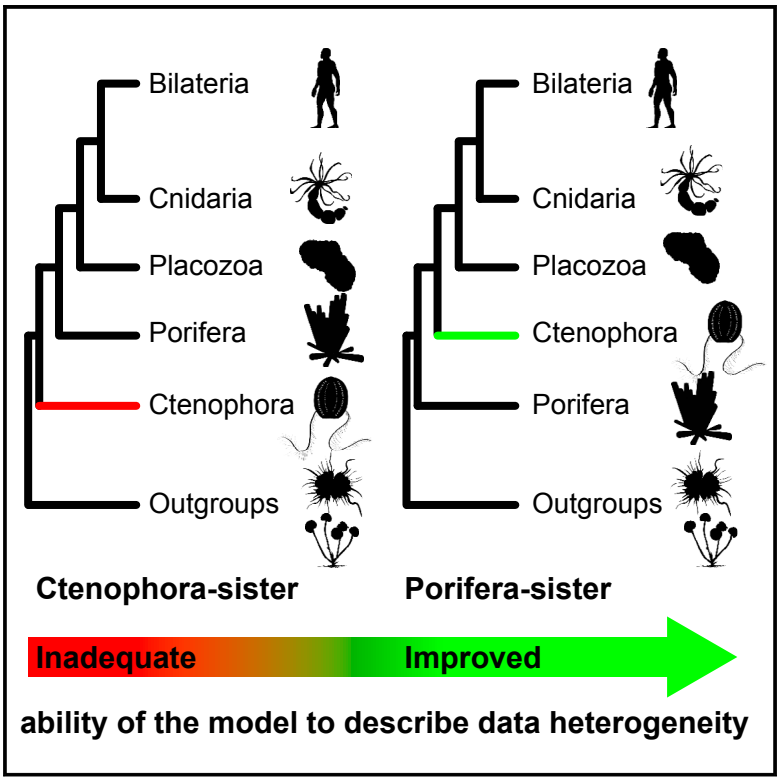
Report

Current Biology

Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals

Graphical Abstract



Highlights

- This is the first comparison of model adequacy in non-bilaterian phylogenomics
- Animal relationships depend on the ability of the models to describe the data
- Sponges are the sister group of the remaining animals
- Trees showing comb jellies as the sister of all the other animals are artifactual

Authors

Roberto Feuda, Martin Dohrmann, Walker Pett, ..., Nicolas Lartillot, Gert Wörheide, Davide Pisani

Correspondence

woerheide@lmu.de (G.W.), davide.pisani@bristol.ac.uk (D.P.)

In Brief

The relationships at the root of the animal tree are debated. Feuda et al. show that comb jellies emerge as the sister of all the other animals when the model inadequately describes the data. As modeling improves, sponges emerge in this position instead, indicating that trees placing the comb jellies at the root of the animals are artifactual.

Feuda et al., 2017, Current Biology *27*, 3864–3870
December 18, 2017
https://doi.org/10.1016/j.cub.2017.11.008

CellPress

Current Biology
Report

Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals

Roberto Feuda,[1,11] Martin Dohrmann,[2,11] Walker Pett,[3] Hervé Philippe,[4,5] Omar Rota-Stabelli,[6] Nicolas Lartillot,[7] Gert Wörheide,[2,8,9,*] and Davide Pisani[10,12,*]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[2]Department of Earth and Environmental Sciences, Palaeontology & Geobiology, Ludwig-Maximilians-Universität München, Munich, Germany
[3]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA
[4]Centre de Théorisation et de Modélisation de la Biodiversité, Station d'Ecologie Théorique et Expérimentale, CNRS, UMR 5321, Moulis 09200, France
[5]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, QC H3C 3J7, Canada
[6]Research and Innovation Centre, Fondazione Edmund Mach (FEM), Via E. Mach 1, 38010 San Michele all'Adige (TN), Italy
[7]Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Lyon, France
[8]GeoBio-Center, Ludwig-Maximilians-Universität München, Munich, Germany
[9]SNSB, Bayerische Staatssammlung für Paläontologie und Geologie, Munich, Germany
[10]School of Biological Sciences and School of Earth Sciences, University of Bristol, Bristol, UK
[11]These authors contributed equally
[12]Lead Contact
*Correspondence: woerheide@lmu.de (G.W.), davide.pisani@bristol.ac.uk (D.P.)
https://doi.org/10.1016/j.cub.2017.11.008

SUMMARY

The relationships at the root of the animal tree have proven difficult to resolve, with the current debate focusing on whether sponges (phylum Porifera) or comb jellies (phylum Ctenophora) are the sister group of all other animals [1–5]. The choice of evolutionary models seems to be at the core of the problem because Porifera tends to emerge as the sister group of all other animals ("Porifera-sister") when site-specific amino acid differences are modeled (e.g., [6, 7]), whereas Ctenophora emerges as the sister group of all other animals ("Ctenophora-sister") when they are ignored (e.g., [8–11]). We show that two key phylogenomic datasets that previously supported Ctenophora-sister [10, 12] display strong heterogeneity in amino acid composition across sites and taxa and that no routinely used evolutionary model can adequately describe both forms of heterogeneity. We show that data-recoding methods [13–15] reduce compositional heterogeneity in these datasets and that models accommodating site-specific amino acid preferences can better describe the recoded datasets. Increased model adequacy is associated with significant topological changes in support of Porifera-sister. Because adequate modeling of the evolutionary process that generated the data is fundamental to recovering an accurate phylogeny [16–20], our results strongly support sponges as the sister group of all other animals and provide further evidence that Ctenophora-sister represents a tree reconstruction artifact.

RESULTS AND DISCUSSION

Data Recoding Reduces Compositional Heterogeneity, and Models That Accommodate Site-Specific Compositional Heterogeneity Most Adequately Describe Recoded Datasets

An adequate modeling of the evolutionary process that generated the data is fundamental to the recovery of accurate phylogenies [16–20]. We used posterior predictive analyses (PPAs) [17] to perform model adequacy tests on two key datasets used in arguments to support Ctenophora as the sister group of all other animals (Ctenophora-sister). These are dataset D20 from Whelan et al. [10], hereafter "WhelanD20," and the dataset of Chang et al. [12], hereafter "Chang" (see STAR Methods). We performed multiple PPA tests and used Z scores to measure the deviation of the test statistics from the null expectation. We showed that for both datasets, models that have previously been used to study the animal phylogeny fell short, to various degrees, of adequately describing among-site amino acid preferences (i.e., site-specific replacement-pattern heterogeneity) and among-lineage compositional heterogeneity. These models are WAG+G [21], LG+G [22], GTR+G [23], dataset-specific PartitionFinder-defined models [24] (hereafter PF-schemes), and CAT-GTR+G [25] (see STAR Methods for a detailed comparison of these models).

To test whether available models could adequately describe among-site amino acid preferences, we used three alternative statistics. The first is site-specific amino acid diversity (hereafter PPA-DIV) [26]. PPA-DIV is a well-established test (e.g., [27]) measuring whether models can adequately estimate a simple but fundamental property of amino acid alignments, the mean number of distinct amino acids observed at each site (STAR Methods for details and [26] for theoretical justifications). To further check our results, we developed two new tests. The first

Table 1. Comparing Model Adequacy

Posterior Predictive Test			
Recoding	Model	Site-Specific Amino Acid Preferences (PPA-DIV)	Across-Taxa Compositional Heterogeneity (PPA-MAX)
Dataset: WhelanD20-Opistho			
none	WAG	127.35	79.99
none	LG	116.17	78.09
none	GTR	92.62	68.99
none	PF-scheme	104.71	44.75
none	CAT-GTR	6.21	34.78
Dayhoff-6	GTR	59.01	33.21
Dayhoff-6	CAT-GTR	−0.94	25.48
S&R-6	GTR	60.52	31.10
S&R-6	CAT-GTR	−1.09	7.09
KGB-6	GTR	56.82	33.13
KGB-6	CAT-GTR	−0.24	13.54
Dataset: Chang			
none	WAG	175.61	24.18
none	LG	161.28	19.00
none	GTR	126.33	18.84
none	PF-scheme	163.06	14.99
none	CAT-GTR	6.17	12.15
Dayhoff-6	GTR	78.06	26.99
Dayhoff-6	CAT-GTR	−1.71	3.64
S&R-6	GTR	87.37	28.48
S&R-6	CAT-GTR	−1.54	3.12
KGB-6	GTR	70.91	8.68
KGB-6	CAT-GTR	−0.71	1.21

Comparing the adequacy of WAG+G, LG+G, GTR+G, the optimal PF-scheme, and CAT-GTR+G when modeling site-specific and lineage-specific compositional heterogeneity for WhelanD20-Opistho (70 taxa, including all the original outgroups, and 46,542 amino acid positions) and Chang (77 taxa and 51,940 amino acid positions). This table presents *Z* values for PPA-DIV and PPA-MAX tests demonstrating how well each model describes site-specific and lineage-specific amino acid preferences, respectively (see STAR Methods and main text for details). Alternative statistics (PPA-MEAN, PPA-CONV, and PPA-VAR, detailed in STAR Methods), are reported in Table S1. For each PPA test in Table 1, the observed (empirical) heterogeneity, the posterior predictive mean heterogeneity, and the SD around the mean are reported in Table S2. Dayhoff-6, S&R-6, and KGB-6 recodings cannot be implemented with 20-states empirical amino acid substitution matrices like LG, WAG, and the matrices used in PF-schemes (see STAR Methods). Therefore, the effect of data recoding has been tested only for GTR+G and CAT-GTR+G. Positive *Z* scores for PPA-DIV indicate that average amino acid diversity is underestimated, whereas negative *Z* scores indicate that it is overestimated. See Figures S1–S4 for the trees inferred for each dataset under each considered model. See Table S3 for the definition of the PF-schemes.

(hereafter PPA-CONV) measures whether models can approximate the site-specific propensity to undergo convergent evolution toward the same amino acid in distantly related taxa. The second (hereafter PPA-VAR) measures the variance in empirical amino acid frequencies across sites (STAR Methods for details). A comparison of the spread of the *Z* scores obtained, across all models and from all of our PPAs (Tables 1 and S1), indicates that PPA-DIV is more discriminatory: it has a broader distribution of *Z* scores across models. Nevertheless, for both datasets, absolute *Z* scores larger than 5 (i.e., $|Z| > 5$) were obtained for all three PPAs (Tables 1 and S1). Absolute *Z* scores of this magnitude indicate a strong rejection of the null hypothesis that the model adequately describes the data (see STAR Methods for guidelines on interpreting $|Z|$ scores), showing that all models fell short of adequately describing site-specific amino acid preferences. However, PPAs also agreed that models varied greatly in their ability to describe replacement-pattern heterogeneity and indicated that CAT-GTR+G, the model of [6], describes site-specific amino acid preferences much better than any other model. This result was not unexpected [26] given that CAT-GTR+G was the only considered model that can explicitly accommodate this form of heterogeneity (see STAR Methods).

The relative ranking of the remaining models changed depending on the statistic implemented in the PPA tests (Tables 1 and S1). However, for both datasets, PPA-DIV [26] identifies GTR+G (the model of [8]) as a distant second best (Table 1), followed by either the PF-scheme [10] or LG+G [11] and WAG+G [9]. PPAs indicate that PF-schemes (strongly advocated by [10, 28, 29]) do not model site-specific amino acid preferences much better than LG+G, with CAT-GTR+G invariably describing the data better.

None of the considered models explicitly accounts for lineage-specific compositional heterogeneity (see STAR Methods). We therefore used two more, well-established [30], PPA tests to evaluate whether the inability of modeling this form of heterogeneity could have affected previous studies that attempted to resolve relationships at the root of the animal tree (see STAR Methods for details). The first test (PPA-MAX) evaluates whether alternative models can estimate the maximal compositional heterogeneity observed across the taxa. The second (PPA-MEAN) evaluates whether models can estimate the observed, lineage-specific mean squared heterogeneity. With both tests, CAT-GTR+G obtained the lowest *Z* scores, and PF-schemes emerged as the second best modeling strategy (Tables 1 and S1). However, all models, including CAT-GTR+G, were strongly rejected—with the smallest absolute *Z* score ($|Z| = 12.15$) obtained from the CAT-GTR+G analysis of Chang. This result was expected, given that the considered models do not account for compositional variation across lineages.

We then investigated whether the use of data-recoding approaches (e.g., [7, 13–15, 31–36]) could help reducing the compositional heterogeneity of WhelanD20 and Chang. We tested three different data transformation strategies: the well-known Dayhoff-6 [37] recoding (e.g., [7, 13–15, 31–36]), and the more recently developed recoding strategies of Susko and Roger [14] (hereafter S&R-6) and Kosiol et al. [15] (hereafter KGB-6). All considered data transformation strategies recode amino acids with similar physicochemical properties into one of six categories [7, 33] (see STAR Methods and Table 2 for details about the definition of the amino acid classes of each recoding strategy). All recoded datasets obtained lower *Z* scores than the original amino acid datasets for both PPA-MAX and PPA-MEAN (Tables 1 and S1), indicating that recoding is effective at reducing lineage-specific compositional heterogeneity. In the case of site-specific heterogeneity, *Z* scores lower than those

Table 2. Data-Recoding Strategies Implemented in This Study

Recoding Strategy	Binning Scheme
Dayhoff-6	(AGPST) (DENQ) (HKR) (MIVL) (WFY) (C)
S&R-6	(APST) (DENG) (QKR) (MIVL) (WC) (FYH)
KGB-6	(AGPS) (DENQHKRT) (MIL) (W) (FY) (CV)

Amino acid binning schemes used in Dayhoff-6, S&R-6, and KGB-6 analyses. These binning schemes were originally presented in Figure 84 of [37] (Dayhoff-6), Figure 1 (left panel) of [14] (S&R-6), and Figure 6A of [15] (KGB-6). The rationale used to define these recoding strategies is reported in STAR Methods. There are some commonalities between the three binning schemes, reflecting shared biochemical properties of the amino acids in each bin.

obtained for the amino acid datasets are invariably observed only for GTR+G (Table 1 and S1). Under CAT-GTR+G, only PPA-DIV invariably recovered reduced *Z* scores from the recoded datasets (Table 1), with *Z* scores from PPA-CONV and PPA-VAR being similar in analyses performed with and without recoding (Table S1). When considered together, these results indicate that only the combined use of data recoding and CAT-GTR+G minimizes inadequacy in the modeling of compositional heterogeneity across both sites and lineages.

Improved Modeling of Heterogeneity Supports Porifera-Sister

We investigated how the results of our phylogenetic analyses change as different evolutionary models were used to analyze either the original amino acid data or the recoded data. Phylogenies obtained under LG+G and WAG+G, as well as trees derived using GTR+G with and without Dayhoff-6, S&R-6, and KGB-6 recoding, supported Ctenophora-sister (posterior probabilities [PPs] ≈ 1; Table 3 and Figures S1–S4). Similarly, phylogenies inferred from Chang using PF-scheme and CAT-GTR+G (see STAR Methods) supported Ctenophora-sister with 100% bootstrap support and PP ≈ 1, respectively (see Table 3 and Figures S4D and S4E). Differently, CAT-GTR+G analyses of WhelanD20 recovered either Ctenophora-sister or Porifera-sister (Porifera as the sister group of all other animals) depending on the outgroups used. WhelanD20-Opistho, the dataset including all original outgroups from [10] (STAR Methods; Table 3; Figure S1D), supported Ctenophora-sister (PP ≈ 1). In contrast, WhelanD20-Holo, the dataset from which fungal outgroups were excluded, and WhelanD20-Choano, the dataset from which all outgroups but the Choanoflagellata were excluded, supported Porifera-sister (PP = 0.68 and PP = 0.77, respectively; Table 3; Figures S2B and S3B).

Support for Ctenophora-sister was severely reduced in CAT-GTR+G analyses of both recoded datasets, with support for Porifera-sister being higher than that for Ctenophora-sister (Table 3; Figures 1 and S1–S4). This was particularly evident for the Dayhoff-6 and S&R-6 recoded versions of both datasets. For KGB-6, the PP for Porifera-sister was 0.78 for Chang and 0.68 for WhelanD20-Holo. The KGB-6 recoded versions of WhelanD20-Opistho and WhelanD20-Choano failed to resolve the relationships at the root of the animal tree. However, Ctenophora-sister received the least support also with these datasets (Table 3; Figures S1J and S3H). Our results show that support for Ctenophora-sister decreases, while at the same time support for Porifera-sister increases, when heterogeneity in the data is better accounted for.

Weighing the Evidence

The recent discussion on the relationships at the root of the animal tree has focused on two alternative scenarios. These are the Porifera-sister hypothesis, which proposes that sponges (phylum Porifera) are the sister group of all other animals, and the Ctenophora-sister hypothesis, which suggests that comb jellies (phylum Ctenophora) are sister to all other animals [1–5]. Discriminating between these hypotheses is key to understanding early animal evolution, including the origin of fundamental innovations like the nervous system, muscles, and a through-gut [40]. However, phylogenomic analyses have supported both Porifera-sister [6, 7, 39, 41–44] and Ctenophora-sister [8–12, 35, 45, 46], with some studies (like [47]) presenting both trees. A key aspect of the debate is that support for both hypotheses is generally recovered from the same dataset when different substitution models and/or taxon- and gene-sampling strategies are used (contrast [6, 10] and [42, 45], and see [7, 43, 47]). Porifera-sister generally [6], but not invariably [12, 38], emerges when site-specific differences in amino acid composition are modeled and distant outgroups (e.g., Fungi) are not included in the analyses. Ctenophora-sister invariably emerges when site-specific amino acid differences are not modeled [10] or when site-specific amino acid differences are modeled but distant outgroups are used to root the tree [6]. These results indicate that phylogenomic datasets convey signal for both hypotheses and that discriminating between Porifera- and Ctenophora-sister requires distinguishing between phylogenetic signal and noise causing systematic errors [43].

To minimize systematic error, it is important to use the evolutionary model that best fits the data [18, 48–50]. Many tools exist to select the best-fit model [24, 51–54], but model fit is a relative concept: given a set of models, a best-fitting one can always be identified. Yet, if the best-fitting model does not describe the evolutionary process that generated the data adequately, there can still be a high probability of recovering a tree with relationships representing systematic error [16–20]. Model adequacy can be tested using posterior predictive tests [17]. Across-site replacement-pattern heterogeneity [26] and compositional heterogeneity across taxa [13, 55] can both have a misleading effect on phylogenetic reconstruction. Whelan et al. [10] attempted to minimize the negative effect of compositional heterogeneity by excluding compositionally heterogeneous proteins when generating WhelanD20. However, our PPAs indicate that WhelanD20 is still too heterogeneous to allow their preferred modeling strategy (a PF-scheme) to adequately describe the data (Tables 1 and S1). More broadly, our results demonstrate that PF-schemes (strongly advocated for by [10, 28, 29, 38]) are much worse than CAT-GTR+G and are not always better than unpartitioned modeling strategies at capturing the heterogeneity of the considered datasets (Tables 1 and S1). There are two, not mutually exclusive, explanations for these observations. The first is that PF-schemes cannot capture within-gene replacement-pattern heterogeneity [56]. The second is that protein-specific amino acid replacement patterns across WhelanD20 and Chang could not be captured by the empirical substitution matrices considered in the PartitionFinder analyses of [10]

Table 3. Support for Porifera-Sister and Ctenophora-Sister from the WhelanD20-Opistho, WhelanD20-Holo, WhelanD20-Choano, and Chang Datasets

Recoding	Hypothesis	Dataset			
		WhelanD20-Opistho	WhelanD20-Holo	WhelanD20-Choano	Chang
Model: GTR+G					
None	Porifera-sister	≈0	≈0	≈0	≈0
None	Ctenophora-sister	≈1	≈1	≈1	≈1
Dayhoff-6	Porifera-sister	≈0	≈0	≈0	≈0
Dayhoff-6	Ctenophora-sister	≈1	≈1	≈1	≈1
S&R-6	Porifera-sister	≈0	≈0	≈0	≈0
S&R-6	Ctenophora-sister	≈1	≈1	≈1	≈1
KGB-6	Porifera-sister	≈0	≈0	≈0	≈0
KGB-6	Ctenophora-sister	≈1	≈1	≈1	≈1
Model: CAT-GTR+G					
None	Porifera-sister	≈0	0.68	0.77	≈0
None	Ctenophora-sister	≈1	0.31	0.2	≈1
Dayhoff-6	Porifera-sister	0.89	0.99	0.95	0.74
Dayhoff-6	Ctenophora-sister	0.09	≈0	0.01	0.25
S&R-6	Porifera-sister	0.98	0.99	0.94	0.62
S&R-6	Ctenophora-sister	≈0	≈0	0.02	0.22
KGB-6	Porifera-sister	0.42	0.68	0.37	0.78
KGB-6	Ctenophora-sister	0.17	0.05	0.29	0.22

Support values are PPs. All analyses but the CAT-GTR+G analysis of the amino acid version of Chang converged well, and all convergence statistics are detailed in the captions of Figures S1–S4. Support values for Porifera-sister and Ctenophora-sister do not always add up to 1 because alternative topologies, e.g., monophyletic Porifera+Ctenophora sister to the remaining animals, are sometime recovered.

(see STAR Methods for a list). Indeed, the results of our PartitionFinder analyses (Table S3) show that the optimal PF-schemes identified LG as the amino acid replacement matrix fitting ~98% of the sites in both datasets best. That is, using PF-schemes to analyze these datasets is not very different from using a single LG matrix. Furthermore, our results demonstrate that, to further improve the adequacy with which compositional heterogeneity is described, CAT-GTR+G analyses of recoded datasets are most effective.

Our results show a correlation between overall model inadequacy, as measured by our PPA tests, and phylogenetic outcomes. WAG+G, LG+G, GTR+G (with or without recoding), and PF-schemes describe the data worse than CAT-GTR+G and invariably found support for Ctenophora-sister. Ctenophora-sister sometimes emerges in CAT-GTR+G analyses of amino acid datasets, too. However, support for this topology under CAT-GTR+G is more prominent when distant outgroups, which represent a natural source of long branches [43], are included in the analyses. Branch length is the product of time and substitution rate, and even slowly evolving outgroups can be long branched if they diverged much earlier than the ingroup, potentially exacerbating long-branch attraction artifacts [18, 57]. Our PPA tests indicate that although CAT-GTR+G describes Chang and WhelanD20 better than other models, overall model adequacy can still be improved using data recoding, and Porifera-sister was favored under all recoding strategies.

The relationships at the root of the animal tree depend on model adequacy, with Ctenophora-sister emerging more prominently when the data are modeled less adequately and Porifera-sister being better supported when the data are more adequately modeled (Figure 1). As phylogenetic methods are less likely to recover artifactual topologies when the models of evolution more adequately describe the data [16–20], we conclude that Porifera-sister better represents the phylogenetic signal in the data.

Shen et al. [11] investigated the distribution of phylogenetic signal in a dataset from Whelan et al. [10] and concluded that the signal for Porifera-sister is limited to a few "outlier genes." However, their results were obtained contrasting the likelihood of Porifera- and Ctenophora-sister under LG+G, a model that poorly describes the datasets of [10]. Indeed, CAT-GTR+G analyses (with and without Dayhoff-6 recoding) of a dataset of [10] from which [11] removed "outlier genes" still finds support for Porifera-sister (PP = 0.98 and PP = 0.99, respectively; Figures S3I and S3J). This result rejects the conclusions of [11] that signal for Porifera-sister is limited to "outlier genes" and indicates that their approach is model dependent: arguments about model adequacy should be considered when implementing it.

Data recoding reduced the impact of heterogeneity on our phylogenetic analyses (see also [7]), and this lent support to Porifera-sister. Notably, this result emerged independently of the outgroups included. Recently, Whelan et al. [38] expanded ctenophoran taxon sampling and found support for Ctenophora-sister using CAT-GTR+G and close outgroups. Based on our conclusions about the relationship between model adequacy and alternative hypotheses of animal relationships, we expected that the dataset of [38] should support Porifera-sister once recoded. Indeed, a CAT-GTR+G analysis of the Dayhoff-6 recoded version of this dataset reached excellent convergence and found strong support for Porifera-sister (PP = 1; Figure S3K), indicating that the analyses of Whelan

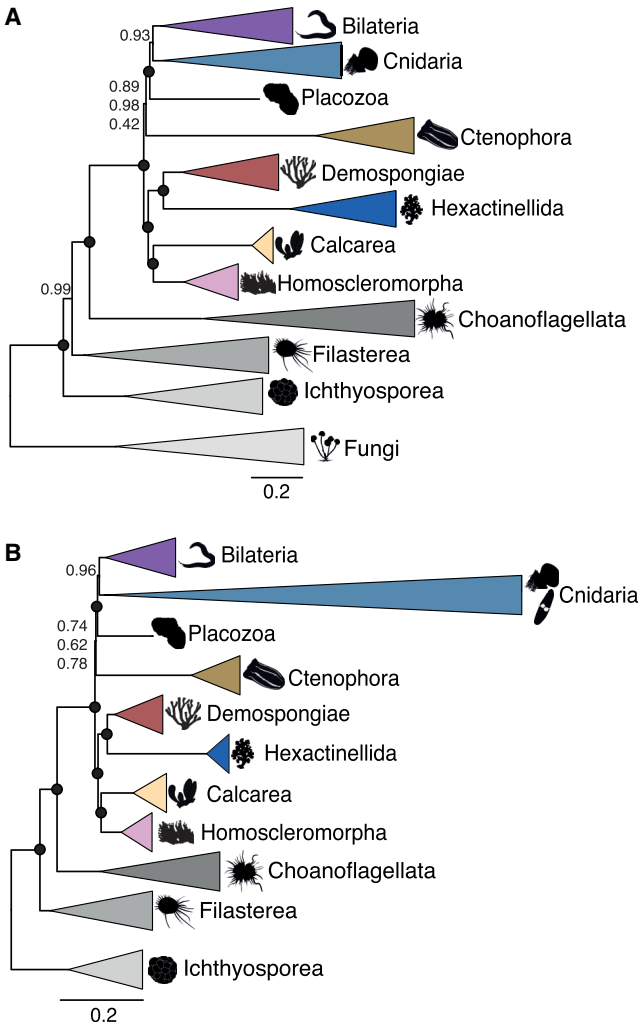


Figure 1. CAT-GTR+G Trees from Dayhoff-6 Recoded Datasets

WhelanD20-Opistho (A) and Chang (B). Numbers at nodes are Bayesian posterior probabilities (PPs). Solid circles indicate nodes with PP ≈ 1. Support for the node identifying Porifera-sister is provided for each of the three recoding strategies that we tested. Top: Dayhoff-6. Middle: S&R-6. Bottom: KGB-6. See Figures S1 and S4 for relationships within the clades. The results in Figure 1 were further corroborated by the analyses of datasets from [11, 38] (Figures S3I, S3J, and S3K). Animal silhouettes are from http://www.phylopic.org, except for the figure of the calcarean sponge, which is from [39]. See the supplemental figures for convergence statistics.

et al. [38], are affected by the same problems that we highlighted in the case of the amino acid datasets of WhelanD20 and Chang.

Ultimately, however, analyses should be conducted under models explicitly accounting for heterogeneity across both sites and lineages (STAR Methods). Such models have been proposed [30] but are still computationally too demanding to be applied to phylogenomic datasets. Simion et al. [7] also showed that the exclusion of sites for which site-specific replacement patterns change over time (heteropecillous sites [58]) uncovers support for Coelenterata (i.e., Ctenophora sister to Cnidaria) [39, 41]. This result suggests that future investigations of early animal relationships should also account for the potentially misleading effect of heteropecilly.

Phylogenetic analyses of early animal relationships using models that do not describe site-specific amino acid differences (WAG+G, LG+G, GTR+G, and optimal PF-schemes) invariably support Ctenophora-sister, whereas analyses using models that can accommodate such preferences (e.g., CAT-GTR+G) predominantly favor Porifera-sister. This model-dependent outcome has been interpreted as reflecting a fundamental lack of robustness [59], and it has been suggested that the relationships at the root of the animal tree might be impossible to reconstruct based on current amino acid datasets [2, 4]. However, sensitivity to the model (and to other analytical factors) does not imply an impossibility to decide which solution is most likely to be correct. Not all models are equal, and there are objective methods to assess which models have a higher fit (model comparison) or most adequately describe important features of the data (posterior predictive analyses). Far from being random, the varied outcomes of phylogenetic analyses of animal relationships show a clear pattern, with Ctenophora-sister being systematically associated with the use of inadequate models and/or taxon-sampling schemes that are most likely to exacerbate systematic errors. In addition, Porifera-sister has been corroborated by analyses of presence/absence of orthologous genes [6], whereas Ctenophora-sister is currently uncorroborated by independent evidence. We conclude that Ctenophora-sister is a tree reconstruction artifact, whereas a placement of the sponges as the sister group of all other animals most likely reflects genuine phylogenetic signal.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Dataset selection
  - PartitionFinder analyses
  - Testing model adequacy
  - Data recoding
  - Phylogenetic analyses
  - Testing the distribution of the signal in favor of Porifera-sister
  - Testing the effect of incrementing the number of ctenophoran lineages on the phylogenetic stability of recoded datasets
  - A discussion of alternative evolutionary models
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and three tables and can be found with this article online at https://doi.org/10.1016/j.cub.2017.11.008.

A video abstract is available at https://doi.org/10.1016/j.cub.2017.11.008#mmc4.

AUTHOR CONTRIBUTIONS

Conceptualization, D.P., G.W., R.F., and O.R.-S.; Methodology, D.P., O.R.-S., W.P., N.L., and H.P.; Software, W.P. and N.L.; Validation, W.P., R.F., M.D.,

D.P., and G.W.; Investigation, R.F., D.P., G.W., and H.P.; Writing – Original Draft, R.F., M.D., D.P., and G.W.; Writing – Review & Editing, M.D., R.F., D.P., G.W., W.P., N.L., H.P., and O.R.-S.; Visualization, R.F.; Supervision, D.P. and G.W.; Funding Acquisition, D.P., G.W., N.L., and H.P.

ACKNOWLEDGMENTS

We would like to thank the University of Bristol ACRC (Advanced Computing Research Center) and Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften for providing access to supercomputing infrastructure. This work was supported by a NERC grant (NE/P013643/1) and a Templeton Foundation grant (ID 60579) to D.P. H.P. acknowledges the French Laboratory of Excellence project "TULIP" (ANR-10-LABX- 41 and ANR-11-IDEX-0002-02). N.L. acknowledges French National Research Agency grant no. ANR-10-BINF-01-01 "Ancestrome." G.W. acknowledges funding by LMU Munich's Institutional Strategy LMUexcellent within the framework of the German Excellence Initiative and German Research Foundation (DFG) grant no. Wo896/15-1. The authors would like to thank the reviewers for their helpful suggestions.

Received: June 21, 2017
Revised: September 19, 2017
Accepted: November 2, 2017
Published: November 30, 2017

REFERENCES

1. Dohrmann, M., and Wörheide, G. (2013). Novel scenarios of early animal evolution–is it time to rewrite textbooks? Integr. Comp. Biol. *53*, 503–511.
2. Dunn, C.W., Giribet, G., Edgecombe, G.D., and Hejnol, A. (2014). Animal phylogeny and its evolutionary implications. Annu. Rev. Ecol. Evol. Syst. *45*, 371–395.
3. Telford, M.J., Budd, G.E., and Philippe, H. (2015). Phylogenomic insights into animal evolution. Curr. Biol. *25*, R876–R887.
4. King, N., and Rokas, A. (2017). Embracing uncertainty in reconstructing early animal evolution. Curr. Biol. *27*, R1081–R1088.
5. Dunn, C.W. (2017). Ctenophore trees. Nat Ecol Evol *1*, 1600–1601.
6. Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., and Wörheide, G. (2015). Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. USA *112*, 15402–15407.
7. Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D.J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., et al. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Curr. Biol. *27*, 958–967.
8. Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A.-D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., Smith, S.A., et al.; NISC Comparative Sequencing Program (2013). The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science *342*, 1242592.
9. Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. Nature *510*, 109–114.
10. Whelan, N.V., Kocot, K.M., Moroz, L.L., and Halanych, K.M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA *112*, 5773–5778.
11. 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.
12. Chang, E.S., Neuhof, M., Rubinstein, N.D., Diamant, A., Philippe, H., Huchon, D., and Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. USA *112*, 14912–14917.
13. Hrdy, I., Hirt, R.P., Dolezal, P., Bardonová, L., Foster, P.G., Tachezy, J., and Embley, T.M. (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature *432*, 618–622.
14. Susko, E., and Roger, A.J. (2007). On reduced amino acid alphabets for phylogenetic inference. Mol. Biol. Evol. *24*, 2139–2150.
15. Kosiol, C., Goldman, N., and Buttimore, N.H. (2004). A new criterion and method for amino acid classification. J. Theor. Biol. *228*, 97–106.
16. Goldman, N. (1993). Statistical tests of models of DNA substitution. J. Mol. Evol. *36*, 182–198.
17. Bollback, J.P. (2002). Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. *19*, 1171–1180.
18. Heath, T.A., Hedtke, S.M., and Hillis, D.M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. *46*, 239–257.
19. Ripplinger, J., and Sullivan, J. (2010). Assessment of substitution model adequacy using frequentist and Bayesian methods. Mol. Biol. Evol. *27*, 2790–2803.
20. Ekman, S., and Blaalid, R. (2011). The devil in the details: interactions between the branch-length prior and likelihood model affect node support and branch lengths in the phylogeny of the Psoraceae. Syst. Biol. *60*, 541–561.
21. Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. *18*, 691–699.
22. Le, S.Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. Mol. Biol. Evol. *25*, 1307–1320.
23. Yang, Z. (1994). Estimating the pattern of nucleotide substitution. J. Mol. Evol. *39*, 105–111.
24. Lanfear, R., Calcott, B., Ho, S.Y.W., and Guindon, S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. *29*, 1695–1701.
25. 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.
26. 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* (*Suppl 1*), S4.
27. Tarver, J.E., Dos Reis, M., Mirarab, S., Moran, R.J., Parker, S., O'Reilly, J.E., King, B.L., O'Connell, M.J., Asher, R.J., Warnow, T., et al. (2016). The interrelationships of placental mammals and the limits of phylogenetic inference. Genome Biol. Evol. *8*, 330–344.
28. Halanych, K.M., Whelan, N.V., Kocot, K.M., Kohn, A.B., and Moroz, L.L. (2016). Miscues misplace sponges. Proc. Natl. Acad. Sci. USA *113*, E946–E947.
29. Whelan, N.V., and Halanych, K.M. (2017). Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. Syst. Biol. *66*, 232–255.
30. Blanquart, S., and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. *25*, 842–858.
31. Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. *56*, 389–399.
32. Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., and Embley, T.M. (2008). The archaebacterial origin of eukaryotes. Proc. Natl. Acad. Sci. USA *105*, 20356–20361.
33. Rota-Stabelli, O., Lartillot, N., Philippe, H., and Pisani, D. (2013). Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst. Biol. *62*, 121–133.
34. Domman, D., Horn, M., Embley, T.M., and Williams, T.A. (2015). Plastid establishment did not require a chlamydial partner. Nat. Commun. *6*, 6421.
35. Borowiec, M.L., Lee, E.K., Chiu, J.C., and Plachetzki, D.C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome

data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics *16*, 987.

36. Schwentner, M., Combosch, D.J., Pakes Nelson, J., and Giribet, G. (2017). A phylogenomic solution to the origin of insects by resolving crustacean-hexapod relationships. Curr. Biol. *27*, 1818–1824.
37. Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978). A model of evolutionary change in proteins. In Atlas of Protein Sequence and Structure, M.O. Dayhoff, ed. (National Biomedical Research Foundation), pp. 345–352.
38. Whelan, N.V., Kocot, K.M., Moroz, T.P., Mukherjee, K., Williams, P., Paulay, G., Moroz, L.L., and Halanych, K.M. (2017). Ctenophore relationships and their placement as the sister group to all other animals. Nat. Ecol. Evol. *1*, 1737–1746.
39. Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Müller, W.E.G., Nickel, M., Schierwater, B., et al. (2013). Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. *67*, 223–233.
40. Jékely, G., Paps, J., and Nielsen, C. (2015). The phylogenetic position of ctenophores and the origin(s) of nervous systems. Evodevo *6*, 1.
41. Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., et al. (2009). Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. *19*, 706–712.
42. Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alié, A., Morgenstern, B., Manuel, M., and Wörheide, G. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. *27*, 1983–1987.
43. 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.
44. Eitel, M., Francis, W., Osigus, H.-J., Krebs, S., Vargas, S., Blum, H., Williams, G.A., Schierwater, B., and Wörheide, G. (2017). A taxogenomics approach uncovers a new genus in the phylum Placozoa. bioRxiv. https://doi.org/10.1101/202119.
45. Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature *452*, 745–749.
46. Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., et al. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. Biol. Sci. *276*, 4261–4270.
47. 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.
48. Posada, D., and Crandall, K.A. (1998). MODELTEST: testing the model of DNA substitution. Bioinformatics *14*, 817–818.
49. Cunningham, C.W., Zhu, H., and Hillis, D.M. (1998). Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution *52*, 978–987.
50. Sullivan, J., and Joyce, P. (2005). Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. *36*, 445–466.
51. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Series B Stat. Methodol. *36*, 111–147.
52. Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*, 2104–2105.
53. Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion. J. R. Stat. Soc. Series B Stat. Methodol. *39*, 44–47.
54. Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286–2288.
55. Foster, P.G. (2004). Modeling compositional heterogeneity. Syst. Biol. *53*, 485–495.
56. Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., and Wörheide, G. (2016). Reply to Halanych et al.: ctenophore misplacement is corroborated by independent datasets. Proc. Natl. Acad. Sci. USA *113*, E948–E949.
57. Hillis, D.M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. *47*, 3–8.
58. Roure, B., and Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. *11*, 17.
59. Maxmen, A. (2017). Big data renews fight over animal origins. Nature News, http://www.nature.com/news/big-data-renews-fight-over-animal-origins-1.21703.
60. Le, S.Q., Dang, C.C., and Gascuel, O. (2012). Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. *29*, 2921–2936.
61. 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.
62. Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. *8*, 275–282.
63. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.
64. Lanave, C., Preparata, G., Saccone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. J. Mol. Evol. *20*, 86–93.
65. Feuda, R., Rota-Stabelli, O., Oakley, T.H., and Pisani, D. (2014). The comb jelly opsins and the origins of animal phototransduction. Genome Biol. Evol. *6*, 1964–1971.
66. Feuda, R., Hamilton, S.C., McInerney, J.O., and Pisani, D. (2012). Metazoan opsin evolution reveals a simple route to animal vision. Proc. Natl. Acad. Sci. USA *109*, 18868–18872.
67. Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. *10*, 1396–1401.
68. Thorne, J.L., Kishino, H., and Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. *15*, 1647–1657.
69. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. *39*, 306–314.
70. Blanquart, S., and Lartillot, N. (2006). A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. *23*, 2058–2071.
71. Groussin, M., Boussau, B., and Gouy, M. (2013). A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst. Biol. *62*, 523–538.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Datasets analyzed, results of all analyses conducted, and output files generated by programs used	This work	https://bitbucket.org/bzxdp/feuda_et_al_2017
Software and Algorithms		
Phylobayes MPI	N/A	https://github.com/bayesiancook/pbmpi
PPA-Conv	N/A	https://github.com/bayesiancook/pbmpi
PPA-Var	N/A	https://github.com/bayesiancook/pbmpi
Phylobayes – PF	N/A	https://github.com/bayesiancook/pbmpi/tree/partition
PartitionFinder	N/A	http://www.robertlanfear.com/partitionfinder/
RAxML	N/A	https://github.com/stamatak/standard-RAxML/releases/tag/v8.2.9

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Davide Pisani (davide.pisani@bristol.ac.uk).

METHOD DETAILS

Dataset selection

We considered two phylogenomic datasets that strongly bear on our understanding of relationships at the root of the animal tree. The first is the dataset of [12] – hereafter referred to as Chang, which scores 77 taxa and 51,940 amino acid positions. The second is a dataset from [10] that had not been reassessed in [6]. The dataset we selected from [10] was dataset D20, their least compositionally heterogeneous dataset, which apparently explicitly excludes proteins of significant compositional heterogeneity (see [10]). This dataset scores 70 taxa and 42,542 amino acid positions. Following [6], dataset D20 (hereafter WhelanD20) was subjected to analyses that used different subsets of outgroups. We name ''WhelanD20-Opistho'' the version of WhelanD20 that includes all the original outgroups (Fungi, Choanoflagellata and several other holozoans); ''WhelanD20-Holo'' the dataset that excludes fungal outgroups; and ''WhelanD20-Choano'' the dataset that excludes all the outgroups but Choanoflagellata.

PartitionFinder analyses

PartitionFinder version 1.1.1 Mac and PartitionFinder version 2.0.0 [24] were used following the protocol of [10] to select, for WhelanD20-Opistho and Chang, the optimal sets of partitions and their associated sets of best-fitting, universal, predefined general time reversible matrices. It is important to note that we followed the protocol of [10] in order to be able to precisely test whether PF-schemes previously used to investigate early animal relationships can adequately describe the considered datasets. Following [10], not all models that have been developed for amino acid datasets, and that could be considered by PartitionFinder (e.g., LG4X [60]), are included in the set of compared models. The complete list of models considered in the protocol of [10] only includes: LG, WAG, mtREV, Dayhoff, DCMut, JTT, VT, Blosum62, CpREV, RtREV, MtMam, MtArt, HIVb, HIVw (see PartitionFinder manual and references therein for details about these models). The selected sets of partitions and their associated sets of models will be referred to as ''PF-schemes.''

Testing model adequacy

Posterior Predictive Analyses were performed on WhelanD20-Opistho and Chang to test whether WAG+G, LG+G, GTR+G, the optimal PF-schemes, or CAT-GTR+G can adequately describe site-specific amino acid preferences and lineage-specific compositional heterogeneity for these datasets. These models were selected because they had previously been used in key studies of early animal evolution that reached contradictory conclusions. For example, GTR+G was used by [8], WAG+G was used by [9], PF-schemes were used by [10], CAT-GTR+G was used by [6] and LG+G was used by [11]. Other important studies [7, 12] used CAT+G, which has not been considered here, as CAT-GTR+G is used as a representative of the CAT-based models more broadly (see below: ''A discussion of alternative evolutionary models'' for a comparison of these models).

Posterior Predictive Analyses (PPA) of site-specific amino acid diversity test how well site-specific amino acid preferences are accounted for by a model. We used three alternative tests statistics: (1) the mean number of distinct amino acids observed at each site

(PPA-DIV [26]). (2) An empirical measure of the site-specific probability of convergent evolution toward the same amino-acid in distantly related taxa (PPA-CONV). This is obtained as follows: for each site, the empirical frequencies of the 20 amino-acids are calculated and the probability of randomly drawing twice the same amino-acid is estimated. This quantity is then averaged across all sites. (3) The variance across sites of the empirical frequency of each amino-acid, averaged over all amino-acids (PPA-VAR).

PPA for compositional heterogeneity test how well lineage-specific compositional preferences are accounted for by the model. We used as a primary measure maximal heterogeneity across taxa (PPA-MAX [30]). We also investigated whether using mean squared heterogeneity across taxa (PPA-MEAN [30]) could have led to different conclusions. Both PPA-MAX and PPA-MEAN use the deviation between specific taxa and the dataset average, measured as the sum over the 20 amino-acids of the absolute differences between the taxon-specific and global empirical frequencies. The maximal heterogeneity across taxa is the maximum deviation over all taxa, while the mean squared heterogeneity across taxa is the sum of the squared deviations (calculated across all the taxa in the dataset). All PPA analyses used a minimum of 100 replicates. All PPA performed under WAG+G, LG+G, GTR+G and CAT-GTR+G used the latest release of Phylobayes MPI [61], which includes PPA-CONV and PPA-VAR (see below for software access). To perform PPA analyses under the optimal PF-schemes we developed a new version of Phylobayes MPI (see below for software access). For the PF analyses, amino acid frequencies and the Gamma distribution used to model across-site rate-heterogeneity were independently inferred from the data for each partition.

Results of PPA analyses were quantified using Z-scores. When computing Z-scores it is assumed that the null distribution is approximately normal, with the Z-score being analogous to a standard normal quantile. Results of PPA can also be quantified using a P value, which is the fraction of simulated replicates with a test statistic less than the observed statistic. Because P values are computed from a discrete set of simulated replicates, they may not be distinguishable from 0 or 1 when a model is strongly rejected (i.e., when there are too few samples to accurately estimate a very small tail probability). Accordingly, Z-scores have the advantage of providing a much greater comparative resolution when a model is strongly rejected, with a very large Z-score (in absolute value) indicating strong model mis-specification. As a general rule, to compare Z-scores and P values, one-tailed tests based on results of PPA where $|Z| > 5$ would obtain a $P \approx 1e\text{-}5$, indicating a strong rejection of the null hypothesis that the model can adequately describe the data. On the other hand, one-tailed tests based on results where $|Z| < 2$ would obtain a P that would not be rejected at the standard 0.05 level. All the output files from our PPA analyses are available (see below for data access).

Data recoding

Amino acid recoding has long been considered as a powerful strategy to reduce compositional heterogeneity and saturation in the data [7, 13–15, 31–36] (see also below: A discussion of alternative evolutionary models). Amino acid recoding strategies identify sets of amino acids where there is a high probability of change within each set and a small probability of change between the sets [15]. The best known recoding strategy is Dayhoff-6, which was originally presented in Figure 84 of [37]. Dayhoff recoding partitions amino acids in six classes or bins (Table 2), based on the log odds matrix of probabilities of pairs of amino acids appearing together in a PAM 250 matrix (see [14, 37] for details). Alternative recoding strategies have been devised, which use different criteria to bin amino acids into classes [14, 15]. To test whether our results were specific to the application of Dayhoff-6 recoding, we used two other amino acid transformation strategies. These are the JTT-based [62] 6-state recoding strategy of Susko and Roger [14] (hereafter S&R-6) and the PAM 120 / PAM 250-based, 6-state recoding strategy of Kosiol et al. [15] (hereafter KGB-6). The bins defined by S&R-6 and KGB-6 are reported in Table 2. The bins in S&R-6 are defined based on the JTT rate matrix. In S&R-6 the bins are defined to maximize the ratio of the expected number of substitutions within bins to the expected number under the Jukes-Cantor model [14]. The scheme of Kosiol et al. [15] is based on the concept of *conductance*: the expected number of changes between amino acids within two sets when a Markov process is close to equilibrium (see [15] for details). Both WhelanD20 and Chang were recoded using each of the three considered strategies.

Phylogenetic analyses

Phylogenetic analyses of Chang and WhelanD20-Opistho were performed under WAG+G, LG+G, GTR+G (with and without Dayhoff-6, S&R6 and KGB-6 recoding), and CAT-GTR+G (with and without Dayhoff-6, S&R6 and KGB-6 recoding). GTR+G and CAT-GTR+G analyses were also performed (with and without Dayhoff-6, S&R6 and KGB-6 recoding) for WhelanD20-Holo and WhelanD20-Choano. Note that 6-state recoded datasets cannot be analyzed using WAG+G, LG+G or a PF-scheme as they require a 6-state GTR matrix, while WAG, LG and the other universal, predefined, matrices used by PartitionFinder are 20-state matrices. All LG+G, WAG+G, GTR+G and CAT-GTR+G analyses were performed using Phylobayes MPI v1.7a [61]. For each analysis two chains were run and convergence was assessed using the *bpcomp* and *tracecomp* tools in PhyloBayes. The number of samples to be discarded as burnin was independently assessed for each analysis by visually checking traces of likelihood and parameter values. Chang was analyzed under its optimal PF-scheme using RAxML version 8.2.9 [63], with support estimated using the rapid bootstrap option (100 replicates). WhelanD20 was not analyzed under its optimal PF-scheme because [10] already showed that this dataset supports Ctenophora-sister when the PF-scheme is used. Results of all our analyses including trace files and output of all the tests performed to investigate convergence are available (see below for data access).

Testing the distribution of the signal in favor of Porifera-sister

While the debate on the relationships at the root of the metazoan tree has been dominated by arguments about model choice [6, 28, 29, 56], a recent study [11] argued that signal for Porifera-sister is limited to a few outlier genes. The study of Shen et al.

[11] used datasets from various studies, including dataset D16 from [10], which is very similar to WhelanD20, as the latter was obtained in [10] by removing genes of significant compositional heterogeneity from D16. To test their results, Shen et al. [11] generated datasets that excluded the few genes that they identified as "outliers," analyzed these datasets under LG+G, and found support for Ctenophora-sister. As datasets without "outlier genes" should not include signal in favor of Porifera-sister, we tested their hypothesis by analyzing one of their purportedly outlier-free, and hence Porifera-sister signal-free, datasets under CAT-GTR+G (with and without Dayhoff recoding). To run this test we selected WhelanD16-OutlierExcluded-Choano – i.e., WhelanD16 with only Choanoflagellata as outgroups and excluding all outlier genes identified by [11]. If support for Porifera-sister emerges from the analysis of this dataset (which should be free of any signal supporting this topology), Shen et al.'s [11] hypothesis would be rejected, and the emergence of Ctenophora-sister in their analyses a consequence of the use of the inadequate (see Table 1 and S1) LG+G model. For each analysis, two chains were run in Phylobayes MPI version 1.7a [61] and convergence was assessed using the *bpcomp* and *tracecomp* tools in PhyloBayes. The number of samples to be discarded as burnin was independently assessed for each analysis by visually checking traces of likelihood and parameter values. Results of all our analyses including tracefiles are available (see below for data access).

Testing the effect of incrementing the number of ctenophoran lineages on the phylogenetic stability of recoded datasets

Whelan et al. [38] have increased the number of ctenophoran species compared to those they used in [10]. Ctenophores are all long branched (see Figure 2 in [38]), and adding more long branched taxa to a dataset decreases the ability of all available models to adequately describe the data (see Results and Discussion for details). Given our key finding, that relationships at the root of the animal tree depend on the relative adequacy with which alternative modeling strategies describe the data, it is not unsurprising that [38] recovered Ctenophora-sister using CAT-GTR+G and closely related outgroups; similarly to the case of [12]. The dataset of [38] provided us with a case to further test the conclusions of our study: that support for alternative hypotheses of animal relationships depends on the ability of the model used in the analyses to describe the data. If our conclusions hold and Ctenophora-sister is an artifact emerging when the data are inadequately modeled, the analysis of a recoded version of the dataset of [38] should support Porifera-sister. We tested this hypothesis performing an analysis of a Dayhoff-6 recoded version of the dataset used by Whelan et al. [38] to recover their Figure 2 (i.e., their Whelan2017_Metazoa_Choano_RCFV_strict dataset). For this analysis, two chains were run in Phylobayes MPI version 1.7a [61] and convergence was assessed using the *bpcomp* and *tracecomp* tools in PhyloBayes. The number of samples to be discarded as burnin was assessed by visually checking traces of likelihood and parameter values. Results of all our analyses including tracefiles are available (see below for data access).

A discussion of alternative evolutionary models

The simplest substitution models are site- and lineage-homogeneous, representing the amino acid replacement process using the same model across sites and lineages [23, 64]. Examples of homogeneous amino acid models include those using empirical matrices that have been derived from large collections of protein families like the JTT [62], WAG [21] and LG [60] matrices, and dataset-specific GTR (General Time Reversible) matrices that are directly inferred from the alignment under study. Empirical matrices are expected to fit most alignments relatively well. However, a dataset-specific GTR matrix is expected to fit the alignment from which it has been inferred best (and better than empirical matrices) [65, 66]. A GTR matrix, if used in isolation, can only describe a site-, lineage-, rate-, composition-, and replacement-homogeneous process. However, protein evolution is not homogeneous. Different sites and lineages accumulate substitutions at different rates (e.g., [67–69]), and different amino acids are preferentially found at different sites or in different lineages (e.g., [26, 55]). Finally, different proteins can have their own protein-specific replacement rates (e.g., [24]).

Variation in substitution rates among sites is typically accounted for by assuming that site-specific substitution rates vary according to a Gamma distribution [67, 69], and the well-known GTR+G and WAG+G models are heterogeneous with respect to site-specific substitution rates, as they relax the site-specific rate homogeneity assumption. Similarly, models have been introduced to account for variation in equilibrium amino acid frequencies among sites, including those based on the CAT mixture, which describes site-specific amino acid profiles using a Dirichlet process [25]. These models were introduced to account for the observation that different amino acid sites in one or many proteins may have comparable amino acid profiles due to biochemical functional constraints [25]. The combined application of a GTR matrix and the CAT mixture (as in the CAT-WAG or CAT-GTR model) thus relaxes the assumption that the same replacement process applies to all sites of the alignment (i.e., that all sites are equally likely to accept all amino acids). Models that use both site-specific substitution rates and site-specific amino acid profiles (e.g., CAT-WAG+G or CAT-GTR+G, respectively) account for both site-specific rates and amino acid preferences. Finally, multiple GTR matrices can be used to model gene-level replacement-process heterogeneity. PartitionFinder [24] is the most commonly used automated approach to partitioning a superalignment and assigning to each of its partitions the best-fitting in a predefined set of empirical matrices, which currently does not include CAT-based models.

In addition to site-specific amino-acid preferences, lineage-specific compositional heterogeneity has been shown to constitute another important confounding factor in phylogenetics (e.g., [7, 13–15, 31–36, 55]). Yet, none of the above discussed models are designed to account for this type of evolutionary heterogeneity. Models that can explicitly take into consideration across-taxon compositional heterogeneity, such as the Breakpoint model [70], the Node Discrete Compositional Heterogeneity (NDCH) model [55] and the Correspondence and Likelihood Analysis (COaLA) model [71] have been developed. However, current implementations of these models are computationally too demanding to allow their application to large phylogenomic datasets. As an alternative to

using a model that captures lineage-specific heterogeneity, one can transform the dataset so that it displays less heterogeneity, leading to weaker violations of the assumptions of lineage-homogeneous models. Dayhoff-6 and other recoding strategies [14, 15, 37] are well-known procedures that can be used to reduce the compositional heterogeneity of an amino acid alignment (e.g., [7, 13, 31–36]). Using this approach, amino acids with similar physicochemical properties are recoded into one of six categories, reducing lineage- and site-specific amino acid preferences [33]. The amino acids within each category are known to interchange more frequently with each other than they do with amino acids from other categories because of their similar properties, which allows maintaining protein function and structure. Because Dayhoff recoding clusters amino acids with comparable properties within the same category, it reduces heterogeneity across both sites and lineages, potentially reducing the impact of systematic errors on phylogenetic estimation under currently available models. Because amino acids are grouped into classes recoding strategies can in some cases cause signal erosion. This however can be monitored because it will result in collapsed nodes that do not provide support for any topology, rather than a change in the topology supported by the data.

DATA AND SOFTWARE AVAILABILITY

Results of all our analyses including tracefiles are available at https://bitbucket.org/bzxdp/feuda_et_al_2017.

Phylobayes MPI which includes PPA-CONV and PPA-VAR is available at https://github.com/bayesiancook/pbmpi.

Partitioned Phylobayes MPI, to perform PPA analyses under the optimal PF-schemes is available at https://github.com/bayesiancook/pbmpi/tree/partition.