Supplementary Materials for

Rooting the animal tree of life

Yuanning Li[1,3], Xing-Xing Shen[2,3], Benjamin Evans[4], Casey W. Dunn[1]* and Antonis Rokas[3]*

* Corresponding authors, antonis.rokas@vanderbilt.edu, casey.dunn@yale.edu

List of Supplementary Materials

Supplementary Text S1 - S3

Tables S1 – S9

Fig. S1 - S9

References

Supplementary Materials

Supplementary Text

S1 Summaries of published analyses

Narrative summaries of the studies considered here.

Dunn *et al.* 2008

Dunn *et al.* (Dunn et al. 2008) added Expressed Sequence Tag (EST) data for 29 animals. It was the first phylogenomic analysis that included ctenophores, and therefore that could test the relationships of both Ctenophora and Porifera to the rest of animals. It was also the first phylogenetic analysis to recover Ctenophora as the sister group to all other animals.

The data matrix was constructed using a semi-automated approach. Genes were translated into proteins, promiscuous domains were masked, all gene sequences from all species were compared to each other with blastp, genes were clustered based on this similarity with TribeMCL (Enright et al. 2002), and these clusters were filtered to remove those with poor taxon sampling and high rates of lineage-specific duplications. Gene trees were then constructed, and in clades of sequences all from the same species all but one sequence were removed (these groups are often due to assembly errors). The remaining gene trees with more than one sequence for any taxon were then manually inspected. If strongly supported deep nodes indicative of paralogy were found, the entire gene was discarded. If the duplications for a small number of taxa were unresolved, all genes from those taxa were excluded. Genes were then realigned and sites were filtered with Gblocks (Castresana 2000), resulting in a 77 taxon matrix. Some taxa in this matrix were quite unstable, which obscured other strongly-supported relationships. Unstable taxa were identified with leaf stability indices (Thorley and Wilkinson 1999), as implemented in phyutility (Smith and Dunn 2008), and removed from the matrix. This resulted in the 64-taxon matrix that is the focus of most of their analyses. Phylogenetic analyses were conducted under the Poisson+CAT model in PhyloBayes, and under the WAG model in MrBayes (Ronquist and Huelsenbeck 2003) and RAxML (Stamatakis 2006).

Regarding the recovery of Ctenophora-sister, the authors concluded:

The placement of ctenophores (comb jellies) as the sister group to all other sampled metazoans is strongly supported in all our analyses. This result, which has not been postulated before, should be viewed as provisional until more data are considered from placozoans and additional sponges.

Note that there was, in fact, an exception to strong support. An analysis of the 40 ribosomal proteins in the matrix recovered Ctenophora-sister with only 69% support. This study did not include the placozoan ingroup *Trichoplax*.

Philippe *et al.* 2009

Philippe *et al.* (Philippe et al. 2009) assembled an EST dataset for 55 species with 128 genes to explore the deepest animal phylogenetic relationships by adding 9 new species. The data matrix was assembled based on the phylogenetic analysis using Poisson+CAT model, which strongly supported Porifera-sister, and ctenophores were recovered as sister to cnidarians forming the "coelenterate" clade. Gene trees were then constructed, and potentially paralogs were removed by a bootstrap threshold of 70. Ambiguously aligned regions were trimmed and only genes sampled for at least two-thirds of species were retained. The phylogenetic analyses were conducted under the Poisson+CAT model in PhyloBayes.

Regarding the recovery of Ctenophora-sister, the authors concluded:

"The resulting phylogeny yields two significant conclusions reviving old views that have been challenged in the molecular era: (1) that the sponges (Porifera) are monophyletic and not paraphyletic as repeatedly proposed, thus undermining the idea that ancestral metazoans had a sponge-like body plan; (2) that the most likely position for the ctenophores is together with the cnidarians in a 'coelenterate' clade".

Hejnol *et al.* 2009

Hejnol *et al.* (Hejnol et al. 2009) added EST sequences from seven taxa, and a total of 94 taxa were included in the final data matrix to explore animal phylogeny, especially the position of acoelomorph flatworms. The orthology inference was largely similar to Dunn *et al.* 2008, with the exception of orthology genes which were clustered by MCL. The final data matrix included 1497 genes, and then subsampled with 844, 330 and 53 genes by different thresholds of gene occupancy. Except for the 53 gene matrix, maximum likelihood analyses from all other datasets strongly supported Ctenophora-sister (models were selected by RaxML perl script).

Pick *et al.* 2010

Pick *et al.* (Pick et al. 2010) sought to test whether Ctenophora-sister was an artefact of insufficient taxon sampling. They added new and additional published sequence data to the 64-taxon matrix of Dunn *et al.* (Dunn et al. 2008). The new taxa included 12 sponges, 1 ctenophore, 5 cnidarians, and *Trichoplax*. They further modified the matrix by removing 2,150 sites that were poorly sampled or aligned. They considered two different sets of outgroups: Choanoflagellatea (resulting in Choanozoa) and the same sampling as Dunn *et al.* (resulting in Opisthokonta).

All their analyses were conducted under the F81+CAT+Gamma model in PhyloBayes, in both a Bayesian framework and with bootstrapping. All analyses have the same ingroup sampling and site removal so it isn't possible to independently assess the impact of these factors. Analyses with Choanozoa sampling recovered Porifera-sister with 72% posterior probability (PP) and 91% bootstrap support (BS). With broader Opisthokonta sampling, support for Porifera-sister is 84% PP. This is an interesting case where increased outgroup sampling leads to increased support for Porifera-sister.

The authors argue that previous results supporting Ctenophora-sister "are artifacts stemming from insufficient taxon sampling and long-branch attraction (LBA)" and that "this hypothesis should be rejected". Although the posterior probabilities supporting Porifera-sister are not strong, they conclude: "Results of our analyses indicate that sponges are the sister group to the remaining Metazoa, and Placozoa are sister group to the Bilateria".

They also investigated saturation and concluded that the Dunn *et al.* (Dunn et al. 2008) data matrix is more saturated than Philippe *et al.* 2009 [Philippe:2009hh]. Note that the Pick *et al.* (Pick et al. 2010) dataset is not reanalyzed here because partition data are not available, and due to site filtering the partition file from Dunn *et al.* (Dunn et al. 2008) cannot be applied to this matrix.

Nosenko *et al.* 2013

Nosenko *et al.* (Nosenko et al. 2013) added Expressed Sequence Tag (EST) data for 9 species of non-bilaterian metazoans (7 sponges). They constructed a novel matrix containing 122 genes and parsed them into two non-overlapping matrices (ribosomal and non-ribosomal genes) and found incongruent results of deep metazoan phylogeny. The other major finding was that ribosomal gene partitions showed significantly lower saturation than the non-ribosomal ones.

Orthologs were constructed using the bioinformatics pipeline OrthoSelect (Schreiber et al. 2009). They also evaluated level of saturation, leaf stability of sampled taxa, compositional heterogeneity, and model comparison of each matrix. By modifying gene sampling, ingroup and outgroup sampling, three major topologies related to the position of animal-root were constructed (including Porifera+Placozoa sister, Ctenophora-sister and Porifera-sister). Phylogenetic analyses were conducted under the Poisson+CAT, GTR+CAT and GTR models in PhyloBayes.

Regarding the recovery of Ctenophora-sister, the authors concluded:

"We were able to reconstruct a metazoan phylogeny that is consistent with traditional, morphology-based views on the phylogeny of non-bilaterian metazoans, including monophyletic Porifera and ctenophores as a sister-group of cnidarians".

Ryan *et al.* 2013

Ryan *et al.* (Ryan et al. 2013) sequenced the first ctenophore genome of *Mnemiopsis leidyi*. With the genome resources of *M. leidyi*, the authors constructed two phylogenomic datasets: a "Genome set" based on 13 animal genomes and a "EST Set" that also included 59 animals. They analyzed both matrices by site-homogeneous GTR+Gamma and site-heterogeneous Poisson+CAT models with three sets of outgroup sampling to evaluate the effect of outgroup selection to the ingroup topology for the Ryan2013_est matrix. The Orthologs were constructed based on the method of Hejnol *et al.* 2009. For the Ryan2013_genome matrix, they performed phylogenetic analyses with both gene content and sequence-based analyses. Overall, their results strongly supported Ctenophora-sister in all datasets they analyzed using a site-homogeneous model. The Poisson+CAT model of the genome dataset strongly supported of a clade of Ctenophora and Porifera as the sister group to all other Metazoa and Bayesian analysis on the EST dataset did not converge after 205 days (but strongly supported Porifera in Choaimalia matrix).

Regarding the recovery of Ctenophora-sister, the authors concluded:

"Our phylogenetic analyses suggest that ctenophores are the sister group to the rest of the extant animals".

Moroz *et al.* 2014

Moroz *et al.* (Moroz et al. 2014) sequenced the second ctenophore genome *Pleurobrachia bachei* to explore the phylogenetic relationship of Metazoa. All phylogenetic analyses strongly supported Ctenophora-sister with different taxon and gene sampling using WAG site-homogeneous model. Two phylogenomic matrices were generated, the first set was represented by two ctenophore species, whereas the other set contained improved ctenophore sampling (10 taxa, Moroz2013_3d). Orthology determination employed in HaMStR (Ebersberger et al. 2009) using 1,032 “model organism” single-copy orthologs. Sequences were then trimmed and aligned. This resulted in a final matrix of 170,871 amino acid positions across 586 genes with 44 taxa for the first matrix, and 114 genes with 60 taxa for the second matrix. All the phylogenetic analyses were analyzed in RAxML under the WAG+CAT+F models (different from CAT models in PhyloBayes) to reduce the computational cost.

Regarding the recovery of Ctenophora-sister, the authors concluded:

“Our integrative analyses place Ctenophora as the earliest lineage within Metazoa. This hypothesis is supported by comparative analysis of multiple gene families, including the apparent absence of HOX genes, canonical microRNA machinery, and reduced immune complement in ctenophores”.

It should be noted that only the Moroz_3d matrix has been reanalyzed in other studies, although the support of Ctenophora-sister is quite low.

Borowiec *et al.* 2015

Borowiec *et al.* (Borowiec et al. 2015) assembled a genome dataset comprising 1080 orthologs derived from 36 publicly available genomes representing major lineages of animals, although only one genome of sponge and ctenophore was included. The orthologs were constructed using OrthologID pipeline (Chiu et al. 2006). After removal of spurious sequences and genes with more than 40% of mission data, the final matrix included 1080 (Total 1080) genes. The authors further filtered the full dataset to 9 sub-datasets by filtering genes with high long-branch scores; genes with high saturation; gene occupancy; fast evolving genes. The main conclusion of the study was largely based on BorowiecTotal_1080 and Borowiec_Best108 matrices. Phylogenetic analyses were conducted under the GTR+CAT model in PhyloBayes in selected matrices, and under the data-partitioning methods in RAxML for all matrices.

Regarding the recovery of Ctenophora-sister, the authors concluded:

“Our phylogeny supports the still-controversial position of ctenophores as sister group to all other metazoans. This study also provides a workflow and computational tools for minimizing systematic bias in genome-based phylogenetic analyses”.

It should be noted that the authors also employed recoding-method in the Borowiec_Best108 matrix and found neither support of Porifera-sister or Ctenophora-sister (Borowiec et al. 2015).

Whelan et al. 2015

Whelan *et al.* 2015 (Whelan et al. 2015) constructed a new phylogenomic data matrix with eight new transcriptomic data and investigated a range of possible sources of systematic error under multiple analyses (*e.g.,* long-branch attraction, compositional bias, fast evolving genes, etc.). Putative orthologs were determined of each species using HaMStR using the model organism core ortholog set (same as Moroz *et al.* 2014) and subsequently removal of genes with too much missing data and potential paralogs. The authors further filtered the full dataset to 24 sub-datasets by filtering genes with high long-branch scores; genes with high RSFV values; genes that are potential paralogs; fast evolving genes and progressive removal of outgroups. All the maximum likelihood analyses with site-homogeneous models and PartitionFinder strongly suggested Ctenophora-sister. GTR+CAT models only used in slow-evolving data matrices 6 and 16 also strongly supported Ctenophora.

Regarding the recovery of Ctenophora-sister, the authors concluded:

“Importantly, biases resulting from elevated compositional heterogeneity or elevated substitution rates are ruled out. Placement of ctenophores as sister to all other animals, and sponge monophyly, are strongly supported under multiple analyses, herein”.

Note that the authors also reanalyzed Philippe2009 matrix (with the removal of ribosomal genes) and recovered Porifera-sister with moderate support (pp=90).

Chang et al. 2015

Chang *et al.* (Chang et al. 2015) was originally used to explore the phylogenetic position of Myxozoa in Cnidaria but also sampled broadly across the breadth of animal diversity. The authors constructed a dataset with 200 protein markers based on Philippe *et al.* 2011 (Philippe et al. 2011) with 51,940 amino acids and 77 taxa. Both site-heterogeneous Poisson+CAT and site-homogeneous GTR models strongly supported Ctenophora-sister.

Pisani *et al.* 2015

Pisani *et al.* (Pisani et al. 2015) reanalyzed representative datasets that supported Ctenophora-sister, including Ryan2013_est, Moroz2014_3d and Whelan2015 datasets. It was the first study showing that progressive removal of more distantly related outgroups could largely affect phylogenomic inference of the position of the root of animal phylogeny. The authors suggested that the inclusion of outgroups very distant from the ingroup can cause systematic errors due to long-branch attraction. Phylogenetic analyses were conducted under the Poisson+CAT and GTR models in PhyloBayes. They found Poisson+CAT models generally had better model-fit than site-homogeneous GTR models in these data matrices. Moreover, they found the support of Ctenophora-sister decreases when the exclusion of distantly related outgroups are excluded and the use of site-heterogeneous CAT models are used.

Regarding the recovery of Porifera-sister, the authors concluded:

"Our results reinforce a traditional scenario for the evolution of complexity in animals, and indicate that inferences about the evolution of Metazoa based on the Ctenophora-sister hypothesis are not supported by the currently available data".

Feuda *et al.* 2017

Feuda *et al.* (Feuda et al. 2017) didn't generate any new data, instead they used the data-recoding methods to reanalyze two key datasets that support Ctenophora-sister (Whelan2015_D20, Chang2015 datasets). It was the first phylogenomic study that suggested recoding methods have better performance than non-recoding methods based on recovering Porifera-sister hypothesis. The authors compared model adequacy using posterior predictive analyses from a set of site-homogeneous (WAG, LG, GTR, data-partitioning) and site-heterogeneous (GTR+CAT) models in non-recoding and recoding datasets. The results showed that data-recoding can significantly reduce compositional heterogeneity in both datasets with GTR+CAT models and strongly supported Porifera-sister hypothesis (see more details in Supplementary Information section S3).

Regarding the recovery of Porifera-sister, the authors concluded:

"Because adequate modeling of the evolutionary process that generated the data is fundamental to recovering an accurate phylogeny, 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".

Whelan and Halanych 2016

Whelan *et al.* (Whelan and Halanych 2016) is the only study to evaluate performance of site-heterogeneous models and site-homogeneous models with data partitioning under the simulation framework. The simulation results suggested that the Poisson+CAT model consistently performed worse than other models in simulation datasets. More importantly, the authors also showed that both Poisson+CAT and GTR+CAT models could overestimate substitutional heterogeneity in almost every case. They also reanalyzed datasets from Philippe2009 and Nosenko2013 using both CAT models and data partitioning with site-homogeneous model. The results indicated that Poisson+CAT model tends to recover less accurate trees and both GTR+CAT and data partitioning strongly supported Ctenophora-sister in reanalyses.

The authors concluded:

“Practices such as removing constant sites and parsimony uninformative characters, or using CAT-F81 when CAT-GTR is deemed too computationally expensive, cannot be logically justified. Given clear problems with CAT-F81, phylogenies previously inferred with this model should be reassessed”.

Whelan *et al.* 2017

Whelan *et al.* (Whelan et al. 2017) added 27 new ctenophore transcriptomic data to explore animal-root position as well as relationships within Ctenophores. It significantly increased ctenophore taxon sampling than other studies. Putative orthologs were determined largely similar to Whelan2015. The subsequent filtering strategy was also similar to the previous study. All analyses using site-homogeneous and site-heterogeneous models strongly supported Ctenophora-sister hypothesis, even with GTR+CAT model in Choanozoa dataset. The main conclusions of this study were based on Whelan2017_full and Whelan2017_strict matrices.

Regarding the recovery of Ctenophora-sister, the authors concluded: “Using datasets with reasonably high ctenophore and other non-bilaterian taxon sampling, our results strongly reject the hypothesis that sponges are the sister lineage to all other extant metazoans”.

Simion *et al.* 2017

Simion *et al.* (Simion et al. 2017) added transcriptomic data for 21 new animals. The data matrix was constructed using a semi-automated approach to comprehensively detect and eliminate potential systematic errors. The resulting dataset comprises 1,719 genes and 97 species, including 61 non-bilaterian species. It was by far the largest phylogenomic dataset in terms of taxon and gene sampling related to the relationship at the root of animal phylogeny.

The final matrix was first analyzed using the Poisson+CAT model. Different from other PhyloBayes analyses, Simion *et al.* used a gene jackknife strategy based on 100 analyses to overcome the computational limitation because of the large data size. Each jackknife is based on a random selection of ~ 25% of the genes. The PhyloBayes with site-heterogeneous model strongly supported the Porifera-sister, whereas site-homogeneous strongly supported Ctenophora-sister in all datasets. Importantly, the authors compared the behavior of long-branch effect between site-heterogeneous and site-homogeneous models by progressively removing taxa and concluded higher sensitivity of site-homogeneous models to LBA than CAT models.

Regarding the recovery of Ctenophora-sister, the authors concluded:

“Our dataset outperforms previous metazoan gene super alignments in terms of data quality and quantity. Analyses with a best-fitting site-heterogeneous evolutionary model provide strong statistical support for placing sponges as the sister-group to all other metazoans, with ctenophores emerging as the second-earliest branching animal lineage”.

It should be noted that all the PhyloBayes runs have not reached convergence due to the computational cost in these large matrices.

S2 Models of molecular evolution

The exchangeability matrix R describes the relative rates at which one amino acid changes to others. Exchangeability matrices have been used in the studies under consideration here include: Poisson (or F81), WAG, LG, GTR. While the exchangeability matrix describes the relative rate of different changes between amino acids, the actual rate can be further scaled. There are a couple approaches that have been used in the studies considered here:

- F81 (Felsenstein 1981) corresponds to equal rates between all states. The F81 matrix is also sometimes referred to as the Poisson matrix. It has no free parameters to estimate since all off-diagonal elements are set to 1.

- WAG (Whelan and Goldman 2001) is an empirically derived exchangeability matrix based on a dataset of 182 globular protein families. It has no free parameters to estimate since all off-diagonal elements are set according to values estimated from this particular sample dataset.

- LG (Le and Gascuel 2008), like WAG, is an empirically derived exchangeability matrix. It is based on a much larger set of genes, and variation in rates across sites was taken into consideration when it was calculated. It has no free parameters to estimate since all off-diagonal elements are set according to values estimated from this particular sample dataset.

- GTR, the General Time Reversible exchangeability matrix, has free parameters for all off-diagonal elements that describe the exchangeability of different amino acids. It is constrained so that changes are reversible, *i.e.* the rates above the diagonal are the same as those below the diagonal. This leaves 190 parameters that must be estimated from the data along with the other model parameters and the phylogenetic tree topology. This estimation requires a considerable amount of data and computational power, but if successful has the advantage of being based on the dataset at hand rather than a different dataset (as for LG and WAG).

While the exchangeability matrix describes the relative rate of different changes between amino acids, the actual rate can be further scaled. There are a couple approaches that have been used in the studies considered here:

- Site homogeneous rates. The rates of evolution are assumed to be the same at all sites in the amino acid alignment.

- Gamma rate heterogeneity. Each site is assigned to a different rate class with its own rate value. This accommodates different rates of evolution across different sites. Gamma is used so commonly that sometimes it isn't even specified, making it difficult at times to know if a study uses Gamma or not.

The vector of equilibrium frequencies Π describes the stationary frequency of amino acids. There are a few approaches that have been used across the studies considered here:

- Empirical site-homogeneous. The frequency of each amino acid is observed from the matrix under consideration and applied to homogeneously to all sites in the matrix.

- Estimated site-homogeneous. The frequency of each amino acid is inferred along with other model parameters, under the assumption that it is the same at all sites.

- CAT site-heterogeneous. Each site is assigned to a class with its own equilibrium frequencies. The number of classes, assignment of sites to classes, and equilibrium frequencies within the data are all estimated in a Bayesian framework.
- C10 to C60 (Si Quang et al. 2008). 10 to 60-profile mixture models as variants of the CAT model under the maximum-likelihood framework.

S3 Data-recoding methods

Feuda *et al.* (Feuda et al. 2017) were concerned that the Ctenophora-sister results Chang *et al.* (Chang et al. 2015) and Whelan *et al.* (Whelan et al. 2015) were artefacts of lineage-specific differences in amino acid frequencies. In an attempt to reduce these differences, they recoded the full set of twenty amino acids into six groups of amino acids. These groups have more frequent evolutionary changes within them than between them, based on empirical observations in large protein datasets (Susko and Roger 2007). The intent is to discard many lineage-specific changes, which are expected to fall within these groups. Rather than model compositional heterogeneity, as their title suggests, this approach discards heterogeneous information so that much simpler models with fewer states can be applied.

Feuda *et al.* (Feuda et al. 2017) report that posterior predictive (PP) analyses (Bollback 2002) indicate 6-state recoded analyses have better model adequacy than 20-state amino acid analyses, and “Porifera-sister was favored under all recoding strategies” in Whelan2015_D20 and Chang2015 data matrices. Here we focus on two aspects of Feuda *et al.* First, we point out that many of their recoded analyses are actually unresolved (*i.e.*, without strong support for either Porifera-sister or Ctenophora-sister) (Fig. S8A). Second, we present new analyses that show the impact of recoding is largely due to discarding information, not accommodating variation in amino acid composition. These findings indicate that recoding can be a problematic method for addressing compositional variation.

Feuda *et al.* examine support for Ctenophora-sister and Porifera-sister under all combinations of two models of molecular evolution, four datasets, and four coding schemes. This provides 32 analyses that they report in their Table 3 and that we present graphically here as Fig. S8A. There is striking variation in support for Ctenophora-sister and Porifera-sister across these analyses (Fig. S8A). Feuda *et al.* accept the results of some analyses and reject others based on posterior predictive (PP) analyses of model adequacy, which assess how well a model explains variation in the data (Bollback 2002). They considered five different posterior predictive statistics that capture different types of variation in the data. From this they conclude that their “results strongly support sponges as the sister group of all other animals”.

This conclusion does not follow from their own presented results. Only a single analysis with posterior predictive scores provides what could be considered strong support > 95 posterior probability) for Porifera-sister. Of the 32 analyses, posterior predictive scores were calculated for 16 (those for the full Whelan and Chang matrices). Based on posterior predictive scores, Feuda *et al.* reject eight of these that were conducted under the GTR+G model (which all have strong support for Ctenophora-sister). This leaves eight GTR+CAT analyses (Fig. S8A). Two of these eight are analyses of the original 20-state amino acid data, both of which provide strong support for Ctenophora-sister. Of the six recoded analyses, five are unresolved. Only a single analysis for which posterior predictive scores are available provides strong support for Porifera-sister, the GTR+CAT analysis of the SR-6 (Susko and Roger 2007) recoded Whelan (N.V. Whelan et al. 2015b) matrix. Furthermore, this analysis does not have the best score according to any of the five posterior predictive statistics they considered (Fig. S8B). The only statistic that stands out for this one analysis is that it has the highest maxdiff (Fig. S8B), indicating that it did not converge as well as other analyses.

Though their study does not provide strong support for Porifera-sister, the sensitivity of their results to recoding provides an opportunity to better understand and evaluate the impact of recoding more generally. This is important given the growing interest in recoding (Feuda et al. 2017). Feuda *et al.* hoped recoding would reduce potential artefacts due to differences across species in amino acid frequencies. They interpreted the fact that their analyses are sensitive to recoding as evidence that such an artefact exists and that they successfully addressed it by recoding. An alternative hypothesis is that recoding impacts phylogenetic analyses because it discards so much information. These two hypotheses can be tested by applying new recoding schemes that also reduce twenty states down to six, but are based on random grouping rather than empirical frequencies of amino acid exchange. Empirical and random recodings both discard the same amount of information, but only empirical recoding reduce the impact of amino-acid frequency as intended. Different results between empirical and random recoding would be consistent with the hypothesis that the empirical approach works as intended to accommodate compositional heterogeneity. Similar results would suggest that the impact of recoding is due to discarding information. Here we focus on their single analysis with a posterior predictive score that supports Porifera-sister, the GTR+CAT analysis of the SR-6 recoded Whelan data. We created four new random recoding schemes by shuffling the amino acids in the SR-6 scheme (see Supplemental Methods and analysis code at https://github.com/caseywdunn/feuda_2017). When we applied each of these randomized codes to the Whelan matrix and analyzed them under the GTR+CAT model with PhyloBayes-MPI, we observed similar results as for the empirical SR-6 recoding. Like SR-6 recoding, random recoding increases support for Porifera-sister and improves the apparent adequacy of models to explain heterogeneity of states across taxa (PP taxon hetero mean and max, Fig. S9).

These analyses suggest that the major impact of recoding on phylogenetic analyses is data reduction, not accommodation of compositional heterogeneity across species. This indicates that recoding may not be an effective tool for accommodating among-species differences in amino acid frequencies. Compositional heterogeneity would be better addressed with models of molecular evolution that explicitly describe such differences (Blanquart and Lartillot 2008, Foster 2004), if progress can be made on the considerable computational challenges of such complex models.

Supplementary Figures

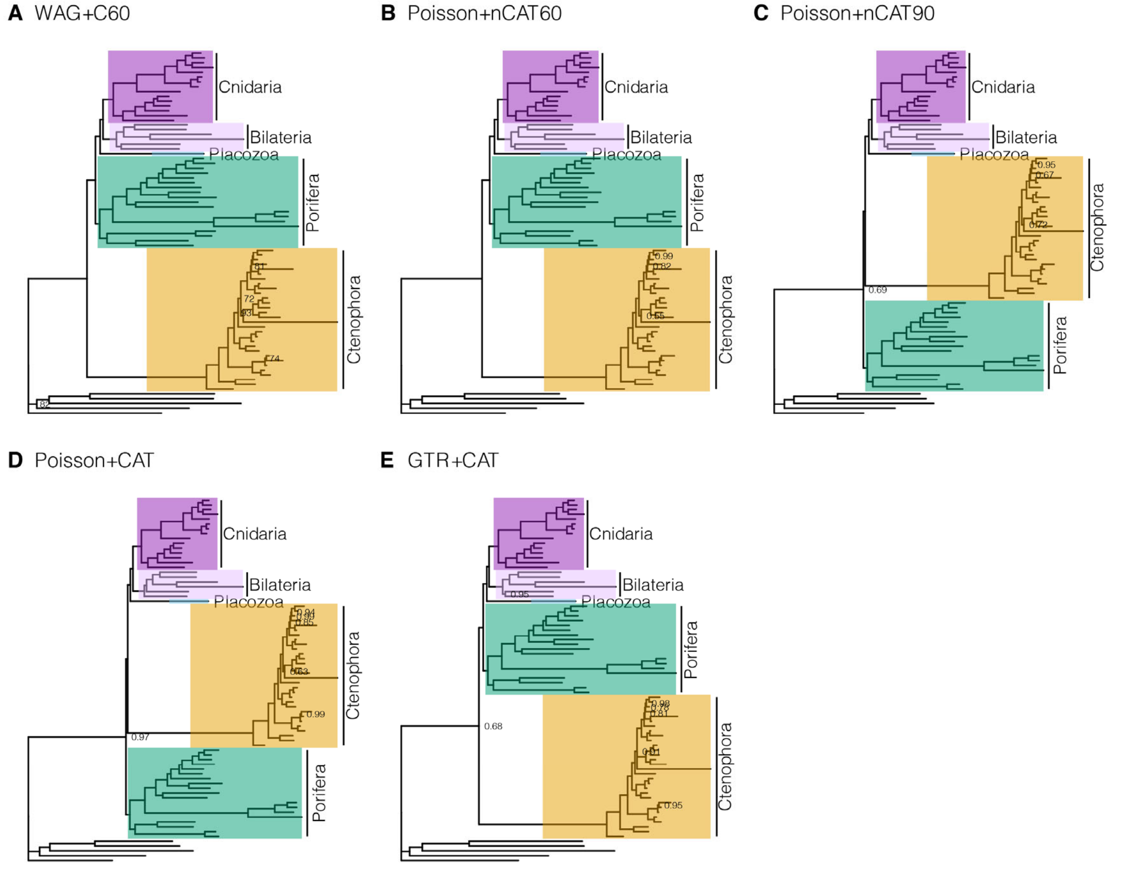


Fig. S1. Comparison of the tree topologies under different substitutional models in Whelan2017_strict data matrix. (A). WAG+C60. (B). Poisson+nCAT60. (C). Poisson+nCAT90. (D). Poisson+CAT. (E). GTR+CAT. All trees were visualized and annotated in R package ggtree (Yu et al. 2018).

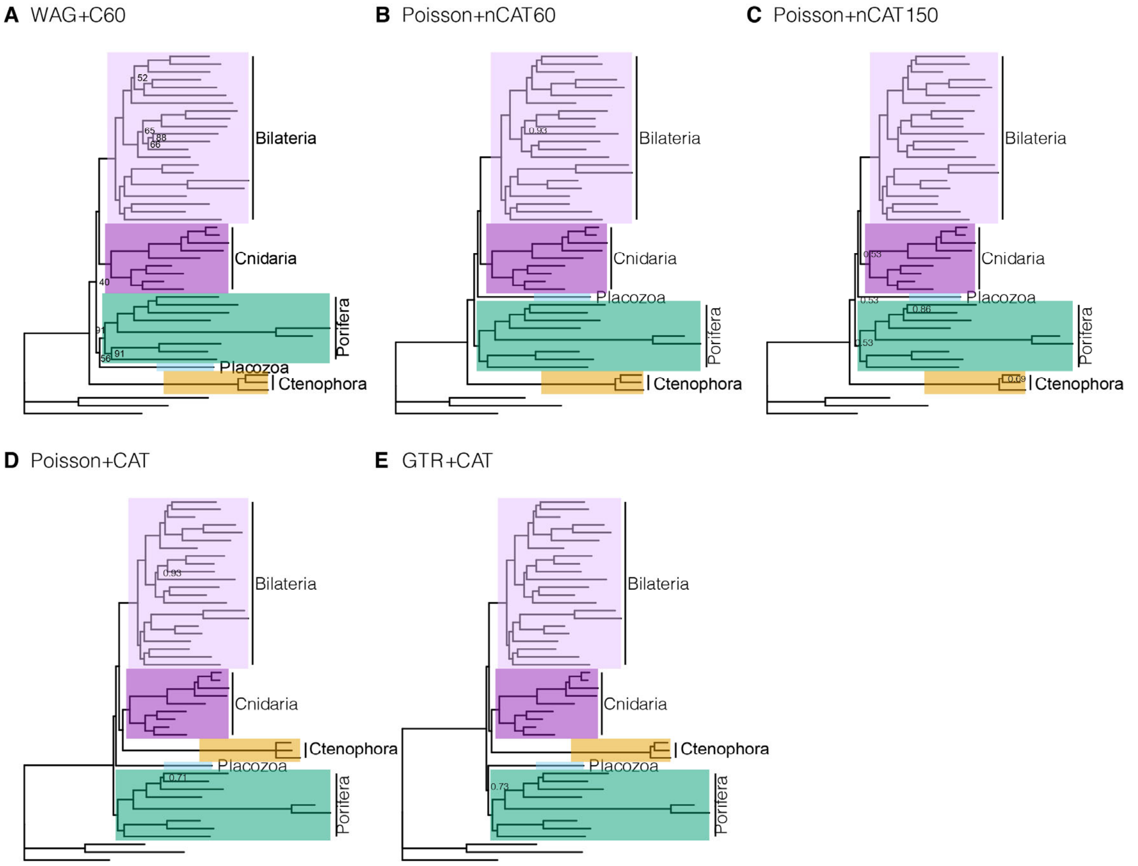


Fig. S2. Comparison of the tree topologies under different substitutional models in Philippe2009_Choanozoa data matrix. (A). WAG+C60. (B). Poisson+nCAT60. (C). Poisson+nCAT150. (D). Poisson+CAT. (E). GTR+CAT. All trees were visualized and annotated in R package ggtree (Yu et al. 2018).

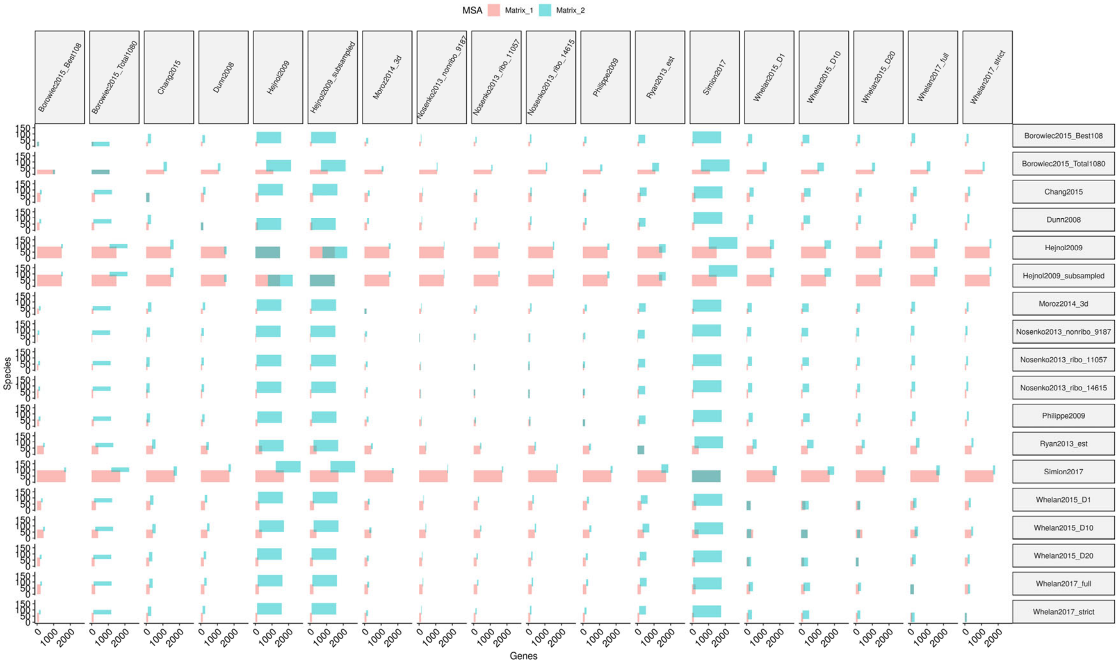


Fig. S3. Pairwise overlap between each of the primary matrices considered here (related to Table S9). Horizontal size is proportional to the number of genes sampled, vertical size to the number of taxa sampled. The horizontal intersection shows the proportions of shared genes, the vertical intersection shows the proportions of shared taxa.

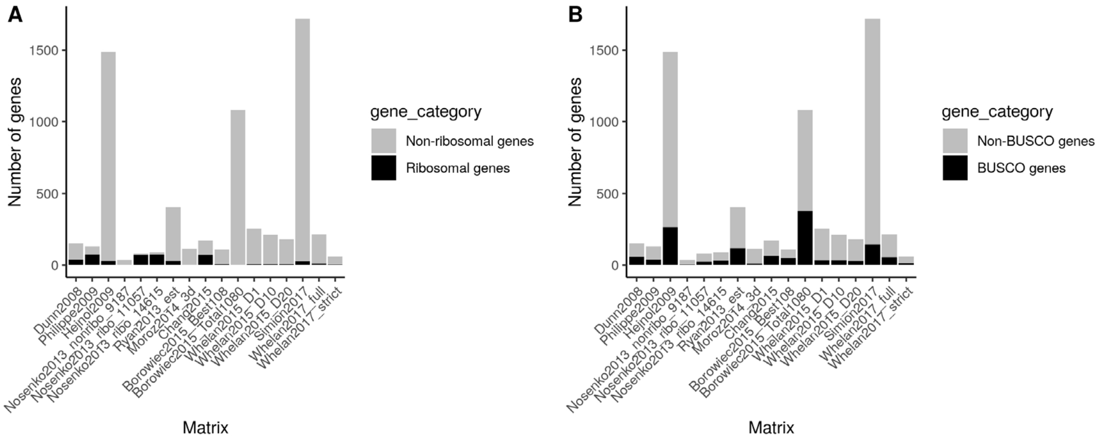


Fig. S4. Annotation and representation of BUSCO and ribosomal genes in each data matrix (related to Table S9). (A). The number of partitions with ribosomal annotations in each matrix, relative to the number of partitions. (C). The number of partitions with annotations of BUSCO genes in each matrix, relative to the number of partitions.

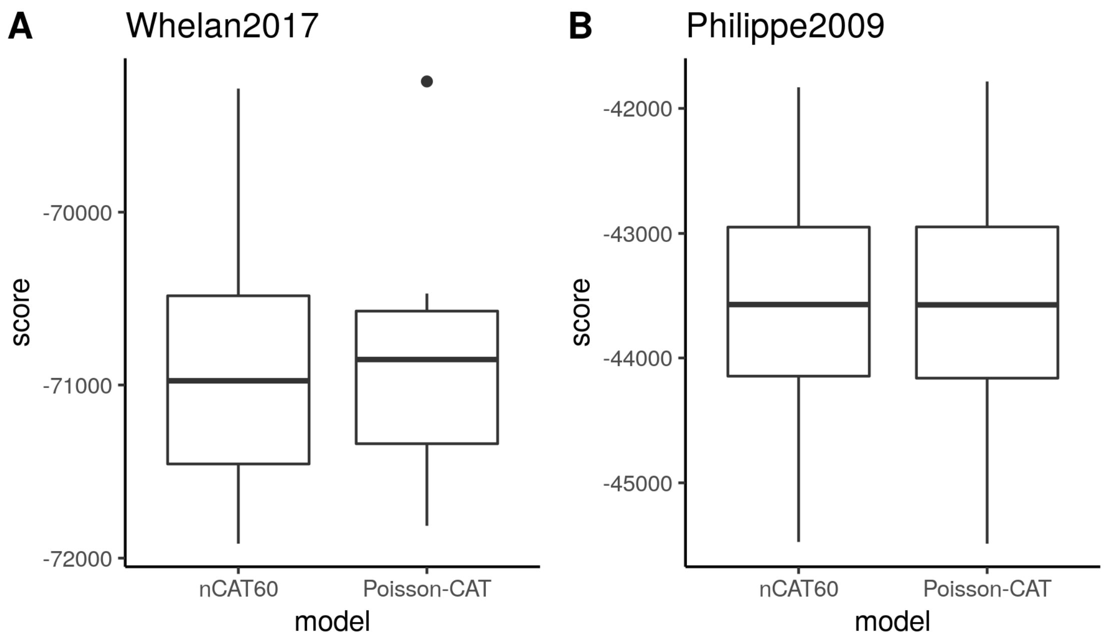


Fig. S5. Box plots of log likelihood score for 10-fold cross-validation between Poisson+CAT and Poisson+nCAT60 models in Whelan2017_strict and Philippe2009_Choanozoa data matrices.

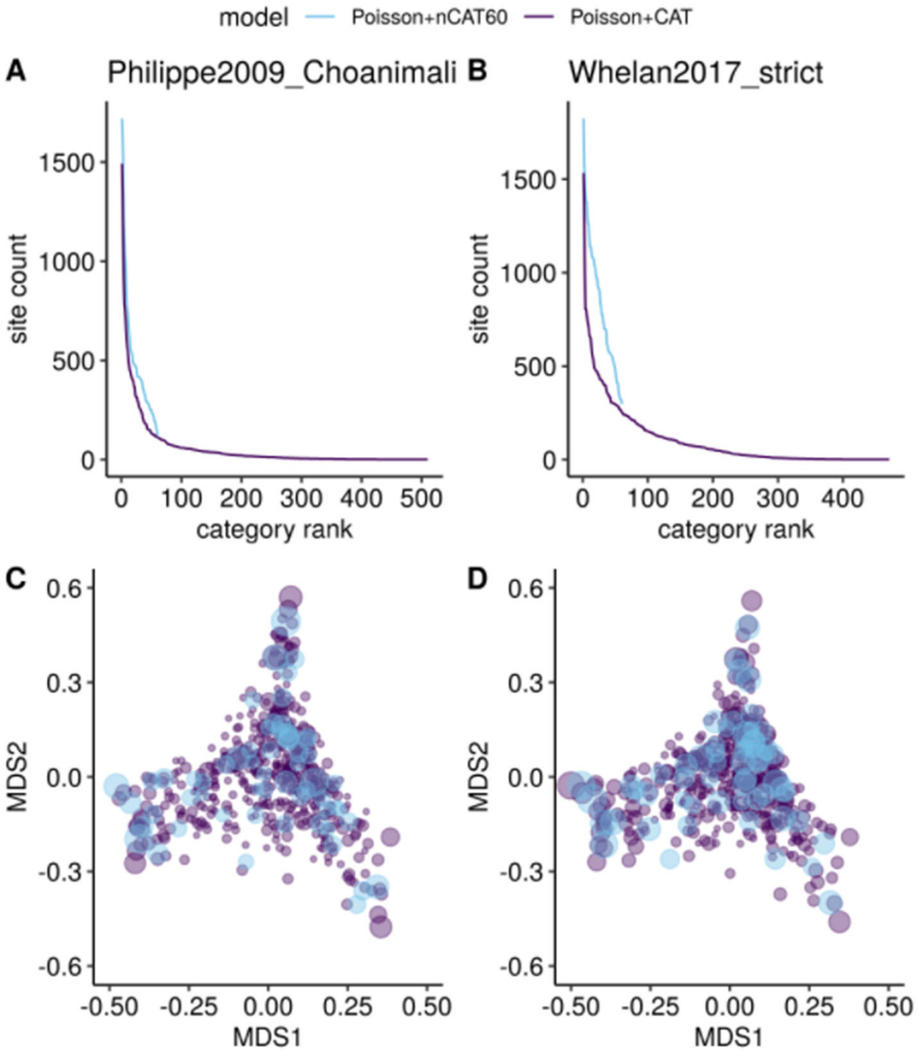


Fig. S6. The allocation of categories across sites in the Philippe2009 matrix (left column) and the Whelan2017_strict matrix (right column) for the constrained Poisson + nCAT60 model and unconstrained Poisson + CAT models (shown in different colors). The plots are from a single PhyloBayes generation. (A, B) The counts of sites allocated to each equilibrium frequency category, with the categories ranked from the most abundant to least abundant along the x axis. The unconstrained CAT analyses have a long tail of categories that are allocated to very few sites, which adds a considerable number of parameters that pertain to a very small fraction of sites. The nCAT60 analyses, which are constrained to 60 sites, have no such long tail and the rarest categories contain far greater numbers of sites than any of the sites in the long tails of the CAT analyses. (C, D) Multidimensional scaling (MDS) plots for the amino acid frequencies of the two data matrices. Each point is for a single category, the area of the point is proportional to the number of sites allocated to that category, and categories with points that are closer to each other on the plot have more similar amino acid frequencies. A single MDS analysis was done for both Poisson + nCAT60 and unconstrained Poisson + CAT analyses of each data matrix, so that the projections are the same in each plot. For both matrices, the much larger number of categories in the unconstrained CAT analyses largely subdivide the amino acid frequency space already covered in the constrained nCAT60 analyses into a far greater number of rare categories. This adds a considerable number of parameters for describing a similar overall space.

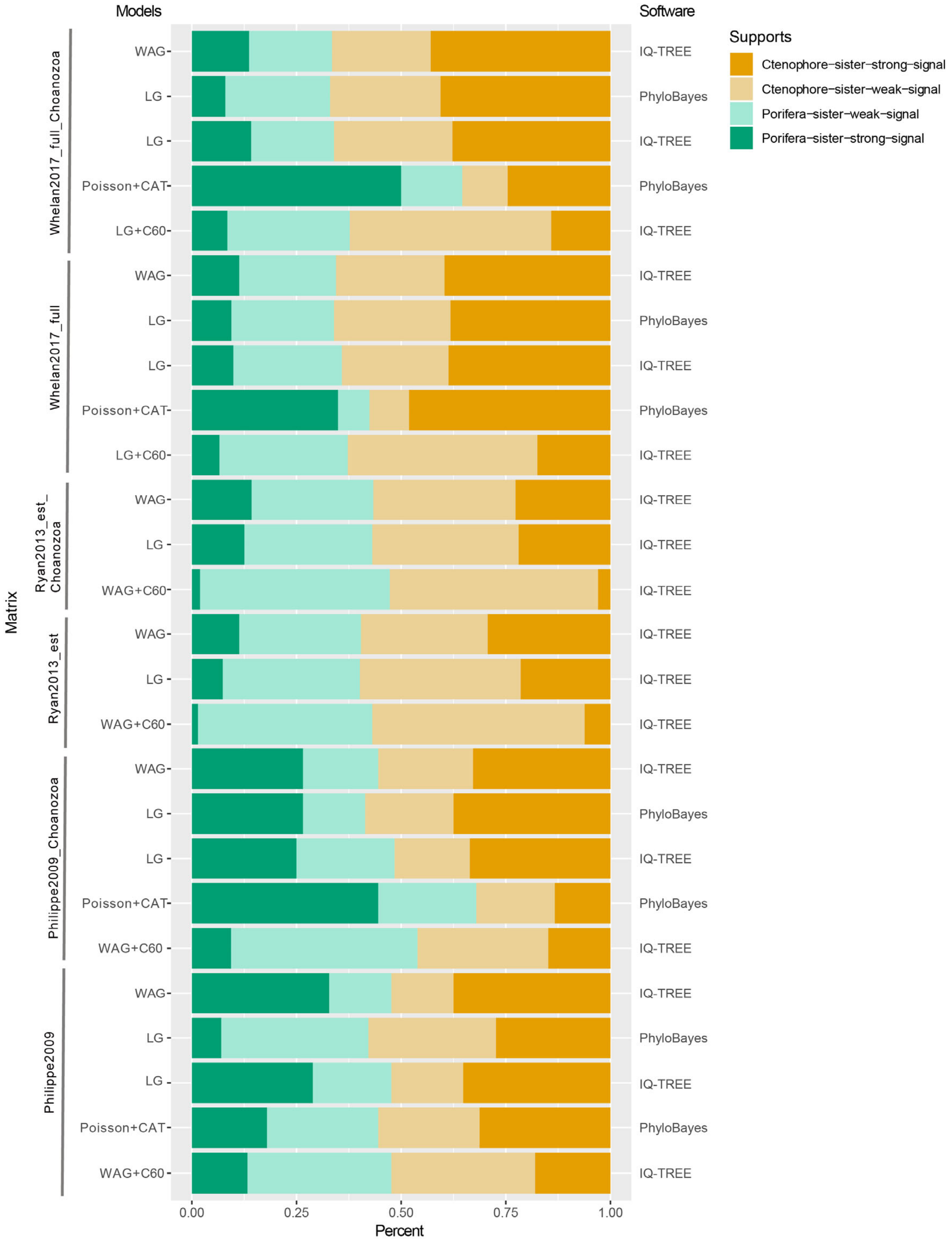
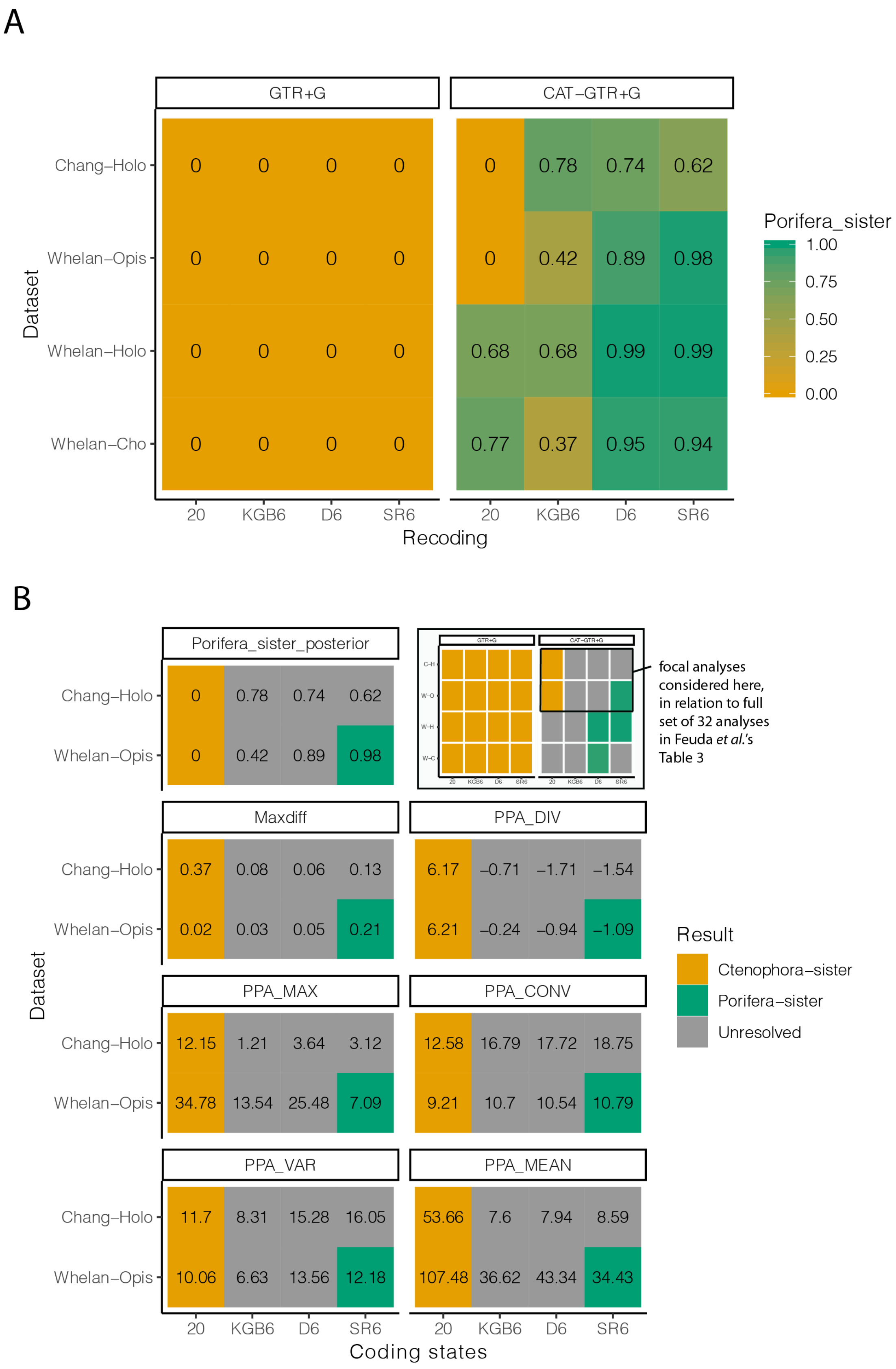


Fig. S7. The distribution of phylogenetic signal for Ctenophora-sister and Porifera-sister with different models and outgroup choice in Philippe2009, Ryan2013 and Whelan2017_full matrices (with two outgroup sampling: Choanozoa and full matrices). The two alternative topological hypotheses are: Ctenophora-sister; T1 (Orange); Porifera-sister T2 (Green). Proportions of

genes supporting each of two alternative hypotheses in the Philippe2009, Ryan2013_est and Whelan2017_full data matrices with different outgroups sampling and substitutional models. The GLS values for each gene in each data matrix are provided in Table S8. We considered a gene with an absolute value of log-likelihood difference of two as a gene with strong ($|\Delta \ln L| \geq 2$) or weak ($0 < |\Delta \ln L| < 2$) phylogenetic signal.



focal analyses considered here, in relation to full set of 32 analyses in Feuda *et al.*'s Table 3

Fig. S8. Graphical representations of the posterior probabilities for the 32 analyses presented by Feuda *et al.* in their Table 3 and the subset of eight GTR+CAT analyses with posterior predictive (PP) scores that is the focus of Feuda *et al.*'s primary conclusions. (A). Cells are color coded by whether posterior probability is > 95 for Porifera-sister, > 95 for Ctenophora-sister, or neither (unresolved). Posterior predictive (PP) statistics were estimated for the 16 analyses in the top two rows of this figure (the Chang and full Whelan matrices), but not the bottom two (the Whelan matrices with reduced outgroup sampling). (B). These are a subset of the 32 analyses presented in their Table 3 and graphically here in the upper right pane. The eight analyses are for two datasets (Chang and Whelan) and four coding schemes. The coding schemes are the original 20 state amino acid data, and three different six state recodings that group amino acids based on different criteria: KGB6, D6, and SR6. Only one of these analyses, the SR6 coding of the Whelan matrix, has > 95 support for Porifera-sister. The 20-state and 6-state points on the plots in Fig. S8A correspond to the 20 and SR6 Whelan cells shown here. The presented statistics for these cells are posterior probability of Porifera-sister, Maxdiff (with lower scores indicating better convergence of runs), and five posterior predictive statistics (where lower absolute value indicates better model adequacy). The only one of these eight analyses that provides strong support for Porifera-sister is not the most adequate analysis by any of the posterior predictive scores, and showed the poorest convergence according to Maxdiff.

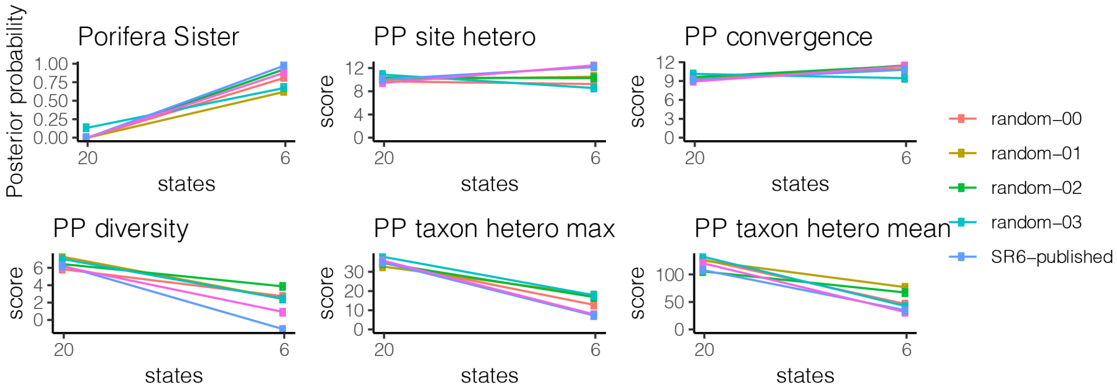


Fig. S9. Each of the five plots presents one statistic, which include Posterior probability of Porifera-sister and the five Posterior Predictive (PP) statistics considered by Feuda *et al.* Within each plot, there are six lines for five different analyses. These five analyses are the published SR-6 analyses presented by Feuda *et al.* (SR6-published), and four analyses based on randomized recoding matrices obtained by shuffling the SR-6 coding scheme (random-00 - random-03). Each analysis includes results for 20 states (the raw amino acid data, shown by the left point) and for 6 states (the 6-state recoded data, shown by the right point). For each

statistic, the results obtained with the random recoding are similar to those of the SR6 recoding. This indicates that the impact of recoding is dominated by discarding data when collapsing from 20 states to 6 states, not accommodating compositional hetezrogeneity across lineages.

Supplementary Tables

Table S1. Summary of a total of 164 phylogenomic analyses were transcribed from the literature (Table is converted from analyses_published in Rdata).

Table S2. Summary of a total of 106 phylogenomic analyses conducted in this study (Table is converted from analyses_new in R data).

Table S3. The models selected by ModelFinder for each matrix in IQ-TREE. (Table is converted from analyses_new in Rdata).

Table S4. Summary of a total of 18 phylogenomic analyses when only Fungi, Holozoa or Choanoflagellatea are used as outgroups.

Table S5. Summary statistics of CAT substitutional categories inferred from different matrices (Table is manualy curated from Tracer result for each PhyloBayes analysis).

Table S6. Summary of sensitive analyses with different number of CAT substitutional categories in representative matrices (Table is converted from analyses_sensitive in Rdata).

Table S7. Summary of amino acid frequencies of 60 categories inferred by C60 model using IQ-TREE (Table is manualy curated from IQtree log file for each analysis).

Table S8. Distribution of phylogenetic signal of different models and outgroup sampling for two alternative hypotheses on the animal-root position in Philippe2009, Ryan2013 and Whelan2017_full matrices (Table is converted from au_tests in Rdata).

Table S9. Summary of annotations (BUSCO, Ensemble, GO-terms, ribosomal protein genes) and network analyses for each partition from all phylogenomic all matrices used in this study (Table is converted from partition_map_globle in Rdata).

References

Blanquart S, Lartillot N. 2008. A Site- and Time-Heterogeneous Model of Amino Acid Replacement. *Molecular Biology and Evolution* 25:842–858. Available from: https://doi.org/10.1093/molbev/msn018

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution* 19:1171–1180. Available from: https://doi.org/10.1093/oxfordjournals.molbev.a004175

Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. 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.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17:540–552. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10742046

Chang ES, Neuhof M, Rubinstein ND, Diamant A, Philippe H, Huchon D, Cartwright P. 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences* [Internet]:1–6. Available from: https://doi.org/10.1073/pnas.1511468112

Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R. 2006. OrthologID: Automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22:699–707.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749. Available from: http://www.nature.com/doifinder/10.1038/nature06614

Ebersberger I, Strauss S, Haeseler A von. 2009. HaMStR: Profile hidden markov model based search for orthologs in ests. *BMC evolutionary biology* 9:157.

Enright A, Van Dongen S, Ouzounis C. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30:1575–1584. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/30.7.1575

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=7288891&retmode=ref&cmd=prlinks

Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, Wörheide G, Pisani D. 2017. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Current Biology* 27:3864–3870.

Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, Wörheide G, Pisani D. 2017. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology* 27:1–12. Available from: https://doi.org/10.1016/j.cub.2017.11.008

Foster PG. 2004. Modeling compositional heterogeneity. *Systematic biology* 53:485–495. Available from: https://doi.org/10.1080/10635150490445779

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences* 276:4261–4270.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* 25:1307–1320. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msn067

Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, et al. 2014. The ctenophore genome and the evolutionary origins of neural systems. *Nature* 510:109.

Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, Maldonado M, Müller WE, Nickel M, Schierwater B, et al. 2013. Deep metazoan phylogeny: When different genes tell different stories. *Molecular phylogenetics and evolution* 67:223–233.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS biology* 9:e1000602.

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, QuEinnec E, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. *Current biology* 19:706–712. Available from: http://dx.doi.org/10.1016/j.cub.2009.02.052

Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, Wiens M, Alie A, Morgenstern B, Manuel M, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Molecular Biology and Evolution* 27. Available from: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msq089

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. *Proceedings of the National Academy of Sciences* 112:15402–15407.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574. Available from: http://www.bioinformatics.oupjournals.org/cgi/doi/10.1093/bioinformatics/btg180

Schreiber F, Pick K, Erpenbeck D, Wörheide G, Morgenstern B. 2009. OrthoSelect: A protocol for selecting orthologous groups in phylogenomics. *BMC bioinformatics* 10:219.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Queinnec E, Ereskovsky A, et al. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology* 27:958–967.

Si Quang L, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24:715–716. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm619

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16928733

Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution* 24:2139–2150. Available from: https://doi.org/10.1093/molbev/msm144

Thorley J, Wilkinson M. 1999. Testing the phylogenetic stability of early tetrapods. *Journal of Theoretical Biology* 200:343–344. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0022519399909992

Whelan NV, Halanych KM. 2016. Who let the cat out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. *Systematic biology* 66:232–255.

Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences* 112:5773–5778.

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. 2017. Ctenophore relationships and their placement as the sister group to all other animals. *Nature ecology & evolution* 1:1737.

Whelan S, Goldman N. 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution* 18:691–699. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a003851

Yu G, Lam TT-Y, Zhu H, Guan Y. 2018. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular biology and evolution* 35:3041–3043.