 doi:10.20944/preprints202101.0390.v1

Article

The roles of protein structure, taxon sampling, and model complexity in phylogenomics: A case study focused on early animal divergences

Akanksha Pandey [1,2] and Edward L. Braun [1,*]

[1] Department of Biology, University of Florida, Gainesville, FL 32611, USA
[2] Current address: Kronos Bio, Inc., Cambridge, MA 02142, USA; aakanksha.vit@gmail.com
* Correspondence: ebraun68@ufl.edu

Abstract: Despite the long history of using protein sequences to infer the tree of life the potential for different parts of protein structures to retain historical signal remains unclear. We propose that it might be possible to improve analyses of phylogenomic datasets by incorporating information about protein structure; we test this idea using the position of the root of Metazoa (animals) as a model system. We examined the distribution of "strongly decisive" sites (alignment positions that support a specific tree topology) in a dataset comprising >1,500 proteins and almost 100 taxa. The proportion of each class of strongly decisive sites in different structural environments was very sensitive to the model used to analyze the data when a limited number of taxa were used but they were stable when taxa were added. As long as enough taxa were analyzed, sites in all structural environments supported the same topology (ctenophores sister to other animals) regardless of whether standard tree searches or decisive sites were used to select the optimal tree. However, the use of decisive sites revealed a difference between the support for minority topologies for sites in different structural environments; buried sites and sites in sheet and coil environments exhibited equal support for the minority topologies whereas solvent exposed and helix sites had unequal numbers of sites supporting the minority topologies. Given the plausible trees equal support for minority topologies is consistent with discordance among gene trees, making it possible the relatively slowly evolving buried (and sheet and coil) sites are giving an accurate picture of the true species tree as well as the amount of conflict among gene trees. Alternatively, the apparent support could reflect currently uncharacterized processes of molecular evolution. Regardless, it is clear that analyses of the deepest branches in the animal tree of life using sites in different structural environments are associated with a subtle data type effect that results in distinct phylogenetic signals.

Keywords: protein structure; relative solvent accessibility; secondary structure; phylogeny; models of sequence evolution; gene tree-species tree discordance; incomplete lineage sorting; Ctenophora; Porifera

Citation: Pandey, A.; Braun, E.L. The roles of protein structure, taxon sampling, and model complexity in phylogenomics: A case study focused on early animal divergences. *Biophysica* **2021**, *1*, submitted for peer review. https://doi.org/10.3390/xxxxx

1. Introduction

The relationship between protein structure and the patterns of sequence evolution has been a topic of interest since the very dawn of molecular evolution as a field [1,2] and, despite the very limited amount of data available at the time of those early studies, many of those early hypotheses have stood the test of time (reviewed by Alvarez-Ponce [3]). Our understanding of the factors that determines rates of amino acid change have certainly expanded since those pioneering studies [4,5] but two important constants have been the idea that the buried residues in globular proteins evolve more slowly [6–8] and are more hydrophobic than solvent exposed residues [8,9]. Likewise, relative amino acid exchangeabilities also differ among structural environments [7,10,11]. However, the relationship

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

between protein structure and phylogenetic signal is less clear. Pandey and Braun [12] found that the position of the metazoan root differed in estimates of phylogeny generated by analyzing the buried vs. solvent exposed sites in a large (231 protein) data matrix. This is an example of a "data-type effect" [13,14], which are cases where different subsets of the genome yield different trees with relatively high support. In the case of Pandey and Braun [12] the data types are sites encoding amino acids in distinct protein structural environments and analyses of those sites yield different topologies for the metazoan tree.

Data type effects can be explained in two different ways [13]. First, at least one data type has a poor fit to the model(s) used for phylogenetic analyses, with that poor fit resulting in an incorrect estimate of phylogeny due to model misspecification. Second, the data types might be associated with different underlying tree(s). If the data types are sites located in different structural environments but selected from many genes they would be expected to reflect the same tree (more accurately, the data types would be expected to reflect the same mixture of gene trees, since the underlying gene trees for different loci are likely to differ due to processes like incomplete lineage sorting [ILS]; see Pamilo and Nei [15] and Edwards [16] for details). Thus, we can exclude the second possibility for protein structure data types and focus on the issue of model fit. Pandey and Braun [12] used early metazoan phylogeny, a phylogenetic question that has received extensive study in the phylogenomic era [17–32], as a model system to explore the relationship between phylogenetic signal and protein structure. The signal examined in that study was the position of the metazoan root, for which there are three plausible positions. First, the root could lie between the sponges (Porifera) and all other extant animals (T1 in Figure 1); this is the "traditional" position of the root [33]. Although some phylogenomic studies have supported T1, many other studies have placed the metazoan root between ctenophores (Ctenophora) (T2 in Figure 1). The third possible tree for these taxa (T3 in Figure 1) is arguably the least plausible, although it also been recovered in some phylogenetic analyses [22,24,31] (recovery of T3 is usually limited to a subset of analyses in the phylogenomic studies). Pandey and Braun [12] found that analyses of solvent exposed sites support T2 and that most analyses of buried sites using "standard models" (e.g., the LG model [34] or the 20-state version of the Tavaré [35] general time reversible [GTR] model) yield T3. This result indicates that protein structure can have a direct influence on the results of phylogenetic analyses, at least for some parts of the tree of life.

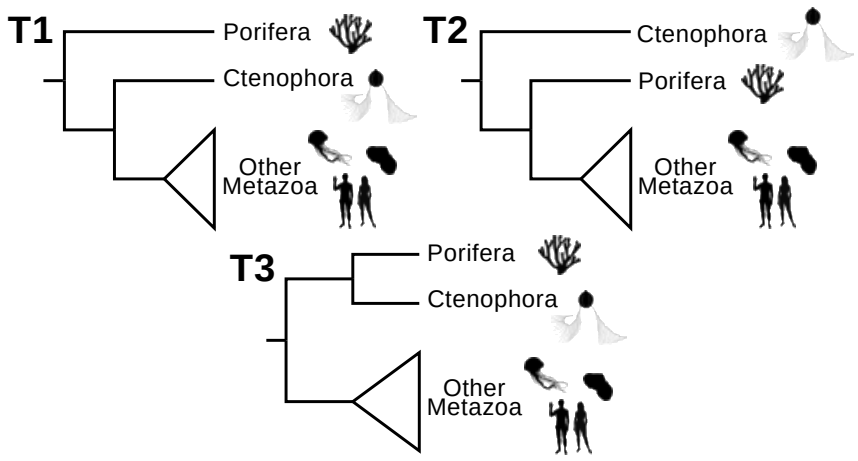

Figure 1. The three plausible topologies for basal Metazoa. The focal taxa are Porifera (sponges), Ctenophora (ctenophores or "comb jellies"), and all other Metazoa. "Other Metazoa" is a clade called Parahoxozoa [36] that comprises Cnidaria (jellyfish, corals, and sea anemones), Placozoa (a phylum of "flat animals" that comprises at least three genera [37]), and Bilateria (all other animal phyla). Animal silhouettes are from http://phylopic.org.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

Analyses of sites in both structural environments conducted after recoding amino acids using several different methods yield T2 [12]. Amino acid recoding involves converting the 20-state data into a reduced alphabet by lumping physicochemically-similar amino acids (e.g., into the six categories proposed by Dayhoff et al. [38]) or reverse translating into nucleotides and coding as purines or pyrimidines. Both approaches are intended to reduce misleading signal due to variation among taxa in amino acid frequencies [39]; if compositional variation is the primary source of biased signal then it is reasonable to conclude that analyses of buried sites are misleading. However, the use of amino acid recoding is somewhat unsatisfying as an analytical approach; ideally, one would use a more "biologically-realistic" model of amino acid evolution.

Models of sequence evolution that incorporate site-specific state frequencies, the development of which can be traced to the framework proposed by Bruno [40] and Halpern and Bruno [41] used to calculate evolutionary distances, have been used to add such biological realism to models of sequence evolution. The Halpern-Bruno framework combined patterns of selection on specific sites in proteins with population genetics to show that there is a direct relationship between the frequencies of amino acids at individual sites and the strength of selection on amino acids, making it possible to convert parameter estimates to estimates of scaled selection coefficients (at least under the special situation where mutation rates are sufficiently low for each new mutation be fixed or lost before the next mutation). Although that condition may not be met in real populations, laboratory deep mutagenesis studies have provided experimental evidence consistent with site specific fitness differences among sites [42–45]. The evidence that the basic Halpern-Bruno framework successfully reconciles protein structure and biophysics with population genetics has led to the development of models suitable for phylogenetic analyses. Examples of these "Halpern-Bruno-type" models (if they are defined broadly) include the CAT [46] model, which can be used in the Bayesian framework, and the CAT-like model of Le et al. [47] (called "LGL mixtures" by Pandey and Braun [12]), which can be used in either a Bayesian or a maximum likelihood (ML) framework. The CAT model appears to reduce phylogenetic artefacts such as long-branch attraction [48] and it seems reasonable to postulate that similar models (like the LGL mixtures) will exhibit similar behaviors.

The biological realism of those models is desirable, but memory and compute time requirements can become prohibitive for the CAT model (and related models) as the number of mixture classes (sets of equilibrium amino acid frequencies) increases. This can make the CAT model intractable for analysis of phylogenomic data sets, even when parallel or multicore implementations are used [49,50]. In general, memory and compute time requirements increase as a multiple of k (the number of mixture classes) for CAT. Wang et al. [50] (2018) proposed the posterior mean site frequency (PMSF) method, improving this to a multiple of ~k/1.5 and using substantially less RAM. PMSF assigns a conditional mean amino acid frequency profile to each site using a preliminary guide tree; those profiles can then be used for in-depth tree searching and per site likelihood calculation in place of the full mixture model. Despite this improvement, PMSF is still computationally burdensome and therefore remains difficult to use in analyses of taxon-rich phylogenomic datasets. Nevertheless, the PMSF approach would appear to provide a way to add biological realism and ameliorate data type effects due to protein structure.

An alternative to the use of complex and realistic models of sequence evolution, with their associated computational burden, has been recognized for more than a decade: adding taxa to phylogenetic analyses [51,52]. Including large numbers of taxa in phylogenetic analyses almost always improves the accuracy of those analyses [53–56], with only a small number of exceptions [57]. The positive effects of adding taxa to phylogenies have often been attributed to the breaking up of long branches; thus, one major impact of adding taxa is expected to be a reduced likelihood of long-branch attraction [58,59]. Adding taxa can also improve phylogenetic analyses by improving parameter estimates in model-based analyses [60,61]. However, it is important to recognize that the original taxon-sampling results [51,52] were obtained for analyses using the parsimony criterion, which has been

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

criticized as unrealistic [62,63]. However, when viewed another way, the ability of increased taxon sampling to improve phylogenetic estimation using the "unrealistic" parsimony criterion could be viewed as a positive: it suggests that analyses of a taxon-rich dataset might yield accurate estimates of phylogeny even if less "realistic" models (like the site-homogeneous LG model [34]) are used for analyses. It is now possible to collect large amounts of data from many different taxa, making it possible to generate taxon-rich phylogenomic trees.

For this study we have used early metazoan phylogeny as a model system to examine which factor (taxon sampling or model realism) is more important to ameliorate the structural data type effect noted by Pandey and Braun [12]. Since phylogenetic estimation using complex models and datasets is a difficult computational problem, even when the relatively efficient PMSF method is used, we focused on a metric of phylogenetic signal that can be calculated in a rapid manner. Specifically, we used the numbers of *strongly decisive sites* supporting each of the three plausible topologies for the base of Metazoa (Figure 1). We define strongly decisive sites in a manner that generalizes the approaches used by Kimball et al. [64] and Francis and Canfield [31]; this definition allows us to identify sites that strongly favor one tree relative to all other trees in a set of candidate trees. By using the strongly decisive sites criterion it was possible to examine protein structure data type effects for a taxon-rich data matrix using simple site-homogeneous and complex site-heterogeneous models. For these analyses we used the Simion et al. [28] dataset, which comprises more than 1,500 orthologous proteins sampled from 97 taxa and includes all major metazoan clades. We examined the number of strongly decisive sites in the two solvent exposure classes (surface residues and buried residues) and in three secondary structural classes (helix, sheet, and coil) given different models and numbers of taxa. These results allowed us to extend the examination of phylogenetic signal and protein structure data types reported by Pandey and Braun [12].

2. Methods

We used the globular proteins in the Simion et al. [28] dataset, subdividing those proteins using structural criteria, and conducting ML analyses of each structural component individually. We began by identifying and eliminating transmembrane proteins; we did this by using the TopCons prediction server [65] to identify a total of 153 transmembrane proteins which we excluded to create the "filtered Simion et al. dataset" (hereafter, the "FSD"). The FSD comprised 1,566 globular proteins with 356,014 aligned sites and 39% missing data. The globular proteins were then subdivided using either relative solvent accessibility (RSA) or secondary structure (SS) as described by Pandey and Braun [11,12]. Briefly, the protein subdivision pipeline (available from (https://github.com/aakanksha12/Structural_class_assignment_pipeline) used SCRATCH-1D [66], a suite of neural network programs for protein structure prediction (ACCpro for RSA and SSpro for SS). A weighted consensus sequence for each protein was used as input for ACCpro and SSpro; the consensus amino acid at each position was the residue with the highest Henikoff and Henikoff [67] weight. This pipeline creates a nexus file [68] for each multiple sequence alignment with a sets block that has charsets for the two RSA classes (Exposed and Buried) and the three SS based classes (Helix, Sheet and Coil). These sets blocks allowed us to use PAUP* 4.0b10 [69] to extract each subset of the data. The annotated nexus files are available in Supplementary File S1.

The FSD included 97 taxa: 25 Choanozoa outgroups, 12 Ctenophora, 25 Porifera), and 33 additional metazoans (23 Cnidaria, nine Bilateria, and the placozoan *Trichoplax adhaerens*). To examine the effects of taxon sampling on tree searches and proportions of strongly decisive sites, we randomly sampled 40 combinations of taxa that included one sponge, one ctenophore, one cnidarian, six outgroups, and all sampled bilateria (yielding data matrices with a total of 19 taxa). We then added taxa to these sparse taxon samples in a manner that is often used by systematists: by adding relatively distant relatives of the taxa that are already included in the data matrix. This strategy will subdivide long

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

branches as much as possible. Specifically, we took each 19-taxon dataset and added four sponges (one from each class), two ctenophores, and two cnidarians (one from each subphylum), yielding 27-taxon datasets. This procedure was repeated three additional times, generating 35, 43, and 51 taxon extensions of each of the 40 19-taxon datasets.

The number of decisive sites (*sensu* Kimball et al. [64]) favoring each topology is likely to provide a useful criterion to examine different topologies. Because this manuscript has an interdisciplinary target audience, we will provide a brief description of the likelihood calculations; interested readers are referred to reviews on the topic (e.g., chapter 8 of Warnow [70]) for additional details. Briefly, the likelihood is proportional to the probability of the data (the site patterns in the alignment) given the tree with branch lengths. If we consider the simplest alignment (two sequences) we can use equation 1 to calculate the probability that a site will be occupied by amino acid j after time t given that the initial state is amino acid i:

$$\boldsymbol{P}(t) = \exp(\boldsymbol{Q}t), \quad (1)$$

Where $\boldsymbol{P}(t)$ is a matrix of probabilities that the site is occupied by amino acid j after time t given that the initial state is amino acid i, $\boldsymbol{Q}$ is the instantaneous rate matrix, and t is the time (the branch length in units in units of substitutions per site; other units would require information about the substitution rate). Most models of sequence evolution used in empirical studies are time reversible; time reversibility makes it possible to write $\boldsymbol{Q}$ as the product of a symmetric matrix of amino acid exchangeabilities and diagonal matrix ($\boldsymbol{\Pi}$) with the equilibrium frequencies of each amino acid [71,72]. The Felsenstein [73] pruning algorithm can be used to calculate the likelihood of the complete alignment is simply the product of the likelihoods of all sites (in practice, the sum of the log likelihoods is used). Since trees for multiple taxa will have unknown internal states (i.e., the ancestral states at each node) all possible states are used for each node and the marginal likelihood is calculated. The branch lengths (and any other free parameters in the model) are obtained by numerical optimization. Amino acids frequencies are part of the $\boldsymbol{Q}$ matrix, so it is possible to calculate likelihoods using the Halpern-Bruno framework either by using a mixture of $\boldsymbol{Q}$ matrices (each generated using a different $\boldsymbol{\Pi}$ matrix and having a specific weight) or by assigning a unique $\boldsymbol{\Pi}$ matrix to each site. In practice, these calculations are performed by efficient ML programs, such as RAxML [74] and IQ-TREE [75].

This review of likelihood calculations makes two points: 1) each site has an associated *site likelihood*; and 2) there is no consideration of flanking sites in the site likelihood calculations. The first property allows us to define a decisive site as a site favoring a specific topology (Ta) relative to one or more other candidate topologies (Tb, Tc, ...) given the likelihood criterion. The degree to which site i supports a specific topology can be calculated using per site likelihoods (obtained from RAxML or IQ-TREE) using equation 2:

$$\Delta\ln L_i^{T\mathrm{a}} = \max\left(\ln L_i^{T\mathrm{b}},\ \ln L_i^{T\mathrm{c}}, \ldots\right) - \ln L_i^{T\mathrm{a}}, \quad (1)$$

Where $\ln L_i^T$ is the likelihood for site i given topology T. $\Delta\ln L_i^{T\mathrm{a}}$ will be positive for sites that favor Ta and negative if one (or more) of the other candidate topologies is favored relative to Ta. We examined three candidate topologies for this study, but equation 1 can be used for any number of candidate topologies. We designated sites with $\Delta\ln L \geq 5$ standard deviations from the mean $\Delta\ln L$ value as *strongly decisive sites*. We note that the second property is undesirable because it discards information about protein structure. However, the fact that we separated the FSD into subsets defined by RSA and SS addresses this issue (at least to some degree).

We used RAxML 8.2.10 for tree searches and site likelihood calculations using simple, site-homogeneous models. Specifically, we used the 20-state GTR model with Γ-distributed rates across sites (i.e., GTR_{20}+Γ). The large number of sites in each dataset (larger than 160,000 sites for each RSA structural class) led us to estimate GTR_{20} model parameters from subsets of the data. Briefly, we generated 10 random datasets with all 97 taxa and 10,000 sites from each structural class. Then we used RAxML to conduct a tree search and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

parameter optimization using a random starting tree for each 10,000-site dataset. The GTR_{20} model parameters (equilibrium amino acid frequencies and the exchangeability matrix) estimated for each site subsample were then averaged; these averages were used as the GTR_{20} model parameters for each structural class. These rate matrices are available in Supplementary File S1.

We used IQ-TREE 1.5.1 to estimate per site likelihoods for site-heterogeneous models. Specifically, we used the PMSF approach with five 25,000 site datasets (sites randomly sampled from the complete data matrix). PMSF requires a starting tree with branch length estimates to generate the site-specific frequency profiles; we used the RAxML $GTR_{20}+\Gamma$ tree for this. All PMSF analyses used the LG exchangeability matrix. We used six different numbers of site categories (i.e., we used LG+C10+PMSF+Γ, LG+C20+PMSF+Γ, and so forth up to C60). We chose this sample size based on preliminary tests with the exposed site data; those tests revealed that compute time and memory requirements (>150 GB per core) became prohibitive for 50,000 site datasets when we used the largest number of profile mixture categories (C50 and C60). We repeated the site likelihood calculations using the site-homogeneous $GTR_{20}+\Gamma$ model in IQ-TREE to determine whether the specific implementation of those models in each program had any impact on our inferences (as might occur, for example, if there are differences between RAxML and IQ-TREE in the details of the numerical optimization).

We examined model fit using the sample-size corrected Akaike information criterion (AIC_c) [76]. In most cases, we simply used the AIC_c value from the '.iqtree' output file. However, determining whether RSA or SS subdivision resulted in a better overall fit to the data required us to sum the likelihoods for each data subset and then calculate the AIC_c. The branch length estimates for each structural partition were unlinked so we included all branch lengths as free parameters in the AIC_c calculation (i.e., we used two times the number of branches for RSA and three times the number of branches for SS).

3. Results

3.1. Tree searches and the use of strongly decisive sites lead to similar conclusions

Because we used the strongly decisive site criterion (i.e., the view that the tree with the largest number of strongly decisive sites is the best-corroborated tree) we believed it was important to establish that it yields conclusions similar to those for standard tree searches. The proportion of strongly decisive sites supporting T2 was much larger than the proportions supporting either T1 or T3 when the complete FSD was examined (Table 1). Tree searches using RAxML resulted in topology T2 with 100% bootstrap support in all cases, indicating that using the proportion of decisive sites yields conclusions that are similar to those of tree searches. However, some strongly decisive sites supporting T1 and T3 were present despite the strong support for T2 in the bootstrap analyses.

Table 1. Percentages of strongly decisive (SD) sites in the complete FSD.

Dataset	T1	T2	T3	SD sites	Total sites
Exposed	19.3	70.8	9.9	1,420	161,897
Buried	16.6	66.1	17.3	1,625	194,117
Helix	19.2	69.3	11.5	1,290	161,117
Sheet	12.5	71.6	16.9	514	60,563
Coil	17.7	68	14.3	1,194	134,334

[1] Calculated using the $GTR_{20}+\Gamma$ model optimized on each subset of the data.

In sharp contrast to analyses using all taxa, tree searches using the most reduced taxon set (40 replicates of 19 taxa subsampled from the FSD dataset; see Methods for details) resulted in all three topologies. In all cases, analyses using either tree searches or the strongly decisive sites criterion indicated T2 was the most common topology (Figure 2). However, the number of times each topology was recovered differed between the exposed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

and buried sites; T3 was more common in the analyses of buried sites, as we expected based on Pandey and Braun [12]. These results corroborate the exposed vs buried site data type effects hypothesis for the root of Metazoa, although they also indicate that the effect is weak. Moreover, we found that the results based on strongly decisive sites were similar to those for tree searches, although they did not yield identical results (Figure 2). Specifically, the strongly decisive sites criterion revealed greater support for T3 than did tree searches when datasets with the small taxon samples were analyzed.

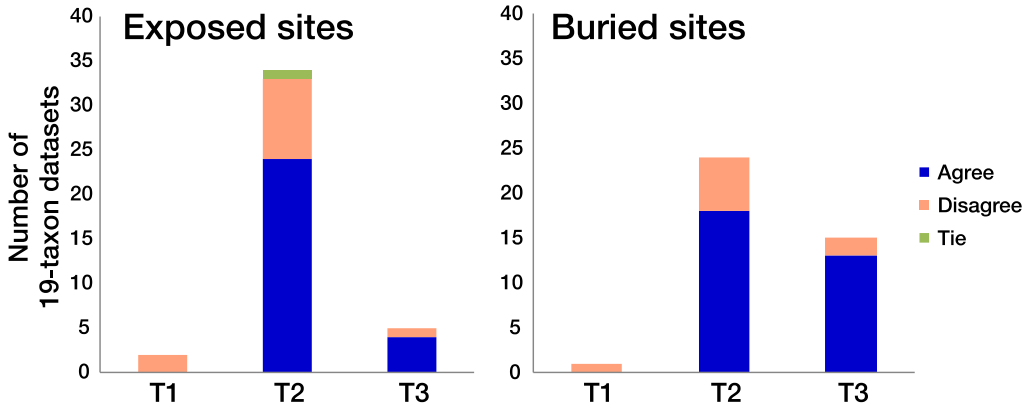

Figure 2. The position of the metazoan root based on tree searches and the strongly decisive sites criterion. These histograms show the number of times each topology was recovered in tree searches using 19-taxon datasets. Each histogram column is subdivided into cases where the decisive sites criterion agrees (blue) or disagrees (salmon) with the result of the tree search. There was one exact tie in the number of strongly decisive sites supporting T2 and T3; this case is indicated in green. Most cases where the tree search and decisive sites differed reflect a shift from T2 to T3.

The somewhat distinct behavior of using tree searches to identify the ML topology and examining the strongly decisive sites emphasizes a potential benefit of the latter method: it provides information about the relative support for multiple contradictory topologies. For the 19-taxon datasets, there were more sites that supported T3 than T1, but this was obscured because most tree searches returned T2 as the optimal tree. Likewise, there were a few tree searches for both site classes that yielded T1 but there was no taxon sample for which T1 was optimal given the strongly decisive sites criterion (Figure 2). These results emphasize that, in contrast to tree searches, consideration of the strongly decisive sites can provide information about the relative support for topologies that are suboptimal given the ML criterion.

In contrast to the results of analysis using the sparse taxon sets, analyses using all 97 taxa the FSD consistently resulted in a large number of strongly decisive sites supporting T2 and nearly equal (but much lower) numbers of sites supporting T1 and T3. One possible interpretation of strongly decisive sites is that they reflect bipartitions in gene trees (for details see Discussion, section 4.1). If true, the simplest expectation is that the number of strongly decisive sites supporting the two minority topologies would be equal. We tested the null hypothesis of equality for the numbers of strongly decisive sites supporting T1 and T3 and we were unable to reject the null hypothesis for buried or sheet residues (χ^2 for buried = 0.262, χ^2 for sheet = 2.22; both $P > 0.1$). In contrast, the null hypothesis could be rejected for exposed and helix residues (χ^2 for exposed = 42.6, χ^2 for helix = 25.3; both $P < 10^{-6}$); however, the number of strongly decisive sites favored T1 relative to T3 for both the exposed and helix sites. The remaining case (coil residues) was not significant after a multiple test correction (χ^2 for coil = 4.19; $P = 0.0407$), but it also favored T1 relative to T3. Thus, taxon addition appears to cause support, as measured using strongly decisive sites, to shift away from T3 toward T2 (the best-supported tree overall) and T1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

3.2. Taxon addition leads to rapid convergence in the proportions of strongly decisive sites observed using the complete taxon sample

When we gradually increased the number of taxa from 19, creating datasets that included 27, 35, 43 and 51 taxa, the proportion of decisive sites supporting T2 increased for all structural subsets. The most dramatic increase in the number of decisive sites supporting T2 topology between 19-taxon and 27-taxon sets for all the structurally defined subsets of the data (Figure 3 and Supplementary Figure S1). Indeed, when we used the strongly decisive sites criterion to analyze the buried sites and sheet residues, we actually observed a shift in support from T3 (for the 19-taxon datasets) to T2 (for all other taxon sets). The exposed, helix, and coil subsets always supported T2, but the proportion of strongly decisive sites supporting T2 was only slightly greater than the proportion supporting T3. The proportion of sites supporting T2 increased sharply as soon as taxa were added. For all structural classes, the increase in the number of strongly decisive sites favoring T2 largely saturated once we included 35 taxa and remained virtually constant until we reached the complete set of 97 taxa.

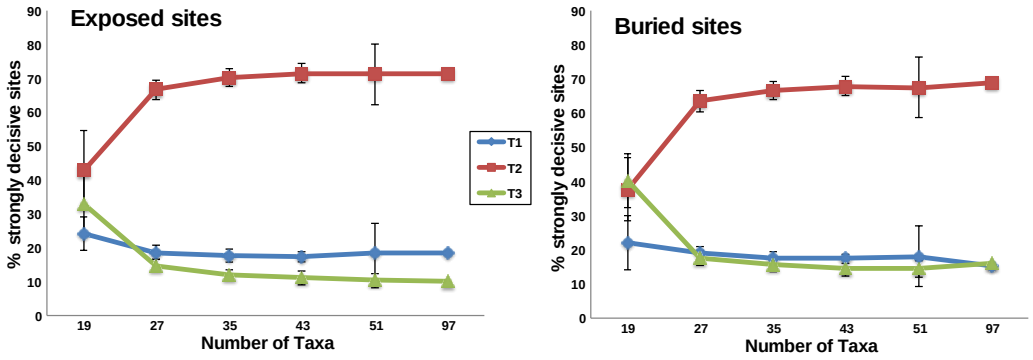

Figure 3. Taxon addition results in rapid convergence of the numbers of strongly decisive sites on the values found using the complete taxon set. The mean percentage of strongly decisive sites supporting each topology for the two subsets of site defined by RSA. Error bars indicate the standard deviation for the 40 taxon sampling replicates. We observed similar results for the SS subsets (Supplementary Figure S1).

3.3. Rich taxon sampling can compensate for low model complexity

It is generally assumed that less-complex (and, presumably, less biologically realistic) models will be more prone to bias than parameter-rich models. This led us to examine the impact of increased taxon sampling on the relatively simple site-homogeneous models (GTR_{20}) and the site-heterogeneous models (LG+PMSF), comparing 19-taxon datasets to those with 97 taxa (in all cases, we used random samples of 25,000 sites). We reduced dataset heterogeneity before conducting any analyses by dividing our data into structurally defined subsets. It is possible to determine the best way to subdivide the data by calculating the AIC_c using the overall dataset likelihoods (i.e., the sums of the likelihoods for the two RSA subsets and the three SS subsets). The overall likelihood for subdivision using RSA was much better than that for subdivision based on SS (ΔlnL favoring subdivision by RSA = 84,985.12); since RSA subdivision actually introduces fewer free parameters the RSA subdivision was also favored by the AIC_c (Supplementary File S2). The better fit of the RSA model could reflect the relatively large difference in mean substitution rates for the two RSA categories (the substitution rate for exposed sites is 1.67x the buried site rate). In contrast, the difference between the SS category with the highest mean rate (helix residues) and the slowest SS rate category (sheet residues) is only 1.21x. We note, however, that the equilibrium amino acid frequencies for each subset of the data also exhibited substantial differences (Supplementary File S2). The amino acid frequencies conformed to our expectations; for example, polar amino acids dominated the solvent exposed positions

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

whereas hydrophobic amino acids were more common in the buried environment. Likewise, the rate matrix exchangeabilities for the exposed and buried sites were similar to those observed by Pandey and Braun [11] (Supplementary File S2).

Adding among-sites heterogeneity in amino acid frequencies using the PMSF approach improved model fit relative to site-homogeneous models for all data subsets. In fact, based on AIC_c scores, the model with the largest number of mixture classes (i.e., LG+C60+PMSF+Γ) was always the best fitting model for both the 19- and 97-taxon datasets (Supplementary File S3). However, the relationship between model complexity and proportions of strongly decisive sites supporting each topology differed substantially for the 19- and 97-taxon datasets. For the 19-taxon dataset the proportions of strongly decisive sites supporting each topology were highly unstable as the complexity of the model increased from LG to C60 (Figure 4 and Supplementary Figure S2). Moreover, the topology with the largest number of strongly decisive sites in the 19-taxon dataset was T3 when the site-homogeneous (i.e., the LG and GTR_{20} models) were used but the support generally shifted to T2 when site-heterogeneous models were used. However, the situation was even more complex, since the support actually underwent another shift (to T1) when the 19-taxon buried site dataset was analyzed using the C50 and C60 mixtures. In sharp contrast to the case for the 19-taxon datasets, all analyses using the 97-taxon dataset supported T2 and the proportions of strongly decisive sites was remarkably stable across models.

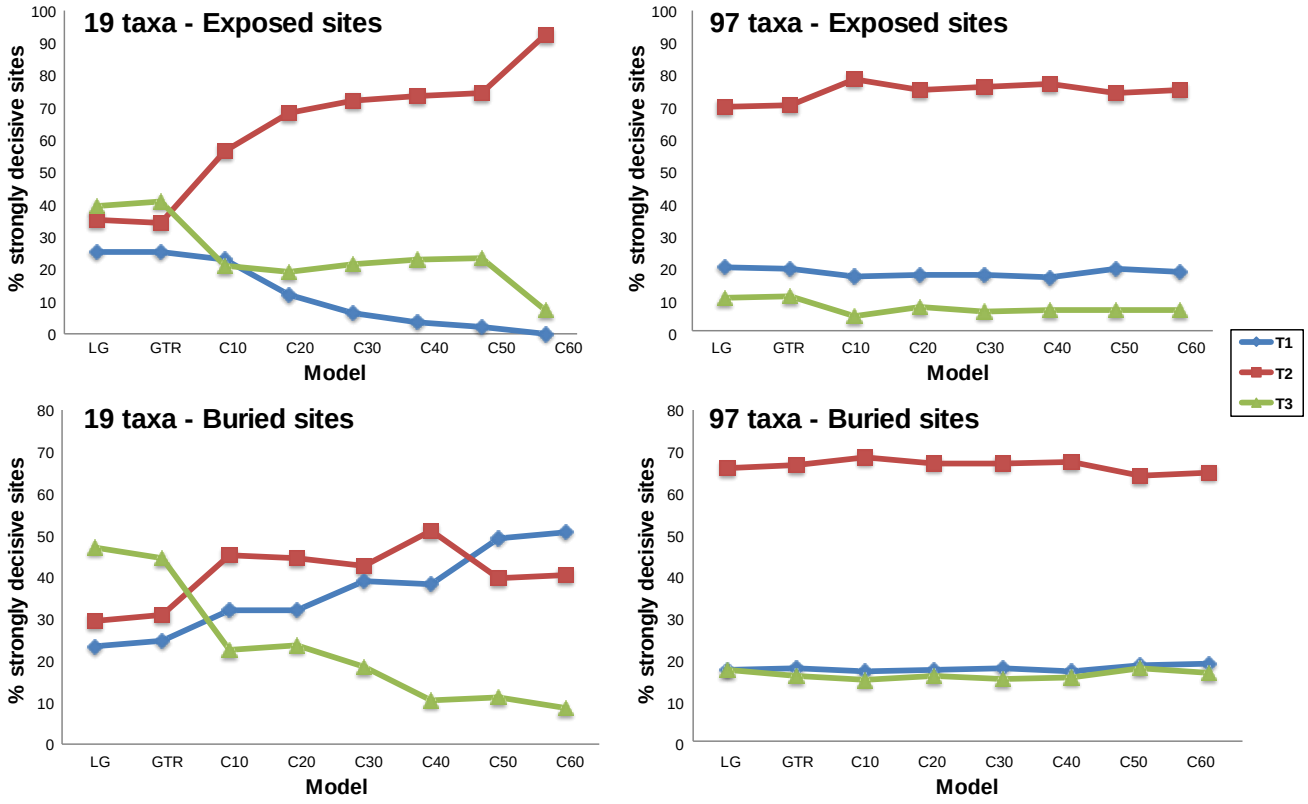

Figure 4. Proportions of strongly decisive sites supporting each topology are stable with respect to model complexity when a rich taxon sample is analyzed. A pair of 25,000 site datasets was chosen to explore the impact of increasing model complexity and adding taxa. The results of analyses using sparse taxon samples (left) were unstable to increasing model complexity, shifting from T3 to T2 for exposed sites and from T3 to T2 and finally to T1 in the case of the buried site sample. The results of analyses using the 97-taxon dataset were robust to changes in model complexity. We observed similar results for the SS subsets (Supplementary Figure S2).

4. Discussion

 doi:10.20944/preprints202101.0390.v1

These analyses showed the protein structure data type effect found by Pandey and Braun [12] could be replicated using another dataset, although the data type effect appeared to be relatively weak. This was not especially surprising since Pandey and Braun [12] noted that conducting analyses after reducing the set of outgroups to choanoflagellates yielded T2 regardless of whether exposed or buried sites were analyzed. The FSD only included choanoflagellate outgroups so it might be reasonable to expect the signal supporting T3 to be especially weak; however, there was a clear difference between exposed and buried sites for small taxon samples (Figure 2). However, there were also conflicting signals in the data and adding taxa rapidly eliminated the weak tendency for analyses of buried sites to support T3 (Figure 3). Our taxon sampling results also provide evidence against T3; after all, it has long been appreciated that adding taxa improves the performance of phylogenetic analyses (reviewed by Heath et al. [59]). The proportion of strongly decisive sites supporting T3 decreased as taxa were added and the proportion of sites supporting T2 (ctenophores sister to other Metazoa) increased (Figure 2); the proportion of strongly decisive sites supporting T1 was always smaller than the proportion supporting T2 and it was relatively stable to taxon addition. It was only necessary to add a small number of taxa for the proportions of strongly decisive sites that support each topology to converge on the values for the complete 97-taxon dataset. We also found that the proportions of decisive sites identified using simple homogeneous models is virtually identical to that proportions found using complex models when taxon-rich datasets were analyzed. The latter finding is highly significant, since the relatively simple site-homogeneous models of sequence evolution impose a limited computational burden relative to analyses using more complex models.

Given the programs currently available for phylogenetic analyses, the strongly decisive sites criterion cannot replace standard tree searches using the ML criterion or commonly used support metrics, like the bootstrap. However, it is useful for several reasons. First, as we have already emphasized, it is a computationally efficient approach that can provide information about support for multiple alternative trees. Examining support for alternative trees is especially straightforward when the plausible tree space is small, as it was for this study (Figure 1), but it would also be possible to perform a tree search and then focus on poorly supported nodes (e.g., by identifying strongly decisive sites for the optimal tree and the two trees produced by nearest-neighbor interchanges for each poorly supported node). Second, focusing on decisive sites highlight unusual genes [64]; genes with anomalously large numbers of strongly decisive sites might reflect very poor model fit or hidden paralogs. No individual genes with an unusually large number of decisive sites were identified in this study (Supplementary File S4). Finally, using strongly decisive sites might limit the impact of model misspecification. Models of sequence evolution are useful tools for inference [62,77] but they never capture "full reality" and there may be cases where no model within a set of candidates is an adequate approximation of the true model [78,79]. Assuming the models used for analyses meet a certain minimum level of "realism" dividing sites into two categories, only one of which is then viewed as informative, could be useful. Overall, we believe that this approach was useful for analyses focused on the metazoan root.

4.1. Taxon sampling has a larger impact than model complexity

To examine the impact of taxon sampling we adopted a taxon addition strategy consistent with the approach that has long been used by practicing systematists [80]. Specifically, we attempted to maximize the subdivision of long branches instead of adding taxa randomly. We began with a set of randomly chosen 19-taxon datasets and then added taxa likely to be relatively divergent from the taxa that were already present. This taxon addition strategy has a sound theoretical basis [81] and empirical systematists typically add taxa in this manner [59,82–84]. The rapid convergence on the proportions of strongly decisive sites that emerges when all 97 taxa are analyzed suggests that extremely dense taxon sampling (e.g., thousands or tens of thousands of taxa) is unlikely to be necessary

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

to make inferences regarding phylogenetic questions similar to the problem we examined. In fact, the recent publication of whole genome assemblies for hundreds of vertebrate taxa [85,86] indicate that the necessary sequencing capacity is available at the time. Addressing questions about the earliest metazoan divergences will present additional challenges, including sample availability and issues of DNA quality [87], but it seems reasonable to assert that the biggest challenges are unlikely to relate to data collection; instead, the most challenging issue is likely to be the analyses,

The positive effects of adding taxa in this study are likely to be a direct outcome of breaking up long branches and reducing the number of unobserved substitutions that need to be inferred as part of the analysis. Breaking up long branches is likely to be especially important for analyses using relatively simple, site-homogeneous models, like GTR_{20} (as well as the Braun [72] models and empirical models such as LG [34] and Dayhoff/PAM [38]). Lakner et al. [88] showed that the potential for protein function (more specifically, sequence-to-structure fit assessed using protein threading) decays rapidly when evolution proceeds according to simple models. However, when Lakner et al. [88] examined the paths between empirical sequences (i.e., cases where both ends of the evolutionary path are known to be functional) they found that the sequences along the paths between real sequences were typically consistent with functional tertiary structures. Adding taxa will reduce the lengths of the paths between sequences and should therefore have a positive impact on inference. The fact that the more complex PMSF model did not perform differently from simpler models (based on the proportions of strongly decisive sites) suggests that rich taxon sampling did not have substantial indirect effects (we define indirect effects as cases where the estimate of phylogeny is improved by the more accurate parameter estimates that are possible when many taxa are analyzed). It is possible that adding taxa did provide better information about the expected amino acid frequencies at each site, but even if that is true the better information about amino acid frequencies does not result in an increased proportion of strongly decisive sites that support a specific topology. One indirect effect that is possible would involve improved estimates of site-specific evolutionary rates [61]; since the simple and the complex models that we used both included among-sites rate heterogeneity parameters all models are likely to exhibit similar behaviors. Overall, it seems most likely that breaking up long branches was responsible for the observed amelioration of the protein structure data type effect and that improved estimates of site-specific residue frequencies were not involved.

4.2. Strongly decisive sites, phylogenetic signal, and protein structure

The use of strongly decisive sites as a metric for phylogenetic signal raises an important question: what exactly do they indicate? A simple possibility is that strongly decisive sites are associated with specific bipartitions in individual gene trees. If true, conflicting decisive sites would reflect genuine discordance among gene trees due to ILS (or other mechanisms that lead to disagreements among true gene trees). Of course, it is necessary to invoke a reasonable amount of intragenic recombination since 60% of genes with at least two strongly decisive sites have sites that support more than one topology (Supplementary File S4). However, intragenic recombination is plausible for deep-branching Metazoa given the large number of introns in the common ancestor of all Metazoa [89]; large numbers of introns lengthen genes and therefore increase the potential for intragenic recombination. The central question is whether ILS provides a plausible explanation for discordance among gene trees at the base of Metazoa. The important variables are the number of generations separating cladogenic events at the base of Metazoa, which is far from clear, and the effective population sizes for the ancestral lineages. Small-bodied organisms (as the metazoan ancestor is likely to have been) often have extremely large effective population sizes [90]. Indeed, Ewing et al. [91] invoked the possibility of gene tree discordance ILS has led to discordance among gene trees for deep metazoan relationships (albeit for the relationships at the base of Bilateria). Likewise, Arcila et al. [92] implicitly assumed ILS was possible for in their analysis of the metazoan root. We say this because

 doi:10.20944/preprints202101.0390.v1

they analyzed gene trees using ASTRAL-II [93], a program that addresses the issue of ILS phylogenetic estimation (we also note that Arcila et al. [92] concluded that ctenophores are sister to other Metazoa, just as we did in this study). Regardless of whether or not strongly decisive sites are reliable indicators of gene tree bipartitions it seems reasonable to assert that analyses of deep metazoan relationships should consider the possibility of gene tree-species tree discordance.

Given a rooted three taxon species tree (e.g., the plausible trees in Figure 1) it is possible to make a fundamental prediction about the expected patterns of discordance among gene trees if that discordance reflects ILS: the majority gene tree topology will match the species tree and the two minority gene tree topologies will be equiprobable [15]. We were unable to reject the null hypothesis of equality for the two minority topologies when we analyzed buried sites (or, if we consider SS rather than RSA, coil and sheet sites). However, we were able to reject that null hypothesis of equality for exposed residues (and for helix residues when SS was used to subdivide alignments). The observation that different structural environments have different proportions of strongly decisive sites would seem to falsify any explanation that for the conflicting strongly decisive sites that rely on true discordance among gene trees. For example, a straightforward explanation for asymmetry in the numbers of strongly decisive sites by invoking a model with substantial gene flow among the earliest stem ctenophores, sponges, and Parahoxozoa (e.g., via horizontal gene transfer) while also hypothesizing that exchanges between stem ctenophores and sponges (which would yield T3) is suppressed. The idea that basal metazoan lineages have experienced horizontal gene transfers from divergent lineages is strongly corroborated [94–96] so the notion that there might have been gene transfers among the stem ctenophores, sponges, and Parahoxozoa (and other related lineages [89,97] like choanoflagellates and Filasterea). However, if the gene transfer hypothesis represents the correct explanation for the observed conflicts, there is no clear mechanism to explain the more limited number of gene transfers between stem ctenophores and sponges then between the other stem lineages. Likewise, there is no mechanism to explain the differential asymmetry for different structural environments.

The observed asymmetry in the minority types of strongly decisive sites ultimately requires some mechanism that can break symmetry for different structurallydefined subsets of the data. The fact that we were only able to reject symmetry of the minority topologies in the more rapidly evolving exposed sites could be an important observation; perhaps the more slowly evolving buried sites are giving the best picture of the true spectrum gene trees. In this context, it is important to note that there is also an apparent correlation between the average rate of evolution and asymmetry for sites categorized by SS: we rejected symmetry for the relatively rapidly evolving helix sites, observed nearly significant asymmetry for the intermediated rate coil sites, and were unable to reject symmetry for the low-rate sites in the sheet environment. The failure to reject symmetry does not reflect power; after all, there more than 1,000 strongly decisive sites for almost all of the structurally defined sets of sites (the only exception was the sheet sites, where there were 514 strongly decisive sites; see Table 1). The evidence for a relationship between the mean rate of evolution in each structural environment does suggest that strongly decisive sites could have more to do with patterns of molecular evolution than with true discordance among gene trees. Regardless of the impact of gene tree-species tree conflicts on phylogenetic estimation at the base of Metazoa one thing is clear: the proportions of strongly decisive sites differ among structural environments and those differences are observed for site-homogeneous and site-heterogeneous models.

As stated in the introduction, the position of the metazoan root has been the subject of many studies; those studies that have recovered T1 have typically invoked the long-branch attraction artefact [98] to explain studies that have recovered T2, typically arguing that site-heterogeneous models are less susceptible to long-branch attraction. However, many of the studies focused on the root of Metazoa - including those that have recovered T2 - have used site-heterogeneous models. This suggests that the use of site-heterogene-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 January 2021 doi:10.20944/preprints202101.0390.v1

ous models are not necessarily the critical variable. Nevertheless, it is important to consider the hypothesis that T1 is the true tree and T2 is a long-branch artefact. The simulations reported by Kapli and Telford [32] arguably provide some of the most compelling support for that hypothesis because they found an asymmetry in the result of tree searches conducted using simulated data. More specifically, analyses of data simulated assuming T2 yielded T2 whereas analyses of data simulated assuming T1 yielded both T1 and T2. Kapli and Telford [32] used site-heterogeneous models but their simulations did not incorporate discordance among gene trees due to ILS. It would be interesting to determine whether that represents an important variable. Indeed, it would be desirable to improve the simulation models in other ways; structurally-constrained models (as in Arenas et al. [99]) would clearly be desirable, as would other ways to integrate protein structure (e.g., long-range constraints [100,101]). The broader point of this discussion of simulation models is to emphasize the challenges of interpreting simulations; it is difficult to interpret simulations unless the models used for those simulations are realistic but it is also difficult to determine whether models are realistic. Strongly decisive sites may be useful for model evaluation - data simulated using a biologically-realistic model should yield alignments with numbers of strongly decisive sites similar to those observed for the empirical data.

5. Conclusions

This study has three fundamental conclusions. The first is specific to the phylogenetic question that we explored, and it is relatively straightforward: T2 is the best-corroborated hypothesis for the position of the metazoan root based on the results of this study (i.e., the best-supported tree places ctenophores sister to all other extant animals). However, we urge readers to view our phylogenetic conclusion with a certain degree of caution; the position of the metazoan root has varied among studies despite the extensive work focused on this question [17–32], making it is difficult to view any position as particularly strongly supported at this point. The second conclusion is methodological: we found evidence that focusing on strongly decisive sites can provide a useful way to examine phylogenetic questions, especially when many taxa and sites are analyzed. We conclude that the decisive sites criterion is useful and that it provided evidence that denser taxon sampling can compensate, at least partially, for model complexity and permit the use of simpler models for analyses. However, we note that conventional tree searches were also consistent with the idea that rich taxon samples are beneficial, as expected based on many previous studies (reviewed by Heath et al. [59]). However, the fact that the strongly decisive sites criterion provides direct evidence for conflicting signals provides a good reason to view them as a tool that can be complementary to standard tree searches.

The third and final conclusion of this study relates to the intersection of phylogenetics and protein structure: different topological signals can be associated with sites in different protein structural environments. Given the extensive evidence that patterns of protein evolution are structure dependent [6–8,10–12] it might not be surprising that structure can have an impact on estimates of phylogeny, especially for studies focused on difficult parts of the tree of life, like the position of the metazoan root, are examined. Wilke [102] lamented that a negative aspect of the efforts to improve models of sequence evolution "...has been that the underlying biophysical objects represented by the sequences, DNA molecules, RNA molecules, and proteins, have taken a back-seat in much computational molecular-evolution work." The Halpern-Bruno model framework is appealing, but it is ultimately a statement regarding variation among alignment positions in the underlying frequencies of amino acids; it does not consider the physicochemical properties of those amino acids or protein structure. The long-standing controversy regarding the position of the root of Metazoa raises a fundamental question: is it possible to conceive of any specific information that would resolve the question in a "satisfying" manner? We believe that our strongly decisive site results can be interpreted in two ways:

1) There gene tree-species tree discordance at the base of Metazoa, possibly reflecting ILS, and a structure dependent bias. The bias leads to the observed

asymmetry in the proportions of minority decisive site types in the relatively rapidly evolving solvent exposed and helix sites.

2) There are processes of sequence evolution that generate conflicting strongly decisive sites without discordance among gene trees. Those process result in equal proportions of minority decisive site types in low-rate structural environments and unequal proportions of minority decisive site types in higher rate structural environments.

We also believe that distinguishing between those hypotheses will require a more detailed understanding of the roles of protein structure and biophysics in models of sequence evolution rather than the continued use of models that fit into the Halpern-Bruno framework. We believe that such an integrative approach will be necessary to obtain convincing resolutions for difficult nodes in the tree of life, like the root of Metazoa.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, File S1: Annotated multiple sequence alignments and rate matrices, File S2: Spreadsheet with model fit statistics for alternative partitioning methods, File S3: Model fit statistics for site-heterogeneous models, File S4: Numbers of strongly decisive sites per sequence alignments, Figure S1: Impact of taxon sampling on analyses of data subsets defined by SS, Figure S2: Impact of model complexity on analyses of data subsets defined by SS.

Author Contributions: Conceptualization, A.P. and E.L.B.; methodology, A.P. and E.L.B.; software, A.P.; formal analysis, A.P. and E.L.B.; investigation, A.P.; data curation, A.P.; writing—original draft preparation, A.P. and E.L.B.; writing—review and editing, A.P. and E.L.B.; visualization, A.P. and E.L.B.; project administration, E.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was not supported by extramural funding.

Acknowledgments: We are grateful to Rebecca Kimball for careful reading of earlier versions of this manuscript and to Gordon Burleigh, Siavash Mirarab, and the members of the Kimball-Braun lab for helpful discussions while drafting this manuscript.

Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.

Data Availability Statement: please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Zuckerkandl, E.; Pauling, L. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*; Bryson, V.; Vogel, H.J., Ed.; Elsevier, **1965**; pp. 97–166 ISBN 9781483227344.
2. Dickerson, R. E. The structures of cytochrome *c* and the rates of molecular evolution. *J. Mol. Evol.* **1971**, 1, 26–45, doi:10.1007/BF01659392.
3. Alvarez-Ponce, D. Richard Dickerson, molecular clocks, and rates of protein evolution. *J. Mol. Evol.* **2020**, Online ahead of print. doi:10.1007/s00239-020-09973-x.
4. Zhang, J.; Yang, J.-R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **2015**, *16*, 409–420, doi:10.1038/nrg3950.
5. Echave, J.; Spielman, S. J.; Wilke, C. O. Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **2016**, *17*, 109–121, doi:10.1038/nrg.2015.18.
6. Gerstein, M.; Sonnhammer, E. L.; Chothia, C. Volume changes in protein evolution. *J. Mol. Biol.* **1994**, *236*, 1067–1078, doi:10.1016/0022-2836(94)90012-4.
7. Goldman, N.; Thorne, J. L.; Jones, D. T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **1998**, *149*, 445–458.
8. Illergård, K.; Ardell, D. H.; Elofsson, A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* **2009**, *77*, 499–508, doi:10.1002/prot.22458.
9. Worth, C. L.; Gong, S.; Blundell, T. L. Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 709–720, doi:10.1038/nrm2762.
10. Le, S. Q.; Gascuel, O. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst. Biol.* **2010**, *59*, 277–287, doi:10.1093/sysbio/syq002.

11. Pandey, A.; Braun, E. L. Protein evolution is structure dependent and non-homogeneous across the tree of life. In *Proceedings of ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB'20)*; ACM, New York, NY, USA, **2020**; Article No.: 28, 11 pages, doi: 10.1145/3388440.3412473.
12. Pandey, A.; Braun, E. L. Phylogenetic analyses of sites in different protein structural environments result in distinct placements of the metazoan root. *Biology (Basel)* **2020**, *9*, 64, doi:10.3390/biology9040064.
13. Reddy, S.; Kimball, R. T.; Pandey, A.; Hosner, P. A.; Braun, M. J.; Hackett, S. J.; Han, K.-L.; Harshman, J.; Huddleston, C. J.; Kingston, S.; Marks, B. D.; Miglia, K. J.; Moore, W. S.; Sheldon, F. H.; Witt, C. C.; Yuri, T.; Braun, E. L. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Syst. Biol.* **2017**, *66*, 857–879, doi:10.1093/sysbio/syx041.
14. Braun, E. L.; Kimball, R. T. Data types and the phylogeny of Neoaves. *Birds* **2021**, *2*, 1–22, doi:10.3390/birds2010001.
15. Pamilo, P.; Nei, M. Relationships between gene trees and species trees. *Mol. Biol. Evol*. **1988**, *5*, 568–583, doi:10.1093/oxfordjournals.molbev.a040517.
16. Edwards, S. V. Is a new and general theory of molecular systematics emerging? *Evolution* **2009**, *63*, 1–19, doi:10.1111/j.1558-5646.2008.00549.x.
17. Dunn, C. W.; Hejnol, A.; Matus, D. Q.; Pang, K.; Browne, W. E.; Smith, S. A.; Seaver, E.; Rouse, G. W.; Obst, M.; Edgecombe, G. D.; Sørensen, M. V.; Haddock, S. H. D.; Schmidt-Rhaesa, A.; Okusu, A.; Kristensen, R. M.; Wheeler, W. C.; Martindale, M. Q.; Giribet, G. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **2008**, *452*, 745–749, doi:10.1038/nature06614.
18. Hejnol, A.; Obst, M.; Stamatakis, A.; Ott, M.; Rouse, G. W.; Edgecombe, G. D.; Martinez, P.; Baguñà, J.; Bailly, X.; Jondelius, U.; Wiens, M.; Müller, W. E. G.; Seaver, E.; Wheeler, W. C.; Martindale, M. Q.; Giribet, G.; Dunn, C. W. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. B* **2009**, *276*, 4261–4270, doi:10.1098/rspb.2009.0896.
19. Philippe, H.; Derelle, R.; Lopez, P.; Pick, K.; Borchiellini, C.; Boury-Esnault, N.; Vacelet, J.; Renard, E.; Houliston, E.; Quéinnec, E.; Da Silva, C.; Wincker, P.; Le Guyader, H.; Leys, S.; Jackson, D. J.; Schreiber, F.; Erpenbeck, D.; Morgenstern, B.; Wörheide, G.; Manuel, M. Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **2009**, *19*, 706–712, doi:10.1016/j.cub.2009.02.052.
20. Pick, K. S.; Philippe, H.; Schreiber, F.; Erpenbeck, D.; Jackson, D. J.; Wrede, P.; Wiens, M.; Alié, A.; Morgenstern, B.; Manuel, M.; Wörheide, G. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol. Biol. Evol.* **2010**, *27*, 1983–1987, doi:10.1093/molbev/msq089.
21. Nosenko, T.; Schreiber, F.; Adamska, M.; Adamski, M.; Eitel, M.; Hammel, J.; Maldonado, M.; Müller, W. E. G.; Nickel, M.; Schierwater, B.; Vacelet, J.; Wiens, M.; Wörheide, G. Deep metazoan phylogeny: when different genes tell different stories. *Mol. Phylogenet. Evol.* **2013**, *67*, 223–233, doi:10.1016/j.ympev.2013.01.010.
22. Ryan, J. F.; Pang, K.; Schnitzler, C. E.; Nguyen, A.-D.; Moreland, R. T.; Simmons, D. K.; Koch, B. J.; Francis, W. R.; Havlak, P.; NISC Comparative Sequencing Program; Smith, S. A.; Putnam, N. H.; Haddock, S. H. D.; Dunn, C. W.; Wolfsberg, T. G.; Mullikin, J. C.; Martindale, M. Q.; Baxevanis, A. D. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* **2013**, *342*, 1242592, doi:10.1126/science.1242592.
23. Moroz, L. L.; Kocot, K. M.; Citarella, M. R.; Dosung, S.; Norekian, T. P.; Povolotskaya, I. S.; Grigorenko, A. P.; Dailey, C.; Berezikov, E.; Buckley, K. M.; Ptitsyn, A.; Reshetov, D.; Mukherjee, K.; Moroz, T. P.; Bobkova, Y.; Yu, F.; Kapitonov, V. V.; Jurka, J.; Bobkov, Y. V.; Swore, J. J.; Kohn, A. B. The ctenophore genome and the evolutionary origins of neural systems. *Nature* **2014**, *510*, 109–114, doi:10.1038/nature13400.
24. Borowiec, M. L.; Lee, E. K.; Chiu, J. C.; Plachetzki, D. C. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics* **2015**, *16*, 987, doi:10.1186/s12864-015-2146-4.
25. Pisani, D.; Pett, W.; Dohrmann, M.; Feuda, R.; Rota-Stabelli, O.; Philippe, H.; Lartillot, N.; Wörheide, G. Genomic data do not support comb jellies as the sister group to all other animals. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 15402–15407, doi:10.1073/pnas.1518127112.
26. Whelan, N. V.; Kocot, K. M.; Moroz, L. L.; Halanych, K. M. Error, signal, and the placement of Ctenophora sister to all other animals. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 5773–5778, doi:10.1073/pnas.1503453112.
27. Feuda, R.; Dohrmann, M.; Pett, W.; Philippe, H.; Rota-Stabelli, O.; Lartillot, N.; Wörheide, G.; Pisani, D. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Curr. Biol.* **2017**, *27*, 3864-3870.e4, doi:10.1016/j.cub.2017.11.008.
28. Simion, P.; Philippe, H.; Baurain, D.; Jager, M.; Richter, D. J.; Di Franco, A.; Roure, B.; Satoh, N.; Quéinnec, É.; Ereskovsky, A.; Lapébie, P.; Corre, E.; Delsuc, F.; King, N.; Wörheide, G.; Manuel, M. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* **2017**, *27*, 958–967, doi:10.1016/j.cub.2017.02.031.
29. Whelan, N. V.; Kocot, K. M.; Moroz, T. P.; Mukherjee, K.; Williams, P.; Paulay, G.; Moroz, L. L.; Halanych, K. M. Ctenophore relationships and their placement as the sister group to all other animals. *Nat. Ecol. Evol.* **2017**, *1*, 1737–1746, doi:10.1038/s41559-017-0331-3.
30. Laumer, C. E.; Fernández, R.; Lemer, S.; Combosch, D.; Kocot, K. M.; Riesgo, A.; Andrade, S. C. S.; Sterrer, W.; Sørensen, M. V.; Giribet, G. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. R. Soc. B* **2019**, *286*, 20190831, doi:10.1098/rspb.2019.0831.
31. Francis, W. R.; Canfield, D. E. Very few sites can reshape the inferred phylogenetic tree. *PeerJ* **2020**, *8*, e8865, doi:10.7717/peerj.8865.

32. Kapli, P.; Telford, M. J. Topology-dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. *Sci. Adv.* **2020**, **6**, eabc5162, doi:10.1126/sciadv.abc5162.
33. Nielsen, C. Early animal evolution: a morphologist's view. *R. Soc. Open Sci.* **2019**, *6*, 190638, doi:10.1098/rsos.190638.
34. Le, S. Q.; Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **2008**, *25*, 1307–1320, doi:10.1093/molbev/msn067.
35. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **1986**, 17, 57–86.
36. Ryan, J. F.; Pang, K.; NISC Comparative Sequencing Program; Mullikin, J. C.; Martindale, M. Q.; Baxevanis, A. D. The homeodomain complement of the ctenophore *Mnemiopsis leidyi* suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *Evodevo* **2010**, *1*, 9, doi:10.1186/2041-9139-1-9.
37. Osigus, H.-J.; Rolfes, S.; Herzog, R.; Kamm, K.; Schierwater, B. *Polyplacotoma mediterranea* is a new ramified placozoan species. *Curr. Biol.* **2019**, *29*, R148–R149, doi:10.1016/j.cub.2019.01.068.
38. Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5; Dayhoff, M.O., Ed.; National Biomedical Research Foundation, Silver Springs, MD **1978**; 345–352.
39. Susko, E.; Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **2007**, *24*, 2139–2150, doi:10.1093/molbev/msm144.
40. Bruno, W. J. Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.* **1996**, *13*, 1368–1374, doi:10.1093/oxfordjournals.molbev.a025583.
41. Halpern, A. L.; Bruno, W. J. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* **1998**, *15*, 910–917, doi:10.1093/oxfordjournals.molbev.a025995.
42. Melamed, D.; Young, D. L.; Gamble, C. E.; Miller, C. R.; Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **2013**, *19*, 1537–1551, doi:10.1261/rna.040709.113.
43. Roscoe, B. P.; Bolon, D. N. A. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *J. Mol. Biol.* **2014**, *426*, 2854–2870, doi:10.1016/j.jmb.2014.05.019.
44. Starita, L. M.; Young, D. L.; Islam, M.; Kitzman, J. O.; Gullingsrud, J.; Hause, R. J.; Fowler, D. M.; Parvin, J. D.; Shendure, J.; Fields, S. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **2015**, *200*, 413–422, doi:10.1534/genetics.115.175802.
45. Mighell, T. L.; Evans-Dutson, S.; O'Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **2018**, *102*, 943–955, doi:10.1016/j.ajhg.2018.03.018.
46. Lartillot, N.; Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **2004**, **21**, 1095–1109, doi:10.1093/molbev/msh112.
47. Le, S. Q.; Gascuel, O.; Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **2008**, *24*, 2317–2323, doi:10.1093/bioinformatics/btn445.
48. Lartillot, N.; Brinkmann, H.; Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **2007**, *7 Suppl 1*, *S4*, doi:10.1186/1471-2148-7-S1-S4.
49. Whelan, N. V.; Halanych, K. M. Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. *Syst. Biol.* **2017**, *66*, 232–255, doi:10.1093/sysbio/syw084.
50. Wang, H.-C.; Minh, B. Q.; Susko, E.; Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **2018**, *67*, 216–235, doi:10.1093/sysbio/syx068.
51. Hillis, D. M. Inferring complex phytogenies. *Nature* **1996**, *383*, 130–131, doi:10.1038/383130a0.
52. Hillis, D. M. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* **1998**, 47, 3–8, doi:10.1080/106351598260987.
53. Pollock, D. D.; Zwickl, D. J.; McGuire, J. A.; Hillis, D. M. Increased taxon sampling is advantageous for phylogenetic inference. *Syst. Biol.* **2002**, *51*, 664–671, doi:10.1080/10635150290102357.
54. Zwickl, D. J.; Hillis, D. M. Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* **2002**, *51*, 588–598, doi:10.1080/10635150290102339.
55. Hillis, D. M.; Pollock, D. D.; McGuire, J. A.; Zwickl, D. J. Is sparse taxon sampling a problem for phylogenetic inference? *Syst. Biol.* **2003**, *52*, 124–126, doi:10.1080/10635150390132911.
56. Hedtke, S. M.; Townsend, T. M.; Hillis, D. M. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* **2006**, *55*, 522–529, doi:10.1080/10635150600697358.
57. Braun, E. L.; Kimball, R. T. Examining basal avian divergences with mitochondrial sequences: Model complexity, taxon sampling, and sequence length. *Syst. Biol.* **2002**, *51*, 614–625, doi:10.1080/10635150290102294.
58. Wiens, J. J. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Syst. Biol.* **2005**, *54*, 731–742, doi:10.1080/10635150500234583.
59. Heath, T. A.; Hedtke, S. M.; Hillis, D. M. Taxon sampling and the accuracy of phylogenetic analyses. *J. Syst. Evol.* **2008**, *46*, 239–257, doi:10.3724/SP.J.1002.2008.08016

60. Sullivan, J.; Swofford, D. L.; Naylor, G. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.* **1999**, *16*, 1347–1356, doi:10.1093/oxfordjournals.molbev.a026045.
61. Pollock, D. D.; Bruno, W. J. Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. *Mol. Biol. Evol.* **2000**, *17*, 1854–1858, doi:10.1093/oxfordjournals.molbev.a026286.
62. Swofford, D. L.; Waddell, P. J.; Huelsenbeck, J. P.; Foster, P. G.; Lewis, P. O.; Rogers, J. S. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* **2001**, *50*, 525–539.
63. Huelsenbeck, J. P.; Ané, C.; Larget, B.; Ronquist, F. A Bayesian perspective on a non-parsimonious parsimony model. *Syst. Biol.* **2008**, *57*, 406–419, doi:10.1080/10635150802166046.
64. Kimball, R. T.; Wang, N.; Heimer-McGinn, V.; Ferguson, C.; Braun, E. L. Identifying localized biases in large datasets: a case study using the avian tree of life. *Mol. Phylogenet. Evol.* **2013**, *69*, 1021–1032, doi:10.1016/j.ympev.2013.05.029.
65. Tsirigos, K. D.; Peters, C.; Shu, N.; Käll, L.; Elofsson, A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **2015**, *43*, W401–W407, doi:10.1093/nar/gkv485.
66. Magnan, C. N.; Baldi, P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* **2014**, *30*, 2592–2597, doi:10.1093/bioinformatics/btu352.
67. Henikoff, S.; Henikoff, J. G. Position-based sequence weights. *J. Mol. Biol.* **1994**, *243*, 574–578, doi:10.1016/0022-2836(94)90032-9.
68. Maddison, D. R.; Swofford, D. L.; Maddison, W. P. NEXUS: an extensible file format for systematic information. *Syst. Biol.* **1997**, *46*, 590–621, doi:10.1093/sysbio/46.4.590.
69. Swofford, D. L. PAUP* (*Phylogenetic Analysis Using PAUP); **2020**; available from https://paup.phylosolutions.com
70. Warnow, T. Computational Phylogenetics (an introduction to designing methods for phylogeny estimation); 1st ed.; Cambridge University Press, **2017**; 394 pages; ISBN 978-1107184718.
71. Whelan, S.; Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **2001**, *18*, 691–699, doi:10.1093/oxfordjournals.molbev.a003851.
72. Braun, E. L. An evolutionary model motivated by physicochemical properties of amino acids reveals variation among proteins. *Bioinformatics* **2018**, *34*, i350–i356, doi:10.1093/bioinformatics/bty261.
73. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **1981**, *17*, 368–376, doi:10.1007/BF01734359.
74. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313, doi:10.1093/bioinformatics/btu033.
75. Nguyen, L.-T.; Schmidt, H. A.; von Haeseler, A.; Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274, doi:10.1093/molbev/msu300.
76. Burnham, K. P.; Anderson, D. R. Model selection and multimodel inference. Springer New York: New York, NY, 2004; ISBN 978-0-387-95364-9.
77. Kelchner, S. A.; Thomas, M. A. Model use in phylogenetics: nine key questions. *Trends Ecol. Evol.* **2007**, *22*, 87–94, doi:10.1016/j.tree.2006.10.004.
78. Sanderson, M. J.; Kim, J. Parametric phylogenetics? *Syst. Biol.* **2000**, *49*, 817–829, doi:10.1080/106351500750049860.
79. Gatesy, J. A tenth crucial question regarding model use in phylogenetics. *Trends Ecol. Evol.* **2007**, *22*, 509–510, doi:10.1016/j.tree.2007.08.002.
80. Graybeal, A. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **1998**, *47*, 9–17, doi: 10.1080/106351598260996
81. Goldman, N. Phylogenetic information and experimental design in molecular systematics. *Proc. R. Soc. B* **1998**, *265*, 1779–1786, doi:10.1098/rspb.1998.0502.
82. Geuten, K.; Massingham, T.; Darius, P.; Smets, E.; Goldman, N. Experimental design criteria in phylogenetics: where to add taxa. *Syst. Biol.* **2007**, *56*, 609–622, doi:10.1080/10635150701499563.
83. Lanier, H. C.; Knowles, L. L. Applying species-tree analyses to deep phylogenetic histories: challenges and potential suggested from a survey of empirical phylogenetic studies. *Mol. Phylogenet. Evol.* **2015**, *83*, 191–199, doi:10.1016/j.ympev.2014.10.022.
84. Tamashiro, R. A.; White, N. D.; Braun, M. J.; Faircloth, B. C.; Braun, E. L.; Kimball, R. T. What are the roles of taxon sampling and model fit in tests of cyto-nuclear discordance using avian mitogenomic data? *Mol. Phylogenet. Evol.* **2019**, *130*, 132–142, doi:10.1016/j.ympev.2018.10.008.
85. Feng, S.; Stiller, J.; Deng, Y.; Armstrong, J.; Fang, Q.; Reeve, A. H.; Xie, D.; Chen, G.; Guo, C.; Faircloth, B. C.; et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature* **2020**, 587, 252–257, doi:10.1038/s41586-020-2873-9.
86. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **2020**, *587*, 240–245, doi:10.1038/s41586-020-2876-6.

87. Panova, M.; Aronsson, H.; Cameron, R. A.; Dahl, P.; Godhe, A.; Lind, U.; Ortega-Martinez, O.; Pereyra, R.; Tesson, S. V. M.; Wrange, A.-L.; Blomberg, A.; Johannesson, K. DNA extraction protocols for whole-genome sequencing in marine organisms. *Methods Mol. Biol.* **2016**, *1452*, 13–44, doi:10.1007/978-1-4939-3774-5_2.
88. Lakner, C.; Holder, M. T.; Goldman, N.; Naylor, G. J. P. What's in a likelihood? Simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. *Syst. Biol.* **2011**, *60*, 161–174, doi:10.1093/sysbio/syq088.
89. Grau-Bové, X.; Torruella, G.; Donachie, S.; Suga, H.; Leonard, G.; Richards, T. A.; Ruiz-Trillo, I. Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife* **2017**, *6*, e26036, doi:10.7554/eLife.26036.
90. Rossberg, A. G.; Rogers, T.; McKane, A. J. Are there species smaller than 1 mm? *Proc. R. Soc. B* **2013**, *280*, 20131248, doi:10.1098/rspb.2013.1248.
91. Ewing, G. B.; Ebersberger, I.; Schmidt, H. A.; von Haeseler, A. Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.* **2008**, *8*, 118, doi:10.1186/1471-2148-8-118.
92. Arcila, D.; Ortí, G.; Vari, R.; Armbruster, J. W.; Stiassny, M. L. J.; Ko, K. D.; Sabaj, M. H.; Lundberg, J.; Revell, L. J.; Betancur-R, R. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat. Ecol. Evol.* **2017**, *1*, 20, doi:10.1038/s41559-016-0020.
93. Mirarab, S.; Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **2015**, *31*, i44-i52, doi:10.1093/bioinformatics/btv234.
94. Jackson, D. J.; Macis, L.; Reitner, J.; Wörheide, G. A horizontal gene transfer supported the evolution of an early metazoan biomineralization strategy. *BMC Evol. Biol.* **2011**, *11*, 238, doi:10.1186/1471-2148-11-238.
95. Boto, L. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc. R. Soc. B* **2014**, *281*, 20132450, doi:10.1098/rspb.2013.2450.
96. Hernandez, A. M.; Ryan, J. F. Horizontally transferred genes in the ctenophore *Mnemiopsis leidyi*. *PeerJ* **2018**, *6*, e5067, doi:10.7717/peerj.5067.
97. Hehenberger, E.; Tikhonenkov, D. V.; Kolisko, M.; Del Campo, J.; Esaulov, A. S.; Mylnikov, A. P.; Keeling, P. J. Novel predators reshape holozoan phylogeny and reveal the presence of a two-component signaling system in the ancestor of animals. *Curr. Biol.* **2017**, *27*, 2043-2050.e6, doi:10.1016/j.cub.2017.06.006.
98. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.* **1978**, *27*, 401–410, doi:10.1093/sysbio/27.4.401.
99. Arenas, M.; Dos Santos, H. G.; Posada, D.; Bastolla, U. Protein evolution along phylogenetic histories under structurally constrained substitution models. *Bioinformatics* **2013**, *29*, 3020–3028, doi:10.1093/bioinformatics/btt530.
100. Sharir-Ivry, A.; Xia, Y. Nature of long-range evolutionary constraint in enzymes: Insights from comparison to pseudoenzymes with similar structures. *Mol. Biol. Evol.* **2018**, *35*, 2597–2606, doi:10.1093/molbev/msy177.
101. Echave, J. Beyond stability constraints: a biophysical model of enzyme evolution with selection on stability and activity. *Mol. Biol. Evol.* **2019**, *36*, 613–620, doi:10.1093/molbev/msy244.
102. Wilke, C. O. Bringing molecules back into molecular evolution. *PLoS Comput. Biol.* **2012**, *8*, e1002572, doi:10.1371/journal.pcbi.1002572.