LETTER

doi:10.1038/nature16520

Xenacoelomorpha is the sister group to Nephrozoa

Johanna Taylor Cannon[1], Bruno Cossermelli Vellutini[2], Julian Smith III[3], Fredrik Ronquist[1], Ulf Jondelius[1] & Andreas Hejnol[2]

The position of Xenacoelomorpha in the tree of life remains a major unresolved question in the study of deep animal relationships[1]. Xenacoelomorpha, comprising Acoela, Nemertodermatida, and *Xenoturbella*, are bilaterally symmetrical marine worms that lack several features common to most other bilaterians, for example an anus, nephridia, and a circulatory system. Two conflicting hypotheses are under debate: Xenacoelomorpha is the sister group to all remaining Bilateria (= Nephrozoa, namely protostomes and deuterostomes)[2,3] or is a clade inside Deuterostomia[4]. Thus, determining the phylogenetic position of this clade is pivotal for understanding the early evolution of bilaterian features, or 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 11 novel xenacoelomorph transcriptomes and using different models of evolution under maximum likelihood and Bayesian inference analyses, strongly corroborate this result. Rigorous testing of 25 experimental data sets designed to exclude data partitions and taxa potentially prone to reconstruction biases indicates that long-branch attraction, saturation, and missing data do not influence these results. The sister group relationship between Nephrozoa and Xenacoelomorpha supported by our phylogenomic analyses implies that the last common ancestor of bilaterians was probably 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 evolution[5]. Acoels have been compared to cnidarian planula larvae because they possess characters such as a blind gut, a net-like nervous system, and they lack nephridia. However, they also share apomorphies with Bilateria such as bilateral symmetry and a mesodermal germ layer that gives rise to circular and longitudinal muscles. Classic systematics placed acoels in Platyhelminthes[6], or as a separate early bilaterian lineage[7,8]. When nucleotide sequence data became available, Acoela were placed as the sister group of Nephrozoa[9]. Nemertodermatida were originally classified within Acoela, but were soon recognized as a separate clade on morphological grounds[10]. Subsequently, nucleotide sequence data fuelled 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 refs 11 and 12. The enigmatic *Xenoturbella* was first placed together with Acoela and Nemertodermatida[13,14]; then an ultrastructural appraisal supported its position as sister group of all other bilaterians[15]. The first molecular study suggested *Xenoturbella* to be closely related to molluscs[16], whereas other analyses proposed a deuterostome affiliation[17,18]. Recent analyses of molecular data reunited *Xenoturbella* with acoels and nemertodermatids[2–4] to form a clade called Xenacoelomorpha (Fig. 1a).

Current conflicting hypotheses suggest that Xenacoelomorpha are the sister group of Deuterostomia[4], are nested within Deuterostomia[4], are the sister group of Nephrozoa[2,3], or are polyphyletic, with *Xenoturbella* included within Deuterostomia and the Acoelomorpha as sister taxon to remaining Bilateria[19] (Fig. 1b–e). The deuterostome affiliation derives support from three lines of evidence[4]: an analysis of mitochondrial gene sequences, microRNA complements, and a phylogenomic data set. Analyses of mitochondrial genes recovered *Xenoturbella* within deuterostomes[18]. However, limited mitochondrial data (typically ~16 kilobase total nucleotides, 13 protein-coding genes) are less efficient in recovering higher-level animal relationships than phylogenomic approaches, especially in long-branching taxa[1]. The one complete and few partial mitochondrial genomes for acoelomorphs are highly divergent in terms of both gene order and nucleotide sequence[19,20]. Analyses of new complete mitochondrial genomes of *Xenoturbella* spp. do not support any phylogenetic hypothesis for this taxon[21]. Ref. 4 proposes that microRNA data support Xenacoelomorpha within the deuterostomes; however, microRNA distribution is better explained by a sister relationship between Xenacoelomorpha and Nephrozoa both under parsimony[4,22] and under Bayesian inference[22].

Phylogenomic analyses recovering xenacoelomorph taxa within Deuterostomia show branching patterns that differ significantly

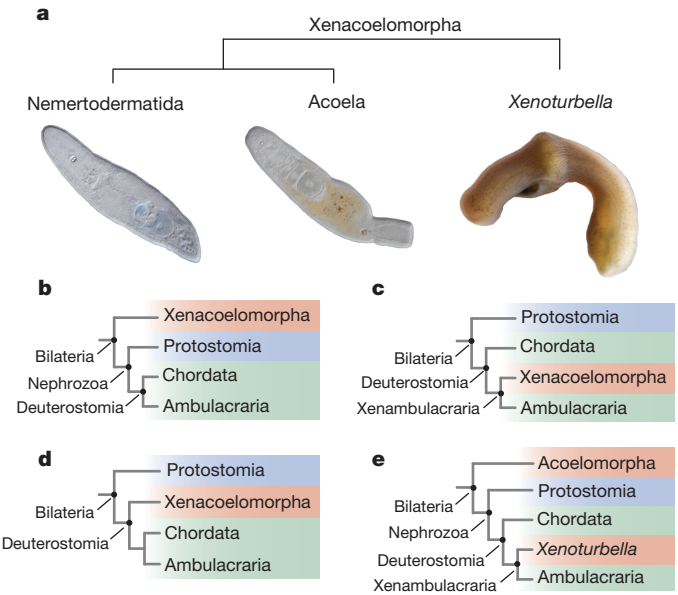


Figure 1 | Phylogenetic hypotheses concerning Xenacoelomorpha from previous molecular studies. a, Relationships among Xenacoelomorpha. *Xenoturbella* is sister to Acoelomorpha (Acoela + Nemertodermatida). Illustrated species from left to right: *Flagellophora apelti*, *Diopisthoporus psammophilus*, *X. bocki*. **b**, Xenacoelomorpha is sister taxon to Nephrozoa (phylogenomic analyses[2,3]). **c**, Xenacoelomorpha is sister taxon to Ambulacraria within deuterostomes (phylogenomic analyses[4]). **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 analyses[11]), or its sister group (some mitochondrial protein analyses[19]). Colours in **b**–**e** indicate Xenacoelomorpha (red), Protostomia (blue), Deuterostomia (green).

[1]Naturhistoriska Riksmuseet, PO Box 50007, SE-104 05 Stockholm, Sweden. [2]Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, 5008 Bergen, Norway. [3]Department of Biology, Winthrop University, 701 Oakland Avenue, Rock Hill, South Carolina 29733, USA.

© 2016 Macmillan Publishers Limited. All rights reserved

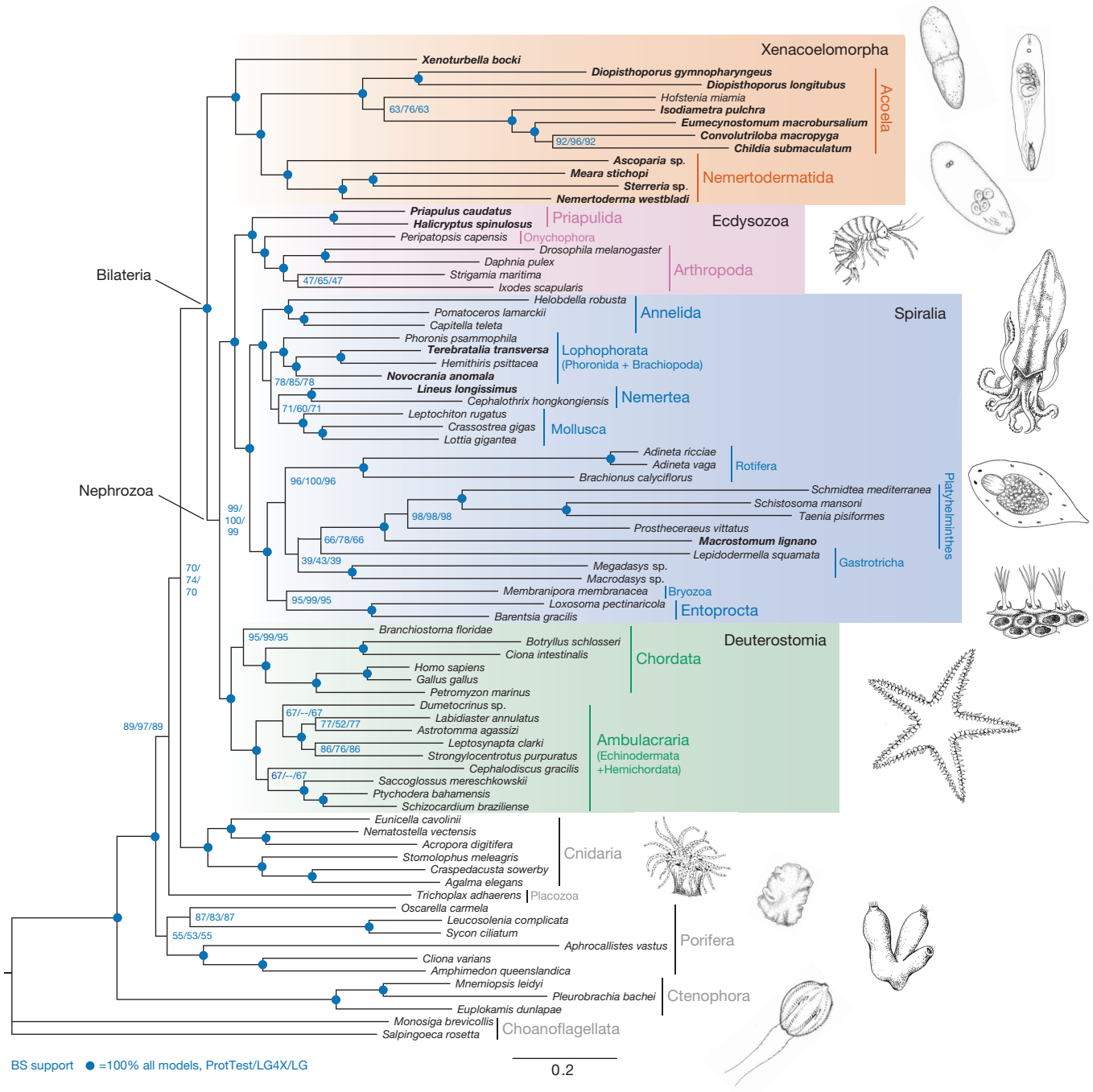


Figure 2 | Maximum likelihood topology of metazoan relationships inferred from 212 genes. Maximum likelihood tree is shown as inferred using the best-fitting amino-acid substitution model for each gene. Bootstrap support values from analyses inferred under alternative models of amino-acid substitution are indicated at the nodes (best-fitting model for each orthologous group 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.

between alternative models of evolution[4]. Conflicting results in studies that used the same expressed sequence tag data for xenacoelomorphs[2,4] suggest some degree of model misspecification, missing data generating positively misleading signal, or long-branch attraction (LBA) in either or both of these studies. Testing of hypotheses under alternative models of evolution, data set partitioning, and taxon selection schemes can identify possible weaknesses of a data set. 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 Table 1). Acoel and nemertodermatid species were selected to broadly represent the diversity of these two clades, including two representatives of the earliest-branching clade of Acoela, Diopisthoporidae[23]. With the exception of *Hofstenia miamia* in ref. 3, previous phylogenomic analyses of acoels have included only representatives of Convolutidae and Isodiametridae, which possess several highly derived morphological characters. Our data sets include 76 diverse metazoan taxa and 2 choanoflagellate outgroups (Supplementary Table 1). Our primary data set consists of 212 orthologous groups, 44,896 amino-acid positions, and 31% missing data (Extended Data Table 1).

© 2016 Macmillan Publishers Limited. All rights reserved

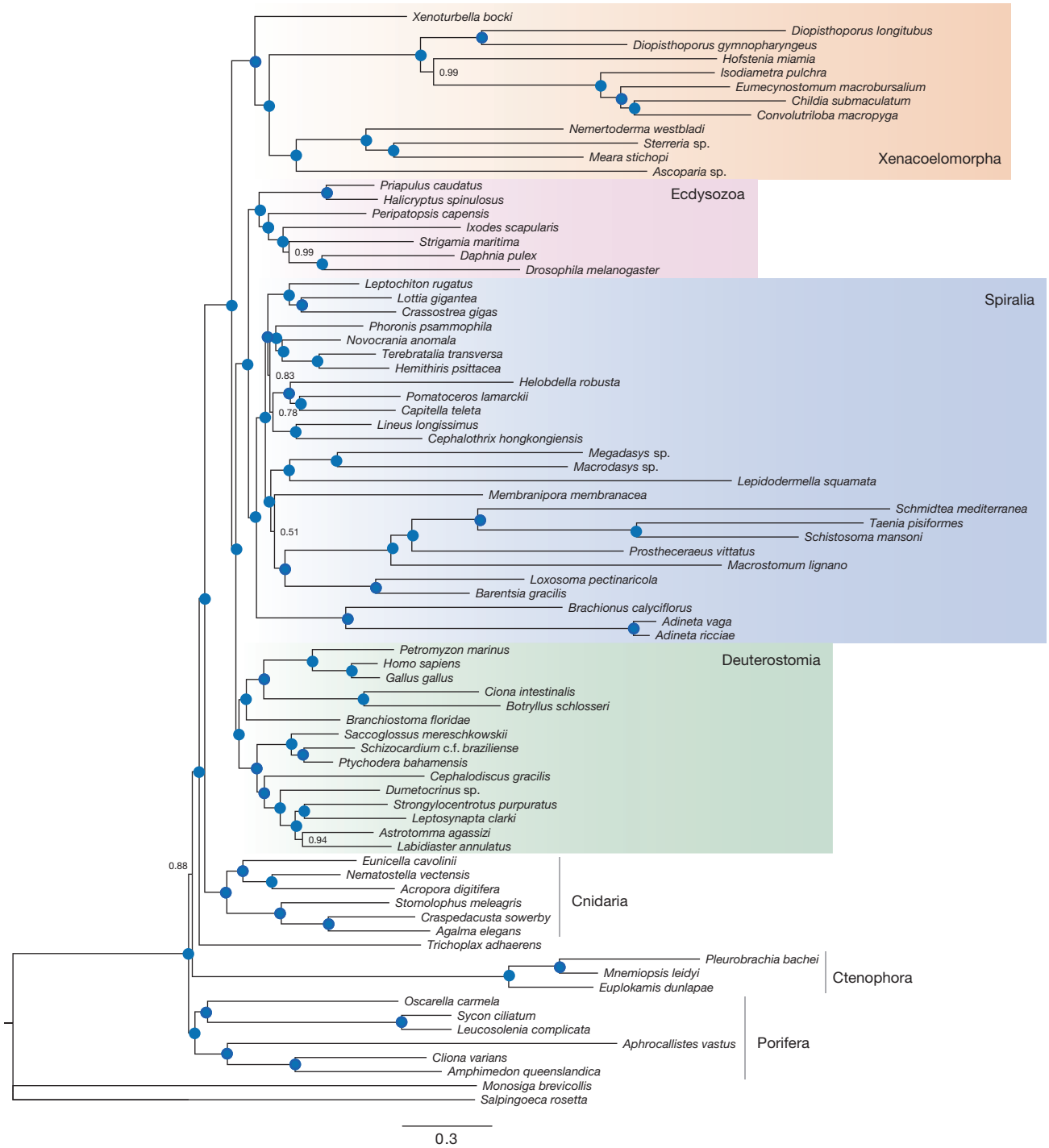


Figure 3 | Bayesian inference topology of metazoan relationships inferred from 212 genes under the CAT + GTR + Γ model. Filled blue circles indicate posterior probabilities of 1.0. Shown is the majority rule consensus tree of two independent chains of > 17,000 cycles each and burn-in of 5,000 cycles. 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 owing to alternative positions of *Trichoplax adhaerens* and *Membranipora membranacea*.

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 Table 2). Notably, gene coverage for key taxa is enhanced over previous phylogenomic analyses: *X. bocki*, six acoels, and two nemertodermatids have > 90% gene occupancy in our 212 orthologous group data set, whereas the best represented acoelomorph terminal in ref. 4 had an occupancy of 63%.

Maximum likelihood analyses were conducted under the best-fitting model for each individual gene partition, or the LG model, or the LG4X model[24] over each independent partition. The LG4X model is composed of four substitution matrixes designed to improve modelling of site heterogeneity[24]. Bayesian analyses were conducted with the site-heterogeneous CAT + GTR + Γ model and GTR + Γ. To further validate the 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 data set derived from our amino-acid alignment. To test whether any

© 2016 Macmillan Publishers Limited. All rights reserved

particular taxon was biasing our analyses owing to artefacts such as LBA, we conducted a series of taxon-pruning experiments. Additional data sets were analysed that minimized missing data, excluded taxa and individual genes identified to be potentially more subject to LBA artefacts, and genes or positions that were more saturated. Using our standard pipeline, for the best-sampled 56 taxa, we also generated a data set with 336 orthologous groups, 81,451 amino acids, and 11% missing data. Lastly, using an independent pipeline for orthologous gene selection, we generated a set of 881 orthologous groups. This larger data set contained 77 operational taxonomic units, 337,954 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 (Figs 2–4, Extended Data Figs 1–4 and Extended Data Table 1). Within Xenacoelomorpha, *Xenoturbella* is the sister taxon of Acoela + Nemertodermatida. Maximum likelihood analyses under all models (Fig. 2), Bayesian analyses under the site-heterogeneous CAT + GTR + Γ model (Fig. 3), as well as analyses of back-translated nucleotides (Extended Data Fig. 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 Figs 4–9); neither does exclusion of taxa or genes more subject to LBA (Supplementary Figs 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 (Fig. 3). Taxon-exclusion analyses, where Acoelomorpha alone (Supplementary Fig. 1) or *Xenoturbella* alone (Extended Data Fig. 3) were included, recovered these taxa as the first branch of Bilateria. Approximately unbiased tests strongly reject the alternative hypothesis constraining Xenacoelomorpha within Deuterostomia. Leaf stability indices for all taxa in the primary 212 orthologous group analysis were $>97\%$ (Supplementary Table 2), suggesting that improved matrix and taxon coverage in our analyses had a positive effect on overall taxon stability compared with ref. 25, where both included acoels had leaf stability indices of 78%. In our own calculations of leaf stability index from the data set of ref. 4, the six representative xenacoelomorph species have the six lowest leaf stabilities of all included taxa, ranging from 88% to 79% (Supplementary Table 3).

To assess gene conflict, we conducted decomposition analyses using ASTRAL[26], 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 Fig. 4). This analysis finds strong support for the position of Xenacoelomorpha (bootstrap 99%). Refs 27 and 28 pointed to issues with incongruence in phylogenomic analyses of ribosomal protein genes versus other protein-coding genes. Notably, in our 212 orthologous group set, only five 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 data set, Xenacoelomorpha, Acoelomorpha, Nemertodermatida, Deuterostomia and Spiralia are all non-monophyletic (Supplementary Fig. 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 Xenacoelomorpha[4], so re-analysis of the data without ribosomal protein genes was not possible. We suggest that insufficient data for key taxa and a reliance on ribosomal protein genes were biasing the results, causing Xenacoelomorpha to group within Deuterostomia.

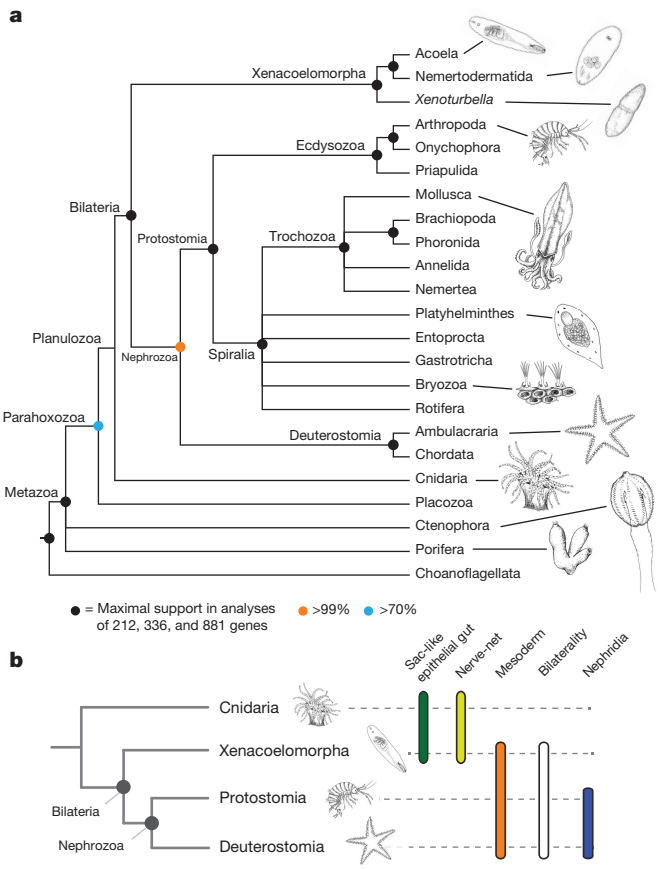


Figure 4 | Summary of metazoan relationships as inferred in this study. **a**, Summary of phylogenomic results based on analyses of 212, 336, and 881 genes. Xenacoelomorpha is a monophyletic clade sister to Nephrozoa with $>99\%$ support in all analyses. **b**, Interrelationships among four major animal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

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 sac-like epithelial gut opening to a simple mouth, a basiepidermal nervous system, and no gonopores or secondary reproductive organs[13]. Nemertodermatida also have an epithelial gut, but the mouth appears to be a transient structure[10]. Furthermore, the position and anatomy of the nervous system and the male copulatory organ are variable. The more than 400 nominal species of Acoela (compared with 18 nemertodermatids and 5 *Xenoturbella* species) exhibit considerable morphological variation: acoels have no intestinal lumen although a mouth opening and sometimes a pharynx is present[23]. The nervous system is highly variable, there are one or two gonopores, and often accessory reproductive organs[23]. 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 probably a soft-bodied, small ciliated benthic worm[5,23,29,30]. Mesoderm and body axis were established before the split between Xenacoelomorpha and Nephrozoa, whereas nephridia evolved in the stem lineage of nephrozoans (Fig. 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.

© 2016 Macmillan Publishers Limited. All rights reserved

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 September; accepted 7 December 2015.

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. R. Soc. B* **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).
4. Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* **470,** 255–258 (2011).
5. Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla* (Oxford Univ. Press, 2012).
6. Ehlers, U. *Das phylogenetische System der Plathelminthes* (G. Fischer, 1985).
7. Smith, J. P. S., III, Tyler, S. & Rieger, R. M. Is the Turbellaria polyphyletic? *Hydrobiologia* **132,** 13–21 (1986).
8. Haszprunar, G. Plathelminthes and Plathelminthomorpha — paraphyletic taxa. *J. Zool. Syst. Evol. Res.* **34,** 41–48 (1996).
9. Ruiz-Trillo, I., Riutort, M., Littlewood, D. T. J., Herniou, E. A. & Baguña, J. Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science* **283,** 1919–1923 (1999).
10. Steinböck, O. Ergebnisse einer von E. Reisinger & O. Steinböck mit Hilfe des Rask-Örsted fonds durchgefuhrten Reise in Grönland 1926. 2. *Nemertoderma bathycola* nov. gen. nov. spec., eine eigenartige Turbellarie aus der Tiefe der Diskobay: nebst einem Beitrag zur Kenntnis des Nemertinenepithels. *Vidensk. Medd. Dan. Naturhist. Foren.* **90,** 47–84 (1930).
11. Paps, J., Baguñà, J. & Riutort, M. Bilaterian phylogeny: a broad sampling of 13 nuclear genes provides a new Lophotrochozoa phylogeny and supports a paraphyletic basal acoelomorpha. *Mol. Biol. Evol.* **26,** 2397–2406 (2009).
12. Jondelius, U., Ruiz-Trillo, I., Baguñà, J. & Riutort, M. The Nemertodermatida are basal bilaterians and not members of the Platyhelminthes. *Zool. Scr.* **31,** 201–215 (2002).
13. Westblad, E. *Xenoturbella bocki* n.g, n.sp, a peculiar, primitive turbellarian type. *Ark. Zool.* **1,** 3–29 (1949).
14. Franzén, Å. & Afzelius, B. A. The ciliated epidermis of *Xenoturbella bocki* (Platyhelminthes, Xenoturbellida) with some phylogenetic considerations. *Zool. Scr.* **16,** 9–17 (1987).
15. Ehlers, U. & Sopott-Ehlers, B. Ultrastructure of the subepidermal musculature of *Xenoturbella bocki*, the adelphotaxon of the Bilateria. *Zoomorphology* **117,** 71–79 (1997).
16. Norén, M. & Jondelius, U. *Xenoturbella*'s molluscan relatives... . *Nature* **390,** 31–32 (1997).
17. Bourlat, S. J. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444,** 85–88 (2006).
18. Bourlat, S. J., Rota-Stabelli, O., Lanfear, R. & Telford, M. J. The mitochondrial genome 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., Baguñà, 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* http://dx.doi.org/10.1038/nature16545 (this issue).
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. USA* **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–i548 (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. USA* **112,** 5773–5778 (2015).
29. Hejnol, A. & Martindale, M. Q. Acoel development supports a simple planula-like urbilaterian. *Phil. Trans. R. Soc. B* **363,** 1493–1501 (2008).
30. Laumer, C. E. *et al.* Spiralian phylogeny informs the evolution of microscopic lineages. *Curr. Biol.* **25,** 2000–2006 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements The Swedish Research Council provided funding for U.J. and J.T.C. (grant 2012-3913) and F.R. (grant 2014-5901). A.H. received support from the Sars Core budget and Marie Curie Innovative Training Networks 'NEPTUNE' (FP7-PEOPLE-2012-ITN 317172) and FP7-PEOPLE-2009-RG 256450. We thank N. Lartillot and K. Kocot for discussions. Hejnol laboratory members K. Pang and A. Børve assisted with RNA extraction; A. Boddington, J. Bengtsen and A. Elde assisted with culture for *Isodiametra pulchra* and *Convolutriloba macropyga*. Thanks to W. Sterrer for collection of *Sterreria* sp. and *Ascoparia* sp., and to R. Janssen for finding *X. 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. S. Baldauf assisted with laboratory space and resources for complementary DNA synthesis. We thank K. Larsson for the original illustrations. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC). Transcriptome assembly, data set 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.S. III collected *Diopisthoporus gymnopharyngeus* specimens. J.T.C. and B.C.V. performed molecular work and RNA sequencing 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 have been deposited in the NCBI Sequence Read Archive under BioProject PRJNA295688. Data matrices and trees from this study are available from the Dryad Digital Repository (http://datadryad.org) under DOI 10.5061/dryad.493b7. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.T.C. (joie.cannon@gmail.com) or A.H. (andreas.hejnol@uib.no).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Molecular methods and sequencing. We generated novel RNA-seq data from six acoels, four nemertodermatids, *X. bocki*, and six additional diverse metazoans (Supplementary Table 1). Total RNA was extracted from fresh or RNA*later* (Ambion) preserved specimens using TRI Reagent Solution (Ambion) or the RNeasy Micro Kit (Qiagen), prepared using the SMART complementary DNA library construction kit (Clontech), and sequenced as 2×100 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 data set including 76 diverse metazoans and 2 choanoflagellate outgroup taxa.

Data set assembly. Both novel RNA-seq data and raw Illumina sequences taken from the NCBI Sequence Read Archive were assembled using Trinity[31]. 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 Orthologues using Reciprocity)[32], as well as an approach designed to generate a broader set of genes for phylogenetic inference, using the software ProteinOrtho[33]. Protocols for gene selection using HaMStR followed refs 34 and 35. Translated unigenes for all taxa were searched against the model organisms core orthologue set of HaMStR using the strict option and *D. melanogaster* as the reference taxon. Sequences shorter than 50 amino acids were deleted, and orthologous groups sampled for fewer than 30 taxa were excluded 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 orthologous groups were then aligned using MAFFT[36] and trimmed using Aliscore[37] 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 FastTree[38] and filtered these using PhyloTreePruner[39]. For 78 taxa, this protocol retained 212 orthologous groups, 44,896 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, 81,451 amino acids, and 11% missing data.

To generate the ProteinOrtho data set, *Sterreria* sp. was excluded owing 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%. 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 infoalign[40]. 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. orthologous groups 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.

Owing to the smaller size of the data set 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 orthologous group alignments were realigned using MAFFT following the removal of selected taxa. TreSpEx[41] 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).

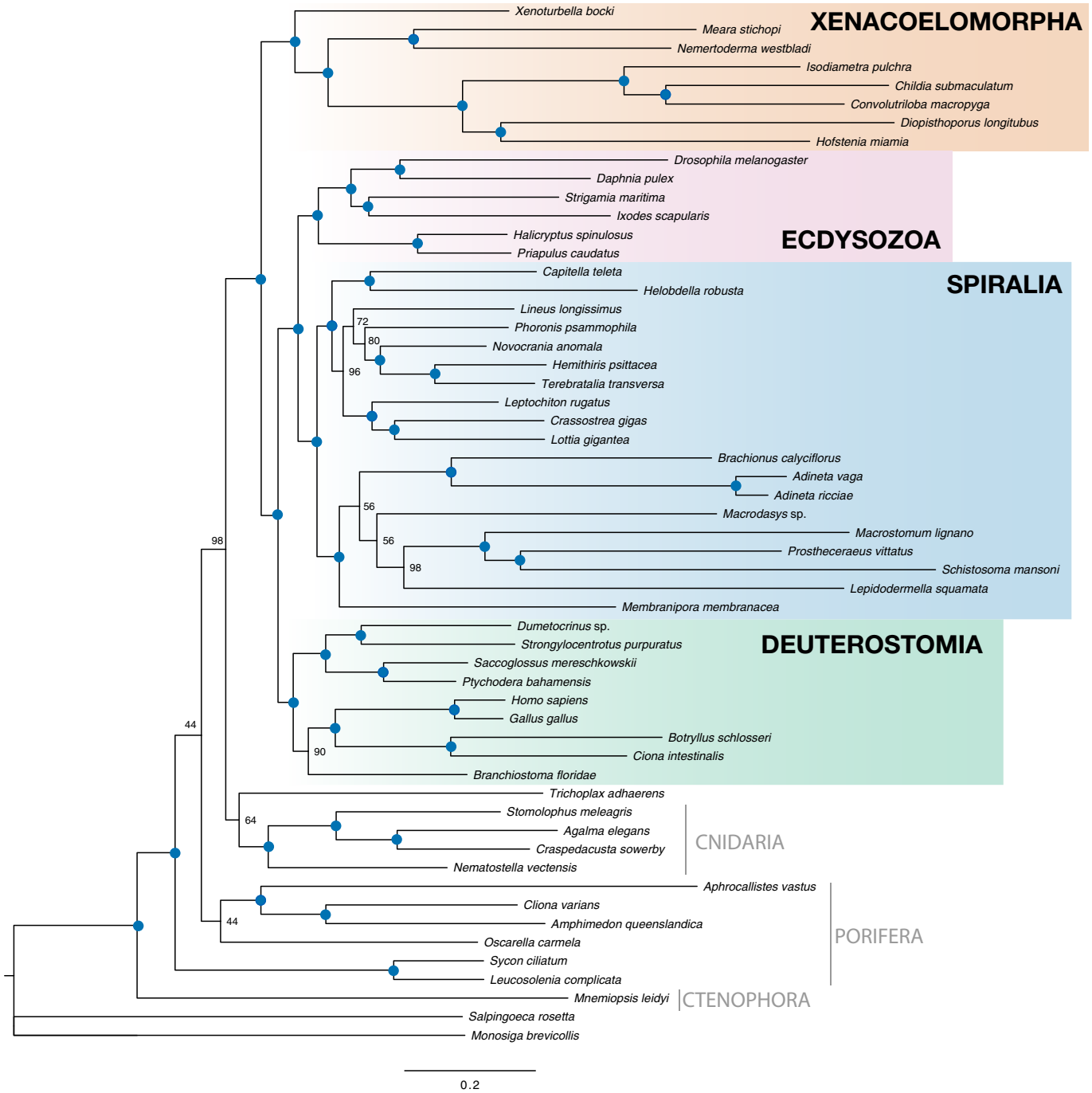
Phylogenetic analysis. Maximum likelihood analyses of the complete 212 orthologous group data matrix were performed using RAxML version 8.0.20-mpi[43] under the best-fitting models for each gene partition determined by ProtTest version 3.4 (ref. 44). The best fitting model for all but 3 of the 212 orthologous groups was LG, so further maximum likelihood analyses were performed using the PROTGAMMAILG option. Bootstrapped trees from the 212-gene data set were used to calculate leaf stability indices of each operational taxonomic unit using the Roguenarok server (http://www.exelixis-lab.org/). Bayesian analyses were conducted using PhyloBayes-MPI[45] 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 3.2 (ref. 46). 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 for 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 with 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, modelling 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 4,000,000 generations, sampling every 500 generations ('mcmcp nrun = 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 at the Dryad Digital Repository.

We additionally used ASTRAL[26] to calculate an optimal bootstrapped species tree from individual RAxML gene trees decomposed into quartets.

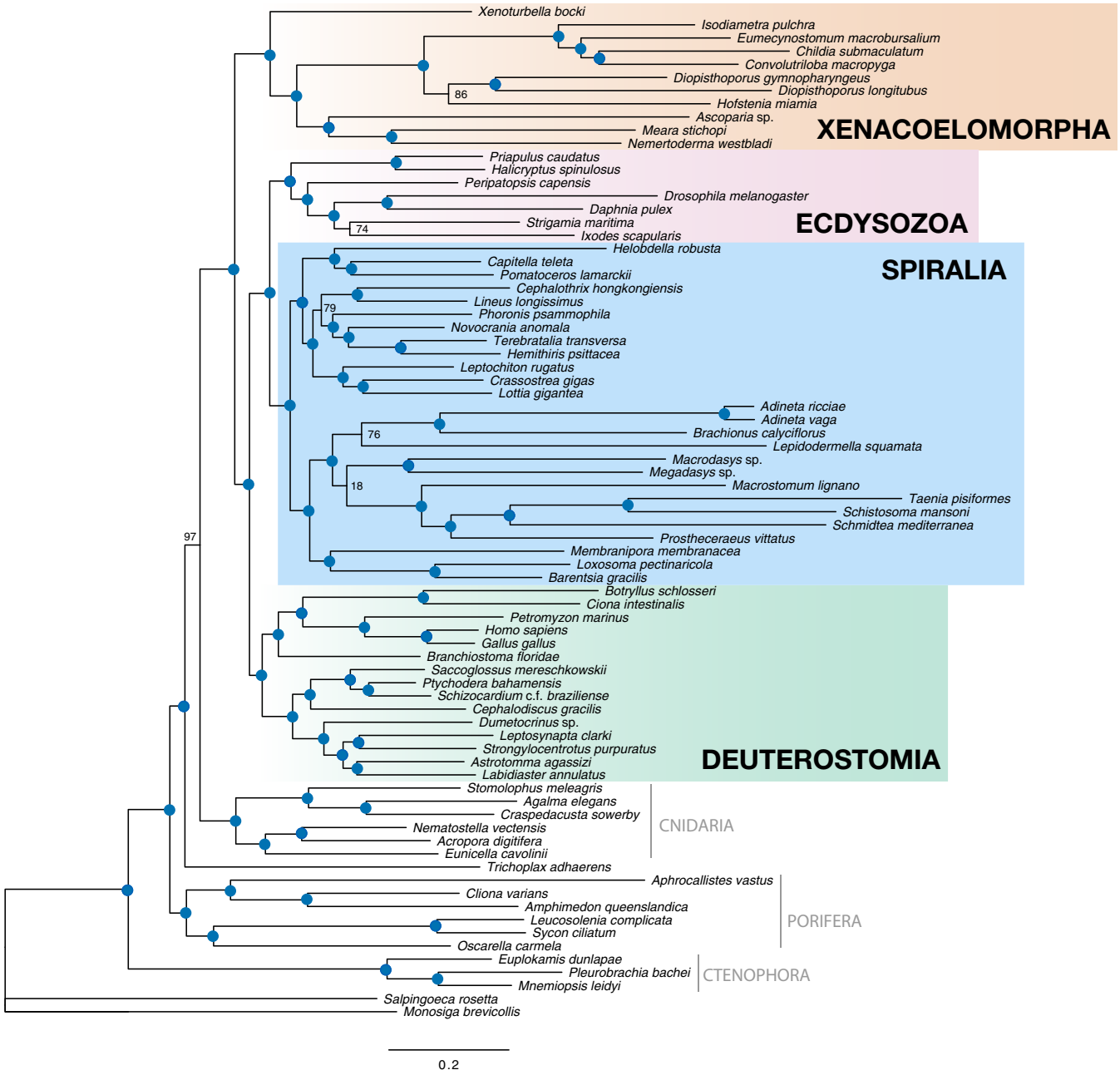
31. Grabherr, M. G.*et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol.* **29,** 644–652 (2011).
32. Ebersberger, I., Strauss, S. & von Haeseler, A. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.* **9,** 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).
43. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
44. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27,** 1164–1165 (2011).
45. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62,** 611–615 (2013).
46. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61,** 539–542 (2012).

© 2016 Macmillan Publishers Limited. All rights reserved



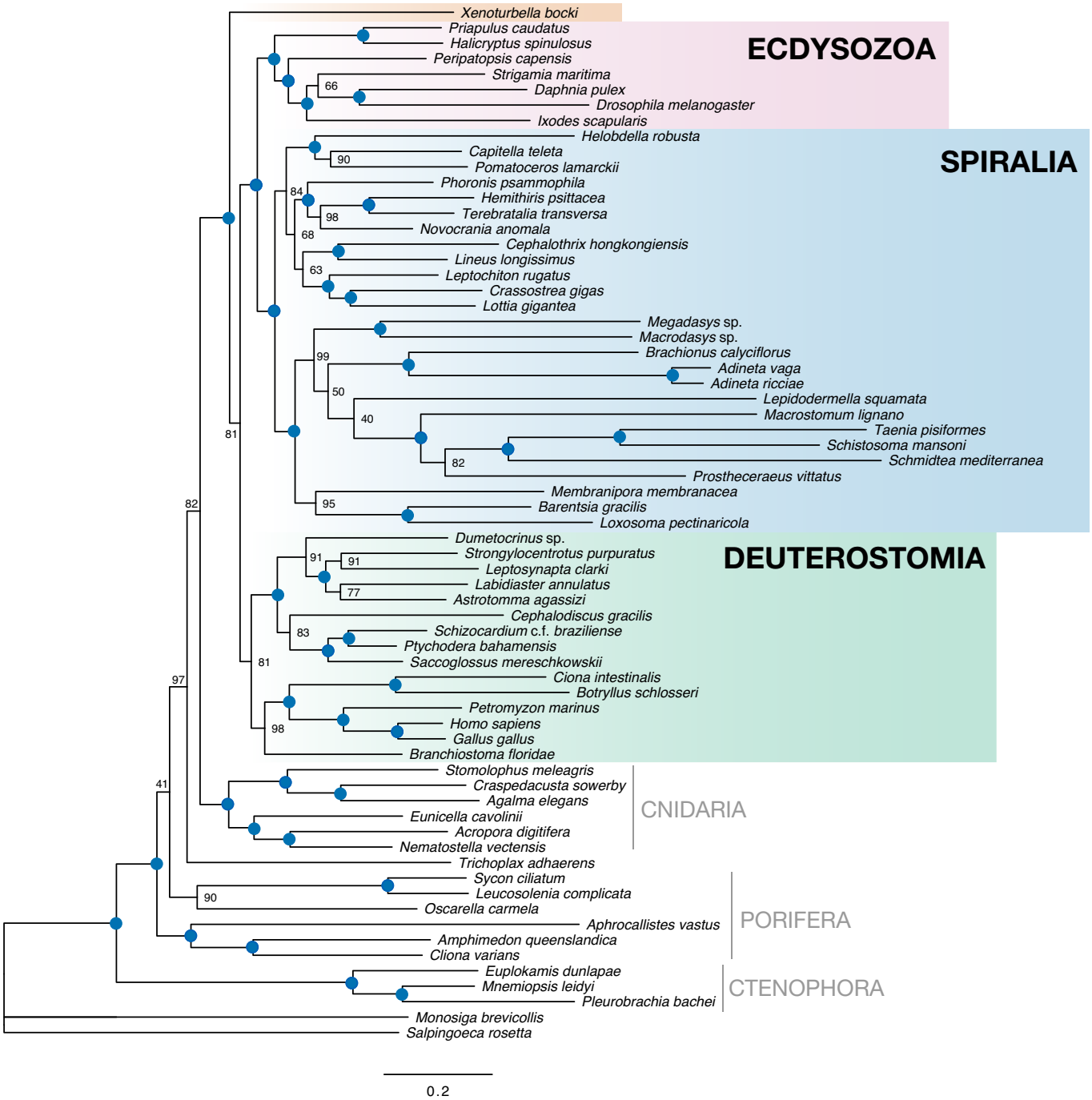
Extended Data Figure 1 | Maximum likelihood topology of metazoan relationships inferred from 336 genes from the best-sampled 56 taxa. Maximum likelihood tree is shown as inferred using the LG + I + Γ model for each gene partition, and 100 bootstrap replicates. Filled blue circles represent 100% bootstrap support. The length of the matrix is 81,451 amino acids and overall matrix completeness is 89%.

© 2016 Macmillan Publishers Limited. All rights reserved



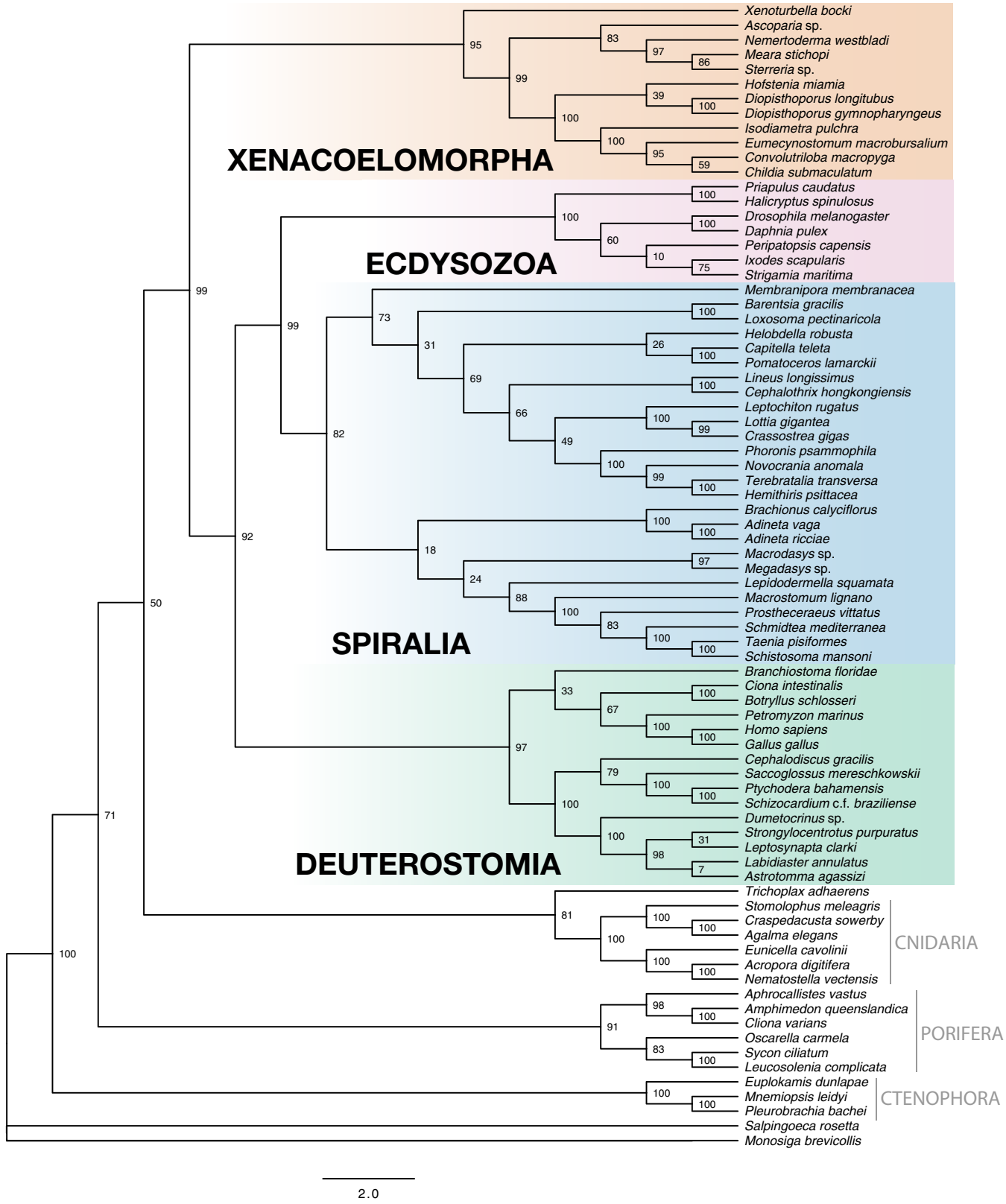
Extended Data Figure 2 | Maximum likelihood topology of metazoan relationships inferred from 881 genes and 77 taxa. Maximum likelihood tree is shown as inferred using the LG + I + Γ model for each gene partition, and 100 bootstrap replicates. Filled blue circles represent 100% bootstrap support. The length of the matrix is 337,954 amino acids and overall matrix completeness is 62%.

© 2016 Macmillan Publishers Limited. All rights reserved



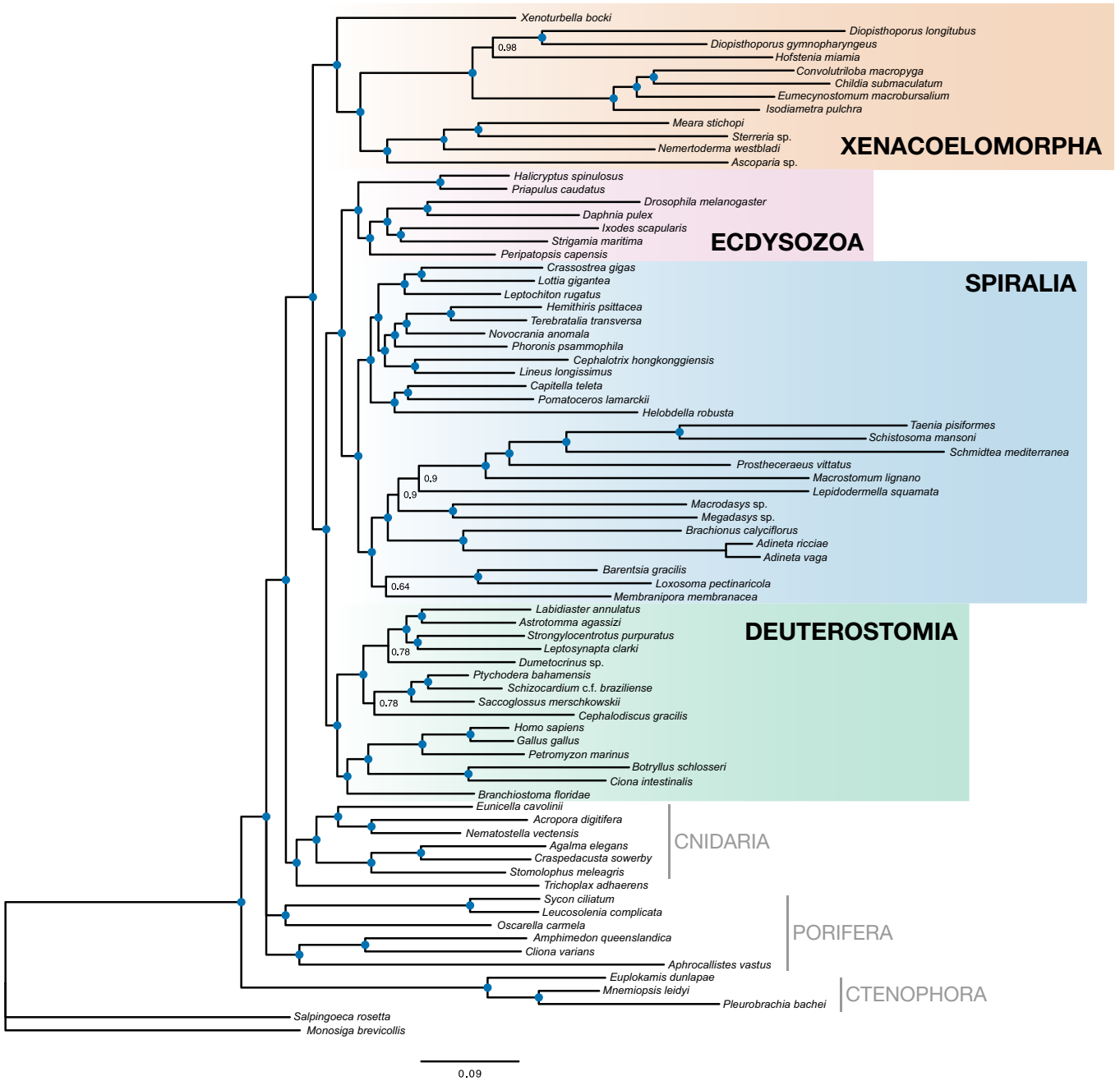
Extended Data Figure 3 | Maximum likelihood topology of metazoan relationships inferred from 212 genes with Acoelomorpha removed. Maximum likelihood tree is shown as inferred using the LG + I + Γ model for each gene partition, and 100 bootstrap replicates. Filled blue circles represent 100% bootstrap support. The length of the matrix is 43,942 amino acids and overall matrix completeness is 70%.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 4 | ASTRAL species tree, constructed from 212 input partial gene trees inferred in RAxML version 8.0.20. Nodal support values reflect the frequency of splits in trees constructed by ASTRAL from 100 bootstrap replicate gene trees.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 5 | Bayesian inference topology of metazoan relationships inferred on the basis of 212 genes and 78 taxa. Results are shown from MrBayes analyses of four independent Metropolis-coupled chains run for 4,000,000 generations, with sampling every 500 generations. Amino-acid data were back-translated to nucleotides and analysed under an independent substitution model.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Summary of data sets analysed in this study and support for monophyly of major groups

Dataset description	Number of OGs	Number of Taxa	AA positions	% Missing Data	Xenacoelomorpha	Nephrozoa	Bilateria
HaMStR all taxa	212	78	44896	31	100/100/100	99/100/100	100/100/100
HaMStR best coverage taxa	336	56	81451	11	100	100	100
ProteinOrtho	881	77	337954	38	100	100	100
Remove Acoelomorpha	212	67	43942	30	N/A	81	100
Remove *Xenoturbella*	212	77	43510	31	(Acoelomorpha 100)	70	100
Remove Acoela	212	71	43451	31	100/100	100/1.0	100/100
Remove Nemertodermatida	212	74	45054	30	100	100	100
Remove Ctenophora	212	75	47011	30	100	100	100
Remove Cnidaria	212	72	44990	31	100	100	100
Remove Porifera	212	72	43829	31	100	100	100
Remove Placozoa	212	77	43940	31	100	100	100
Porifera only non-bilaterian Metazoa	212	68	47115	30	100	100	100
Remove non-metazoans	212	76	43764	31	100	100	100
Reduce deuterostomes	210	74	46101	29	100	100	100
Taxa >80% gene occupancy only	212	52	43868	16	100	99	100
Taxa >90% gene occupancy only	212	40	42840	11	100	99	100
Remove taxa with LB score >13	212	59	43247	30	100	98	100
Remove taxa with LB score >30	212	73	44260	30	100	100	100
Genes with best LB scores	106	78	22295	30	100	71	100
Genes with poor LB scores	106	78	22601	32	100	99	100
Genes with lowest saturation	106	78	23414	29	100	95	100
Genes with highest saturation	106	78	21482	34	100	100	100
Only non-ribosomal protein genes	207	78	44715	32	100	100	100
Ribosomal protein genes, LG all partitions	53	78	9010	19	non-monophyletic	non-monophyletic	88
BMGE trimming	Merged	78	33323	34	100	100	100

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 data set derived from our amino-acid alignment, and PhyloBayes analyses under the CAT + GTR + Γ model.

© 2016 Macmillan Publishers Limited. All rights reserved