is the Turbellaria polyphyletic? *Hydrobiologia*
192 **132**, 13–21 (1986).
193
- 194 8. Haszprunar, G. Plathelminthes and Plathelminthomorpha — paraphyletic taxa. *J. Zool. Syst.*
195 *Evol. Res.* **34**, 41–48 (1996).
196
- 197 9. Ruiz-Trillo, I., Riutort, M., Littlewood, D. T. J., Herniou, E. A. & Baguña, J. Acoel
198 flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science*
199 **283**, 1919–1923 (1999).
200
- 201 10. Steinböck, O. Ergebnisse einer von E. Reisinger & O. Steinböck mit Hilfe des Rask-Örsted
202 fonds durchgeführten Reise in Grönland 1926. 2. *Nemertoderma bathycola* nov. gen. nov.
203 spec., eine eigenartige Turbellarie aus der Tiefe der Diskobay: nebst einem Beitrag zur
204 Kenntnis des Nemertinenepithels. *Vidensk Medd Dan Naturhist Foren* **90**, 47–84 (1930).
205

- 206 11. Paps, J., Baguñà, J. & Riutort, M. Bilaterian phylogeny: a broad sampling of 13 nuclear
207 genes provides a new Lophotrochozoa phylogeny and supports a paraphyletic basal
208 Acoelomorpha. *Mol. Biol. Evol.* **26**, 2397–2406 (2009).
209
- 210 12. Jondelius, U., Ruiz-Trillo, I., Baguñà, J. & Riutort, M. The Nemertodermatida are basal
211 bilaterians and not members of the Platyhelminthes. *Zool. Scr.* **31**, 201–215 (2002).
212
- 213 13. Westblad, E. *Xenoturbella bocki* n.g, n.sp, a peculiar, primitive turbellarian type. *Ark. Zool* **1**,
214 3–29 (1949).
215
- 216 14. Franzén, Å. & Afzelius, B. A. The ciliated epidermis of *Xenoturbella bocki*
217 (Platyhelminthes, Xenoturbellida) with some phylogenetic considerations. *Zool. Scr.* **16**, 9–
218 17 (1987).
219
- 220 15. Ehlers, U. & Sopott-Ehlers, B. Ultrastructure of the subepidermal musculature of
221 *Xenoturbella bocki*, the adelphotaxon of the Bilateria. *Zoomorphology* **117**, 71–79 (1997).
222
- 223 16. Norén, M. & Jondelius, U. *Xenoturbella* 's molluscan relatives... *Nature* **390**, 31–32 (1997).
224
- 225 17. Bourlat, S. J. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new
226 phylum Xenoturbellida. *Nature* **444**, 85–88 (2006).
227
- 228 18. Bourlat, S. J., Rota-Stabelli, O., Lanfear, R. & Telford, M. J. The mitochondrial genome
229 structure of *Xenoturbella bocki* (phylum Xenoturbellida) is ancestral within the

deuterostomes. *BMC Evol. Biol.* **9**, 107 (2009).

19. Mwinyi, A. *et al.* The phylogenetic position of Acoela as revealed by the complete mitochondrial genome of *Symsagittifera roscoffensis*. *BMC Evol. Biol.* **10**, 309 (2010).

20. Ruiz-Trillo, I., Riutort, M., Fourcade, H. M., Baguña, J. & Boore, J. L. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Mol. Phylogenet. Evol.* **33**, 321–332 (2004).

21. Rouse, G, Wilson, N.G., Carvajal, J.I. & Vrijenhoek, R.C. New deep-sea species of *Xenoturbella* and the position of Xenacoelomorpha. *Nature* (accepted).

22. Thomson, R. C., Plachetzki, D. C., Mahler, D. L. & Moore, B. R. A critical appraisal of the use of microRNA data in phylogenetics. *Proc. Natl. Acad. Sci.* **111**, E3659–E3668 (2014).

23. Jondelius, U., Wallberg, A., Hooge, M. & Raikova, O. I. How the worm got its pharynx: phylogeny, classification and Bayesian assessment of character evolution in Acoela. *Syst. Biol.* **60**, 845–871 (2011).

24. Le, S. Q., Dang, C. C. & Gascuel, O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.* **29**, 2921–2936 (2012).

25. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749 (2008).
26. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–548 (2014).
27. Bleidorn, C. *et al.* On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *BMC Evol. Biol.* **9**, 150 (2009).
28. Whelan, N. V., Kocot, K. M., Moroz, L. L. & Halanych, K. M. Error, signal, and the placement of Ctenophora sister to all other animals. *Proc. Natl. Acad. Sci.* **112**, 5773–5778 (2015).
29. Hejnol, A. & Martindale, M. Q. Acoel development supports a simple planula-like urbilaterian. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 1493–1501 (2008).
30. Laumer, C. E. *et al.* Spiralian phylogeny informs the evolution of microscopic lineages. *Curr. Biol.* **25**, 2000–2006 (2015).

Figure Legends

Figure 1 Phylogenetic hypotheses concerning Xenacoelomorpha from previous molecular

studies. a, Relationships among Xenacoelomorpha. *Xenoturbella* is sister to Acoelomorpha

(Acoela + Nemertodermatida). Illustrated species from left to right: *Flagellophora apelti*,

Diopisthoporus psammophilus, *Xenoturbella bocki*. **b**, Xenacoelomorpha is sister taxon to

Nephrozoa (phylogenomic analyses^{2,3}), **c**, Xenacoelomorpha is sister taxon to Ambulacraria

within deuterostomes (phylogenomic analyses⁴), **d**, Xenacoelomorpha is sister taxon to

Ambulacraria + Chordata (mitochondrial protein analyses^{4,19}), **e**, *Xenoturbella* is within

Deuterostomia, while Acoelomorpha form two separate clades outside Nephrozoa (molecular

systematic analyses¹¹), or its sister group (some mitochondrial protein analyses¹⁹). Colors in b–e

indicate Xenacoelomorpha (red), Protostomia (blue), Deuterostomia (green).

Figure 2 Maximum likelihood topology of metazoan relationships inferred from 212 genes.

ML tree is shown as inferred using the best-fitting AA substitution model for each gene.

Bootstrap support values from analyses inferred under alternative models of AA substitution are

indicated at the nodes (Best-fitting model for each OG selected by ProtTest / LG4X across all

partitions / LG+I+Γ across all partitions, 100 bootstrap replicates). Filled blue circles represent

100% bootstrap support under all models of evolution. Species indicated in bold are new

transcriptomes published with this study.

Figure 3 Bayesian inference topology of metazoan relationships inferred from 212 genes

under the CAT+GTR+ Γ model. Filled blue circles indicate posterior probabilities of 1.0.

Shown is the majority rule consensus tree of two independent chains of >17,000 cycles each and

burnin of 5,000 cycles. Convergence of the two chains was indicated by a “maxdiff” value of 0.25. Position of Xenacoelomorpha was unchanged in two additional independent chains, which did not converge with the chains shown above due to alternate positions of *Trichoplax adhaerens* and *Membranipora membranacea*.

Figure 4 Summary of metazoan relationships as inferred in this study. **a**, Summary of phylogenomic results based on analyses of 212, 336, and 881 genes. Xenacoelomorpha is a monophyletic clade sister to Nephrozoa with >99% support in all analyses. **b**, Interrelationships among four major animal clades Cnidaria, Xenacoelomorpha, Protostomia and Deuterostomia, with selected morphological characters mapped onto the tree as ancestral states for each of the four clades.

METHODS

Molecular methods and sequencing. We generated novel RNAseq data from 6 acoels, 4 nemertodermatids, *Xenoturbella bocki*, and six additional diverse metazoans (Supplementary Table 1). Total RNA was extracted from fresh or RNAlater (Ambion) preserved specimens using TRI Reagent Solution (Ambion) or the RNeasy Micro Kit (Qiagen), prepared using the SMART cDNA library construction kit (Clontech), and sequenced as 2x100 paired end runs with Illumina HiSeq 2000 at SciLifeLab (Stockholm, Sweden) or GeneCore (EMBL Genomics Core Facilities). Illumina data were supplemented with publically available RNAseq and genome data (Supplementary Table 1) to generate a final dataset including 76 diverse metazoans and 2 choanoflagellate outgroup taxa.

Dataset assembly. Both novel RNAseq data and raw Illumina sequences taken from NCBI's Sequence Read Archive (SRA) were assembled using Trinity³¹. Assembled data were translated using Transdecoder (<http://transdecoder.sf.net>). To determine orthologous genes, we used two methods; a more restrictive and standard approach using HaMStR (Hidden Markov Model based Search for Orthologs using Reciprocity)³², as well as an approach designed to generate a broader set of genes for phylogenetic inference, using the software ProteinOrtho³³. Protocols for gene selection using HaMStR followed Kocot *et al.*³⁴ and Cannon *et al.*³⁵. Translated unigenes for all taxa were searched against the model organisms core ortholog set of HaMStR using the strict option and *Drosophila melanogaster* as the reference taxon. Sequences shorter than 50 amino acids were deleted, and OGs sampled for fewer than 30 taxa were excluded in order to reduce missing data. To trim mistranslated ends, if one of the first or last 20 characters of sequences was an X, all characters between that X and the end of the sequence were removed. The OGs were then aligned using MAFFT³⁶ and trimmed using Aliscore³⁷ and Alicut (<https://www.zfmk.de/en/research/research-centres-and-groups/utilities>). At this stage sequences that were greater than 50% gaps and alignments shorter than 100 amino acids were discarded. To remove potentially paralogous genes, we generated single gene trees using FastTree³⁸ and filtered these using PhyloTreePruner³⁹. For 78 taxa, this protocol retained 212 OGs, 44896 amino acids, with 31% missing data. This protocol was repeated with the 56 taxa with highest percentage of gene coverage, resulting in a data matrix of 336 genes, 81451 amino acids, and 11 percent missing data.

To generate the ProteinOrtho dataset, *Sterreria* sp. was excluded due to its small library size. Translated assemblies were filtered to remove mistranslated ends as described above, and only sequences longer than 50 amino acids were retained for clustering in ProteinOrtho. In

ProteinOrtho, we used the steps option, the default E-value for BLAST, and minimum coverage of best BLAST alignments of 33 percent. Resulting clusters were filtered to include only putative orthologous groups containing greater than 40 taxa, then aligned as above with MAFFT. For each alignment a consensus sequence was inferred using the EMBOSS program infoalign⁴⁰. Infoalign's "change" calculation computes the percentage of positions within each sequence in each alignment that differ from the consensus. Sequences with a "change" value larger than 75 were deleted, helping to exclude incorrectly aligned sequences. OGs were then realigned with MAFFT, trimmed with Aliscore and Alicut, and processed as above. After filtering for paralogous groups with PhyloTreePruner, 881 orthologous groups were retained.

Due to the smaller size of the dataset and amount of computational resources required, taxon pruning and signal dissection analyses were performed solely on the primary HaMStR gene set. For taxon exclusion experiments, individual OG alignments were realigned using MAFFT following the removal of selected taxa. TreSpEx⁴¹ was used to assess potential sources of misleading signal, including standard deviation of branch-length heterogeneity (LB) and saturation. Sites showing evidence of saturation and compositional heterogeneity were removed using Block Mapping and Gathering with Entropy (BMGE)⁴², using the "fast" test of compositional heterogeneity (-s FAST) and retaining gaps (-g 1).

Phylogenetic analysis. Maximum likelihood analyses of the complete 212 OG data matrix were performed using RAxML Version 8.0.20-mpi⁴³ under the best-fitting models for each gene partition determined by ProtTest version 3.4⁴⁴. The best fitting model for all but 3 of the 212 OGs was LG, so further maximum likelihood analyses were performed using the PROTGAMMAILG option. Bootstrapped trees from the 212-gene dataset were used to calculate leaf stability indices of each OTU using the Roguenarok server (<http://www.exelixis-lab.org/>).

Bayesian analyses were conducted using PhyloBayes-MPI⁴⁵ version 1.5a under the CAT+GTR+
 Γ model or GTR+ Γ with four independent chains per analysis. Analyses ran for >12,000 cycles,
 until convergence of at least two chains was reached as assessed by maxdiff. Further Bayesian
 analyses were conducted in MrBayes version 3.2⁴⁶. For the MrBayes analyses, we back-
 translated the aligned amino acid data to nucleotides for first and second codon positions using
 the universal genetic code. Third codon position data were ignored. When the back translation
 was ambiguous, we preserved the ambiguity in the nucleotide data. For instance, serine is coded
 by TC{A, C, G, T} or AG{T, C}, where {...} denotes alternative nucleotides for a single codon
 site. Thus, for Serine the back translation is {A,T}{C, G}. This is the only back translation that is
 ambiguous both for the first and the second codon positions. The back translation for arginine
 and leucine are also ambiguous but only for the first codon position. All other back translations
 are unambiguous for both the first and second codon sites. Thus, the back translation of first and
 second codon sites results in negligible information loss compared to the original nucleotide
 data.

We analysed the resulting nucleotide data in MrBayes 3.2.6-svn(r1037)⁴⁶ using a model with two
 partitions, one for first codon positions and one for second codon positions. For each partition we
 employed an independent substitution model, modeling rate variation across sites using a
 discrete gamma distribution (four categories) with a proportion of invariable sites (“lset rates =
 invgamma”), and nucleotide substitutions with independent stationary state frequencies and a
 reversible-jump approach to the partitioning of exchangeability rates (“lset nst=mixed”). We also
 uncoupled the partition rates (“prset ratepr=variable”). All other settings were left at their
 defaults. For each analysis, we used four independent runs with four Metropolis-coupled chains
 each and ran them for 4 M generations, sampling every 500 generations (“mcmcprun=4 nch=4

ngen=4000000 samplefreq=500”). The analyses finished with an average standard deviation of split frequencies of 0.033 or less, and a potential scale reduction factor of 1.003 or less. The MrBayes data files and run scripts are provided as supplementary material.

We additionally used ASTRAL²⁶ to calculate an optimal bootstrapped species tree from individual RAxML gene trees decomposed into quartets.

METHODS REFERENCES

31. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

32. Ebersberger, I., Strauss, S. & Haeseler, A. von. HaMStR: Profile hidden Markov model based search for orthologs in ESTs. *BMC Evol. Biol.* **9**, 157 (2009).

33. Lechner, M. *et al.* Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).

34. Kocot, K. M. *et al.* Phylogenomics reveals deep molluscan relationships. *Nature* **477**, 452–456 (2011).

35. Cannon, J. T. *et al.* Phylogenomic resolution of the hemichordate and echinoderm clade. *Curr. Biol.* **24**, 2827–2832 (2014).

36. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
37. Misof, B. & Misof, K. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst. Biol.* **58**, 21–34 (2009).
38. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
39. Kocot, K. M., Citarella, M. R., Moroz, L. L. & Halanych, K. M. PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol. Bioinform. Online* **9**, 429–435 (2013).
40. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
41. Struck, T. H. TreSpEx-Detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol. Bioinform. Online* **10**, 51–67 (2014).
42. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

- 436 43. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
437 large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
438
- 439 44. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit
440 models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
441
- 442 45. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI. Phylogenetic
443 reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**,
444 611–615 (2013).
445
- 446 46. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian phylogenetic inference and model
447 choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
448
449
450

451 **Acknowledgements**

452 The Swedish Research Council provided funding for U.J. and J.T.C. (grant 2012-3913) and F.R.
453 (grant 2014-5901). A.H. received support from the Sars Core budget and Marie Curie ITN
454 “NEPTUNE” (FP7-PEOPLE-2012-ITN 317172) and FP7-PEOPLE-2009-RG 256450. We’d like
455 to thank Nicolas Lartillot and Kevin Kocot for helpful discussions. Hejnol lab members Kevin
456 Pang & Aina Børve assisted with RNA extraction; Anlaug Boddington, Jonas Bengtsen and
457 Anette Elde assisted with culture for *Isodiametra pulchra* and *Convolutriloba macropyga*.
458 Thanks to Wolfgang Sterrer for collection of *Sterreria* sp. and *Ascoparia* sp., and to Ralf Janssen

for finding *Xenoturbella bocki*. The Sven Lovén Centre of Marine Sciences Kristineberg, University of Gothenburg and the Interuniversity Institute of Marine Sciences in Eilat provided logistical support for field collection. Sandra Baldauf assisted with lab space and resources for cDNA synthesis. Thank you to Karolina Larsson for the original illustrations. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC). Transcriptome assembly, dataset construction, RAxML and PhyloBayes analyses were performed using resources provided through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project b2013077, and MrBayes analyses were run under project snic2014-1-323.

Author Contributions

J.T.C., U.J., B.C.V. and A.H. conceived and designed the study. U.J. and A.H. collected several specimens and J.P.S.S.III collected *Diopisthoporus gymnopharyngeus* specimens. J.T.C. and B.C.V. performed molecular work and RNAseq assembly. J.T.C. assembled the datasets and performed phylogenetic analyses. F.R. conducted Bayesian phylogenetic analyses using MrBayes. All authors contributed to writing the manuscript.

Author Information

Sequence data deposited at SRA under BioProject PRJNA295688. Data matrices and trees from this study are available from the Dryad Digital Repository (<http://datadryad.org>) with the DOI #####. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to J.T.C. (joie.cannon@gmail.com) and A.H. (andreas.hejnol@uib.no).

Extended Data Figure Legends

Extended Data Figure 1. Maximum likelihood topology of metazoan relationships inferred from 336 genes from the best-sampled 56 taxa. ML tree is shown as inferred using the LG+I+ Γ model for each gene partition, and 100 bootstrap replicates. Filled blue circles represent 100% bootstrap support. The length of the matrix is 81451 AAs and overall matrix completeness is 89%.

Extended Data Figure 2. Maximum likelihood topology of metazoan relationships inferred from 881 genes and 77 taxa. ML tree is shown as inferred using the LG+I+ Γ model for each gene partition, and 100 bootstrap replicates. Filled blue circles represent 100% bootstrap support. The length of the matrix is 337954 AAs and overall matrix completeness is 62%.

Extended Data Figure 3. Maximum likelihood topology of metazoan relationships inferred from 212 genes with Acoelomorpha removed. ML tree is shown as inferred using the LG+I+ Γ model for each gene partition, and 100 bootstrap replicates. Filled blue circles represent 100% bootstrap support. The length of the matrix is 43942 AAs and overall matrix completeness is 70%.

Extended Data Figure 4. ASTRAL species tree, constructed from 212 input partial gene trees inferred in RAxML v8.0.20. Nodal support values reflect the frequency of splits in trees constructed by ASTRAL from 100 bootstrap replicate gene trees.

Extended Data Figure 5. Bayesian inference topology of metazoan relationships inferred based on 212 genes and 78 taxa. Results are shown from MrBayes analyses of four independent Metropolis-coupled chains run for four million generations, with sampling every 500 generations. Amino acid data were back-translated to nucleotides and analysed under an independent substitution model.

Extended Data Table Legend

Extended Data Table 1. Summary of datasets analysed in this study and support for monophyly of major groups. Bootstrap support values given from RAxML analyses inferred with the LG+I+ Γ model from 100 rapid bootstrap replicates. Bayesian posterior probabilities are listed from MrBayes analyses inferred under an independent substitution model using a back-translated nucleotide dataset derived from our amino acid alignment, and PhyloBayes analyses under the CAT+GTR+ Γ model.

● = Maximal support in analyses of 212, 336, and 881 genes ● >99% ● >70%