# Filtering artifactual signal increases support for Xenacoelomorpha and Ambulacraria sister relationship in the animal tree of life

## Highlights

- High rate of violation of incontrovertible splits in animal phylogenomic datasets

- Filtering gene families accordingly increases phylogenetic information

- Increased, but not conclusive, support for the existence of Xenambulacraria

## Authors

Peter O. Mulhair,
Charley G.P. McCarthy,
Karen Siu-Ting,
Christopher J. Creevey,
Mary J. O'Connell

## Correspondence

mbzmjo@nottingham.ac.uk

## In brief

Mulhair, McCarthy et al. explore the signal for alternative topologies relating to the placement of Xenacoelomorpha. They determine that by filtering out genes whose trees violate incontrovertible clans in the animal phylogeny, model fit improves, and the support for a relationship between Xenacoelomorpha and Ambulacraria increases.

# Current Biology

## Report

# Filtering artifactual signal increases support for Xenacoelomorpha and Ambulacraria sister relationship in the animal tree of life

Peter O. Mulhair,[1,2,4] Charley G.P. McCarthy,[1,4] Karen Siu-Ting,[3] Christopher J. Creevey,[3] and Mary J. O'Connell[1,2,5,*]
[1]Computational and Molecular Evolutionary Biology Research Group, School of Life Sciences, Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham NG7 2RD, UK
[2]Computational and Molecular Evolutionary Biology Research Group, School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK
[3]Institute for Global Food Security, School of Biological Sciences, Queen's University Belfast, Belfast BT9 5DL, UK
[4]These authors contributed equally
[5]Lead contact
*Correspondence: mbzmjo@nottingham.ac.uk
https://doi.org/10.1016/j.cub.2022.10.036

## SUMMARY

Conflicting studies place a group of bilaterian invertebrates containing xenoturbellids and acoelomorphs, the Xenacoelomorpha, as either the primary emerging bilaterian phylum[1–6] or within Deuterostomia, sister to Ambulacraria.[7–11] Although their placement as sister to the rest of Bilateria supports relatively simple morphology in the ancestral bilaterian, their alternative placement within Deuterostomia suggests a morphologically complex ancestral bilaterian along with extensive loss of major phenotypic traits in the Xenacoelomorpha. Recent studies have questioned whether Deuterostomia should be considered monophyletic at all.[10,12,13] Hidden paralogy and poor phylogenetic signal present a major challenge for reconstructing species phylogenies.[14–18] Here, we assess whether these issues have contributed to the conflict over the placement of Xenacoelomorpha. We reanalyzed published datasets, enriching for orthogroups whose gene trees support well-resolved clans elsewhere in the animal tree.[16] We find that most genes in previously published datasets violate incontestable clans, suggesting that hidden paralogy and low phylogenetic signal affect the ability to reconstruct branching patterns at deep nodes in the animal tree. We demonstrate that removing orthogroups that cannot recapitulate incontestable relationships alters the final topology that is inferred, while simultaneously improving the fit of the model to the data. We discover increased, but ultimately not conclusive, support for the existence of Xenambulacraria in our set of filtered orthogroups. At a time when we are progressing toward sequencing all life on the planet, we argue that long-standing contentious issues in the tree of life will be resolved using smaller amounts of better quality data that can be modeled adequately.[19]

## RESULTS

Where the Xenacoelomorpha fall in the animal tree of life has significant implications for understanding the evolution of complexity within Bilateria. Under the early branching hypothesis, this group represent a key intermediate branch to understand the transition from non-bilaterian lineages (e.g., Cnidaria) to animals with bilateral symmetry. However, if nested within Deuterostomia, these organisms have undergone significant loss of morphological characters, possessing a simple brain, blind gut, and lacking excretory and vascular systems.[3,4] Both hypotheses point to this group of enigmatic worms and their position in the animal tree of life as being of great importance to the evolutionary biology field. Resolving the placement of Xenacoelomorpha is challenging due to the high rates of sequence evolution and gene loss within the phylum, short and deep bilaterian branches with little signal, and inadequate taxon sampling.[13] In this study, we shed new light on this question by exploring the

impact of inadvertent paralog inclusion and poor phylogenetic signal in orthologous gene families used for phylogenetic reconstruction and ask how this may have affected previous interpretations of the phylogeny. We assess the phylogenetic signal within three previously published datasets with conflicting placements for the Xenacoelomorpha[5,6,10] (Figures 1A–1C), each of which differ in their taxon sampling, data types and size, orthology assignment methods, and phylogenomic methods applied (Figures 1D and 1E; Table S1). Each of these factors is known to be important to consider and may lead to biases in phylogenetic inference.[20–23]

### Frequent violation of known monophyletic clans suggests misleading signal within gene families used for phylogenetic inference

For each of the three datasets,[5,6,10] we used Clan_Check[16] to assess whether the constructed gene trees could recapitulate known, incontestable clans (*sensu* Wilkinson et al.[24])
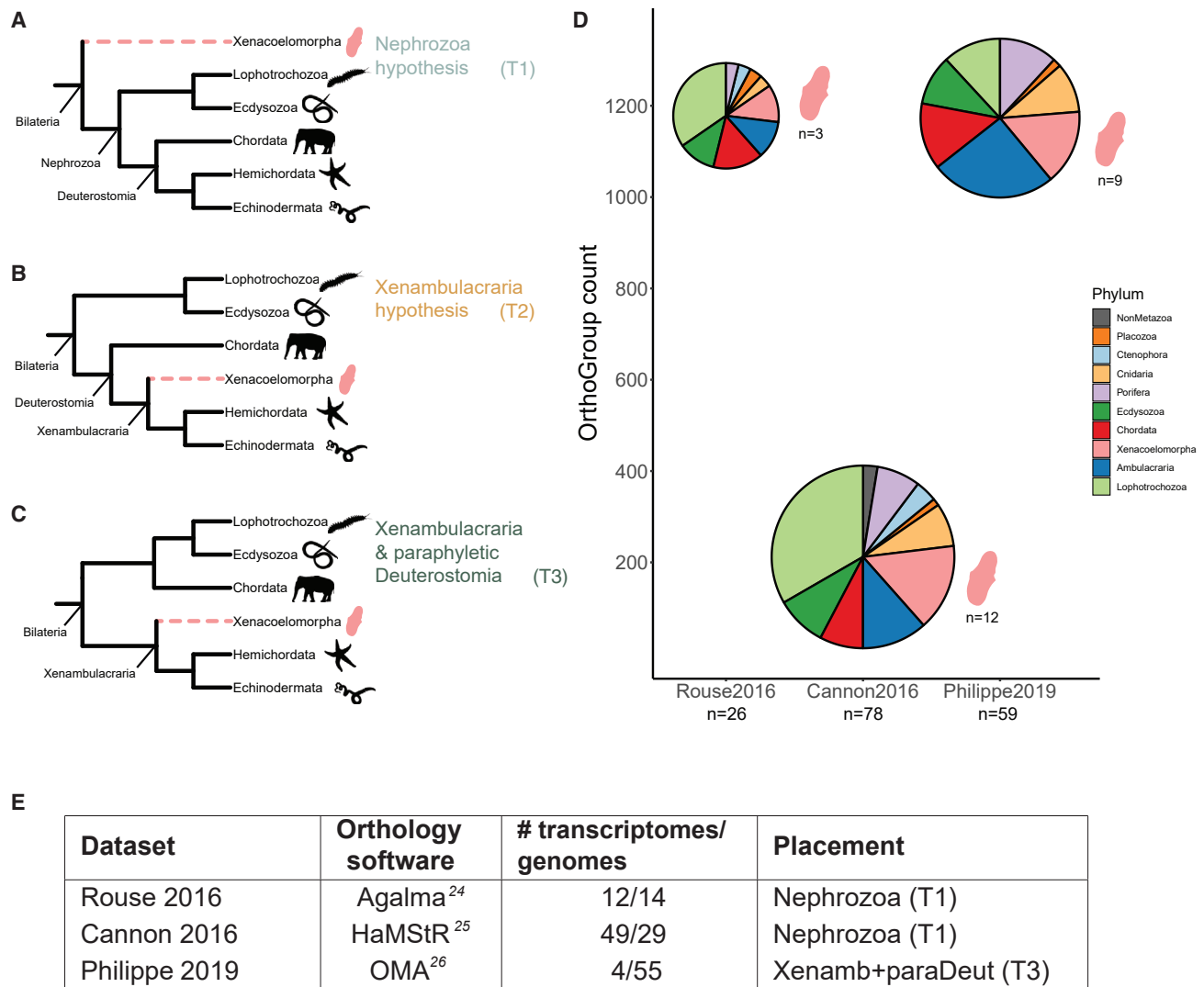
**Figure 1. Conflicting placements for Xenacoelomorpha and distribution of phylum sampling in three major datasets applied to resolve their position**

(A) Nephrozoa hypothesis (T1) posits that Xenacoelomorpha are sister to the remaining bilaterian phyla.

(B) Xenambulacraria hypothesis (T2) places Xenacoelomorpha as sister to Ambulacraria, a clade consisting of Hemichordata and Echinodermata.

(C) Xenambulacraria has also been suggested, along with non-monophyletic Deuterostomia, with Chordata placed sister to Protostomia.

(D) Overview of previous phylogenomic studies addressing the placement of Xenacoelomorpha. Three studies are shown, with circle sizes based on the number of taxa sampled for the analysis. Each of the wedges in the circle are colored based on a given animal phylum (legend) and sized based on the number of species in that phylum that were used in the study. The y axis represents the number of orthogroups used in the phylogenomic analysis.

(E) Summary table for the three datasets, Rouse et al.,[5] Cannon et al.,[6] and Philippe et al.,[10] included in this study, showing in each case the orthology detection software used, the proportion of the data from transcriptome or genome sequencing, and the placement resolved in the original publication. Note: Xenamb+-paraDeut = Xenambulacraria and paraphyletic Deuterostomes.

See also Table S1.

(Figure S1A). Clan_Check filters datasets used for phylogenetic inference, removing inferred gene families that violate predefined splits/clans. Under the assumption that gene trees inferred from orthologous genes should recapitulate a set of predefined incontestable clans (Figure S1B), a gene tree that violates these assumptions at a high rate may indicate inadvertent paralog selection or other spurious signal.[16] We found widespread violation of the tested clans in gene trees from each of the three datasets (Figure 2A). For example, in the

Philippe et al.[10] dataset, each clan was violated by between 60% and 99% of the gene trees (Figure 2A). The clan most often violated was Deuterostomia, with 1,155/1,164 (99%) of the gene trees failing to recapitulate the clan. Similar patterns are observed in each of the other two datasets (i.e., Rouse et al.[5] and Cannon et al.[6]), with widespread violation of incontestable clans in the constructed gene trees (Figure 2A). We classified gene families as those with extensive violation of clans by Clan_check ("CC fail") and those with low or no

# Current Biology
## Report
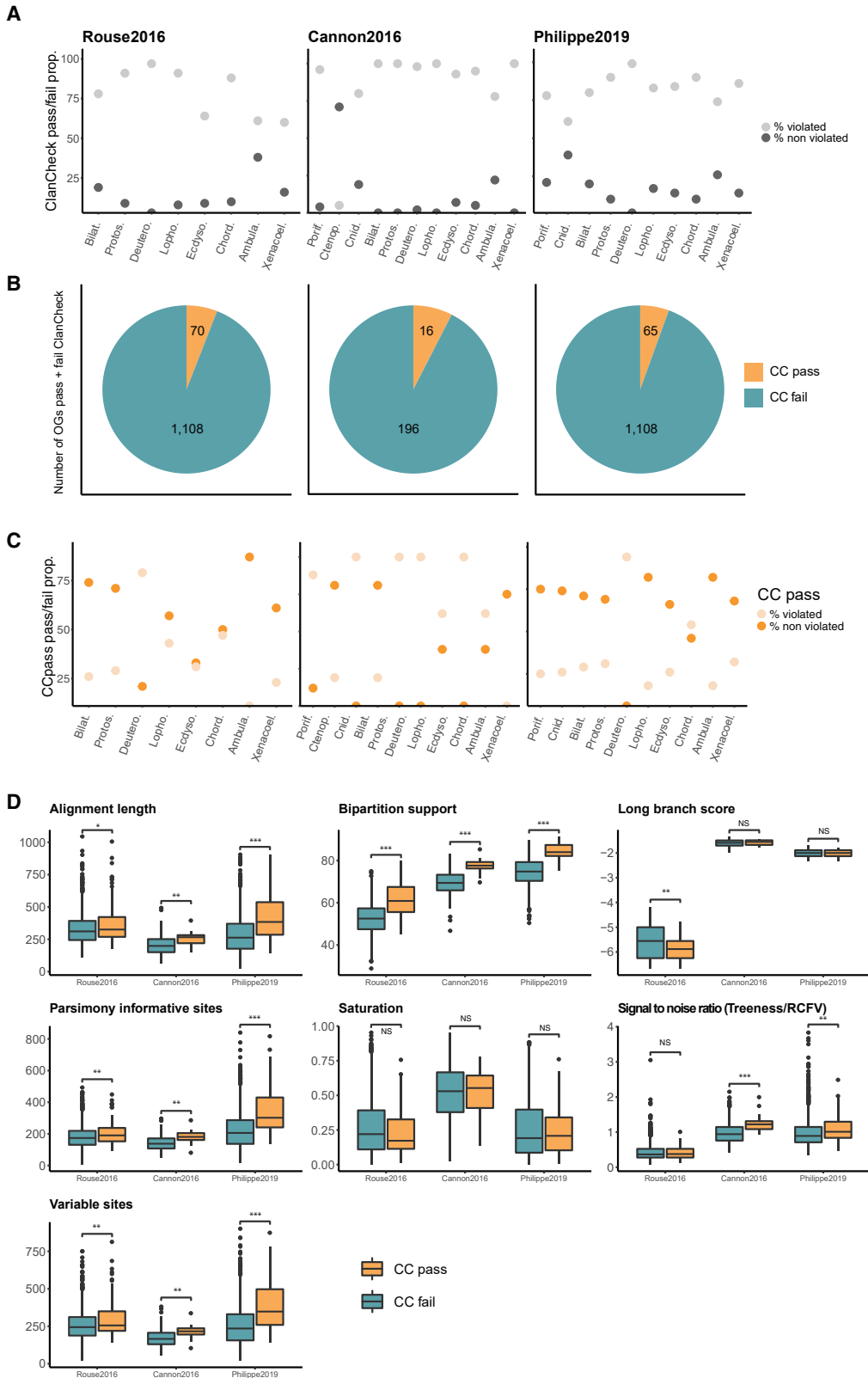
violations ("CC pass"). As the level of violation of the predefined clans differed between the datasets, we filtered for genes with low levels of violation in a dataset-specific manner, i.e., 4/10, 3/10, and 5/10 clans recapitulated for Rouse et al.,[5] Cannon et al.,[6] and Philippe et al.,[10] respectively (Table S2; STAR Methods). This resulted in three "CC pass" datasets consisting of 70, 16, and 65 genes for Rouse et al.,[5] Cannon et al.,[6] and Philippe et al.,[10] respectively (Figure 2B). Using these dataset-specific criteria retained a similar proportion of the total number of orthogroups for each of the reduced datasets (Figure 2B), while still ensuring a high rate of clan recovery. Indeed, when we reassess the rate at which the CC pass genes can recapitulate the tested monophyletic clans as above, we see a much-improved rate of clan retention (Figure 2C). Interestingly, however, Deuterostomia appears to be violated at a similar rate between the full gene sets and the CC pass gene sets (Figure 2C).

### Clan_Check filtering improves overall phylogenetic signal without biasing for compositional heterogeneity, rates of sequence evolution, or function

We compared overall branch lengths, compositional heterogeneity, and gene function between the CC fail and CC pass gene sets for each of the three datasets. We found no significant difference between the two sets of genes for any of the traits tested in any of the three datasets (Table S3). This suggests that our filter does not generate gene sets with biases in branch length, compositional heterogeneity, or function.[16] We also measured taxon sampling between the gene sets and found no evidence for subsampling of genes with missing species or biases for species in certain phyla (Figure S2).

Additionally, we used seven different gene and tree-based metrics (STAR Methods) shown to reflect the accuracy of species tree inference, to assess whether the dataset that can recapitulate incontrovertible clans (CC pass), differs significantly from the dataset that failed the Clan_Check test (CC fail).[25,26] We found that for all three datasets, and across almost all tests, the smaller dataset of genes that can recover known clades (CC pass) contained a significantly greater amount of phylogenetic signal based on the different metrics (Wilcoxon rank-sum test, p < 0.05) (Figure 2D). Some deviations were noted in the "long branch score," "saturation," and "treeness divided by relative composition variability" in some datasets. Overall, however, when we filter each of the three datasets based on their ability to recapitulate known clans, phylogenetic signal is increased, suggesting a possible resolution of conflicting signals and providing justification for reducing the data matrices for phylogenetic inference.

### Signal at single gene level varies between datasets and support for a single topology increases in filtered datasets

The three currently debated positions for the Xenacoelomorpha are abbreviated to T1, T2, and T3 (Figures 1A–1C). We determined the proportion of gene trees supporting each hypothesis (T1–T3). First, gene-wise log likelihood values[27] for the full set of genes in each dataset showed T1 (Nephrozoa hypothesis) to be supported by 58%, 28%, and 47% of genes in Cannon et al.,[6] Philippe et al.,[10] and Rouse et al.,[5] respectively, with the remainder supporting T2 and T3 (the two Xenambulacraria hypotheses). When filtered for violation of incontestable clans, support for T1 (Nephrozoa hypothesis) increased in both Rouse et al.[5] (57% of genes) and Cannon et al.[6] (69% of genes) (Figure 3A) but decreased in Philippe et al.[10] (12% of genes), where an increase in support for both T2 (43% of genes) and T3 (45% of genes) was observed.

In order to investigate the signal at single gene level, we used the approximately unbiased (AU) test[28] to identify genes with enough phylogenetic signal to significantly reject all but one of the tested topologies (see STAR Methods for details). In theory, enriching for putative true orthologs should increase support for a single phylogeny.[16] However, we observed no major difference in the subset of genes with the ability to reject all but one topology between CC fail and CC pass datasets (Figure 3B). The 46 gene trees in the Rouse et al.[5] dataset capable of rejecting all but one topology predominantly support T1 (Nephrozoa hypothesis), with support for T1 from 37/42 (88%) of the CC fail dataset and from 4/4 of the CC pass dataset. Of the 12 gene trees in the Cannon et al.[6] dataset, we find 100% support for T1 in both the CC fail (11 gene trees) and CC pass (1 gene tree) datasets. For the Philippe et al.[10] dataset (41 gene trees; 37 gene trees in CC fail set and 4 in CC pass set), we find the same pattern of majority support, but this time for T3 (Xenambulacraria and paraphyletic Deuterostomia).

### Increased support for Xenambulacraria hypothesis using genes that recover incontestable clans and contain comparable outgroups

Next, we carried out phylogenomic analyses on the concatenated filtered, ortholog-enriched datasets to determine the placement of Xenacoelomorpha (i.e., genes from the CC pass set only). Our three CC pass subsets of Rouse et al.,[5] Cannon et al.,[6] and Philippe et al.[10] consisted of 70, 16, and 65 genes. These represented a total of 27,183, 4,080, and 27,448 aligned amino acid sites, significantly reduced compared with the original studies (Table S1). Phylogenomic reconstruction was carried out on each concatenated supermatrix using PhyloBayes-MPI,[29] applying the CAT-GTR+G4 model. Two independent chains

---

**Figure 2. Results of Clan_Check filtering and comparison of pass and fail gene sets**
(A) The proportion of gene trees violating each of the clans tested for each of the three datasets.
(B) The number of genes from the original dataset that passed or failed the Clan_Check filter. "CC pass" (yellow) genes are those gene families that could recapitulate above the threshold number of the splits, and "CC fail" (blue) are those that failed to recapitulate enough clans above the threshold.
(C) Proportion of gene trees from the CC pass gene set that violate each of the clans tested.
(D) Summary of seven tree-based traits of signal between "CC pass" (yellow) and "CC fail" (blue) orthogroups (OGs) for each of the three datasets. Metrics tested included alignment length, bipartition support, long branch score, number of parsimony informative sites, level or saturation, treeness divided by relative compositional variability (RCFV), and the number of variable sites. Wilcoxon rank-sum test was then carried out between CC pass and CC fail gene sets for each metric tested for each of the three datasets (NS, p > 0.05; *p < 0.05, **p < 0.01, and ***p < 0.001).
See also Figures S1 and S2 and Tables S2 and S3.

# Current Biology
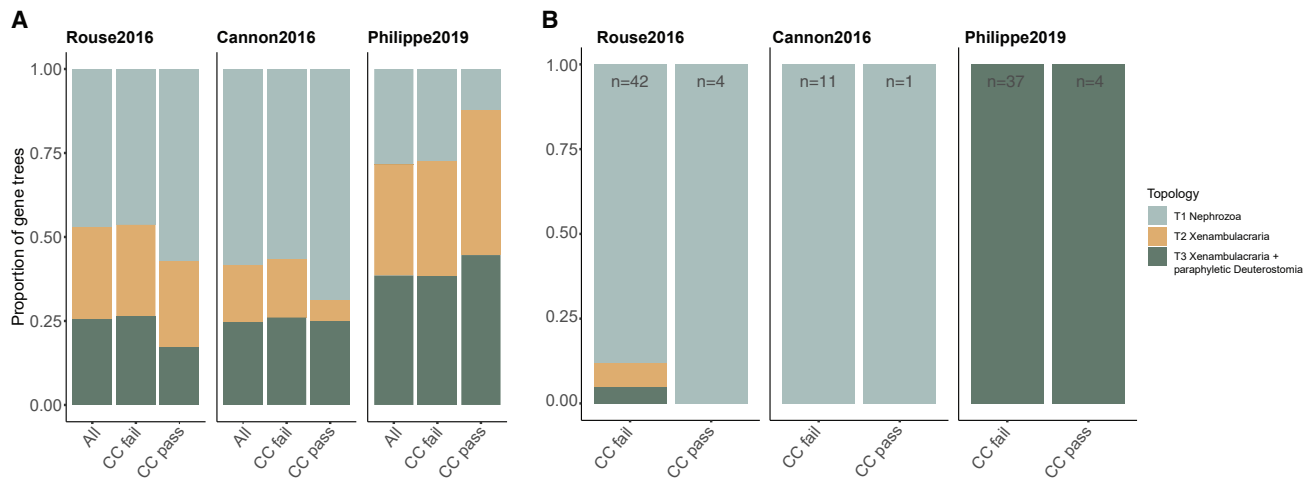## Report

CellPress
OPEN ACCESS



**Figure 3. Gene-level support for alternative topologies from gene trees that pass or fail the Clan_Check filter for each of the three datasets**

(A) Proportion of support for each topology from gene-wise log likelihood test. The x axis represents all genes, along with the CC fail or CC pass subsets of the data for each of the three datasets, Rouse et al.,[5] Cannon et al.,[6] and Philippe et al.[10] (left to right). The y axis represents the proportion of the data that supports each of the three conflicting positions for Xenacoelomorpha: T1 = Nephrozoa hypothesis (yellow), T2 = Xenambulacraria (red), and T3 = Xenambulacraria with paraphyletic Deuterostomes (blue).

(B) Distribution of signal in genes that reject all but one of the topologies. Support for alternative topologies in gene trees based on the AU test. The x axis contains columns for genes from CC fail and CC pass gene sets for each of the three datasets, Rouse et al.,[5] Cannon et al.,[6] and Philippe et al.[10] The y axis shows the proportion of genes supporting one of the three alternative topologies for the placement of Xenacoelomorpha.

were run for at least 10,000 iterations and convergence in all cases was assessed using the bpcomp function in PhyloBayes. All three analyses reached convergence between chains (with observed maxdiff < 0.3).

The CC pass subset of Rouse et al.[5] recovered a topology consistent with the Xenambulacraria hypothesis (T2), with Xenacoelomorpha placed sister to Ambulacraria with high support (posterior probabilities [PP] = 0.94) (Figures 4 and S3A). Deuterostomia was recovered as monophyletic, with a clade grouping Xenambulacraria + Chordata, albeit with lower support (PP = 0.7). These findings conflict with the original study, where Xenacoelomorpha was sister to the remaining bilaterian lineages.[5] The CC pass subset of Cannon et al.[6] supported the Nephrozoa hypothesis (T1), with Xenacoelomorpha sister to the remaining bilaterian lineages (PP = 0.81) (Figures 4 and S3B), albeit with a paraphyletic Deuterostomia where Chordata were sister to Ambulacraria + Protostomia (PP = 0.99). This topology agrees with the placement of Xenacoelomorpha in the original study, although with lower support, and highlights conflicting support previously found for Deuterostomia monophyly.[10,13] Finally, the CC pass subset of Philippe et al.[10] achieved the same result as the original study supporting the (T3) Xenambulacraria hypothesis placing Xenacoelomorpha sister to Ambulacraria (Xenambulacraria hypothesis) (PP = 1.0), with a paraphyletic Deuterostomia (PP = 1.0) (Figures 4 and S3C). In the original study, Philippe et al.[10] supported their assertion that Deuterostomia is non-monophyletic, using a jackknifing approach; we confirm that result, achieving full support with the CC pass subset of Philippe et al.[10] using the Bayesian analysis.

Across the datasets in our study, the outgroup species are variable, with the Cannon et al.[6] dataset containing the non-metazoan Choanoflagellate species as well as three Ctenophora species, while Rouse et al.[5] contains one Porifera and one

Ctenophore outgroup and Philippe et al.[10] contains only the Porifera species as an outgroup. Distant outgroup lineages can lead to systematic bias in species tree construction, particularly given that the Xenacoelomorpha are a fast-evolving group.[11] Therefore, we generated two new forms of the Cannon et al.[6] CC pass dataset to assess the impact of outgroup selection on retrieved topology. In the first of these "outgroup reduced sets," we removed the Choanoflagellate outgroup species, and in the second we removed both Choanoflagellate and Ctenophora outgroup species. For both outgroup reduced datasets, the reconstructed species phylogenies demonstrated that the removal of these outgroups had no impact on the overall species topology retrieved, i.e., Xenacoelomorpha remain as sister to the other bilaterian lineages. However, when both the Choanoflagellate and Ctenophore outgroup species were removed, we observe that support for the Nephrozoa (T1) hypothesis is reduced considerably (from PP = 0.81 in the original tree to PP = 0.52). We conclude that the inclusion of the distant outgroup species may be driving support for T1 (Nephrozoa) in the original study (Figure S3D) and may be linked to the large proportion of transcriptomic data incorporated in that study (Figure 1E).

## Filtering for orthologous signal improves model fit

We applied posterior predictive analysis (PPA) to assess the adequacy of model fit to each of the three concatenated CC pass datasets, using the output from the previous species tree inference step.[30] We used PhyloBayes-MPI[29] to infer statistics designed to test model adequacy. Specifically, we have included one test for among-site amino acid preferences (PPA-DIV) and one for among-lineage compositional heterogeneity (PPA-MAX) (Figure S4). The null hypothesis ($H_0$) states that the model is capable of adequately describing the data, and $H_1$ that the
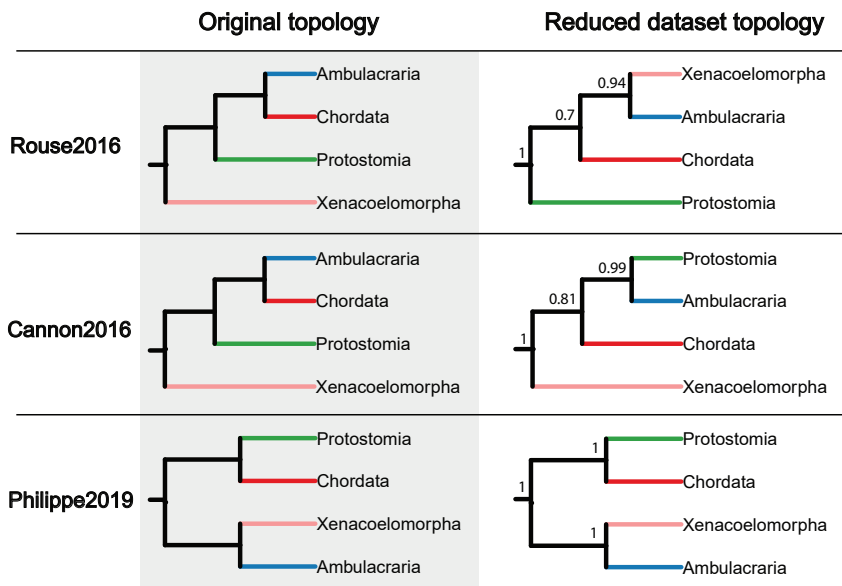
**Figure 4. Phylogenetic trees generated from the CC pass subsets of Rouse et al., Cannon et al., and Philippe et al.**
Original topologies are shown on the left in simplified form, showing just the major bilaterian groups. The species topology from the dataset of genes that recovered known clans (i.e., CC pass), subsets of each dataset are shown on the right. Posterior probabilities (PP) are shown for the nodes of major lineages.
See also Figures S3 and S4 and Table S4.

model is not adequately describing the data, where a $Z$ score $< 2$ indicates acceptance of the null.[31,32]

Analyzing whether the model adequately describes site specific amino acid preferences, i.e., PPA-DIV,[33] we retrieve the following $Z$ scores: $Z < 2$ for the Rouse et al.[5] dataset indicating appropriate model fit in this statistic, and $Z = 3.39$ for Cannon et al.[6] and $Z = 2.35$ for Philippe et al.,[10] respectively (Table S4; Figure S4). The main test of among-lineage compositional heterogeneity, PPA-MAX, shows a similar pattern, with a $Z$ score of $< 2$ for Rouse et al.,[5] and $Z$ scores $> 2$ for both Cannon2016 (2.77) and Philippe2019 (8.61) (Table S4; Figure S4). The larger $Z$ score for the reduced Philippe et al.[10] dataset suggests that the model is not adequately capturing the maximal compositional heterogeneity observed across taxa in this dataset. This may be because in the original analysis the genes were not filtered for variation in rates of compositional heterogeneity.[10]

To determine whether the improvement in model fit was simply because we decreased the size of the data matrix, we subsampled by jackknifing the CC fail datasets to produce datasets of similar size to the CC pass datasets and directly compared the model fit values. The $Z$ scores for PPA-DIV and PPA-MAX statistics on the subsampled CC fail datasets reflect similar results to the model fit results on the CC pass dataset (Figure S4). However, there are some key differences, namely PPA-MAX for Cannon et al.[6] decreased to 0.9 (as compared with 2.27 for Cannon et al.[6] CC pass dataset) and PPA-MAX increased to 24.13 for Philippe et al.[10] (as compared with 8.61 for Philippe et al.[10] CC pass dataset). These results show that while improved model fit does correlate with dataset size, differences in how compositional heterogeneity is accommodated is not accounted for in all datasets by simply reducing dataset size.

## DISCUSSION

The placement of Xenacoelomorpha has profound implications for our understanding and interpretation of animal evolution. Data previously applied to this problem have patterns indicative of low levels of signal (Figure 3B), suggesting a small proportion of genes with strong signal are contributing to the placement of Xenacoelomorpha.[15,17,27,34] The issue of inadequate or misleading signal within phylogenomic studies is often overlooked in favor of appropriate model selection or the size of data matrices.[22] The work presented here and elsewhere[16,27,35] demonstrates that without assessment of the underlying molecular data we risk misinterpreting the signal present. This, in turn, may lead to substantial issues downstream in the phylogenetic pipeline.[15,16,23,27,35,36] Our approach reduces the data matrices to include only those genes which can recover "incontrovertible" relationships *sensu* Siu Ting et al.,[16] thus enriching for orthologous signal. Although inferring the "true" species tree remains difficult, even with a filtered dataset (Figure 4), accounting for genes with potentially misleading signal does provide sound criteria for species tree inference.[15,16,25,27,37–44] We have shown that in most cases, reducing the dataset to include only those genes that can recapitulate known splits in the animal tree increases overall quality of signal. Importantly, these reduced datasets have improved model fit and provide tractable time scales for reaching convergence in Bayesian phylogenomic analyses.

Model fit analysis using PPA was carried out in the original study by Philippe et al.,[10] comparing the model fit for CAT and CAT-GTR models in the best and worst genes (based on a metric for phylogenetic signal within the gene tree) in each of the three datasets.[10] In the "best genes" pool, they observed improved model fit with the CAT-GTR model, as determined by the $Z$ score from PPA-DIV and PPA-MAX, and we show further improvement in observed $Z$ scores in our reduced datasets, implying increased model fit in CC pass datasets (Table S4; Figure S4). However, it is important to note that while model fit is improved by applying highly parameterized models like CAT-GTR on these reduced data matrices, the issue of model inadequacy can still remain. Although the heterogeneous models described represent significantly better fit than site-homogeneous models, future work to develop models that account for compositional heterogeneity[45] and/or the use of data recoding shows promise in addressing some of the inadequacies of model fit.

Whether it is hidden paralogy alone driving these misleading effects in species tree inference also requires further investigation. Although the approach we have taken to identify genes whose trees violate specific known relationships/splits has previously been shown to enrich for orthologous signal and to be insensitive

# Current Biology
## Report

**CellPress**
OPEN ACCESS

to incomplete lineage sorting (ILS),[16] violations may also occur for reasons other than hidden paralogy such as gene tree estimation error, or simply due to poor or low phylogenetic signal. Nevertheless, investigation of the distribution between the CC pass and CC fail datasets of traits related to these problems, such as compositional heterogeneity, rates of evolution, and function, confirmed that the approach is not generating biased datasets, reducing the likelihood that these factors are driving the results obtained. Interestingly, we find that bipartition support (overall bootstrap support in a gene tree) is significantly lower in the CC fail gene sets (Figure 2D), suggesting that, in addition to hidden paralogy (which remains a sporadic issue across the animal tree[16,46]), genes with lower overall phylogenetic signal may bias inference of the correct species topology, particularly for studies addressing deep evolutionary branching patterns in this group. Finally, the influence of dataset size is another aspect that requires further investigation. There were only 16 genes that passed the filtering step for Cannon et al.,[6] which perhaps unsurprisingly resulted in polytomies in deeper parts of the topology (i.e., among Placozoa, Cnidaria, and Bilateria). Caution is recommended when reducing data matrices to such small sizes as there may be inadvertent effects due to low levels of total signal.[47]

Overall, following filtering for genes that violate known clades, the majority of the studies analyzed showed increased support for Xenambulacraria and/or a paraphyletic Deuterostomia. It is clear that studies focused on the placement of Xenacoelomorpha and other challenging nodes, would benefit from tractable and consistent methods as well as from a greater exploration of the signal within the data matrices. Currently, there is a clear deficit in the signal-to-noise ratio in our phylogenomic studies assembled to resolve this particular branch in the animal tree. Here, we attempted to increase the strength of true phylogenetic signal by using only those genes that we know can recapitulate uncontroversial splits and have succeeded in generating datasets that are modeled more adequately. Moving forward, the generation of more complete genomes for groups such as Xenacoelomorpha is of key strategic importance, as is consilience across alternative data types such as rare genomic changes.[48]

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Measuring rate of violation of known monophyletic clans with Clan_Check
  - Metrics to test for biases and signal between CC pass and CC fail datasets
  - Testing gene family support for alternative topologies between pass and fail datasets
  - Phylogenomic analyses using reduced datasets
  - Posterior predictive analysis to assess model fit
- QUANTIFICATION AND STATISTICAL ANALYSIS

## REFERENCES

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

2. Paps, J., Baguñà, J., and 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.

3. Ruiz-Trillo, I., and Paps, J. (2016). Acoelomorpha: earliest branching bilaterians or deuterostomes? Org. Divers. Evol. 16, 391–399.

4. Hejnol, A., and Pang, K. (2016). Xenacoelomorpha's significance for understanding bilaterian evolution. Curr. Opin. Genet. Dev. 39, 48–54.

5. Rouse, G.W., Wilson, N.G., Carvajal, J.I., and Vrijenhoek, R.C. (2016). New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. Nature 530, 94–97.

6. Cannon, J.T., Vellutini, B.C., Smith, J., 3rd, Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. Nature 530, 89–93.

7. Bourlat, S.J., Nielsen, C., Lockyer, A.E., Littlewood, D.T.J., and Telford, M.J. (2003). Xenoturbella is a deuterostome that eats molluscs. Nature 424, 925–928.

8. Bourlat, S.J., Juliusdottir, T., Lowe, C.J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E.S., Thorndyke, M., Nakano, H., Kohn, A.B., et al. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444, 85–88.

9. Philippe, H., Brinkmann, H., Copley, R.R., Moroz, L.L., Nakano, H., Poustka, A.J., Wallberg, A., Peterson, K.J., and Telford, M.J. (2011). Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470, 255–258.

10. Philippe, H., Poustka, A.J., Chiodin, M., Hoff, K.J., Dessimoz, C., Tomiczek, B., Schiffer, P.H., Müller, S., Domman, D., Horn, M., et al.

(2019). Mitigating anticipated effects of systematic errors supports sister-group relationship between Xenacoelomorpha and Ambulacraria. Curr. Biol. *29*, 1818–1826.e6.

11. Kapli, P., and Telford, M.J. (2020). Topology-dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. Sci. Adv. *6*, eabc5162.

12. Marlétaz, F., Peijnenburg, K.T.C.A., Goto, T., Satoh, N., and Rokhsar, D.S. (2019). A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. Curr. Biol. *29*, 312–318.e3.

13. Kapli, P., Natsidis, P., Leite, D.J., Fursman, M., Jeffrie, N., Rahman, I.A., Philippe, H., Copley, R.R., and Telford, M.J. (2021). Lack of support for Deuterostomia prompts reinterpretation of the first Bilateria. Sci. Adv. *7*, eabe2741.

14. Doolittle, W.F., and Brown, J.R. (1994). Tempo, mode, the progenote, and the universal root. Proc. Natl. Acad. Sci. USA *91*, 6721–6728.

15. Brown, J.M., and Thomson, R.C. (2017). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. *66*, 517–530.

16. Siu-Ting, K., Torres-Sánchez, M., San Mauro, D., Wilcockson, D., Wilkinson, M., Pisani, D., O'Connell, M.J., and Creevey, C.J. (2019). Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. Mol. Biol. Evol. *36*, 1344–1356.

17. Walker, J.F., Shen, X.-X., Rokas, A., Smith, S.A., and Moyroud, E. (2020). Disentangling biological and analytical factors that give rise to outlier genes in phylogenomic matrices. Preprint at bioRxiv. https://doi.org/10.1101/2020.04.20.049999.

18. Natsidis, P., Kapli, P., Schiffer, P.H., and Telford, M.J. (2021). Systematic errors in orthology inference and their effects on evolutionary analyses. iScience *24*, 102110.

19. Tihelka, E., Cai, C., Giacomelli, M., Lozano-Fernandez, J., Rota-Stabelli, O., Huang, D., Engel, M.S., Donoghue, P.C.J., and Pisani, D. (2021). The evolution of insect biodiversity. Curr. Biol. *31*, R1299–R1311.

20. Zwickl, D.J., and Hillis, D.M. (2002). Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. *51*, 588–598.

21. Wilberg, E.W. (2015). What's in an outgroup? The impact of outgroup choice on the phylogenetic position of Thalattosuchia (Crocodylomorpha) and the origin of Crocodyliformes. Syst. Biol. *64*, 621–637.

22. Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. *9*, e1000602.

23. Spillane, J.L., LaPolice, T.M., Macmanes, M.D., and Plachetzki, D.C. (2021). Signal, bias, and the role of transcriptome assembly quality in phylogenomic inference. BMC Ecol. Evol. *21*, 43.

24. Wilkinson, M., McInerney, J.O., Hirt, R.P., Foster, P.G., and Embley, T.M. (2007). Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends Ecol. Evol. *22*, 114–115.

25. Shen, X.-X., Salichos, L., and Rokas, A. (2016). A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol. Evol. *8*, 2565–2580.

26. Shen, X.-X., Steenwyk, J.L., and Rokas, A. (2021). Dissecting incongruence between concatenation- and quartet-based approaches in phylogenomic data. Syst. Biol. *70*, 997–1014.

27. Shen, X.-X., Hittinger, C.T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. *1*, 126.

28. Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. Syst. Biol. *51*, 492–508.

29. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. *62*, 611–615.

30. Bollback, J.P. (2002). Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. *19*, 1171–1180.

31. Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., and Pisani, D. (2017). Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Curr. Biol. *27*, 3864–3870.e4.

32. Lartillot, N. (2020). The bayesian approach to molecular phylogeny. https://hal.archives-ouvertes.fr/hal-02535330.

33. Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. *7*, S4.

34. Di Franco, A., Poujol, R., Baurain, D., and Philippe, H. (2019). Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol. Biol. *19*, 21.

35. Francis, W.R., and Canfield, D.E. (2020). Very few sites can reshape the inferred phylogenetic tree. PeerJ *8*, e8865.

36. Fernández, R., Gabaldon, T., and Dessimoz, C. (2020). Orthology: definitions, prediction, and impact on species phylogeny inference. Phylogenet. Genom. Era, 2–4.

37. Salichos, L., and Rokas, A. (2011). Evaluating ortholog prediction algorithms in a yeast model clade. PLoS One *6*, e18755.

38. Doyle, V.P., Young, R.E., Naylor, G.J.P., and Brown, J.M. (2015). Can we identify genes with increased phylogenetic reliability? Syst. Biol. *64*, 824–837.

39. Edwards, S.V. (2016). Phylogenomic subsampling: a brief review. Zool. Scr. *45*, 63–74.

40. Kocot, K.M., Struck, T.H., Merkel, J., Waits, D.S., Todt, C., Brannock, P.M., Weese, D.A., Cannon, J.T., Moroz, L.L., Lieb, B., and Halanych, K.M. (2017). Phylogenomics of Lophotrochozoa with consideration of systematic error. Syst. Biol. *66*, 256–282.

41. Molloy, E.K., and Warnow, T. (2018). To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. *67*, 285–303.

42. Dornburg, A., Su, Z., and Townsend, J.P. (2019). Optimal rates for phylogenetic inference and experimental design in the era of genome-scale data sets. Syst. Biol. *68*, 145–156.

43. Smith, S.A., Walker-Hale, N., Walker, J.F., and Brown, J.W. (2020). Phylogenetic conflicts, combinability, and deep phylogenomics in plants. Syst. Biol. *69*, 579–592.

44. Koch, N.M. (2021). Phylogenomic subsampling and the search for phylogenetically reliable loci. Mol. Biol. Evol. *38*, 4025–4038.

45. Foster, P.G. (2004). Modeling compositional heterogeneity. Syst. Biol. *53*, 485–495.

46. Smith, M.L., and Hahn, M.W. (2022). The frequency and topology of Pseudoorthologs. Syst. Biol. *71*, 649–659.

47. Smith, M.L., and Hahn, M.W. (2021). New approaches for inferring phylogenies in the presence of paralogs. Trends Genet. *37*, 174–187.

48. Rokas, A., and Holland, P.W. (2000). Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. *15*, 454–459.

49. Katoh, K., Kuma, K.-I., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. *33*, 511–518.

50. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

51. Löytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. USA *102*, 10557–10562.

52. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534.

53. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: Genome-scale protein function classification. Bioinformatics *30*, 1236–1240.

54. Steenwyk, J.L., Buida, T.J., Labella, A.L., Li, Y., Shen, X.-X., and Rokas, A. (2021). PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Bioinformatics *37*, 2325–2331.

55. Creevey, C.J., and McInerney, J.O. (2005). Clann: investigating phylogenetic information through supertree analyses. Bioinformatics *21*, 390–392.

56. Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. *7*, S2.

57. Blackburne, B.P., and Whelan, S. (2012). Measuring the distance between multiple sequence alignments. Bioinformatics *28*, 495–502.

58. Thompson, J.D., Plewniak, F., Ripp, R., Thierry, J.C., and Poch, O. (2001). Towards a reliable objective function for multiple sequence alignments. J. Mol. Biol. *314*, 937–951.

59. Muller, J., Creevey, C.J., Thompson, J.D., Arendt, D., and Bork, P. (2010). AQUA: automated quality improvement for multiple sequence alignments. Bioinformatics *26*, 263–265.

60. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

61. R Core Team (2013). R: a language and environment for statistical computing.

62. Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., et al. (2018). GOATOOLS: a Python library for Gene Ontology analyses. Sci. Rep. *8*, 10872.

63. Steenwyk, J.L., Buida, T.J., 3rd, Li, Y., Shen, X.-X., and Rokas, A. (2020). ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. *18*, e3001007.

64. Struck, T.H. (2014). TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. Evol. Bioinform. Online *10*, 51–67.

65. Phillips, M.J., and Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogenet. Evol. *28*, 171–185.

66. Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502–504.

67. Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. *8*, 28–36.

68. Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. *21*, 1095–1109.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Deposited data** | | |
| Alignments, software and trees | Zenodo/Github | https://github.com/PeterMulhair/Xenaceol_Paralogy; https://doi.org/10.5281/zenodo.6412148 |
| **Software and algorithms** | | |
| Clan_Check | Siu-Ting et al.[16] | https://github.com/ChrisCreevey/clan_check |
| PhyloBayes-MPI | Lartillot et al.[29] | https://github.com/bayesiancook/pbmpi |
| Mafft | Katoh et al.[49] | https://mafft.cbrc.jp/alignment/software/ |
| Muscle | Edgar[50] | https://www.drive5.com/muscle/ |
| Prank | Löytynoja and Goldman[51] | http://wasabiapp.org/software/prank/ |
| IQ-Tree | Minh et al.[52] | http://www.iqtree.org/ |
| InterProScan | Jones et al.[53] | https://github.com/ebi-pf-team/interproscan |
| PhyKIT | Steenwyk et al.[54] | https://github.com/JLSteenwyk/PhyKIT |
| Clann | Creevey and McInerney[55] | http://chriscreevey.github.io/clann/ |
| SCaFoS | Roure et al.[56] | http://megasun.bch.umontreal.ca/Software/scafos/scafos.html |
| **Other** | | |
| Rouse et al.[5] dataset | [5] | https://doi.org/10.5061/dryad.79dq1 |
| Cannon et al.[6] dataset | [6] | https://doi.org/10.5061/dryad.493b7 |
| Philippe et al.[10] dataset | [10] | https://github.com/MaxTelford/Xenacoelomorpha2019 |

### RESOURCE AVAILABILITY

#### Lead contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Mary J. O'Connell (mbzmjo@nottingham.ac.uk).

#### Materials availability
This study did not generate new unique reagents.

#### Data and code availability

- All data used in this study have been deposited at Zenodo & Github and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code used in this study has been deposited at Zenodo & Github and is publicly available as of the date of publication. DOIs are listed in the key resources table. Public software that was used in this study are cited in the STAR Methods section and listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

### METHOD DETAILS

#### Measuring rate of violation of known monophyletic clans with Clan_Check
Our aim in this study is to test whether hidden paralogy and/or low phylogenetic signal in phylogenomic datasets, is driving alternative topologies in the placement of Xenacoelomorpha. We employed Clan_Check (https://github.com/ChrisCreevey/clan_check) which is a tool designed to test if putative single copy orthologous gene families violate incontestable monophyletic clans.[16] The demonstration of how this tool works can be seen in Siu Ting et al.,[16] where it was used to investigate the signal from orthologous and hidden paralogous gene families in a group of vertebrates. Phylogenomic datasets for all three studies were downloaded from online data depositories: https://doi.org/10.5061/dryad.79dq1,[5] https://doi.org/10.5061/dryad.493b7,[6] and https://github.com/MaxTelford/Xenacoelomorpha2019.[10] These datasets are referred to as Rouse et al.,[5] Cannon et al.,[6] and Philippe et al.[10] throughout the main text. Each dataset consisted of a concatenated matrix of each of their gene families in amino acid format (see Table S1 for

# Current Biology
## Report

**CellPress**
OPEN ACCESS

details on the number of gene families and species in each dataset). This concatenated matrix was split into constituent gene family fasta files and each one was aligned using three alignment software methods; Mafft,[49] Muscle[50] and Prank.[51] In order to remove any potential bias on downstream analyses from the alignment step, we selected the best alignment using MetAl.[57] MetAl calculates metric distances between alignments of the same sequence, where a score of < 0.15 between a pair of alignments were considered to be in agreement while a score of > 0.15 were considered discordant. If a pair of alignments were found to be discordant, alignment quality was assessed using norMD and the alignment with the highest similarity score was retained for subsequent tree inference analyses.[58,59] If there was no alignment with a greater norMD score than the other two, the Mafft gene alignment was selected. Next, we used IQ-TREE,[52] applying ModelFinder[60] to find the model of best fit and carrying out 1000 ultrafast bootstrap replicates to construct gene trees for all gene family alignments in each dataset. For each dataset, we annotated a number of clans which were to be tested using Clan_Check. These were assigned based on the phylogenetic spread of species within each of the dataset. The clans/splits assigned for testing included Porifera, Ctenophora, Cnidaria, Bilateria, Protostomia, Deuterostomia, Xenacoelomorpha, Ambulacraria, Lophotrochozoa, Ecdysozoa, and Chordata (Figure S1C). Note that the Deuterostomia split is controversial, and whilst we have included it in our list of clans we are not requiring that it is always retrieved in our gene trees. Thus, genes that violate the Deuterostomia split are permitted to contribute to the species phylogeny. In total 16/16, 40/70, 61/65 of the genes in Cannon et al.,[6] Rouse et al.,[5] and Philippe et al.[10] respectively violated the Deuterostomia clan (Figure 2C) and were included in the species phylogeny. Monophyly of Ctenophora was only tested on Cannon et al.,[6] as there was just one species sampled for this clade in Rouse et al.[5] and no species for Philippe et al.[10] For a gene family to pass the Clan_Check filter, we required the corresponding gene tree to recapitulate a given number of clans. For Rouse et al.[5] and Philippe et al.[10] this cut off was 4 and 5 clans, respectively. For the Cannon et al.[6] dataset, the cut off was set to 3 clans. This was due to the fact that none of the gene trees in this dataset could recapitulate 5 clans, and only 4 gene trees could recapitulate 4 clans (Table S2). Using these dataset specific cutoffs retained a similar proportion of the overall number of orthogroups for each of the reduced datasets.

It should be noted that the study by Philippe et al.[10] employed a gene filtering approach, where each gene was stratified according to the phylogenetic accuracy of its gene tree. The top 25% of these genes were then used for additional species tree inference. Whilst this approach is similar to Clan_Check, there are two significant differences: (1) Philippe et al.[10] uses rooted gene trees, whereas Clan_Check uses unrooted gene trees, thus defining relationships/clans that would otherwise be paraphyletic in a rooted tree. This is an important point as rooting a gene tree on a given branch may result in an inability to recover ingroup clades as monophyletic. (2) Philippe et al.[10] allows a gene tree to pass with some degree of violation of clades and by averaging the rate at which each clade in the tree is recovered, i.e., a gene tree placed in the top 25% may not fully recapitulate any of the key clades defined. Conversely, Clan_Check requires complete recapitulation of that clan (or split) in the unrooted tree. These differences are significant when assessing a gene tree's ability to reconstruct known speciation patterns. We also show that our method selects for genes with increased phylogenetic signal, thus proving the utility of this approach for selecting genes for phylogenetic inference.

### Metrics to test for biases and signal between CC pass and CC fail datasets

For each of the three datasets we compared the sets of genes that passed and failed the Clan_Check filter step, first to ensure that the sampled set of genes for downstream analyses were not biased towards branch length, compositional heterogeneity or function. Here we followed the protocol used in Siu Ting et al.[16] To check for bias in branch length between the CC_Fail and CC_Pass gene sets, we used the inferred gene trees to compare distributions of branch lengths between the two sets. Average branch lengths were calculated by summing all branch lengths in the gene tree and dividing by the total number of branches. A Wilcox signed-rank test in R[61] was used to check whether there was a significant difference in average branch lengths between the two datasets. We used the same statistical test to check whether there were differences between the compositional heterogeneity of the taxa sampled in each dataset. We generated gene trees for each of the alignments and obtained the proportion of taxa that passed or failed the compositional heterogeneity test in IQTree.[52] Finally, to test whether there was a bias in the functions of genes which passed or failed the Clan_Check filter, we ran gene ontology analysis on the CC pass and CC fail gene sets. Gene annotation for each orthogroup was carried out using the representative human sequence, where possible, using InterProScan.[53] Gene ontologies were then extracted for each orthogroup and were grouped into one of the three major GO categories; "Molecular Function", "Biological Process" or "Cellular Component". This hierarchical annotation of GO terms was carried out using the GOATools python library.[62] Finally, the putative orthologous and paralogous gene sets were compared using a Pearson's chi-squared test to determine whether either of the gene sets had statistically significantly different distributions of functional categories.

We also ran several gene and tree-based metrics to compare the level of phylogenetic signal between the two gene sets. This was to ensure that the gene families enriched for orthologous genes actually displayed greater levels of phylogenetic signal than those that were filtered out. In total seven sets of tests were applied to the pass and fail gene sets for each of the three datasets. These included alignment length, bipartition support, long branch score, number of parsimony informative sites, level or saturation, treeness divided by relative compositional variability (RCV), and the number of variable sites. Each of these tests were measured using the PhyKIT software.[54] Briefly, alignment length has been shown to correlate with accurate tree inference[25]; gene trees which display higher bipartition support have been found to display greater certainty among bipartitions[25,37]; a larger number of parsimony informative sites is associated with stronger phylogenetic signal[25,63]; the average long branch score in a tree may give insights into the level of heterogeneity within the gene[64]; saturation is driven by sites with multiple substitutions thus underestimating the genetic distance among taxa, with a score of 1 showing no saturation and a score of 0 showing complete saturation; genes with higher treeness (proportion of tree distance on internal branches[65]) divided by RCV have been found to be associated with lower compositional and

other biases and higher phylogenetic signal[22]; and finally, genes with a higher number of variable sites often display higher phylogenetic signal.[25]

### Testing gene family support for alternative topologies between pass and fail datasets

To examine the level of support for each of the three alternative topologies for the placement of Xenacoelomorpha (T1, T2, and T3), we carried out two sets of analyses. First, we performed a gene-wise likelihood test, as in Shen et al.,[27] to measure the distribution of signal for each of the three topologies between the genes that failed and pass the Clan_Check filter. We constructed constrained species trees based on each of the three alternative topologies. Then, site-wise log likelihood estimations were inferred by comparing the data matrix of genes to each alternative topology using RAxML and Phylogenetic_signal_parser perl script from Shen et al.[27] (https://figshare.com/articles/dataset/Contentious_relationships_in_phylogenomic_studies_can_be_driven_by_a_handful_of_genes/3792189/4).

Secondly, to compare between the gene sets that passed and failed the Clan_Check filter, we calculated the number of genes capable of statistically significantly rejecting all but one of three alternative topologies. This was done using an AU test.[28] First, we constructed idealised gene trees consistent with each of the three alternative topologies for each dataset using Clann.[55] IQTree was used to construct a ML tree for each of the gene family alignments, and then calculate the log-likelihood of each of the three alternative species tree topologies based on the estimated parameters inferred for the ML gene tree. For each alternative topology the AU test returns a p-value, where a tree is rejected with a p-value $< 0.05$. Thus, if a gene tree can confidently reject all but one of the alternative topologies, the supported topology was recorded.

### Phylogenomic analyses using reduced datasets

After filtering each of the genesets from the three studies using Clan_Check, concatenated matrices of aligned amino acid sequences in Phylip format were constructed using SCaFoS and TREE-PUZZLE with default options.[66,56] For each resulting supermatrix, phylogenetic reconstruction was carried out using PhyloBayes-MPI.[29] After constant sites were removed (-dc option) the CAT-GTR model was applied, along a gamma distribution consisting of four rate categories. Two independent chains were run until convergence between the runs was reached. Convergence between chains was assessed using the bpcomp function in PhyloBayes with a burn-in of 5,000 iterations and sampling every 10 iterations. A maxdiff score of below 0.3 indicated convergence. Trees were visualised using the ggtree package in R (Figure S3).[67]

### Posterior predictive analysis to assess model fit

Posterior predictive analysis (PPA) was applied in PhyloBayes-MPI to assess how well the CAT-GTR model fit each of the reduced datasets.[29–31,68] The allppred flag in the readpb_mpi module was used to perform PPAs on both chains used for phylogenomic reconstruction. We tested two statistics to measure model fit, including one to test among-site amino acid preferences (PPA-DIV) and one for lineage-specific heterogeneity (PPA-MAX). For each statistic, a Z-score was computed with $|Z|$ representing standard deviations of the simulated data from the observed mean for each statistic.[31] The resulting Z scores were then compared to those calculated for the CAT-GTR model in Philippe et al.[10]

### QUANTIFICATION AND STATISTICAL ANALYSIS

We compared a number of gene alignment and tree traits between genes that passed and failed our Clan_Check filter for each of the three datasets (Figure 2D; Table S3). In each case, Wilcoxon rank-sum tests using R[61] were carried out between pass and fail gene sets, where p-value $< 0.05$ indicated significant difference between the datasets. To test whether there was significant difference in functional classes between the gene sets, gene ontologies were annotated using InterProScan[53] and GOATools,[62] and compared using a Pearson's chi-squared test.

For the Approximately Unbiased test (AU test) to measure support for alternative topologies (Figure 3A), gene trees were first constructed in IQTREE[52] using the model of best fit as determined by ModelFinder.[60] The AU test was then performed under the model of best fit for each gene against the three alternative hypotheses for the placement of Xenacoelomorpha (T1, T2 and T3; Figure 1) using IQTREE.[52]

Posterior predictive analyses (PPA) were implemented using PhyloBayes-MPI[29] to test model fit to our data, using the CAT-GTR model (Figure S4). For each statistic of model fit (PPA-DIV and PPA-MAX) a Z-score is computed and used to test whether the null hypothesis (the model adequately fits the data) can be rejected or not (where a Z-score of $< 2$ accepts the null hypothesis).