# Rooting the animal tree of life

Yuanning Li1,3, Xingxing Shen3, Benjamin Evans2, Antonis Rokas3, and Casey W. Dunn1\*

1Department of Ecology and Evolutionary Biology, Yale University

2Yale Center for Research Computing, Yale University

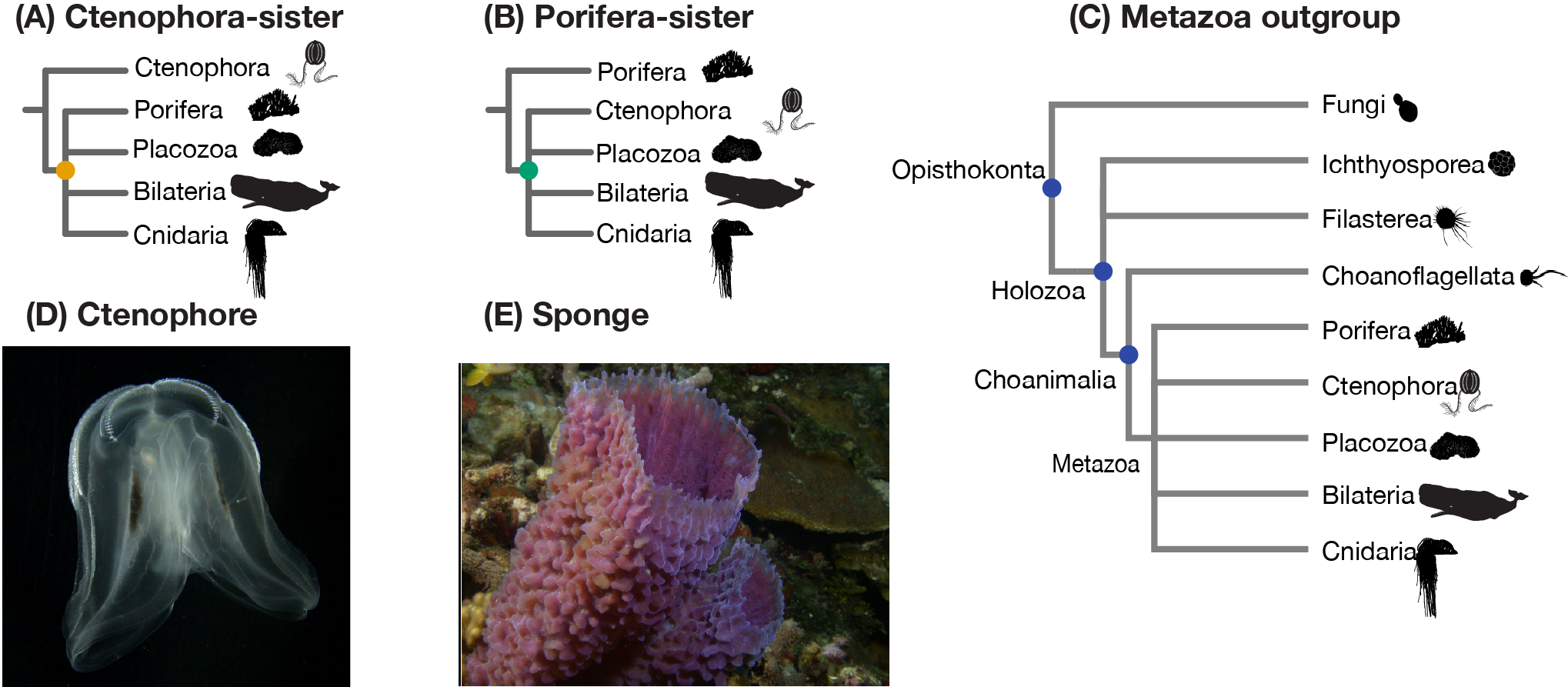
3Department of Biological Sceince, Vanderbilt University \* Corresponding author, [casey.dunn@yale.edu](mailto:casey.dunn@yale.edu)

## Abstract

## Introduction

Phylogenomic analyses with genome-scale data have the potential to revolutionize our understanding of the tree of life. However, over the past decade there has been considerable debate about the position of the root of the animal phylogeny, with Ctenophora-sister (comb jelly) and Porifera-sister (sponges) (Fig XXOverview) emerging as the two primary hypotheses. Historically, there was little debate about the root of the animal tree of life and Porifera-sister was widely accepted though rarely tested since . In contrast to the lack of debate about the position of Porifera, there has long been uncertainty about the relationship of Ctenophora to other animals [1]. An understanding of the first diverging animal lineage is critical for scientists to reconstruct events that occurred early in animal evolution, especially for complex structures such as cell types, nervous systems, and developmental patterns. However, a wide consensus of the position of animal has not been reached despite multiple efforts.

The first phylogenomic study to include ctenophores [2] suggested a new hypothesis, now referred to as Ctenophora-sister, that ctenophores are our most distant animal relative. Since then many more studies have been published, some supporting Ctenophora-sister, some Porifera-sister, and some neither. As it has become clear that this is a very difficult phylogenetic challenge, and the problem has become better characterized, it has become an interesting test-case to phylogenetic biologists beyond those concerned with this particular biological problem. Work has been hindered, though, because it has been difficult to directly compare results across studies and synthesize findings to understand the broader patterns of support. Here we synthesize data and results from all previous phylogenomic analyses that tested Ctenophora-sister and Porifera-sister, and reanalyze these data using standardized methods, and perform new analyses to characterize differences between studies. We hope that this effort provides an integrative overview of the challenge and provides direction for future studies. We also hope that the work we have conducted here, including consolidating all the datasets in one place with consistent formats and species names, will enhance the technical value of this interesting question to methods-focused investigators that look to develop methods to address difficult phylogenetic problems.

 **Fig XXOverview.** (A) The Ctenophora-sister hypothesis posits that there is a clade (designated by the orange node) that includes all animals except Ctenophora, and that Ctenophora is sister to this clade. (B) The Porifera-sister hypothesis posits that there is a clade (designated by the green node) that includes all animals except Porifera, and that Porifera is sister to this clade. Testing these hypotheses requires evaluating the support for each of these alternative nodes. (C) Outgroups of animals. (D) A ctenophore. (E) A sponge.

## Variation across studies

### Models of molecular evolution

Models of molecular evolution have several components that each consider different aspects of the evolutionary process. The models that have been used to model protein evolution in studies of the animal root have largely differed according to three components: the exchangeability matrix , the rate of evolution, and the state equilibrium frequencies .

The exchangeability matrix describes the rate at which one amino acid changes to another. Exchangeability matrices have been used in the studies under consideration here include:

* F81 [3] 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 [4] 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 [5], 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 long 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 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 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 [6]. 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 [**???**]. 10 to 60-profile mixture models as variants of the CAT model under the maximum-likelihood framework.
* Six-state amino acid recoding. Amino acids are recoded into six groups based on function to account for both compositional heterogeneity and substitution saturation. Several recoding strategies have been proposed, including Dayhoff 6-state recoding, S&R 6-state recoding, KGB 6-state recoding based on exchangeability matrix of models.

Models can be assembled by selecting different options for all these different components. The models that are applied in practice area heavily influenced by engineering and computational costs, as well as convention. For example, on the questions considered here F81 and GTR exchangeability matrices have only been used in combination with CAT site heterogeneous models of equilibrium frequency. LG and WAG exchangeability matrices have only been used with site homogeneous estimates of equilibrium frequency. This is further confused by the abbreviations that are used for models. Papers often discuss CAT and WAG models as if they are exclusive, but these particular terms apply to non-exclusive model components– CAT refers to variation across sites and WAG a particular exchangeability matrix. CAT is generally shorthand for F81+CAT and WAG is shorthand for WAG+homogeneous equilibrium frequency estimation. One could, though, run a WAG+CAT model.

To avoid confusion on this point, we always specify the exchangeability matrix first, followed by modifiers that describe accommodation of heterogeneity in equilibrium frequencies (*e.g.*, CAT) or rate (*e.g.*, Gamma). If there are no modifiers, then it is implied that site homogeneous models are used.

### Gene sampling

High-throughput sequencing is leading to exciting new data matrices for phylogenomic analyses, potentially including hundreds or thousands of genes. Although the number of sampled genes might be different across studies, it is reasonable to expect that the sampled genes should be conserved across animal genomes and most data matrices should share a core set of orthology genes (OGs) (*e.g.* BUSCO genes). Surprisingly, gene composition and matrice overlap across different studies has been rarely assessed, since most studies relied on their own bioinformatic pipeline for determination of OGs.

### Outgroup and ingroup taxon sampling

Outgroup selection (the taxa used to root the ingroup taxa) can potentially affect phylogenetic inference of the root position of animal phylogeny. Several studies have progressively removed more distantly related outgroup taxa to test the effect of outgroup selection to ingroup topology [**???**,**??**]. Three progressively more inclusive clades have often been investigated, including Choanomalia (Metazoa plus most closely related Choanoflagellata), Holoza (Choanimalia plus more distantly related holozoans) and Opisthokonta (Holozoa + Fungi). It has been suggested the inclusion of more distant related (e.g. Opisthokonta) outgroup taxa tend to recover Ctenophora-sister using both site-homogeneous and site-heterogeneous models, whereas high support of Porifera-sister often recovered when site-heterogeneous models were used. In contrast of outgroup selection, sensitivity to ingroup sampling has received less attention than sensitivity to outgroup sampling. This may be because results have tended to be more sensitive to outgroup sampling.

**Fig XXOutgroup.** The animals and their outgroups, showing the three progressively more inclusive clades Choanimalia, Holozoa, and Opisthokonta.

### Data-recoding

Compositional heterogeneity (lineage-specific differences in amino acid frequencies) was long thought has a major effect in resolving the animal tree of life, especially for the position of animal root. Feuda et al. 2017 [7] has proposed to recode the full set of twenty amino acids into six groups of amino aicds that have more frequent evolutionary changes within groups than between groups based on different exchangeability matrices. By discarding information on which specific amino acid is found at each site in each species and instead focusing on which groups those amino acids belong to, they sought to reduce the impact of differences between species in which particular amino acids within those groups are most frequent. By reanalyzing Chang et al. 2015 [8] and Whelan et al. 2015 [9], they found that Porifera-sister was strongly favored under all recoding strategies using GTR-CAT models [7]. However, the performance of the recoding methods has never been tested empirically against non-recoding method, and a recent simulation study has suggested that non-recoding approaches significantly outperformed 6-state recoding strategies [10]. Thus, it is important to evaluate the performance of data recoding method verses non-recoding approaches in empritical animal-root matrices .

## Overview of published analyses

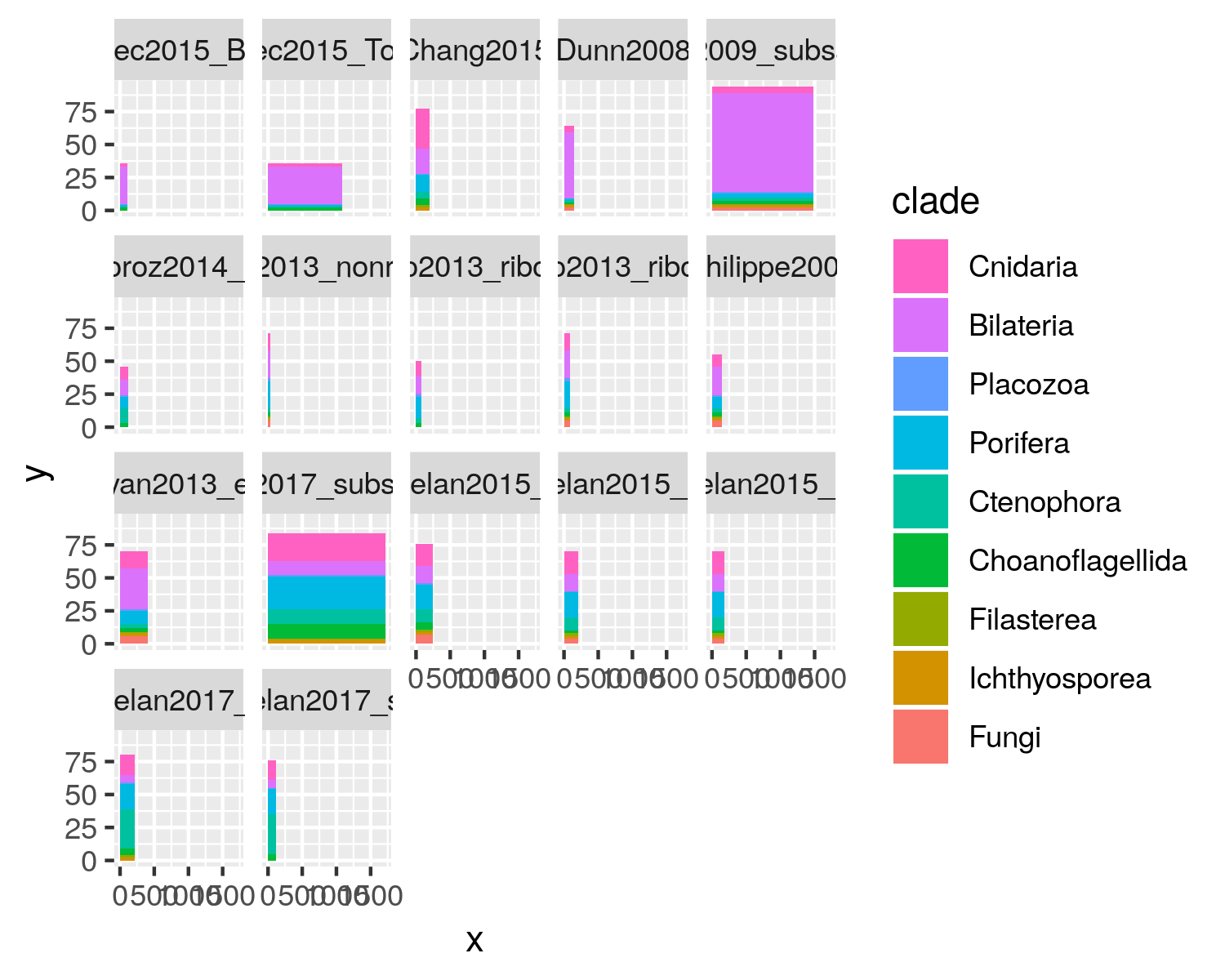
Here, we synthesize data and results from all previous phylogenomic analyses that tested Ctenophora-sister and Porifera-sister, including 17 data matrices from 12 studies (Table 1).

Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| paths | manuscript | matrix | construction\_method | notes |
| data\_raw/Dunn2008\_Dunn2008.nex | Dunn2008 | Dunn2008 | NA | NA |
| data\_raw/Philippe2009\_Philippe2009.nex | Philippe2009 | Philippe2009 | NA | NA |
| data\_raw/Hejnol2009\_Hejnol2009.nex | Hejnol2009 | Hejnol2009 | NA | NA |
| data\_raw/Nosenko2013\_nonribosomal\_9187\_smatrix.nex | Nosenko2013 | Nosenko2013\_nonribo\_9187 | NA | NA |
| data\_raw/Nosenko2013\_ribosomal\_11057\_smatrix.nex | Nosenko2013 | Nosenko2013\_ribo\_11057 | NA | NA |
| data\_raw/Nosenko2013\_ribosomal\_14615\_smatrix.nex | Nosenko2013 | Nosenko2013\_ribo\_14615 | NA | NA |
| data\_raw/Ryan2013\_est.opisthokonta.nex | Ryan2013 | Ryan2013\_est | NA | NA |
| data\_raw/Ryan2013\_genome.opisthokonta.nex | Ryan2013 | Ryan2013\_genome | NA | NA |
| data\_raw/Moroz2014\_ED3d.nex | Moroz2014 | Moroz2014\_3d | NA | NA |
| data\_raw/Whelan2015\_Dataset1\_FullData.nex | Whelan2015 | Whelan2015\_D1 | NA | NA |
| NA | Whelan2015 | Whelan2015\_D6 | NA | NA |
| data\_raw/Whelan2015\_Dataset10\_CertainPruned\_LBAtaxa\_LBAandHeteroGenesPruned.nex | Whelan2015 | Whelan2015\_D10 | NA | NA |
| data\_raw/Whelan2015\_Datset20\_Uncertain\_LBAgenesHeteroRemoved\_LBAtaxa.nex | Whelan2015 | Whelan2015\_D20 | NA | NA |
| NA | Whelan2015 | Whelan2015\_Philippe2009\_no\_ribo | NA | Ribosomal genes were removed |
| data\_raw/Borowiec2015\_Best108.nex | Borowiec2015 | Borowiec2015\_Best108 | NA | NA |
| data\_raw/Borowiec2015\_Total1080.nex | Borowiec2015 | Borowiec2015\_Total1080 | NA | NA |
| data\_raw/Chang2015\_Chang2015.nex | Chang2015 | Chang2015 | NA | NA |
| NA | Pisani2015 | Pisani2015\_Ryan2013\_est\_no\_Xenoturbella | NA | Xenoturbella was removed |
| NA | Pisani2015 | Pisani2015\_Moroz2014\_3d\_no\_Xenoturbella | NA | Xenoturbella was removed |
| data\_raw/Simion2017\_supermatrix\_97sp\_401632pos\_1719genes.nex | Simion2017 | Simion2017 | NA | NA |
| data\_raw/Whelan2017\_Metazoa\_full.nex | Whelan2017b | Whelan2017b\_full | NA | NA |
| data\_raw/Whelan2017\_Metazoa\_Choano\_RCFV\_strict.nex | Whelan2017b | Whelan2017b\_strict | NA | NA |
| NA | NA | NA | NA | NA |

### Matrix taxon composition

Several data matrices have relatively skewed taxon sampling, with more than 50% taxa sampled in bilaterians compared to other major animal lineages (Fig XXTaxon\_composition). These skewed data matrices includes two studies constructed from ESTs (Dunn2008 [2] and Henjol2009 [11]) and one study from whole genome data (Borrowiec2015 [12] ). The rest data matrices have relative balanced taxon composition, although the total number of sampled taxon varies a lot. In general, increasing taxon sampling is broadly accepted to improve phylogenetic inference and aid in the placement of animal root position, it should be noted that Whelan2017 [13] and Simion2017 [14] have much broader taxon sampling than other data matrices: Simion2017 has the most taxon sampling, and Whelan2017 contains greater ctenophore taxon sampling.



**Fig XXTaxon\_composition.** Each of the primary matrices considered here, color coded by taxon sampling. Horizontal size is proportional to the number of genes (XXOr should it be sites?) sampled, vertical size to the number of taxa sampled.

### Matrix gene composition and overlap

Gene sampling plays an essential role in phylogenomic analyses, and most often rely on the concatenation of sampled genes to tackle the incongruence of the topology among individual gene trees. The basic idea is that the concatenation assumes that phylogenetic signal from genes that do not share the species phylogeny will be overwhelmed by the signal from the majority of genes whose genealogy mirrors that of the species evolutionary history [15]. However, large phylogenomic data matrices potentially contain genes with conflicting or weak phylogenetic signal that can lead to incongruence phylogenomic results, especially when analyzing deeper node represents ancient divergence in the tree [16]. Thus, it is important to evaluate the gene composition and matrix overlap across difference data matrices to explore.

We first used the Metazoa BUSCO dataset to evaluate gene composition from all data matrices and the result is unexpected. The animal BUSCO dataset contains 978 single-copy OGs (identified from 64 animal taxa) that are most likely shared by most animal genomes, and are commonly used to assess the completeness of genome or transcriptomic assemblies. The BUSCO genes also are frequently used as phylogenomic markers in a varietly of studies. Thus, it is reasonable to expect that these conserved genes should present at a higher percentage than non-BUSCO genes in most matrices given their nature of conserveness and high-expression properties. However, to our surprise, we found that gene partitions from all matrices comprised less than 50% of BUSCO genes, even with several data matrices contain less than 20 percent.

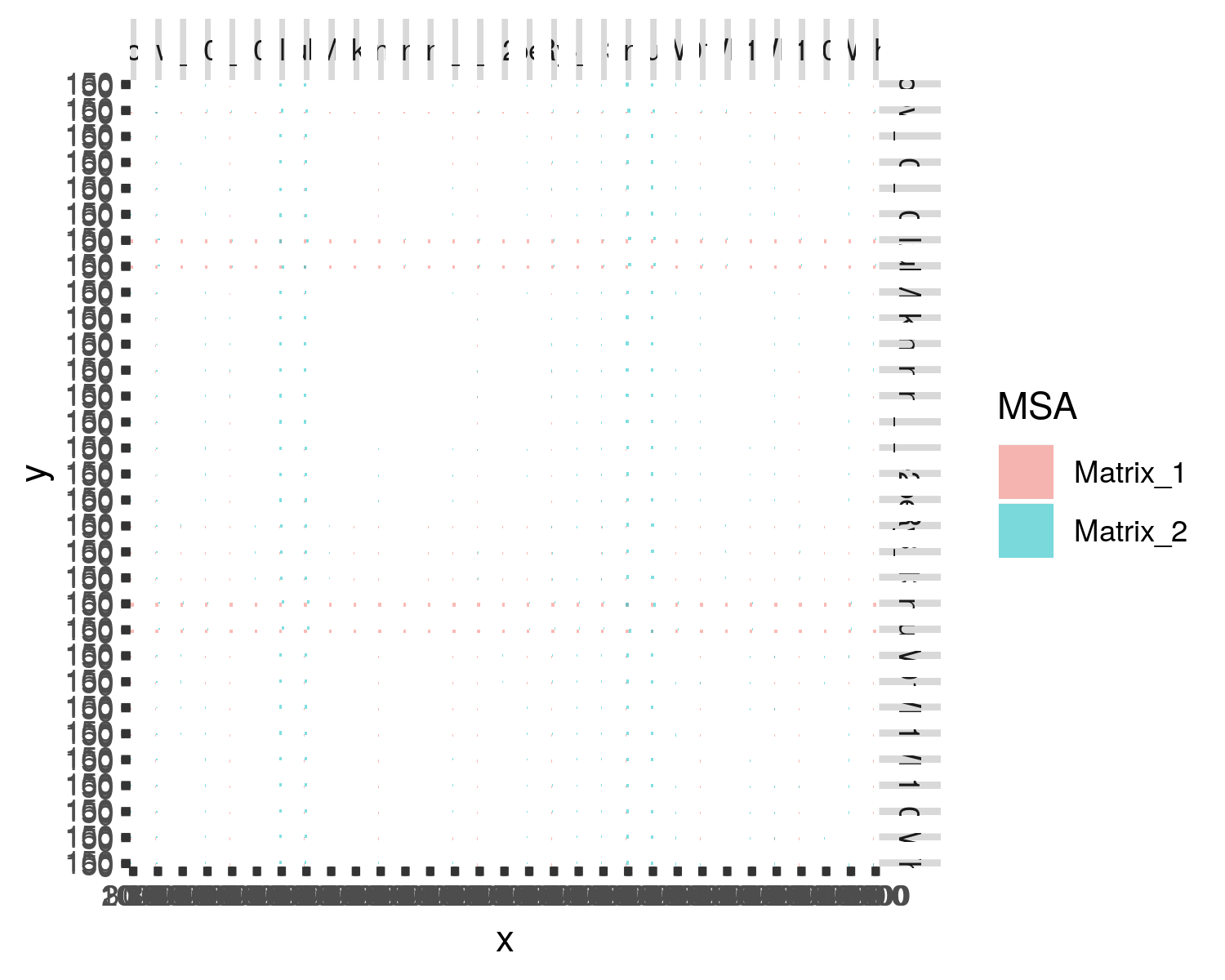
We furthered BLAST gene partitions to SwissProt database

Previous reanalyses have mainly focused on ourgroup selection and/or model fits in different data matrices, and the matrix overlap across different studies has been rarely investigated. Our initial intent of comparing matrix overlap across matrices was to constructed a new matrice with shared taxon and gene sampling to test specific hypotheses about differences in support. The hypothesis is that although the bioinformatics pipeline used for OG determination is different, most larger data matrices should share a core set of OGs. However, our results show that surprisingly little overlap across all the data matrices from different studies (Fig XXAlignment overlap). These results indicated that

## # A tibble: 11 x 3  
## # Groups: matrix [6]  
## matrix partition n  
## <chr> <chr> <int>  
## 1 Borowiec2015\_Total1080 OG621.fasta.aln\_gene687 2  
## 2 Chang2015 l12e\_01 2  
## 3 Philippe2009 rpl2 2  
## 4 Simion2017 V2META12905-42-Calc\_Hmm10-BMGE05 2  
## 5 Simion2017 V2META14417-42-Calc\_Hmm10-BMGE05 2  
## 6 Simion2017 V2META14523-42-Calc\_Hmm10-BMGE05 2  
## 7 Whelan2017\_full Subset18\_01 2  
## 8 Whelan2017\_strict Subset19\_01 2  
## 9 Whelan2017\_strict Subset4\_01 3  
## 10 Whelan2017\_strict Subset5\_01 2  
## 11 Whelan2017\_strict Subset9\_01 2

**Fig XXBUSCO\_annotations.** The number of partitions with BUSCO annotations in each matrix, relative to the number of partitions.

**Fig XXGene\_composition.** Each of the primary matrices considered here, color coded by the types of genes sampled (XX Ribosomal proteins, etc). Horizontal size is proportional to the number of genes sampled, vertical size to the number of taxa sampled.

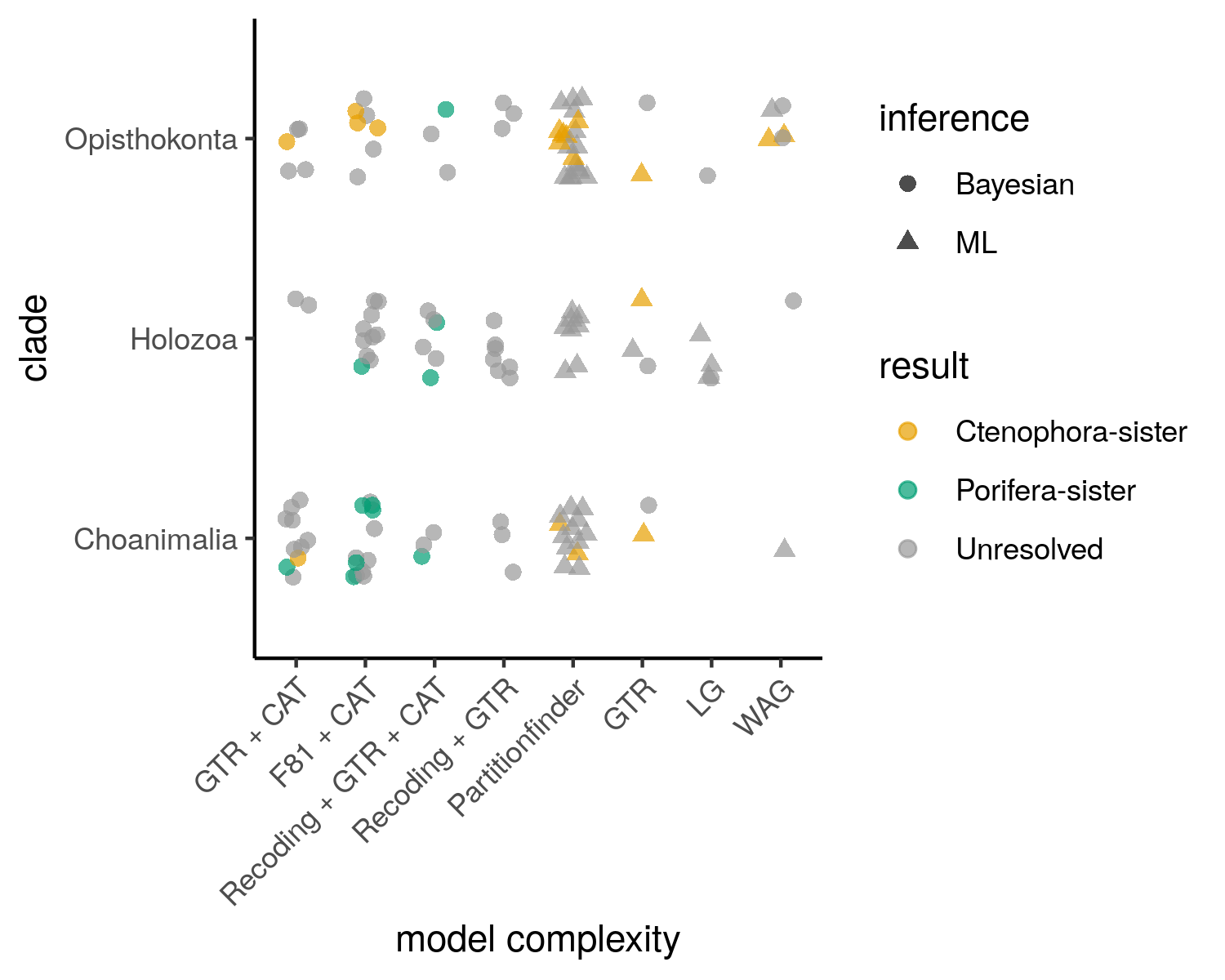


**Fig XXAlignment overlap.** Pairwise overlap between each of the primary matrices considered here. Horizontal size is proportional to the number of genes sampled, vertical size to the number of taxa sampled. The horizontal intersection shows the proportions of shared genes, the vertical intersection shows the proportions of shared taxa.

### Support for Porifera-sister and Ctenophora-sister

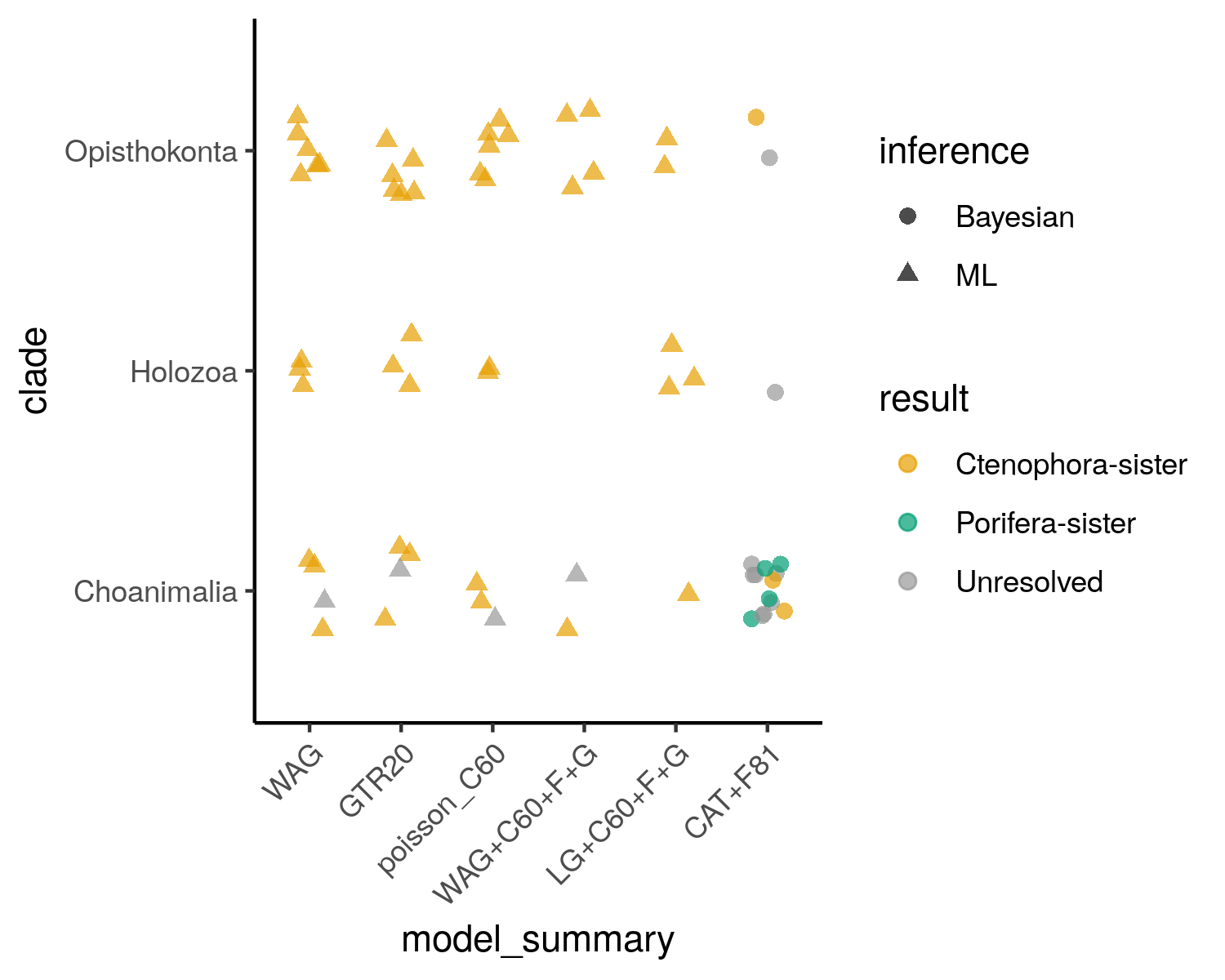
Here we summarized all the 167 previous phylogenetic analyses from all studies (Fig. support\_published\_analyses and Table perviouly\_published\_analyses). We found that all past phylogenomic analyses strongly support of Ctenophora-sister once site-homogeneous models were used, no matter which outgroup choice, phylogenetic inference, recoding or nonrecoding methods are used. We found that support for Porifera-sister increases with the exclusion of more distantly related outgroups and the use of site-heterogeneious CAT models. It should be noted that the convergence has not been reached or posterior probability is not well supported in many BI analyses. Moreover, the strong support of Porifera-sister are mainly from analyses using Poisson + CAT and recoding methods with GTR + CAT site-heterogeneous models. It has been suggested that GTR+CAT model has much higher model fit and better performance than Poisson+CAT model on data matrices with high substitution heterogeneity that may susceptible to LBA. However, GTR+CAT model is extremely computational demanding and not feasible in large empritical matrices. Only a few analyses with the most complex GTR + CAT model are converged and have mixed support for Porifera-sister or Ctenophora-sister. The main reason the GTR + CAT model has only been employed in a few analyses is because it is computational untractable and can only be used in smaller or reduced data matrices. Interestingly, a CAT-like, site-heterogeneous C10 - C60 (C) models implemented in the maximum likelihood (ML) framework has never been tested and compared to other models.

The use of site-heterogeneous CAT is an infinite mixture model has long been suggested as the best approach to account for systematic errors, particular for long-branch attraction and saturation. The CAT model often received the best-model fit from cross validation or posterior predictive analyses compared to site-homogeneous models with or without data partitioning. In contrast, a recent study has suggested that Poisson + CAT model did not outperform site-homogeneous with data partitioning or GTR + CAT models and in some cases can infer incorrect topology based on the simulation analyses. The author also raised the issue of the potential overparameterization in CAT model in some cases. Nevertheless，the debate over the site-heterogeneous vs. site-homogeneous models and/or inclusion or exclusion more distantly related outgroups, constitutes the probably one of the most controversial node in the animal tree of life.



**Fig support\_published\_analyses.** A total of 131 analyses were transcribed from the literature (Table perviouly\_published\_analyses.tsv).

## New analyses of published matrices



One of the challenges of interpreting support for the placement of the animal root across studies is that different programs, software versions, and settings have been used across studies, and phylogenetic analysis decisions have been approached in very different ways. Here we reanalyze the primary matrices from each study under consistent conditions with IQtree under a panel of evolutionary models. We selected this tool because it has greater model flexibility, computational time and accuracy than other tools in maximum likelihood (ML) framework.

We first tested a variety of models for each matrix and compared the relative fit of site-homogeneous and site-heterogeneous C models using ModelFinder in IQtree. We found in all cases, C60 + WAG/LG +F +G models fit these matrices better than the site-homogeneous models in under BIC criteria and inferred support under the best-fit model from IQtree. We didn’t include GTR + C models since it needs to estimate over 10,000 free parameters in ModelFinder in some large datasets compared to several hundred in site-homogeneous + C models. then analyzed every matrix under a panel of standard site-heterogeneous and site-homogeneous models, including WAG+G, GTR+G and Poisson+C60+F+G. Moreover, we also used the the best-fit model identified above with the removal of all distantly related outgroups and only include Choanozoa as outgroup.

With the exception of Moroz2014 matrix supported Placozoa-sister, every analysis conducted herein with different models or outgroup choice strongly supported the ctenophore-sister hypothesis under maximum likelihood framework in IQtree.

The site-heterogeneous CAT model has been introduced in PhyloBayes and it is a non-parametric modeling based on a Dirichlet process mixture, which is often used in Bayesian studies (Lartillot and Philippe, 2004). It was designed to better accommodate data saturation or tackle long-branch attraction for deep phylogenetic question but only used in Bayesian inference (BI). In contrast, C models is 10 to 60-profile mixture models as variants of the CAT model can be used in both ML or BI framework. Thus, it is interesting to compare results between C model to CAT models since it has never been used in any matrices analyzed here.

|  |  |  |  |
| --- | --- | --- | --- |
| matrix | clade | result | model\_summary |
| Borowiec2015\_Best108 | Choanimalia | Ctenophora-sister | WAG+C60+F+G |
| Chang2015 | Holozoa | Ctenophora-sister | LG+C60+F+G |
| Dunn2008 | Opisthokonta | Ctenophora-sister | WAG+C60+F+G |
| Moroz2014\_3d | Choanimalia | Unresolved | WAG+C60+F+G |
| Ryan2013\_est | Opisthokonta | Ctenophora-sister | WAG+C60+F+G |
| Whelan2015\_D1 | Opisthokonta | Ctenophora-sister | LG+C60+F+G |
| Whelan2015\_D10 | Opisthokonta | Ctenophora-sister | WAG+C60+F+G |
| Whelan2015\_D20 | Opisthokonta | Ctenophora-sister | WAG+C60+F+G |
| Whelan2017\_full | Holozoa | Ctenophora-sister | LG+C60+F+G |
| Whelan2017\_strict | Choanimalia | Ctenophora-sister | LG+C60+F+G |

Table XXModelfinder. The models selected by modelfinder for each matrix.

### Comparison of iqtree and phylobayes results

Site heterogeneity in equilibrium frequency has been a major concern in tests of Ctenophora-sister and Porifera-sister. This has been addressed with CAT models. iqtree provides a new family of C models that also address site heterogeneity. Given the extensive computational cost and concerns about overparameterization of CAT models, we compared iqtree C results to CAT results for a subset of matrices to see if they give consistent results. This would be of technical interest because it would reduce the cost of accommodating compositional heterogeneity in future analyses. For computational efficiency, we first only applied Poisson+CAT models in every matrice with only Choanoflagellata as outgroup. It should be noted that most Phylobayes runs were converged, although several large matrices have not reached convergence after at least a month’s computational time.

Interestingly, we found a mixed support when using Poisson-CAT model in Choanimalia matrices. Consistent with previous study, we recovered Porifera-sister in Philippe2008, Ryan2013, Nosenko2013 matrices with only. Moreover, we also found a strong support of Porifera-sister in that Whelan2017 matrices. However, with the exception of Moroz2013, we found strong support for Ctenophora-sister in all other matrices.

## The current state of understanding

### Interpretting variation support

External criteria, eg posterior predictive scores, model fit etc

## Next steps

## Conclusion

## Methods

All files associated with this analysis are available at <https://github.com/caseywdunn/animal_root>

### Data selection and wrangling

We retreived matrices from each publication (Table XX), storing the raw data in this manuscript’s version control repository. We manually edited some minor formatting to make the batch processing of the matrices *en masse*, e.g. standardizing the formatting of charset blocks. All changes made are tracked with git.

### Matrix comparison and annotation

#### Taxon name reconciliation

We programatically queried the NCBI Taxonomy database to standardize names of samples in each matrix. We also use a table where manual entries were needed (Supp Table XX, reconciliation/taxonomy\_info/manual\_taxonomy\_map.tsv), e.g. authors of original matrix indicate species name in original manuscript. For a table summarizing all samples and their new or lengthened names, see Table XX(reconciliation/taxonomy\_info/taxon\_table.tsv).

#### Sequence Comparisons

Using the partition files for each matrix, we isolated each sequence for each taxon from each partition. Because many of the matrices had been processed by the original authors to remove columns that are poorly sampled or highly variable, these matrix-derived sequences can have deletions relative to the actual gene sequences.

We used DIAMOND [17] to compare each sequence to all others using default diamond blastp parameters. We further filtered DIAMOND results such that we retained hits for 90% of partitions (pident > 78.0, eValue < 1e-15, no self->self). We ran BUSCO with default parameters for all sequences against the provided metazoa gene set. We also ran a BLAST+ blastp search against the SwissProt [cite] database, filtering such that we retain at least one hit for ~97% of partitions (pident > 50.0, eValue < 1e-15).

#### Partition Network

We used the sequence similarity comparisons described above to compare partitions.

We constructed a network with Python and NetworkX [18] v2.2 where each node is a partition and each edge represents a DIAMOND sequence-to-sequence match between sequences in the partitions. We extracted each connected component from this network. We further split these components if the the most connected node (i.e. most edges) had two times more the standard deviation from the mean number of edges in the component it is a member of and if removing that node splits the component into two or more components. We then decorated every node in the partition network with the most often found SwissProt BLAST+ result and BUSCO results to see which components contain which classes and families of genes. See Table XX [partition\_network\_summary table] for a summary tally of each part of the comparison.

### Phylogenetic analyses

#### Full data matrices

To investigate the phylogenetic hypotheses and distribution of phylogenetic signal in studies aiming to elucidate the root of animal position, we considered 16 data matrices from recent studies that were constructed from transcriptomic or genomic data (Table). Because different choices of substitution models could largely influence phylogenetic inference of the placement of the root of animal (e.g. site heterogeneous CAT model and site homogeneous model), we first investigated four different sets of substitution models in IQtree [19], including C10 to C60 profile mixture models as variants of the CAT model in ML framework. Phylogenomic analyses of all the datasets were first conducted under WAG+G, GTR+G, associated best-fit substitution model (C10-C60 model were included for model comparison via -madd option) with ModelFinder [20] and Poisson + C60 models using IQtree v [19]. Nodal support was assessed with 1000 ultrafast bootstrap replicates for each analysis.

#### Outgroup taxa sampling with C10-C60 and CAT models

Because different choices of outgroups could also affect phylogenetic inference as suggested in previous analyses, we parsed the full data matrices into three different types of outgroups: Choanimalia , Holozoa and Opisthokonta. These datasets include the same set of genes but differ in the composition of outgroup species. Choanozoa only includes choanofagellates as outgroup; Holozoa also includes more distantly related holozoans; Opistokonta also includes Fungi. For each Choanozoa data matrice, both C10-C60 models in ML and CAT models in PhyloBayes were conducted. The maximum likelihood analysis was performed using the best-fit substitution model identified as above and nodal support was assessed with 1000 ultrafast bootstrap replicates using IQtree. Moreover, bayesian inference with the site-heterogeneous CAT-Poisson substitution model was done with PhyloBayes MPI. To minimize computational burden, CAT-GTR models were not in this study.

For results of choanozoa matrices indicated a strong support that sponges are the sister group to the remaining Metazoa using CAT model, bayesian inference with CAT-Poisson model was also conducted to Holozoa and Opisthokonta data matrices with the same settings as above. For CAT-poisson analyses on all the data matrices, two independent chains were sampled every generation. Tracer plots of MCMC runs were visually inspected in Tracer v1.6 to assess stationarity and appropriate burn-in. Chains were considered to have reached convergence when the maxdiff statistic among chains was below 0.3 (as measured by bpcomp) and effective sample size > 50 for each parameter (as measured by tracecomp). A 50% majority‐rule consensus tree was computed with bpcomp, and nodal support was estimated by posterior probability. All analyses in PhyloBayes were run for at least one month computational time.

#### phylogenetic signal in conflicted dataset

To investigate the distribution of phylogenetic signal in data matrices, we considered four data matrices from four studies that had different topology between ML and BI using CAT model in our reanalysis, including Philippe2008, Ryan2013, and Whelan\_2017 data matrices.We examined two hypotheses: Ctenophora-sister and Porifera-sister to rest of metazoans, under both ML and BI frameworks with different outgroup schemes (Choanozoa and Opisthokonta). For ML analysis in each dataset, site-wise likelihood scores ere inferred for both hypotheses using IQtree (option -g) with the same best-fit model identified above. The two different phylogenetic hypotheses passed to IQtree (via -z) were the corresponding tree that the ctenophore as the sister lineage tree and the corresponding tree that was modified to have sponges as the sister to all other metazoans. The constraint trees were conducted by R package. The numbers of genes and sites supporting each hypothesis were calculated with RAxML output and Perl scripts from Shen et al. For BI analysis, we only considered the Whelan\_2017 dataset.

## Ackowledgements

We thank the Yale Center for Research Computing for use of the research computing infrastructure, specificaly the Farnam cluster.

## Author contributions

# Supplemental Information

## Details of published analyses

### Dunn *et al.* 2008

Dunn *et al.* [2] 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 [21], 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 a small number of taxa were unresolved, all genes from those taxa were excluded. Genes were then realigned and sites were filtered with Gblocks [22], 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 [23], as implemented in phyutility [24], 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 F81+CAT model in phylobayes [6], and under the WAG model in MrBayes [25] and RAxML [26].

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 *Trichoplax*.

### Philippe *et al.* 2009

Philippe *et al.* 2009 [Philippe:2009hh] assembled an EST dataset for 55 species with 128 genes to explore phylogenetic relationship of early diverging animals. The phylogenetic analysis using Poisson-CAT model strongly supported Porifera-sister, and ctenophores was sister to cnidarians to form the “coelenterate” clade. 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.* 2009 added EST sequences from seven taxa, and a total of 94 taxa were included in the final data matrix. The orthology inference was largely simialr to Dunn et al. 2008. Two datsets were constructed, one with 844 genes and the other with 330 genes. Maximum likelihood analyses from both datasets strongly supported Ctenophora-sister.

### Pick *et al.* 2010

Pick *et al.* [27] 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.* [2]. 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: Choanoflagellata (resulting in Choanimalia) and the same sampling as Dunn *et al.* (resulting in Opisthokonta).

All their analyses were conducted under the F81+CAT+Gamma model in phylobayes [6], 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 Choanimalia 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 conclude that Dunn *et al.* [2] is more saturated than Philippe *et al.* 2009 [Philippe:2009hh]. Note that the Pick *et al.* [27] dataset is not reanalyzed here because partition data are not available, and due to site filtering the partition file from Dunn *et al.* [2] cannot be applied to this matrix.

### Ryan *et al.* 2013

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 model and Poisson-CAT model with three sets of outgroup sampling to evaluate the effect of outgroup selection to the ingroup topology. The results strongly supported Ctenophora-sister in all datasets they analysed using 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.

### Moroz *et al.* 2014

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.

### Whelan et al. 2015

Wehlan *et al.* 2015 constructed a new phylogenomic dataset by eight new transcripomic 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 (~ 1000 orthologs) and subsequetly 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 progressively removal of outgroups. All the maximum likelihood analyses with site-homogeneous model and PartitionFinder strongly suggested Ctenophora-sister. CAT-GTR models only used in least saturated dataset 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.

### Chang et al. 2015

Chang *et al.* 2015 was originally used to explore 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 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.* 2015 reanalyzed representative datasets that supported Ctenophora-sister, including Ryan2013, Moroz2014 and Whelan2015 datasets. It was the first study showing that progressively removal of more distantly related outgroups could largely affect phylogenomic inference of animal-root position. The authors suggested that the inclusion of outgroups very distant from the ingroup can cause systematic errors due to long-branch attraction. They found strong support of porifera-sister when only include Choanozoa as outgroup in these datasets using site-heterogeous CAT model. 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.* 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 study phylogenomic analysis using recoding strategy to explore animal-root position. The authors compared model adequacy using posterior predictive analyses from a set of site-homogeneous (WAG, LG, GTR, PartitionFinder)and site-heterogeneous (CAT-GTR) models with non-recoding and recoding datasets. The results showed that data-recoding can significant reduce compositional heterogeneity in both datasets with CAT-GTR models and strongly supported Porifera-sister hypothesis. 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 Halalnych 2017a

Whelan *et al.* is the only study to evaluate performance of site-heterogeneous models and site-homogeneous model with data partitioning under the simulation framework. The simulation results suggested that 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 overestimated substitutional heterogeneity in many cases. They also reanalyzed datasets from Philippe 2009 and Nosenko 2013 using both CAT models and data partitioning with site-homogeneious 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 also 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.* 2017b

Whelan *et al.* 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, with the difference of a core Ctenophora core datasets were constructed here with more than 2000 genes. 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 CAT-GTR model in Choanoazoa dataset. 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

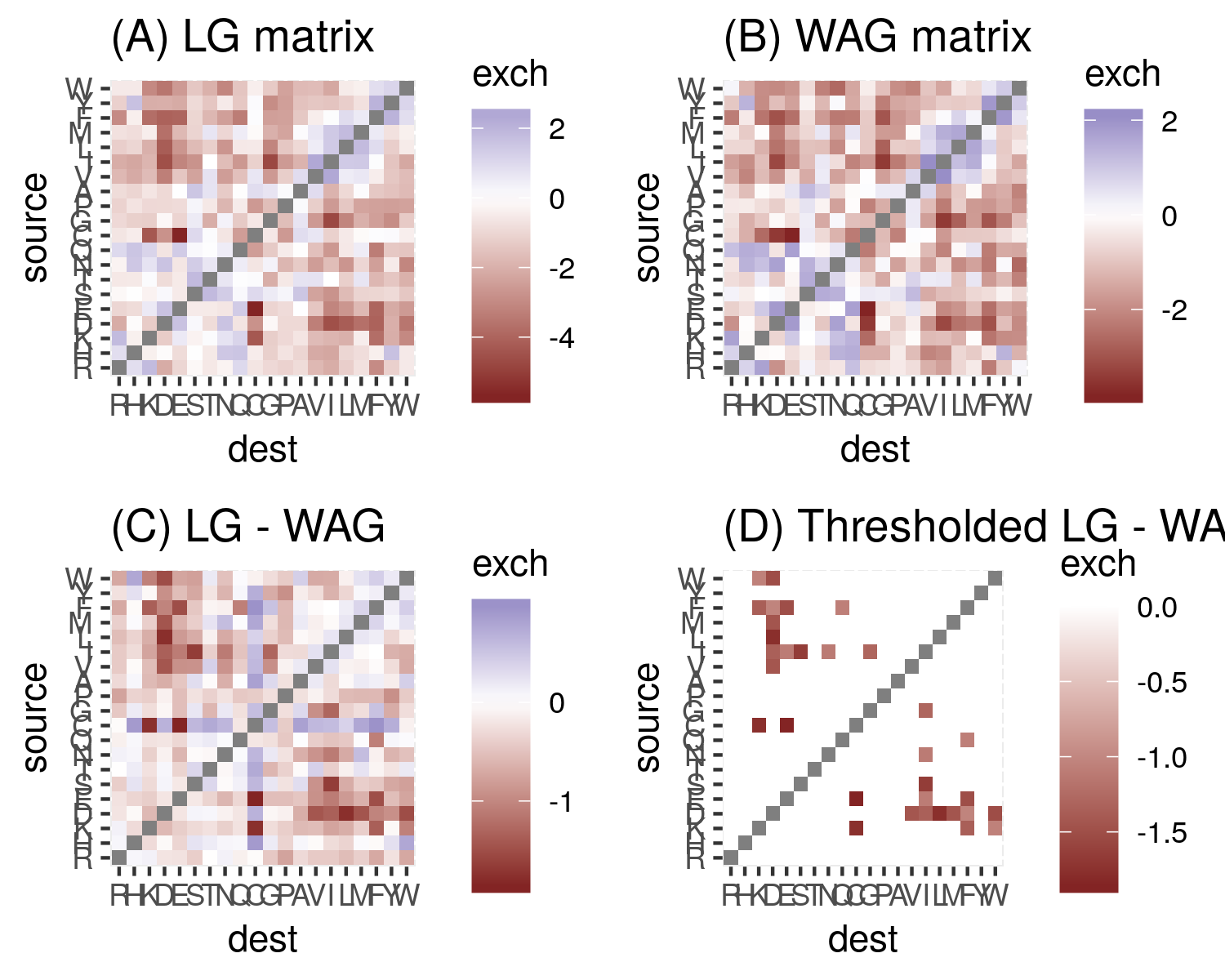
Simion *et al.* [2] added transcriptomic data for 21 new animals. The data matrix was constructed using a semi-automated approach to comprehensively detect and eliminate potential systematic errors. The resulting dataset comprises 1,719 genes and 97 species, including 61 non-bilaterian metazoan species. It was by far the largest phylogenomic dataset in terms of taxon and gene sampling related to animal-root question.

The supermatrix was first analyzed using the Poisson-CAT model. Different that 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 baed 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 the supermatrix dataset. Importantly, the authors compared the behavior of long-branch effect between site-heterogenious and site-homogeneous models by progressively removal 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 superalignments 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.

## Model matrix comparison

WAG and LG are both fixed exchange matrices. Their differences are largely limitted to a few amino acid changes.



## Matrix mapping

Taxa and partition correspondence across manuscripts was assessed by comparing all sequences for each taxon in each partition across all matrices with diamond blast. Based on inspection of sequence similarity, we excluded all comparisons with less than 99% identity and greater than 10-25 e-value.

### Taxa comparison across matrices

The primary intent of comparing taxa across matrices was to validate our taxon name reconciliation across studies.

We first considered pairwise similarity between the same species from different matrices in different studies.

### Partition comparison across matrices

The count for a partition pair can be much larger than the number of genes in the matrix, which suggests that the count is the number of hsps rather than the number of sequences with hits.

There are 17 matrices. A gene that is perfectly sampled would form a cluster with this size. To our surprise, very few clusters, though, are this size. This suggests that intersection of genes between matrices is extreme low. This finding suggested genes used for exploring the position of the root of the animal phylogeny

## References

1. Wallberg A, Thollesson M, Farris J, Jondelius U. The phylogenetic position of the comb jellies (Ctenophora) and the importance of taxonomic sampling. Cladistics. 2004;20: 558–578. Available: <http://onlinelibrary.wiley.com/doi/10.1111/j.1096-0031.2004.00041.x/full>

2. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 2008;452: 745–749. doi:[10.1038/nature06614](https://doi.org/10.1038/nature06614)

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

4. Whelan S, Goldman N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution. 2001;18: 691–699. doi:[10.1093/oxfordjournals.molbev.a003851](https://doi.org/10.1093/oxfordjournals.molbev.a003851)

5. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Molecular Biology and Evolution. 2008;25: 1307–1320. doi:[10.1093/molbev/msn067](https://doi.org/10.1093/molbev/msn067)

6. Lartillot N. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Molecular Biology and Evolution. 2004;21: 1095–1109. doi:[10.1093/molbev/msh112](https://doi.org/10.1093/molbev/msh112)

7. Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, et al. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Current Biology. Elsevier; 2017;27: 3864–3870.

8. Chang ES, Neuhof M, Rubinstein ND, Diamant A, Philippe H, Huchon D, et al. Genomic insights into the evolutionary origin of myxozoa within cnidaria. Proceedings of the National Academy of Sciences. National Acad Sciences; 2015;112: 14912–14917.

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

10. Hernandez AM, Ryan JF. Six-state amino acid recoding is not an effective strategy to offset the effects of compositional heterogeneity and saturation in phylogenetic analyses. BioRxiv. Cold Spring Harbor Laboratory; 2019; 729103.

11. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, et al. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proceedings of the Royal Society B: Biological Sciences. The Royal Society; 2009;276: 4261–4270.

12. Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the ctenophora as sister to remaining metazoa. BMC genomics. BioMed Central; 2015;16: 987.

13. Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, et al. Ctenophore relationships and their placement as the sister group to all other animals. Nature ecology & evolution. Nature Publishing Group; 2017;1: 1737.

14. Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, et al. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Current Biology. Elsevier; 2017;27: 958–967.

15. Lanier HC, Knowles LL. Is recombination a problem for species-tree analyses? Systematic Biology. Oxford University Press; 2012;61: 691–701.

16. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. Nature Publishing Group; 2013;497: 327.

17. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using diamond. Nature Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. SN -; 2014;12: 59 EP. Available: <https://doi.org/10.1038/nmeth.3176>

18. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using networkx. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th python in science conference. Pasadena, CA USA; 2008. pp. 11–15.

19. Nguyen L-T, Schmidt HA, Haeseler A von, Minh BQ. IQ-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. Oxford University Press; 2014;32: 268–274.

20. Kalyaanamoorthy S, Minh BQ, Wong TK, Haeseler A von, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature methods. Nature Publishing Group; 2017;14: 587.

21. Enright A, Van Dongen S, Ouzounis C. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. Oxford University Press; 2002;30: 1575–1584. doi:[10.1093/nar/30.7.1575](https://doi.org/10.1093/nar/30.7.1575)

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

23. Thorley J, Wilkinson M. Testing the phylogenetic stability of early tetrapods. Journal of Theoretical Biology. 1999;200: 343–344. doi:[10.1006/jtbi.1999.0999](https://doi.org/10.1006/jtbi.1999.0999)

24. Smith SA, Dunn CW. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics. Oxford University Press; 2008;24: 715–716. doi:[10.1093/bioinformatics/btm619](https://doi.org/10.1093/bioinformatics/btm619)

25. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19: 1572–1574. doi:[10.1093/bioinformatics/btg180](https://doi.org/10.1093/bioinformatics/btg180)

26. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22: 2688–2690. doi:[10.1093/bioinformatics/btl446](https://doi.org/10.1093/bioinformatics/btl446)

27. Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, et al. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Molecular Biology and Evolution. 2010;27: 1983–1987. doi:[10.1093/molbev/msq089](https://doi.org/10.1093/molbev/msq089)