Novel genomic approaches support Xenacoelomorpha as sister to all Bilateria

Marta Álvarez-Presas

marta.alvarez-presas@bristol.ac.uk

University of Bristol https://orcid.org/0000-0002-4825-9965

Iñaki Ruiz-Trillo

Institute of Evolutionary Biology (IBE, CSIC-Universitat Pompeu Fabra) https://orcid.org/0000-0001-6547-5304

Jordi Paps

University of Bristol https://orcid.org/0000-0003-2636-6636

Research Article

Keywords: phylogenomics, genome-wide analyses, gene content, Nephrozoa hypothesis, Long-Branch Attraction, DTL-based marker selection, single-copy taxa (SCT), evolutionary transitions, systematic errors, novel gene families, Acoelomorpha, Xenambulacraria, Xenoturbella

Posted Date: November 29th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5529390/v1

Additional Declarations: The authors declare no competing interests.

Abstract

One of the most contentious debates in animal evolution is whether the first bilaterian animal was simple or complex. At the centre of this debate is the phylogenetic placement of the enigmatic lineage known as Xenacoelomorpha. For the last 25 years, different phylogenies have shown Xenacoelomorpha to be either a sister group to the rest of the bilaterians or inside the deuterostomes, with some analyses challenging the monophyly of deuterostomes. To solve this conundrum, we mined thousands of gene families from 91 complete genomes and analysed them using various alignment-free genome-wide evolutionary processes. We also applied these approaches to inform the construction of a novel metazoan gene matrix 15 times larger than previous datasets. This dataset was also used to reconstruct the protein complement present in the genome of the last common ancestor of bilaterians. These datasets were thoroughly analysed with approaches and evolutionary models to reduce systematic artifacts. Our results strongly support Xenacoelomorpha as the sister group to all other bilaterians and the monophyly of deuterostomes, pointing to a simple bilaterian, while illustrating the phylogenetic potential of genomic-level approaches.

Introduction

Most animals, 98% of metazoan species, are bilaterians, the lineage that exhibits bilateral symmetry. Discerning the nature of the first bilaterian is essential for reconstructing the evolution of their biology, including genomes, life cycles, or organ systems. However, the uncertainty around the root of the bilaterians remains a major challenge. There are four main hypotheses about early Bilaterian relationships (Fig. 1A). Twenty-five years ago, molecular phylogenies placed Acoelomorpha (acoel flatworms plus nemertodermatids; Fig. 1B) (*1-4*) as sister to the other bilaterians, backing a simple bilaterian ancestor (*1–8*); meanwhile, the flatworm-like *Xenoturbella* (Fig. 1B) (*5*) was often recovered as a sister to Deuterostomia (Fig. 1A, *hypothesis 4*) (*7*, *9*). Later analyses united Acoelomorpha and *Xenoturbella*[10] under a new group named Xenacoelomorpha [11], initially placed as a sister to Nephrozoa, the clade encompassing all other bilaterians (the Nephrozoa hypothesis; Fig. 1A, *hypothesis 1*) [10, 12–14]. However, recent analyses have provided two additional hypotheses. One that positions Xenacoelomorpha as the sister group to Ambulacraria (echinoderms and hemichordates; Xenambulacraria hypothesis; Fig. 1A, *hypothesis 2*) [11], and another that challenges the monophyly of Deuterostomia [15–17] (Fig. 1A, *hypothesis 3*). Thus, despite extensive studies, the evolutionary branching of Xenacoelomorpha and early bilaterians remains controversial.

Phylogenomic approaches infer evolutionary trees from alignments of a large number of genes. Phylogenomics has been key to improving our knowledge of the animal Tree of Life. However, some nodes remain contentious due to limited phylogenetic resolution. This lack of resolution in phylogenomic trees may partly stem from systematic errors [18–24], which occur when evolutionary models are mis-specified due to heterogeneity in molecular evolutionary rates among sequences. These result in the so-called long-branch attraction artifact (LBA)(*25*), in which fast-evolving taxa with long branches, such as

Xenacoelomorpha, artifactually branch together close to the root of the tree. Systematic errors can be minimised by using models that account for the heterogeneous rates across sites in a protein sequence (e.g., the CAT model used in Bayesian inference (BI) approaches) [22], amino acid recoding [22, 26, 27] or using closer outgroups (*28*). Complete genomes, whose numbers have increased in recent years, allow the use of genome-level evolutionary processes to infer phylogenies; this is an alternative source of phylogenetic information to alignment-based approaches, and may minimise some of the biases of the latter. The full genomic content provides a comprehensive view of evolutionary history of animals and opens the door to applying more advanced and complex methods for phylogenetic inference. A recent example is the study of early animal relationships using ancient linkage groups (*29*). Unfortunately, the lack of chromosome-level genomes for Xenacoelomorpha and most invertebrate lineages prevents this type of analyses by the moment; for example, a recent study applying this approach to one xenoturbellid genome was not able to solve its position within bilaterians (*30*), reinforcing the need for approaches like the ones presented here.

Here we combined the potential of complete protein-coding gene complements and approaches to reduce systematic errors to elucidate the phylogenetic position of Xenacoelomorpha. Starting from more than 100 complete genomes, we employed three innovative approaches to infer trees based on genome-level processes: gene content, modes of gene family evolution (gene duplications, transfers, and losses, DTL), and the multispecies coalescence model (MSC). These are recently developed methods that have applied state-of-the-art approaches to a range of phylogenetic problems, but never thoroughly applied to the Bilateria. Furthermore, we constructed a novel phylogenomic dataset that not only minimises missing data and orthology assignment problems, but its dimensions provide a robust foundation for more comprehensive phylogenetic analyses (Fig. 1C). Importantly, we included other fast-evolving lineages (e.g., nematodes, flatworms), which serve as “canaries in the coal mine”, helping to diagnose possible artifacts affecting our results; these taxa are prone to LBA grouping together in some studies (e.g., *31*) and excluded in others (*12*, *13*, *17*). These phylogenomic datasets were analysed with approaches aimed at reducing LBA, and “canaries” were used to test their effectiveness. Overall, our comprehensive and diverse analyses robustly support the placement of Xenacoelomorpha as the sister group to all the other bilaterians and the monophyly of Deuterostomia. Furthermore, leveraging this large dataset, we reconstructed the protein-coding content of the genome of the last common ancestor of bilaterians.

Gene content supports the monophyly of Nephrozoa and Deuterostomia

Briefly, we mined proteomes from more than 100 whole genomes of animals and their close relatives (see Materials and Methods in SM; table S1). We removed genomes with more than 15% missing BUSCO genes, retaining 91 complete genomes (table S1) which include 81 holozoans (animals and their close relatives) and 10 non-holozoan eukaryotes (e.g., plants, fungi, etc.). The proteins in all the genomes were clustered into homology groups (HGs), sets of genes that share a common gene ancestor and include

orthologs and paralogs. We used three distinct taxon samplings in our datasets with different outgroups: the original Holozoa (75 metazoans and six unicellular holozoans), Metazoa (75 taxa), and Planulozoa (64 taxa, 14 cnidarians and 50 bilaterians).

First, we inferred species trees using gene content. The presence/absence of HGs in different taxa is used to build a binary phylogenetic matrix (1 for genes present in a taxon, a 0 if absent). The matrix is analysed using Bayesian methods under both the Mk model, which uses a continuous-time Markov process to model state changes, and the Dollo model, which assumes that lost genes can never be regained. We analysed three different datasets (e.g., the one for holozoans containing 549,544 HGs) using *RevBayes* (*32*) under the continuous-time Markov chain (CTMC) for both models, following the analyses in (*33;* see Materials and Methods in SM). All the datasets and models recovered the same topology (Fig. 2A and fig. S1), showing all the major bilaterian lineages with maximum support, including Nephrozoa (Xenacoelomorpha sister to Bilateria) and monophyletic Deuterostomia. Despite the presence of other long-branched taxa (e.g., *Oikopleura dioica*, four flatworms, and four nematodes), there is no evidence of LBA within nephrozoans (e.g., flatworms and nematodes are not grouped together). These results are consistent with recent gene content studies, which supported Xenacoelomorpha as an early divergent lineage within Bilateria (*33, 34*). However, gene length, substitution rates, and other systematic errors may impact homology assignment (*20, 23, 35*) and therefore gene content analyses. Future approaches using structure-based tools like Foldseek (*36*) could help to improve homology assignment and reduce these issues.

Modes of gene evolution support Nephrozoa and the monophyly of Deuterostomia

The likelihood of gene duplication, gene transfer, and gene loss (DTL) across all gene families can be used as source of phylogenetic data. While gene content analyses use presence/absence of HGs, DTL approaches use the evolutionary history within an HG as a source of phylogenetic information. Starting from individual gene trees for each HG, the likelihood of every gene tree is calculated using a model with parameters describing the rate of gene duplication, transfer, and loss. Then, the joint reconciliation likelihood of all the gene trees can be calculated for a given species tree. (*37, 38*). We used two approaches based on DTL, the *amalgamated likelihood estimation (ALE)* (*38*) and *SpeciesRax* module implemented in *GeneRax* software (*37*). *ALE* accounts for gene tree incongruence due to DTL events and infers reliable DTL parameter estimates through conditional clade probabilities (*39*), while *SpeciesRax* estimates the optimal rooted species tree that maximises the joint reconciliation likelihood of all the gene trees. We filtered the HGs in the 81 holozoan dataset (see Materials and Methods in SM) to exclude those with less than 4 sequences or 3 species, yielding a final dataset with no orphans or singletons. We aligned and trimmed each of the remaining 31,120 HGs and further filtered by length and removed the shortest 25%. Then we inferred gene trees for each of the resulting 23,263 holozoan HGs. The species tree based on the likelihood of DTL evens estimated by *SpeciesRax* (*37*) supports Nephrozoa and deuterostomes monophyly but shows LBA between protostome lineages (fig. S2).

Due to the low incidence of lateral gene transfers within animals, we argue that most transfers identified by *ALE* are in fact the result of phylogenetic noise (*39*). We tested the effect of removing the gene markers with most transfers in all four competing trees from the Holozoan dataset (fig. S3); in one dataset we removed 25% of the markers with most transfers (*ALE25T,* 17,260 HGs), and in another we removed all the markers with transfers (*ALEnoT*, 8,439 HG). *CONSEL* (*40*) was used to perform topology comparison tests on the four hypotheses using the *ALE25T* markers (table S2), rejecting *hypotheses 2, 3*, and *4* but not *hypothesis 1* (Nephrozoa, Fig. 1A). The *ALE25T* dataset was also analysed to infer a holozoan phylogeny based on DTL using *SpeciesRax* (Fig. 2B). The results placed Xenacoelomorpha as a sister group to the other bilaterians (Nephrozoa) and recovered deuterostomes monophyly, all with positive EQPIC values in most nodes (indicating no high incongruence between the gene trees). In contrast to the *SpeciesRax* tree inferred using unfiltered markers (fig. S2), this tree shows no indication of LBA between fast-clock taxa and recovers the major animal clades within bilaterians.

Instead of estimating the joint likelihood of a species tree based on the DTL parameters in all gene trees derived from whole HGs, which comprise orthologs and paralogs, an alternative genome-level approach is to summarise gene family trees from orthologs employing the *multispecies coalescent model* (MSC). The MSC reconciles discrepancies between gene trees and species trees caused by processes such as *incomplete lineage sorting* (ILS), which arises when the allelic variation present in an ancestral population is not sorted into descendant lineages according to the species tree. We analysed the holozoan, metazoan, and planulozoan datasets, inferring gene family phylogenies via Maximum Likelihood (ML; table S3). These gene trees were analysed with *ASTRAL-Pro* (*41*), designed to handle multicopy gene trees while accounting for ILS to infer a species tree in the presence of complex evolutionary histories. However, *ASTRAL-Pro* can't handle datasets with high number of paralogs; for these datasets, we used standard *ASTRAL* software (*42*) after extracting the orthologs from these datasets with *PhyloTreePruner* (PTP), *Partition Finder* (PF), and *Harris' script* (see Materials and Methods in SM). Most of the supertrees inferred (figs. S4 and S5) show low phylogenetic signals and display Nephrozoa LBA. Only the dataset filtered with PTP (fig. S6) recovered all the main groups as monophyletic, except for Xenacoelomorpha, which was split into *Xenoturbella* as a sister to Deuterostomia, and Acoelomorpha as early-branching Bilateria (Fig. 1A, *hypothesis 4*). We also analysed the *ALE25T* dataset with *ASTRAL-Pro*, and the resulting trees recovered the major clades of the animals (Fig. 3A and fig. S5), with no evidence of LBA and high statistical support for most of the clades. In these resulting trees, Deuterostomes are monophyletic but with no support, and Xenacoelomorpha is placed as a sister to Protostomia. In contrast, a single previous study that used transcriptome data with *ASTRAL* supported the Nephrozoa (*13*). Our datasets may comprise genes with complex evolutionary histories for which the MSC is not able to resolve ancient divergences.

Genome-informed phylogenomic datasets support the monophyly of Nephrozoa and Deuterostomia

We constructed a novel phylogenomic alignment from complete genome data to overcome the orthology issues caused by missing data (orthologs and paralogs both captured during sequencing or annotation, *43*) while maximising taxon and gene sampling (see Materials and Methods in SM). From the 81

holozoan dataset (49,419 HGs), we extracted orthologs with the three methods used above (*ParaFilter* (PF), *PhyloTreePruner* (PTP)(*44*), and a custom script from *Harris et al.* (*45*); table S3; see Materials and Methods in the SM). The orthologs comprising at least four sequences were concatenated to construct a holozoan phylogenomic dataset. For instance, the matrix from *ParaFilter* orthologs comprises 14,185 gene markers and more than five million positions, the largest protein-coding alignment ever assembled for bilaterians, surpassing previous ones by a factor of 15 (*HolPF* in Fig. 1C).

ML trees revealed consistent support for the Nephrozoa hypothesis and the monophyly of Deuterostomia (fig. S7). However, all the trees showed signs of LBA between the other bilaterian fast-evolving taxa (e.g., nematodes, flatworms, and urochordates). To minimise the effect of LBA, we applied different recoding strategies (SR4, Dayhoff6, Dayhoff9, and Dayhoff12) *(46*) aimed at reducing molecular heterogeneity, but this approach also failed to effectively reduce the artifacts (fig. S8A). Unfortunately, these datasets are too large for the application of BI approaches using CAT-derived models.

Our first approach to minimise LBA issues capitalised on our genome-level processes. We used the datasets with reduced noise (*ALE25T and ALEnoT*), to extract orthologs with *PhyloPyPruner (PpP)* (*47*) see Materials and Methods in SM, table S3). *PhyloPyPruner* by default extracts orthologs by retaining the largest subtree with branches showing at least 50% bootstrap support and removing branches exceeding five times the standard deviation in length (Fig. 3B, table S3, 6,246 orthologs from *ALE25T* and 2,720 orthologs from *ALEnoT*). Furthermore, we built an *ALE25Tstrict* dataset applying more restrictive parameters (fig. S9 and table S3): using Cnidaria as outgroup, a minimum of 10 taxa per HG; this resulted in a smaller dataset of 624 orthologs with reduced missing data (from 90% to 70%) and producing consistent results with the less restrictive dataset (*ALE25T*). While some gene families are shared by only a few lineages, introducing ambiguous characteristics (fig. S10 and table S3), the ML trees show no internal LBA. Interestingly, the trees based on the ALE25T dataset are the only ML phylogenies in all this study showing no traces of the nephrozoan LBA (Fig. 3B). We also applied recoding to these datasets (*ALE25T* and *ALE25Tstrict*; fig. S8B), and the ML analyses show no internal LBA (except for *ALE25T* recoded with SR4). Unfortunately, the BI analyses for these datasets did not converge. These ML-based results demonstrate that missing data do not negatively impact the placement or statistical support of phylogenies (*48, 49*), as suggested by some previous studies (*50*), nor exacerbate the effect of LBA. This highlights the potential of using genome-scale evolutionary processes, such as DTL, to guide marker selection for reducing systematic errors in phylogenetic inference, even in the presence of missing data. The trees inferred using this approach, curating gene trees based on their evolutionary histories, support Nephrozoa and the monophyly of deuterostomes with maximum support (*hypothesis 1*, Fig. 1A).

Our second procedure aimed to minimise the impact of paralogy. Starting from 14,416 HGs found in all metazoans, we filtered HGs in which genes show a single copy for most taxa; then, taxa with more than one gene copy were removed, thus generating datasets enriched for single-copy taxa (SCT) at the expense of removing taxa that were no single copies. The *SCT50* dataset comprises HGs in which at least 50% of the taxa contained a single-copy gene (3,091 HGs, 39% missing data), while the *SCT60* dataset included genes in which at least 60% of the taxa had a single copy (1,687 HGs, 62% missing

data; Fig. 4), the *SCT70* dataset (640 HGs, 26% missing data), and the *SCT75* dataset (231 HGs, 22% missing data). For these four HG categories, we generated two datasets, one from metazoans and another from planulozoans. All eight ML trees recovered the major bilaterian superclades and supported Nephrozoa but also exhibited the protostome LBA (fig. S11 and table S4). However, in contrast with the previous phylogenomic datasets presented here, we were able to analyse the datasets *Metazoa_SCT70*, *Planulozoa_SCT60*, and *Planulozoa_SCT75* using a Bayesian approach with the CAT + GTR model (*51*). BI trees show Nephrozoa and deuterostomes with maximum support with no nephrozoan LBA (Fig. 4, fig. S12, and table S4). We also tested the effects of recoding approaches (SR4, Dayhoff6, Dayhoff9, and Dayhoff12) on these datasets (table S3 and fig. S8), and conducted BI analyses of the recoded datasets. Among these, only the *Planulozoa_SCT50* dataset reached convergence. Interestingly, this analysis supported Xenambulacraria as sister to protostomes and chordates (*hypothesis 3*, Fig. 1A, and fig. S13), although with no statistical support in the key nodes. Next, we performed a statistical comparison of competing tree topologies, which allows the rejection of trees based on the observed data (alignment); topology tests were applied to each datasets (*Metazoa_SCT50*, *Metazoa_ SCT60*, *Metazoa_ SCT70*, and *Metazoa_ SCT75*), and all of them rejected all the hypotheses except for Nephrozoa in *hypothesis 1* (Fig. 1B, table S5).

Finally, for each of the four categories of HGs enriched in single-copy taxa, we generated two further datasets to rule out potential interactions between different long-branched lineages (e.g., fast-evolving nephrozoans affecting the position of xenacoelomorphs): a metazoan dataset that removed all the long-branched nephrozoans (the “canaries”) but kept all xenacoelomorphans, and the same dataset but comprising the shortest acoel, shortest nemertodermatid, and *Xenoturbella*. This sampling strategy, the removal of other fast-clock, has been widely used in previous studies to reduce LBA, but comes at the cost of losing the canaries. This approach could potentially allow to minimise the possible LBA in xenacoelomorphans and place them within deuterostomes. However, these eight ML trees robustly support Nephrozoa and Deuterostomia (fig. S11).

Reconstruction of the ancestral bilaterian genome

A major open question is the nature of the bilaterian ancestor; here, we used our dataset to reconstruct the protein-coding content present in the genome of the LCA of Bilateria. The minimal gene complement of the first bilaterian comprised ~19,000 gene families, depending on the topology (Fig. 5A and fig. S14, table S6). Predominant molecular functions were related to catalytic metabolism, ribonucleoside triphosphate phosphatase, or lyase activity (fig. S15). Interestingly, only one novel family, the zeta finger homeobox *zfhx*, was present in all the bilaterian genomes analysed (novel core genes). Remarkably, the *zfhx* gene is expressed in the mesoderm and the central nervous system in the fruit fly, two hallmarks of the evolution of bilaterians (*52*). Under the Nephrozoa hypothesis (Fig. 1A, *hypothesis 1*), a high number of novel families are found in the bilaterian (+2,158 new HGs) and nephrozoan (+1,578) nodes (Fig. 5A and fig. S14, table S6), while in *hypothesis 2* (Xenambulacraria), these numbers were even greater in the bilaterian LCA (+3,077). The planulozoan node (bilaterians plus cnidarians) showed unprecedented levels

of innovation across all hypotheses (~3,600 novel gene families). However, neither Xenambulacraria (*hypotheses 2* and *3*) nor Nephrozoa (*hypothesis1*) presented novel core families (Fig. 5A), while in *hypothesis 4,Xenoturbella* and Ambulacraria shared three novel core families (calcium-binding proteins, the lectin domain family and GATA transcription factor). These low values cannot provide any statistical significance to favour one topology over the others. Rare genomic changes also illuminated potential molecular synapomorphies for Bilateria *(53*), such as the presence of one miRNA supporting Xenacoelomorpha (*11*) or lineage-specific gene families like NADH ubiquinone oxidoreductase in deuterostomes (*hypothesis 1)* and sperm acrosome-associated protein 9 (rsb66) in *hypothesis 2*.

Despite the draft nature of the seven Xenacoelomorpha genomes used in this study, they collectively constitute a complete representation of the proteins found in animals with combined BUSCO value of 91%. Moreover, it is unlikely that the same gene families, consisting of multiple paralogs, were not sequenced or annotated in any of the seven proteomes included here. All the Xenacoelomorpha species shared 24 novel core genes, which provided strong support for the monophyly of this lineage and the joint completeness of their genomes. Recent studies have shown the role of gene loss in major evolutionary transitions (*35, 54*), a pattern supported here by the number of gene losses in the LCA of bilaterians, which is 5-10 greater than that of previous metazoan ancestors (Fig. 5A and B, and fig. S14, table S6). These results also confirmed massive gene losses in the LCA of deuterostomes (-4,579 HGs in *hypothesis 1*; Figs. 1A and 5A, table S6). Including Xenacoelomorpha in the deuterostomes (the Xenambulacraria hypothesis, *hypothesis 2*; Fig. 1A) reduced the gene loss in the LCA of deuterostomes (-2,594) while doubling it in ambulacrarians (from -2,045 to -5,146) and chordates (from -3,552 to -7,704, Fig. 5A, table S6), with a high number of gene families (-1,019 HG) convergently lost by ambulacrarians and chordates (Fig. 5B). Remarkably, the LCA of Xenacoelomorpha lost 4-6k gene families, depending on the evolutionary hypothesis, suggesting that this lineage might have undergone further genome simplification after diverging from the other bilaterians; the top three categories of lost genes included transcription factors, enzymes, and structural proteins (fig. S16).

Discussion

Our analyses —including those with no internal LBA — strongly support Xenacoelomorpha as a sister group to other bilaterians (Nephrozoa) and confirm the monophyly of deuterostomes (Fig. 1A, *hypothesis 1*). We present five lines of evidence: 1) gene content analysis for three datasets using two different models; 2) modes of gene family evolution (gene duplications, transfers, and losses; DTLs), with trees and comparisons of topologies from filtered markers (*ALE25T*); 3) Multispecies Coalescence Model (MSC) analyses from curated markers (*ALE25T*); 4) phylogenomic analyses based on curated datasets targeting markers with low phylogenetic noise; and 5) phylogenomic analyses based on BI approaches to reduce systematic errors. The latter included a large gene matrix, a dataset with markers curated for phylogenetic noise using DTLs (*ALE25T*), and gene alignments enriched for single-copy taxa.

Previous studies have suggested that the Nephrozoa hypothesis might be a result of LBA (11, 15–17). Here, we employed genome-wide approaches and new phylogenomic datasets, including fast-evolving

taxa to detect LBA. We used a combination of methods — gene filtering, single-copy taxa enrichment, BI, recoding, and alternative outgroup selections (e.g., non-animal holozoans, non-bilaterian metazoans, or cnidarians) — to mitigate these artifacts. These strategies demonstrated the ability to eliminate LBA between other fast-clock taxa (e.g., nematodes, flatworms). However, if Xenacoelomorpha are truly sister to nephrozoans, proving that this topology is not affected by LBA might be epistemologically impossible: it always might be counterargued that any measures to mitigate LBA were not enough, and their position is artefactual. The only way to falsify the Nephrozoa hypothesis is to overcome putative LBA by using empirical data drawn from a broad taxon and effective gene sampling derived from complete genomes, placing them elsewhere in the tree with high statistical confidence. We attempted this in this study, but xenacoelomorphans were still robustly placed as sister to nephrozoans.

Our datasets were analysed with methods designed to mitigate systematic errors. This included evaluating datasets with alternative outgroups and including fast-evolving taxa typically excluded to test for LBA independently of Xenacoelomorpha. Outgroup selection, however, had no impact on the occurrence of internal LBA. Notably, genome-level approaches such as DTL-based marker selection (*ALE25T*, Fig. 3A) and gene content analyses (Fig. 2A), proved less sensitive to systematic errors compared to ML analyses of datasets maximising gene sampling, which continued to display internal LBA even with strategies such as recoding. While these ML datasets were intractable for BI, phylogenomic datasets comprising genes enriched for single-copy taxa (SCT) analysed with BI showed no nephrozoan LBA. Among the methods tested, the markers curated for noise (*ALE25T*), produced the only ML trees in this study free from LBA effects (Fig. 3B). While it has been argued that missing data introduce systematic errors (e.g., LBA) and reduce statistical support (50), we observed no such effects in our *ALE25T* tree. Thus, we propose that DTL-based marker selection targeting low phylogenetic noise can effectively mitigate systematic errors, even if some markers are not present in all taxa.

Limitations in our analyses include the use of draft genomes. Xenacoelomorphan genomes showed variable levels of missing eukaryotic BUSCO genes (5.4% in *Hofstenia miamia* to 36.1% in *Meara stichopi*, table S1). However, the combined gene complement of all Xenacoelomorpha genomes showed only 9% missing BUSCO eukaryotic genes, indicating that genes absent in one genome were present in others; collectively, these genomes capture the lineage gene diversity required for robust genome-level analyses. Moreover, some of our approaches should not be affected by missing genes (e.g., phylogenomic markers present in all metazoans and enriched for single-copy taxa are also present in all Xenacoelomorpha).

Another issue is the conflicting results on the root of animals and within xenacoelomorphans, although solving these were out of the scope of this study. Ctenophores-first is supported by DTL trees (Fig. 2B), the MSC (Fig. 3A), and *ALE25T* ML tree (Fig. 3B); meanwhile, sponges-first is supported by gene content analyses (Fig. 2A, although sponges are paraphyletic), and the *ALE25T* phylogenomic analyses recoded with Dayhoff6. One possible explanation is that the number of genomes available for sponges and ctenophores used here might not be sufficient to properly profile genome-scale evolutionary processes. This lack of taxon sampling for sponges and comb jellies, represented each only by two genomes, may

explain their instability. Similarly, the limited availability of high-quality xenacoelomorphan genomes at the time of our analyses, displaying different levels of completeness, made it difficult to fully resolve or confirm the internal relationships within Xenacoelomorpha. A recent study has argued that acoelomorphs are result of LBA, and when this is corrected, acoels are sister to xenoturbellids (17, 55). Acoels show derived nervous systems compared to nemertodermatida and xenoturbellids (56), while the latter share digestive system structures (*57*). Our analyses generally support Acoelomorpha using gene content (Fig. 2A), DTLs (Fig. 3A), as well as ML and BI analyses from datasets minimising paralogy (Fig. 4). In contrast, the *ALE25T* dataset recovers Acoela + *Xenoturbella* (Fig. 3B), and Nemertodermatida sister to *Xenoturbella* in DTL and MSC analyses (Fig. 2B). These inconsistencies may stem from internal LBA, uneven genome quality, or limited taxon sampling (our dataset includes only one Xenoturbellida and two Nemertodermatida genomes). Resolving these relationships will require improved genome quality and expanded taxon sampling, particularly for underrepresented groups like xenoturbellids, nemertodermatids, sponges, and ctenophores.

The placement of Xenacoelomorpha as the sister group to Nephrozoa would, potentially, support the hypothesis of a simple first bilaterian animal, with no coelom or complex organs, direct development, and a meiofaunal lifestyle. However, organisms are mosaics of characters, and different traits could have undergone different evolutionary trajectories. For example, high levels of gene loss in the xenacoelomorphans' LCA, which comprises several transcription factors (fig. S16), and the lower gene content in their ancestor compared to other nodes, seem to indicate simplification in the LCA of xenacoelomorphans. However, we found genes typically linked to nephridial systems (eya, six 1/2, pou3, sall, lhx1/5, osr (58)) in the xenacoelomorphan genomes in our analyses. This would support the hypothesis that xenacoelomorphans most likely present the bilaterian ancestral state for genes linked to this trait and that the first bilaterian had no nephridia.

The genome of the LCA of extant bilaterians was abundant in new molecular functions (fig. S15). We report a novel bilaterian core gene family, the *zfhx* gene, which had not been identified in previous studies (35, 59), likely due to differences in taxon sampling and methods. The *zfhx* genes are zeta finger homeobox genes involved in the development of the mesoderm and central nervous system (52), two key innovations in bilaterians. However, it is important to recognise that no single gene family drives evolutionary transitions. Instead, the evolution of complex traits like the mesoderm and nervous system is shaped by the regulation of multiple gene families, including *zfhx*. Patterns of gene turnover highlight the complexity of the modes of gene evolution: early nodes (planulozoans, bilaterians, and nephrozoans) showed high levels of gene gain, while internal nodes of bilaterians (deuterostomes and protostomes) were marked by a dramatic increase in gene loss. This highlights the dynamic nature of gene evolution during these major transitions, with novelty being more predominant in early nodes and gene loss being more predominant in later nodes.

The increasing availability of high-quality genomes from a broader range of organisms has advanced the use of genome-level characters, such as gene content (33, 34), DTLs (37, 38), or ancient linkage groups (29), in phylogenetics. While the use of syntenies holds promises, the limited availability of chromosome-

level genomes currently restricts its application and may not always provide phylogenetic signal. For instance, a recent study using syntenies with a *Xenoturbella* genome was not able to resolve its position within Bilateria using ancient linkage groups (30). Here, we demonstrate how genome-wide approaches can be harnessed to address recalcitrant phylogenetic nodes and inform traditional gene-alignment-based methods. Furthermore, we used these analyses to produce new metazoan phylogenomic matrices that we hope will prove valuable to the scientific community. The combination of these analytical methods introduces fresh and promising avenues for reconstructing the evolutionary history of organisms.

Declarations

Acknowledgments

This work used the ACRC HPC University of Bristol, the genomics server of the Genomics facility (School of Biological Sciences, University of Bristol), and the Isambard 2 UK National Tier-2 HPC Service. Callum Wright, Francisca Segers, and Tom Green are thanked for their IT assistance. Mattia Giacomelli is thanked for providing the ParaFilter script. We thank Pere Martinez and Xavier Bailly for providing images of the organisms. We thank Tom Williams and Davide Pisani for their comments and suggestions on the analyses. Meritxell Antó is thanked for her contribution to the design of the figures. Peter Holland, Phil Donoghue, Tom Williams, Marta Riutort, Davide Pisani and Jaume Baguñà are thanked for their comments on an early draft of the manuscript. The three authors would like to personally thank Full Professors Marta Riutort and Jaume Baguñà, who started us down this path 25 years ago.

Funding:

MÁ-P and JP are supported by the Wellcome Trust (210101/Z/18/Z) and the School of Biological Sciences (University of Bristol). MÁ-P is supported by a fellowship from the Fundación General CSIC´s ComFuturo programme which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101034263.

IR-T is funded by the European Union (ERC, MISSINGRELATIVES, 101097659). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The authors are also supported by the grant PID2020-120609GB-I00 funded by MICIU/AEI/10.13039/501100011033/and by “ERDF A way of making Europe”.

Authors' contributions:

Conceptualization:MÁ-P, JP, IR-T

Methodology: MÁ-P, JP, IR-T

Data analysis: MÁ-P

Writing: MÁ-P, JP, IR-T

Competing interests: The authors declare no competing interests.

Data availability: Data and scripts are available at: https://github.com/MartonaAlvarez/XAN, and bigger files will be available at figshare or upon request.

References

1. Ruiz-Trillo I et al (2002) A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proc. Natl. Acad. Sci. U. S. A. 99, 11246–11251
2. Carranza S, Baguñà J, Riutort M (1997) Are the Platyhelminthes a monophyletic primitive group? An assessment using 18S rDNA sequences. Mol Biol Evol 14:485–497
3. Ruiz-Trillo I, Riutort M, Littlewood DTJ, Herniou EA, Baguñà J (1999) Acoel Flatworms: Earliest Extant Bilaterian Metazoans, Not Members of Platyhelminthes. Science 283:1919–1923
4. Ruiz-Trillo I, Paps J (2016) Acoelomorpha: earliest branching bilaterians or deuterostomes? Org. Divers Evol 16:391–399
5. Ruiz-Trillo I, Riutort M, Fourcade HM, Baguñà J, Boore JL (2004) Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. Mol Phylogenet Evol 33:321–332
6. Wallberg A, Thollesson M, Farris JS, Jondelius U (2004) The phylogenetic position of the comb jellies (Ctenophora) and the importance of taxonomic sampling. Cladistics 20:558–578
7. Paps J, Baguñà J, Riutort M (2009) 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
8. Paps J, Baguñà J, Riutort M (2009) Lophotrochozoa internal phylogeny: new insights from an up-to-date analysis of nuclear ribosomal genes. Proc. R. Soc. B Biol. Sci. 276, 1245–1254
9. Bourlat SJ, Nielsen C, Lockyer AE, Littlewood DTJ, Telford M (2003) J. *Xenoturbella* is a deuterostome that eats molluscs. Nature 424:925–928
10. Hejnol A et al (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. R. Soc. B Biol. Sci. 276, 4261–4270
11. Philippe H et al (2011) Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470:255–260
12. Rouse GW, Wilson NG, Carvajal JI, Vrijenhoek RC (2016) New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. Nature 530:94–97
13. Cannon JT et al (2016) Xenacoelomorpha is the sister group to Nephrozoa. Nature 530:89–93
14. Laumer CE et al (2019) Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc. R. Soc. B Biol. Sci. 286, 20190831

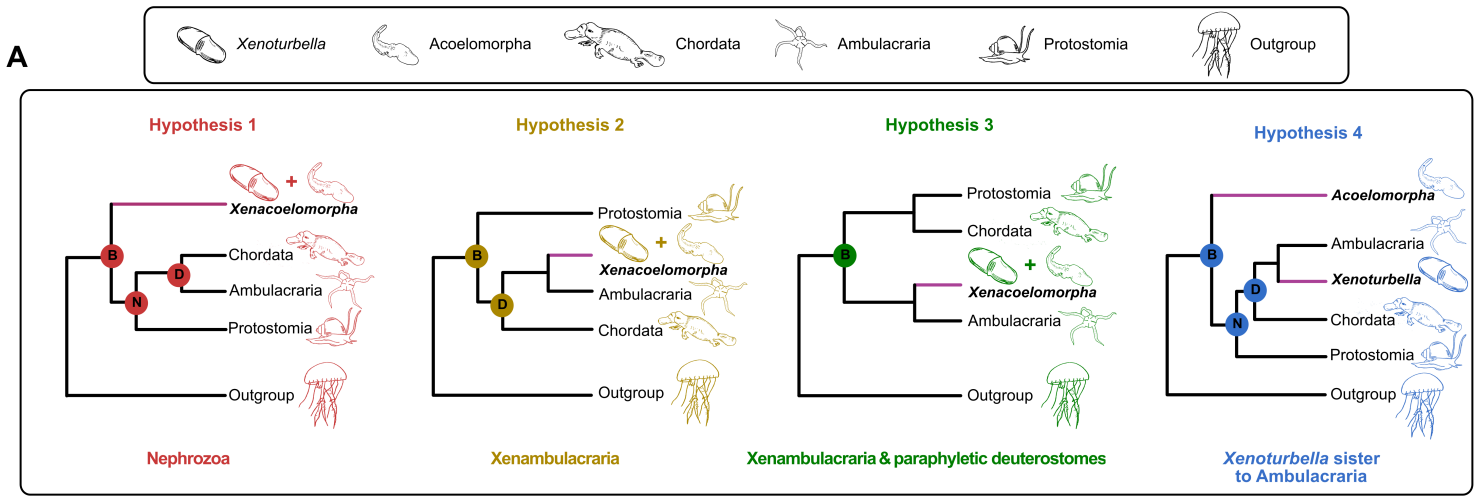
15. Simakov O et al (2015) Hemichordate genomes and deuterostome origins. Nature 527:459–465
16. Marlétaz F, Peijnenburg KTCA, Goto T, Satoh N, Rokhsar DS (2019) A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans. Curr Biol 29:312–318e3
17. Philippe H et al (2019) Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. Curr Biol 29:1–9
18. Kapli P, Telford MJ (2020) Topology dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. Sci Adv 6:eabc5162
19. Kapli P, Flouri T, Telford MJ (2021) Systematic errors in phylogenetic trees. Curr Biol 31:R59–R64
20. Natsidis P, Kapli P, Schiffer PH, Telford MJ (2021) Systematic errors in orthology inference and their effects on evolutionary analyses. iScience 24:102110
21. Kapli P et al (2021) Lack of support for Deuterostomia prompts reinterpretation of the first Bilateria. Sci Adv 7:eabe2741
22. Simion P, Delsuc F, Philippe H (2020) To What Extent Current Limits of Phylogenomics Can Be Overcome? in Phylogenetics in the Genomic Era (eds. Scornavacca, C., Delsuc, F. & Galtier, N.) 2.1:1-2.1:34No commercial publisher
23. Weisman CM, Murray AW, Eddy SR (2020) Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol 18:1–24
24. Baurain D, Brinkmann H, Philippe H (2007) Lack of resolution in the animal phylogeny: Closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol 24(1):6–9
25. Felsenstein J (1978) Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Syst Zool 27:401–410
26. Giacomelli M, Rossi ME, Lozano-Fernandez J, Feuda R, Pisani D (2022) Resolving tricky nodes in the tree of life through amino acid recoding. iScience 25:105594
27. Hernandez AM, Ryan JF (2021) Six-state Amino Acid Recoding is not an Effective Strategy to Offset Compositional Heterogeneity and Saturation in Phylogenetic Analyses. Syst Biol 70:1200–1212
28. Bergsten J (2005) A review of long-branch attraction. Cladistics 21:163–193
29. Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O (2023) Ancient gene linkages support ctenophores as sister to other animals. Nature
30. Schiffer PH et al (2024) Insights into early animal evolution from the genome of the xenacoelomorph worm *Xenoturbella bocki*. elife
31. Telford MJ (2008) Resolving Animal Phylogeny: A Sledgehammer for a Tough Nut? Dev Cell 14:457–459
32. Höhna S et al (2016) RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. Syst Biol 1–29
33. Pett W et al (2019) The role of homology and orthology in the phylogenomic analysis of metazoan gene content. Mol Biol Evol 36:643–649

34. Juravel K, Porras L, Höhna S, Pisani D, Wörheide G (2023) Exploring genome gene content and morphological analysis to test recalcitrant nodes in the animal phylogeny. PLoS ONE 18:e0282444
35. Guijarro-Clarke C, Holland PWH, Paps J (2020) Widespread patterns of gene loss in the evolution of the animal kingdom. Nat Ecol Evol 4:519–523
36. van Kempen M et al (2024) Fast and accurate protein structure search with Foldseek. Nat Biotechnol 42:243–246
37. Morel B et al (2022) SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. Mol Biol Evol 39
38. Szöllosi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V (2013) Efficient exploration of the space of reconciled gene trees. Syst Biol 62:901–912
39. Williams TA et al (2023) Parameter Estimation and Species Tree Rooting Using ALE and GeneRax. Genome Biol Evol 15:1–14
40. Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247
41. Zhang C, Scornavacca C, Molloy EK, Mirarab S (2020) ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. Mol Biol Evol 37:3292–3307
42. Mirarab S et al (2014) ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548
43. Spillane JL, LaPolice TM, MacManes MD, Plachetzki DC (2021) Signal, bias, and the role of transcriptome assembly quality in phylogenomic inference. BMC Ecol Evol 21:1–17
44. Kocot KM, Citarella MR, Moroz LL, Halanych KM (2013) PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol. Bioinforma. 429–435 (2013)
45. Harris BJ et al (2022) Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. Nat Ecol Evol 6:1634–1643
46. Susko E, Roger AJ (2007) On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol 24:2139–2150
47. Thálen F (2019) PhyloPyPruner. https://github.com/fethalen/phylopypruner
48. Philippe H et al (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21:1740–1752
49. Wiens JJ (2006) Missing data and the design of phylogenetic analyses. J Biomed Inf 39:34–42
50. Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol Biol Evol 30:197–214
51. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109
52. Lai Z, Fortini ME, Rubin GM (1991) The embryonic expression patterns of zfh-1 and zfh-2, two Drosophila genes encoding novel zinc-finger homeodomain proteins. Mech Dev 34:123–134

53. Rokas A, Holland PWH (2000) Rokas Holland Rare Genomic Changes Phylp Review 2000. Trends Ecol Evol 15:454–459
54. O'Malley MA, Wideman JG (2016) Ruiz-Trillo, I. Losing Complexity: The Role of Simplification in Macroevolution. Trends Ecol Evol 31:608–621
55. Redmond AK (2024) Acoelomorph flatworm monophyly is a severe long branch-attraction artefact obscuring a clade of Acoela and Xenoturbellida. Proceedings Royal Society B, 291, 20240329
56. Perea-Atienza E et al (2015) The nervous system of Xenacoelomorpha: A genomic perspective. J Exp Biol 218(4):618–628
57. Gavilán B, Sprecher SG, Hartenstein V, Martinez P (2019) The digestive system of xenacoelomorphs. Cell Tissue Res 377(3):369–382
58. Gąsiorowski L et al (2021) Molecular evidence for a single origin of ultrafiltration-based excretory organs. Curr Biol 31:3629–3638e2
59. Paps J, Holland PWH (2018) Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat Commun 9:1–8

Figures

Figure 1



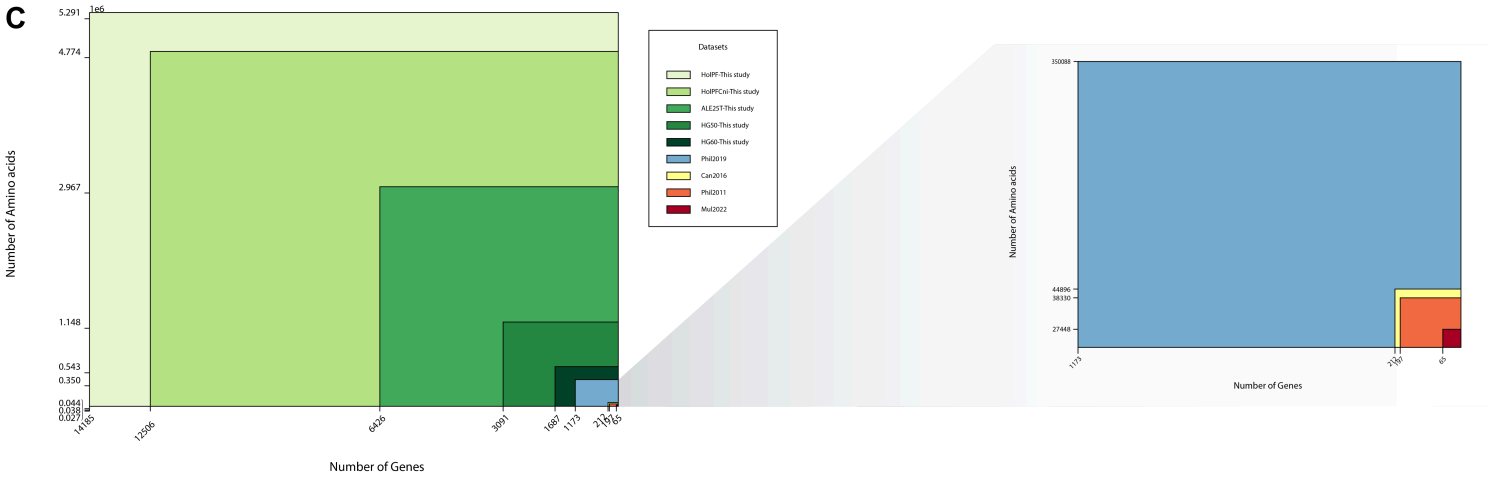


Figure 1

Legend not included with this version.

Figure 2

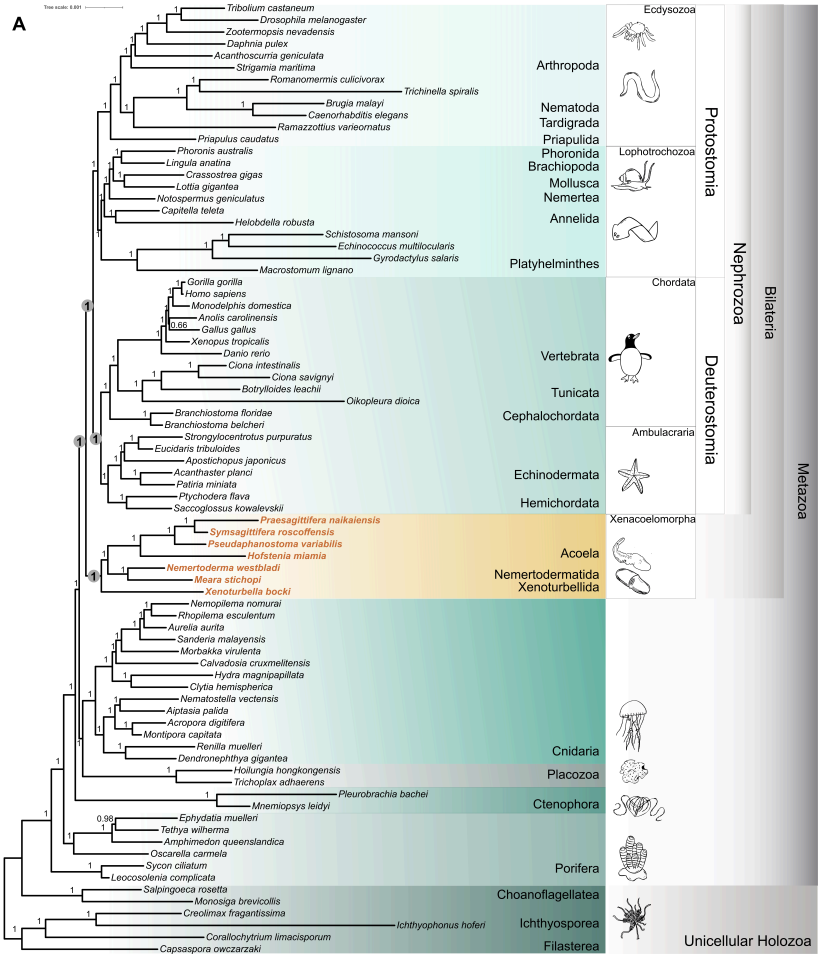


Figure 2

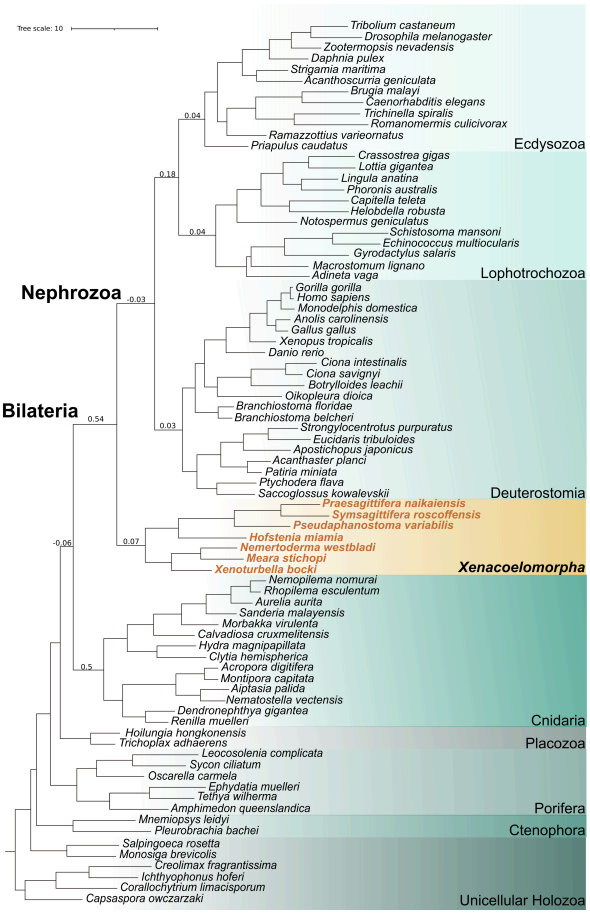


Figure 2

Legend not included with this version.

Figure 3

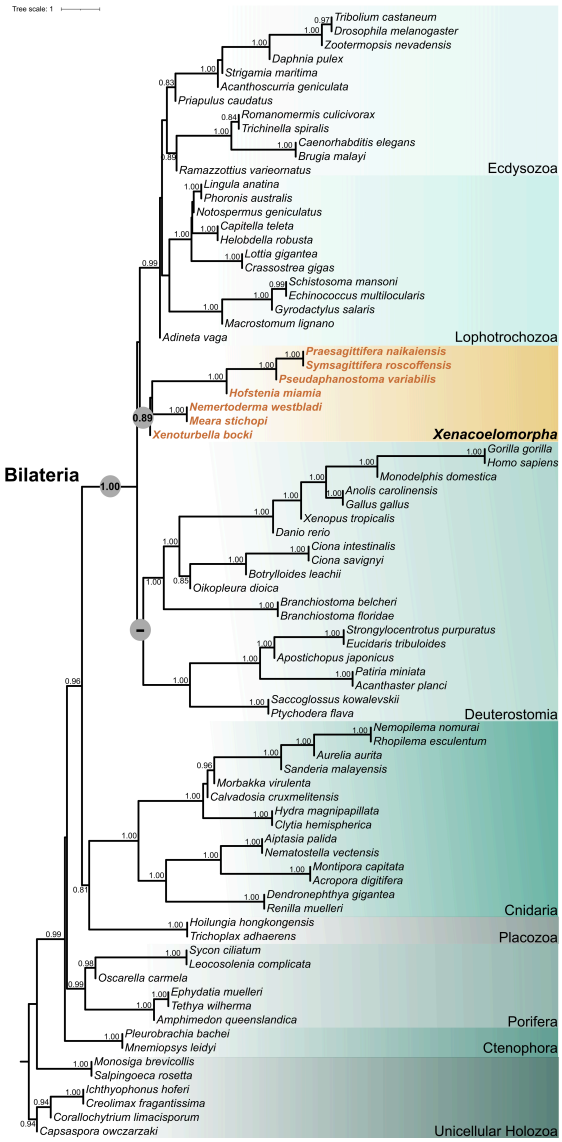


Figure 3

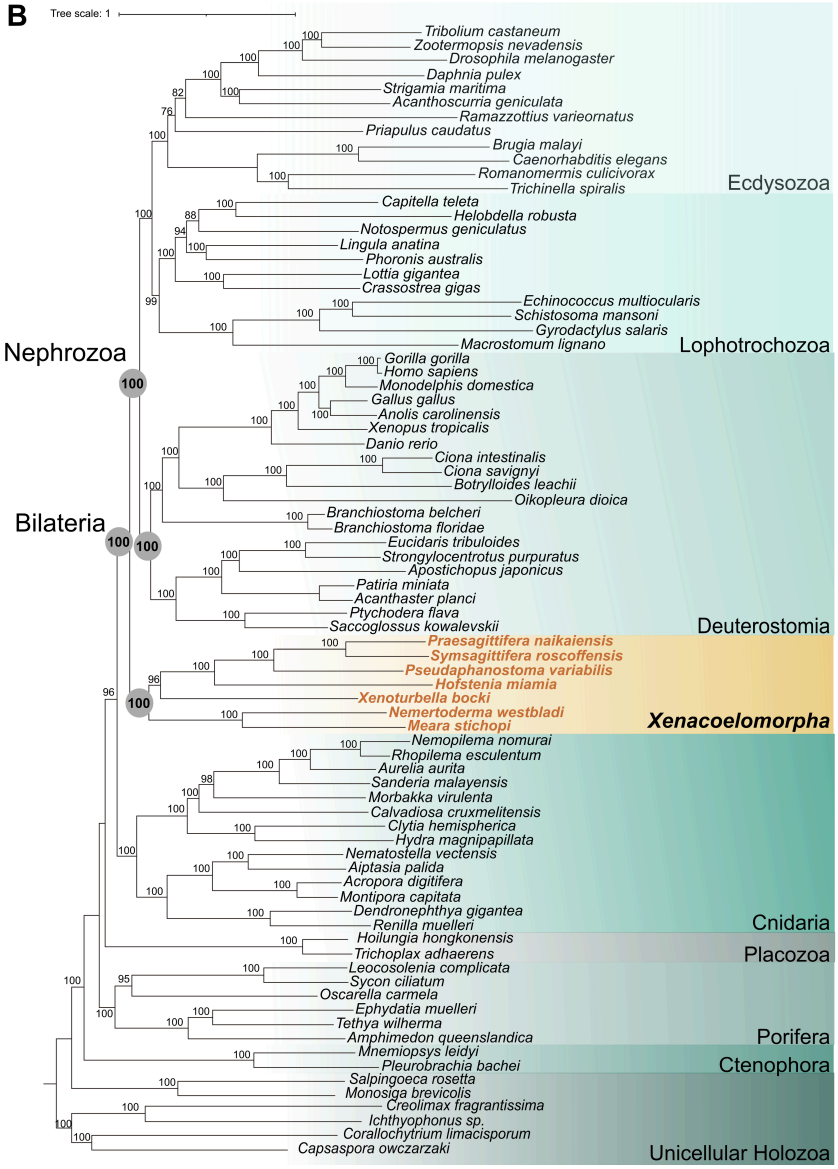


Figure 3

Legend not included with this version.

Figure 4

Figure 4

Legend not included with this version.

Figure 5
A

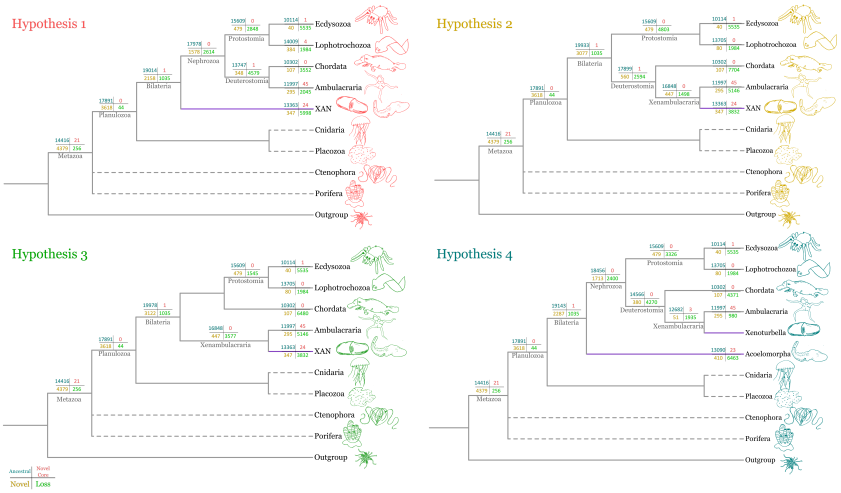


Figure 5
B

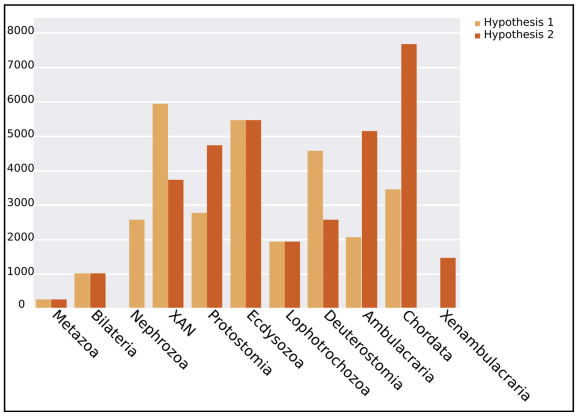


Figure 5

Legend not included with this version.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytablespreprint.pdf
- SuppFiguresprepprint.pdf
- SupplementaryTablesLegends.pdf
- SupplementaryFiguresLegends.pdf
- SupplementaryMaterialSM.pdf