Syst. Biol. 66(2):232–255, 2017
© The Author(s) 2016. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syw084
Advance Access publication September 14, 2016

Who Let the CAT Out of the Bag? Accurately Dealing with Substitutional Heterogeneity in Phylogenomic Analyses

Nathan V. Whelan* and Kenneth M. Halanych

Department of Biological Sciences, Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, 101 Life Sciences Building, Auburn, AL 36849, USA.
**Correspondence to be sent to: Warm Springs Fish Technology Center, U.S. Fish and Wildlife Service, 5308 Spring ST, Warm Springs, GA 31830, USA; Email: nathan_whelan@fws.gov*

Received 19 September 2015; reviews returned 29 August 2016; accepted 4 September 2016
Associate Editor: Lars Jermiin

Abstract.—As phylogenetic datasets have increased in size, site-heterogeneous substitution models such as CAT-F81 and CAT-GTR have been advocated in favor of other models because they purportedly suppress long-branch attraction (LBA). These models are two of the most commonly used models in phylogenomics, and they have been applied to a variety of taxa, ranging from *Drosophila* to land plants. However, many arguments in favor of CAT models have been based on tenuous assumptions about the true phylogeny, rather than rigorous testing with known trees via simulation. Moreover, CAT models have not been compared to other approaches for handling substitutional heterogeneity such as data partitioning with site-homogeneous substitution models. We simulated amino acid sequence datasets with substitutional heterogeneity on a variety of tree shapes including those susceptible to LBA. Data were analyzed with both CAT models and partitioning to explore model performance; in total over 670,000 CPU hours were used, of which over 97% was spent running analyses with CAT models. In many cases, all models recovered branching patterns that were identical to the known tree. However, CAT-F81 consistently performed worse than other models in inferring the correct branching patterns, and both CAT models often overestimated substitutional heterogeneity. Additionally, reanalysis of two empirical metazoan datasets supports the notion that CAT-F81 tends to recover less accurate trees than data partitioning and CAT-GTR. Given these results, we conclude that partitioning and CAT-GTR perform similarly in recovering accurate branching patterns. However, computation time can be orders of magnitude less for data partitioning, with commonly used implementations of CAT-GTR often failing to reach completion in a reasonable time frame (i.e., for Bayesian analyses to converge). 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. [Data partitioning; phylogenomics, simulation, site-heterogeneity, substitution models.]

Nucleotide and amino acid substitution models are integral components of modern-day phylogenetic algorithms (Philippe et al. 2005; Rokas and Carroll 2006). Commonly used substitution models include the general time-reversible (GTR) model for both nucleotides and amino acids (Tavaré 1986) and amino acid substitution models with fixed exchange matrices such as WAG and LG (Whelan and Goldman 2001; Le and Gascuel 2008). Many amino acid substitution models either use fixed amino acid frequencies that are unique to each model or amino acid frequencies derived from the data being analyzed (e.g., WAG+F, LG+F). Substitution models can also be extended to incorporate rate heterogeneity across sites by incorporating a gamma distribution or modeling invariant sites (e.g., WAG+Γ, WAG+Γ+F, WAG+I+Γ, WAG+I+Γ+F; Yang 1994; Gu et al. 1995). The aforementioned models are homogeneous across sites in the sense that the same underlying model of amino acid substitution applies to all sites. These models also assume that there are no lineage-specific differences in the substitution process, which is also an assumption that may be violated in large datasets (Jayaswal et al. 2014).

The onset of phylogenomics and increasingly large sequence datasets has led to development of site-heterogeneous substitution models (Lartillot and Philippe 2004; Le et al. 2008) and new methods that employ heuristic algorithms that use objective criteria for partitioning data for use with site-homogeneous models (Lanfear et al. 2012, 2014). CAT models, implemented in PhyloBayes under a Bayesian framework (Lartillot and Philippe 2004), are by far the most widely used site-heterogeneous models (Brinkmann and Philippe 2008; Delsuc et al. 2008; Liu et al. 2009; Philippe et al. 2009; Finet et al. 2010; Philippe et al. 2011a; Timme et al. 2012; Nesnidal et al. 2013; Nosenko et al. 2013; O'Hara et al. 2014; Siu-Ting et al. 2014; Chang et al. 2015; Kenny et al. 2015; Luo 2015; Whelan et al. 2015b; Cannon et al. 2016; Rouse et al. 2016), but site-heterogeneous models can also be implemented in a maximum likelihood framework (Le et al. 2008). Site-heterogeneous CAT models implemented in PhyloBayes should not be confused with the CAT estimation for modeling rate heterogeneity in RAxML as these are two different models with different objectives (Stamatakis 2006; unless otherwise noted, when we refer to CAT models we mean site-heterogeneous substitution models implemented in PhyloBayes). CAT models employ a Dirichlet process prior to allow for multiple categories of substitution profiles with different nucleotide or amino acid frequencies coupled with a single set of exchange rates for the entire sequence dataset that is either fixed to flat values (i.e., CAT-F81; this model is often referred to simply as "CAT") or inferred from data (i.e., CAT-GTR). These multiple substitution categories are used to account for substitutional heterogeneity across a phylogenetic data matrix. As infinite mixture models, CAT models are considerably more complex than site-homogeneous, fixed amino acid substitution models. However, model complexity should not be taken as a guarantee of quality in empirical analyses because sometimes data are of insufficient quality or quantity to allow for accurate parameter estimation (Ludwig and Walters 1986; Gan et al. 1997).

An alternative method to site-heterogeneous models for handling substitutional heterogeneity in a dataset is to apply different site-homogeneous models (e.g., WAG) to different partitions of a supermatrix. One common approach is to assign a different model to each gene (e.g., Moore et al. 2010; Cannon et al. 2014; Moroz et al. 2014), but this may result in overpartitioning and poor parameter estimation (Kainer and Lanfear 2015). An objective measure for grouping genes into partitions is testing best-fit partitions and associated substitution models with a program like PartitionFinder (Lanfear et al. 2012, 2014). The downside to partitioning with site-homogeneous models is that genes, or other partitions such as codon positions, must be specified *a priori* and substitutional heterogeneity within pre-defined data segments cannot be modeled as with site-heterogeneous models. Both partitioning with site-homogeneous models and CAT models assume that the substitution process is homogeneous across lineages and that all taxa and genes share similar compositions.

CAT models have often been described as more “realistic” than other models (Lartillot and Philippe 2004; Liu et al. 2009; Tsagkogeorga et al. 2009; Finet et al. 2010; Philippe et al. 2011b; Nosenko et al. 2013), and they have also been described as the most significant area of recent progress in mitigating systematic error in phylogenetics (Telford et al. 2015). In some previous studies, CAT models have inferred different trees than other models, and such trees inferred with CAT models have often matched preconceived notions of animal phylogeny (e.g., sponges as the sister group to other animals, Philippe et al. 2009, 2011a; Pick et al. 2010; platyzoan paraphyly, Egger et al. 2015). Such results have yielded conclusions that CAT models suppress long-branch attraction (LBA) better than other models (Lartillot et al. 2007; Brinkmann and Philippe 2008; Philippe et al. 2011b; Roure et al. 2013). However, these statements about CAT model performance are based on assumptions about the true tree, many of which are tenuous. For example, both Brinkmann and Philippe (2008) and Philippe et al. (2011a) argued that CAT models perform better than site-homogeneous models in suppressing LBA because of where *Xenoturbella* and acoels were recovered by CAT models. However, these papers recovered *Xenoturbella* and acoels in different positions, both of which conflict with that of more recent studies placing *Xenoturbella* and acoels in a clade sister to all other bilaterians (Simakov et al. 2015; Cannon et al. 2016). The tendency of CAT models to recover sponges as the sister group to all other animals and cnidarians as the sister group to ctenophores has also been used to argue that CAT models are more accurate at inferring phylogeny and that they suppress LBA better than other models (Philippe et al. 2009, 2011b; Roure et al. 2013). However, the underlying assumption from these studies that the supposed accuracy of the CAT model relies on—sponges as the sister group to other animals and cnidarians as the sister group to ctenophores—is a hypothesis that has now been contradicted by numerous studies including some that employed CAT models (Dunn et al. 2008; Hejnol et al. 2009; Nesnidal et al. 2013; Ryan et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Chang et al. 2015; Whelan et al. 2015b).

Part of the argument concerning superior CAT model performance made by Philippe et al. (2011b) included the observation that CAT models appeared to fit data better than a single WAG model applied to the entire dataset, but a model test or analysis comparing CAT models to partitioning with site-homogeneous models was not done. Other authors have also noted that CAT models appear to fit empirical datasets better than any single site-homogeneous model applied to an entire dataset (Philippe et al. 2009, 2011b; Nosenko et al. 2013; Pisani et al. 2015; Tarver et al. 2016), but such tests of model fit ignore the fact that most studies that use site-homogeneous models couple their use with partitioning. Therefore, we view such tests of model fit as unhelpful for assessing model performance, especially because using a single site-homogeneous model on a large sequence dataset can result in less accurate trees than when partitioning is employed (Brown and Lemmon 2007; Kainer and Lanfear 2015). Rather than using assumptions about animal phylogeny to assess the effectiveness of CAT models for inferring relationships, simulation analyses should be used to compare performance of CAT models and partitioning with site-homogeneous models. Performance comparisons between CAT-F81 and CAT-GTR are also needed. CAT-GTR is probably a more reasonable model than CAT-F81 because CAT-F81 applies equal exchange rates for nucleotides or amino acids across all categories, but analyses done with CAT-GTR require much more CPU time than CAT-F81 (Lartillot et al. 2013). This is presumably why many studies have used CAT-F81 instead of CAT-GTR (e.g., Philippe et al. 2009, 2011b; Pisani et al. 2012; Nesnidal et al. 2013; Nosenko et al. 2013; Chang et al. 2015). Nevertheless, practical implications of applying equal exchange rates are unclear, and they warrant investigation. Furthermore, studies are needed to explore whether CAT models accurately assess substitutional heterogeneity in a dataset. Such analyses can only be done in simulation when the true number of substitutional categories is known. Simulation studies are a powerful tool for testing model and software performance (Hillis and Bull 1993; Huelsenbeck 1995a, 1995b; Guindon and Gascuel 2003; Williams and Moret 2003; Stamatakis et al. 2004), but only CAT-F81 has been tested in simulation (Holder et al. 2008). Comparisons of how well data partitioning with site-homogenous models, CAT-F81, and CAT-GTR infer accurate phylogenies in simulation are needed to better characterize model performance.

We tested the performance of CAT models, as currently implemented in PhyloBayes MPI (Lartillot et al. 2013), compared to data partitioning with site-homogeneous models (herein referred to as simply "partitioning") in simulation with known trees to determine if current implementations of CAT models actually perform better than other techniques at inferring accurate trees. We chose to explore performance of CAT models instead of other site-heterogeneous models (e.g., Pagel and Meade 2004; Le et al. 2008) because CAT models implemented in PhyloBayes are widely used, and they may be the only available site-heterogeneous models computationally efficient enough to run on phylogenomic-scale datasets. In particular, we explored model performance on datasets susceptible to LBA and assessed how well each method characterizes substitutional heterogeneity. Simulated data may never be as complex as empirical data, and so we also characterized the performance of CAT models versus data partitioning by reanalyzing the metazoan datasets of Philippe et al. (2009) and Nosenko et al. (2013), both of which have been hypothesized to suffer from LBA.

Methods

All datasets generated for this study consist of amino acid sequences. We utilized amino acid sequences because most studies using CAT models have employed amino acids. Simulated datasets were generated using site-homogeneous and site-heterogeneous models. Relatively small-sized simulated datasets were used to explore model performance over a large region of tree space, whereas large-sized datasets were utilized to more be representative of empirical data susceptible to LBA. Although areas of tree and model space not covered by analyses herein surely exist (e.g., variable rate heterogeneity among lineages), this study is an important step in better characterizing the relative performance of CAT models and dataset partitioning in tree reconstruction. In total, 53 datasets were generated and analyzed with an unpartitioned amino acid GTR model, partitioning, CAT-F81, and CAT-GTR. Over 670,000 CPU hours (~76.5 CPU years) were used for this study. Although analyses we performed may not have been exhaustive, they explored model performance under a wide variety of conditions that are applicable to empirical studies. All datasets, partitioning schemes, and inferred trees have been placed on Dryad (available at http://dx.doi.org/10.5061/dryad.85b2m). All code used in this study can be found at github.com/nathanwhelan. Table 1 provides an overview of the 53 simulated datasets and naming conventions.

Generating Small, Simulated DataSets

The general approach to simulations is shown in Figure 1. Using the phytools R package (Revell 2012; R Core Development Team 2015), we generated two stochastic, pure-birth trees with 10 taxa (Fig. 2) and two pure-birth trees with 50 taxa (Fig. 3). Six datasets with different starting gene lengths (ranging from 500 to 3500 amino acids; Table 1) for each 10-taxon tree and four datasets for each 50-taxon tree with different starting gene lengths (ranging from 500 to 1500 amino acids; Table 1) were simulated using the programs indel-Seq-Gen 2.0 (Strope et al. 2009) and INDELible (Fletcher and Yang 2009). Initially, we planned on only using indel-Seq-Gen because it has a well-described indel model (Strope et al. 2009), includes a variety of amino acid models implemented in commonly used phylogenetics programs (e.g., RAxML), and it has been recently updated. However, initial analyses showed that indel-Seq-Gen 2.0 failed to incorporate rate heterogeneity in amino acid datasets, hence datasets were also generated using INDELible. Thus, two datasets were simulated for each tree, starting gene length, and associated indel probability (Table 1), but only datasets simulated with INDELible had a gamma distribution of rate heterogeneity incorporated into data. Datasets generated with indel-Seq-Gen are labeled with "_iSG," and datasets generated with INDELible are labeled with "_I" (Table 1).

Each small dataset consisted of five genes simulated under a different substitution model [i.e., WAG (Whelan and Goldman 2001), MTREV (Adachi and Hasegawa 1996), BLOSUM62 (Henikoff and Henikoff 1992), JTT (Jones et al. 1994), and DCmut (Kosiol and Goldman 2005)]. This resulted in datasets with five different substitutional categories. Five genes were simulated to create a reasonable amount of heterogeneity while being small enough to allow for reasonable computational times. All genes were simulated with identical amino acid composition. For datasets simulated with INDELible, a discrete gamma distribution with four categories ($\alpha = 1.0$) of rate heterogeneity were incorporated into each model. An alpha value of 1.0 was chosen because it represents a moderate level of rate heterogeneity that both CAT models implemented in PhyloBayes and partitioning in RAxML should be able to model similarly as they both use a gamma distribution to model rate heterogeneity. Two of the six datasets generated on each 10-taxon tree (Fig. 2) and two of the four datasets generated on each 50-taxon tree (Fig. 3) were simulated with indels using indel probabilities of either 0.01 or 0.1 and a maximum length of 5 amino acids following the indel model described by (Strope et al. 2009) for indel-Seq-Gen or a Zipfian distribution with an alpha value of 0.01 or 0.1 and maximum length of 5 amino acids following the indel model described by (Fletcher and Yang 2009, see Table 1). Generating some datasets with indels was undertaken to explore model performance under conditions similar to empirical datasets. Furthermore, indels introduced missing data into dataset, which may also affect model performance. Any amino acid matrix column output by indel-Seq-Gen or INDELible that contained only gaps was removed. The respective simulation programs

TABLE 1. Simulation conditions and dataset characteristics

Dataset[a,b]	Tree[c]	Number of taxa	Gene starting length	Indel probability	Number of genes	Number of sites	Gaps/missing data (%)
S10.T1.1_iSG	10T_1	10	500	0	5	2,500	0
S10.T1.2_iSG	10T_1	10	500	0.01	5	2,463	0.93
S10.T1.3_iSG	10T_1	10	500	0.1	5	2,516	10.25
S10.T1.4_iSG	10T_1	10	1,500	0	5	7,500	0
S10.T1.5_iSG	10T_1	10	2,500	0	5	12,500	0
S10.T1.6_iSG	10T_1	10	3,500	0	5	17,500	0
S10.T2.1_iSG	10T_2	10	500	0	5	2,500	0
S10.T2.2_iSG	10T_2	10	500	0.01	5	2,500	0.86
S10.T2.3_iSG	10T_2	10	500	0.1	5	2,498	10.47
S10.T2.4_iSG	10T_2	10	1,500	0	5	7,500	0
S10.T2.5_iSG	10T_2	10	2,500	0	5	12,500	0
S10.T2.6_iSG	10T_2	10	3,500	0	5	17,500	0
S50.T1.1_iSG	50T_1	50	500	0	5	2,500	0
S50.T1.2_iSG	50T_1	50	500	0.01	5	2,687	8.17
S50.T1.3_iSG	50T_1	50	500	0.1	5	4,153	46.63
S50.T1.4_iSG	50T_1	50	1,500	0	5	7,500	0
S50.T2.1_iSG	50T_2	50	500	0	5	2,500	0
S50.T2.2_iSG	50T_2	50	500	0.01	5	2,841	14.68
S50.T2.3_iSG	50T_2	50	500	0.1	5	4,775	59.69
S50.T2.4_iSG	50T_2	50	1,500	0	5	7,500	0
L13.TA.1_iSG	A	13	Random	0.01	200	86,097	5.63
L13.TA.2_iSG	A	13	Random	0.01	200	45,957	5.45
L13.TA.3_iSG	A	13	Random	0.01	200	23,232	5.46
L13.TA.4_iSG	A	13	Random	0.1	200	69,390	5.65
L19.TB.1_iSG	B	19	Random	0.01	200	76,625	4.16
L19.TB.2_iSG	B	19	Random	0.1	200	86,343	29.23
S10.T1.1_I	10T_1	10	500	0	5	2,500	0
S10.T1.2_I	10T_1	10	500	0.01	5	2,481	1.25
S10.T1.3_I	10T_1	10	500	0.1	5	2,479	10.12
S10.T1.4_I	10T_1	10	1,500	0	5	7,500	0
S10.T1.5_I	10T_1	10	2,500	0	5	12,500	0
S10.T1.6_I	10T_1	10	3,500	0	5	17,500	0
S10.T2.1_I	10T_2	10	500	0	5	2,500	0
S10.T2.2_I	10T_2	10	500	0.01	5	2,505	0.99
S10.T2.3_I	10T_2	10	500	0.1	5	2,511	10.09
S10.T2.4_I	10T_2	10	1,500	0	5	7,500	0
S10.T2.5_I	10T_2	10	2,500	0	5	12,500	0
S10.T2.6_I	10T_2	10	3,500	0	5	17,500	0
S50.T1.1_I	50T_1	50	500	0	5	2,500	0
S50.T1.2_I	50T_1	50	500	0.01	5	2,681	8.21
S50.T1.3_I	50T_1	50	500	0.1	5	4,234	48.52
S50.T1.4_I	50T_1	50	1,500	0	5	7,500	0
S50.T2.1_I	50T_2	50	500	0	5	2,500	0
S50.T2.2_I	50T_2	50	500	0.01	5	2,868	15.02
S50.T2.3_I	50T_2	50	500	0.1	5	4,838	60.49
S50.T2.4_I	50T_2	50	1,500	0	5	7,500	0
L13.TA.1_I	A	13	Random	0.01	200	81,099	0.62
L13.TA.2_I	A	13	Random	0.01	200	40,313	0.64
L13.TA.3_I	A	13	Random	0.01	200	21,546	0.62
L13.TA.4_I	A	13	Random	0.1	200	69,362	5.78
L19.TB.1_I	B	19	Random	0.01	200	76,792	4.53
L19.TB.2_I	B	19	Random	0.1	200	86,667	30.23

[a] Datasets are named with their dataset size (“S” or “L”), number of taxa, tree identity, and dataset number.

[b] _iSG and _I indicate datasets simulated with indel-seq-Gen and INDELBILE, respectively.

[c] See Figs. 2–4.

generated alignments, so alignment uncertainty was not a factor in downstream analyses. All five genes for each dataset were subsequently concatenated into a single supermatrix using FASconCAT (Kück and Meusemann 2010).

The above datasets were designed to have substitutional heterogeneity across genes, but not within genes. In empirical datasets, intra-gene heterogeneity may be pervasive (Betts and Russell 2003). This is the primary argument for using site-heterogeneous models instead of partitioning (Lartillot and Philippe 2004). However, how accounting for only inter-gene heterogeneity, compared to accounting for inter- and intra-gene heterogeneity, affects tree inference is unclear. Thus, we generated another suite of datasets for which intra-gene substitutional heterogeneity was

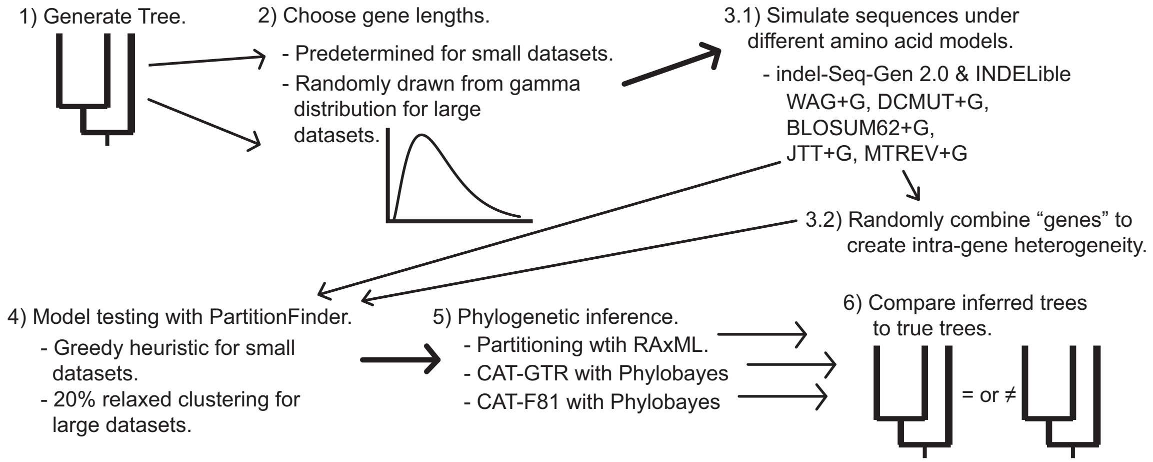
1) Generate Tree.
2) Choose gene lengths.
- Predetermined for small datasets.
- Randomly drawn from gamma distribution for large datasets.
3.1) Simulate sequences under different amino acid models.
- indel-Seq-Gen 2.0 & INDELible
WAG+G, DCMUT+G, BLOSUM62+G, JTT+G, MTREV+G
3.2) Randomly combine "genes" to create intra-gene heterogeneity.
4) Model testing with PartitionFinder.
- Greedy heuristic for small datasets.
- 20% relaxed clustering for large datasets.
5) Phylogenetic inference.
- Partitioning wtih RAxML.
- CAT-GTR with Phylobayes
- CAT-F81 with Phylobayes
6) Compare inferred trees to true trees.
= or ≠

FIGURE 1. Visual schematic of how data were simulated and analyzed.

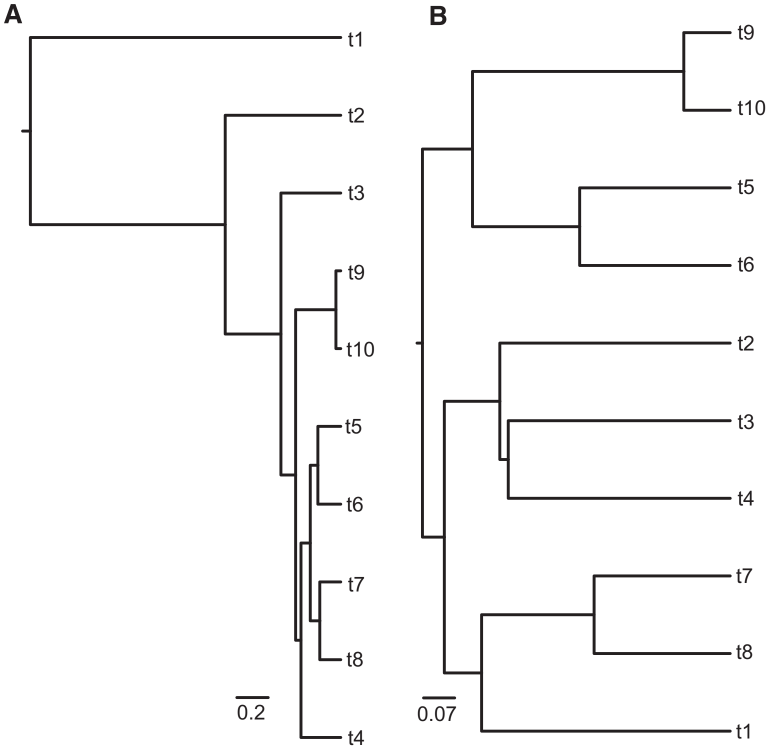

FIGURE 2. Topologies used for simulating 10-taxon amino acid sequence datasets. A) Datasets S10.T1.1-6. B) Datasets S10.T2.1-6 (see Table 1).

simulated by taking each dataset generated above and twice randomly combining two genes simulated under different models. This resulted in three genes: two that were the result of combining two genes simulated under different models and one leftover gene without intra-gene substitutional heterogeneity. The combined genes were treated as a single data block for model testing and phylogenetic inference with partitioning (see below). Amino acid positions in these datasets are identical to datasets that they were generated from, but starting gene blocks used in ParitionFinder were combined. Datasets with intra-gene heterogeneity are scenarios where CAT models would be expected to perform better than partitioning because partitioning is blind to heterogeneity within genes. At the same time, by not creating datasets from scratch we saved on computational time, and we could directly compare partitioning when there was only inter-gene heterogeneity versus when intra-gene heterogeneity existed.

Generating Large, Simulated DataSets

In addition to small datasets, we simulated datasets designed to be more representative of phylogenomic sized studies (i.e., many genes of different lengths

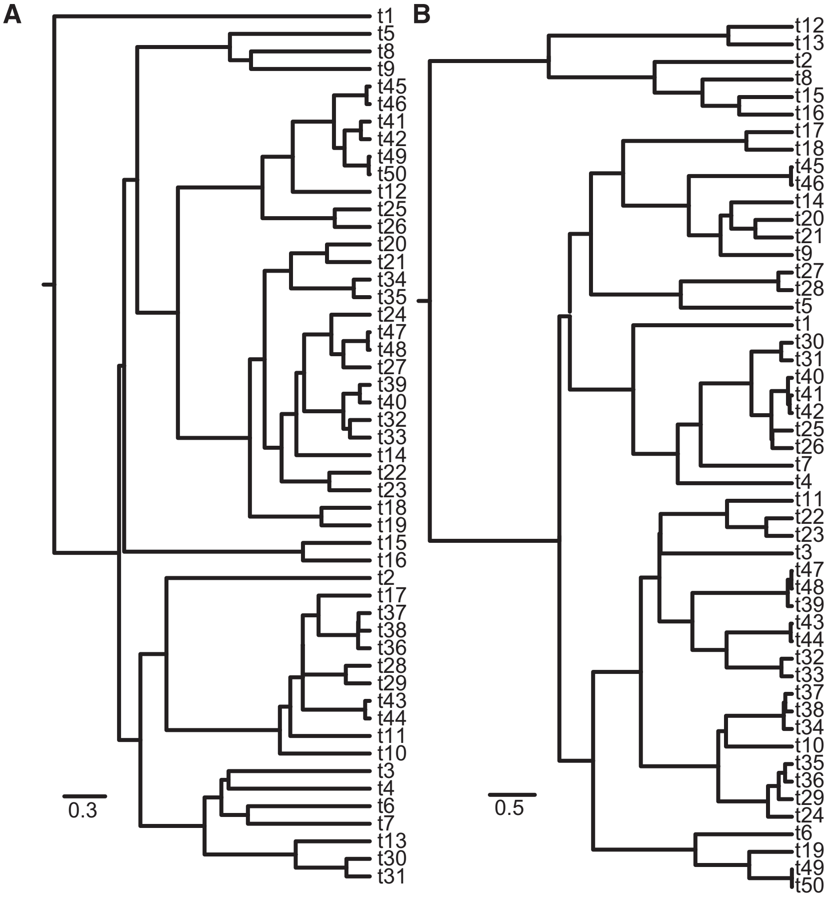

FIGURE 3. Topologies used for simulating 50-taxon amino acid sequence datasets. A) Datasets S50.T1.1-4. B) Datasets S50.T2.1-4 (see Table 1).

and tens of thousands of amino acids) because CAT models have been forcefully advocated for use with large datasets (Brinkmann and Philippe 2008; Philippe et al. 2009, 2011a,b; Nosenko et al. 2013; Pisani et al. 2015; Tarver et al. 2016). Trees and datasets generated for this second set of simulations were designed to test the supposed superior performance of CAT models in suppressing LBA (Lartillot et al. 2007; Philippe et al. 2011b). The first tree used for large dataset simulations, herein referred to as "Tree A" (Fig. 4A), was generated by reconstructing a tree from the insect 18S rRNA dataset of Huelsenbeck (1997), which suffers from LBA when analyzed under maximum parsimony (MP). 18S sequences were retrieved from GenBank (Supplementary Table S1 available on Dryad), and a tree was inferred with PhyML 3.0 (Guindon et al. 2010) using the HKY model as in Huelsenbeck (1997). To avoid confusing real 18S sequences with simulated data, we mapped terminal node names to single letter identifiers as in Figure 4A, but corresponding species names and 18S GenBank numbers are in Supplementary Table S1, available on Dryad. For the second tree, herein referred to as "Tree B," we created a nonrandom, hypothetical phylogeny with 19 tips, a long-branched outgroup, short internode branches, and two long internal branches (Fig. 4B).

Amino acids were separately simulated for datasets L13.TA.1, L13.TA.4, L19.TB.1, and L19.TB.2 with both indel-Seq-Gen and INDELible (Table 1). Datasets L13.TA.1 and L13.TA.4 were simulated on Tree A, and datasets L13.TA.2 and L13.TA.3 were generated by randomly removing 50% and 75% of sequences from dataset L13.TA.1, respectively. This was done to explore model performance with different sized datasets, and to further assess whether PartitionFinder and CAT models accurately categorize substitutional heterogeneity or if the number of partitions/categories determined by PartitionFinder and CAT models simply scales with dataset size. One characteristic of empirical datasets we wished to mimic was variable gene length. To do this, 200 "gene" lengths were randomly drawn from a gamma distribution with a shape parameter of 2.5 and scale parameter of 150. Individual sequence lengths were placed into corresponding input files that were formatted for indel-Seq-Gen 2.0 and INDELible. Both steps were done with a custom Python script (github.com/nathanwhelan/generateSequences), and they were done for each of the four full-sized, simulated datasets (Table 1).

As with small datasets, indel-Seq-Gen 2.0 and INDELible were used to simulate amino acid sequences. For each full-sized, large dataset, five substitution

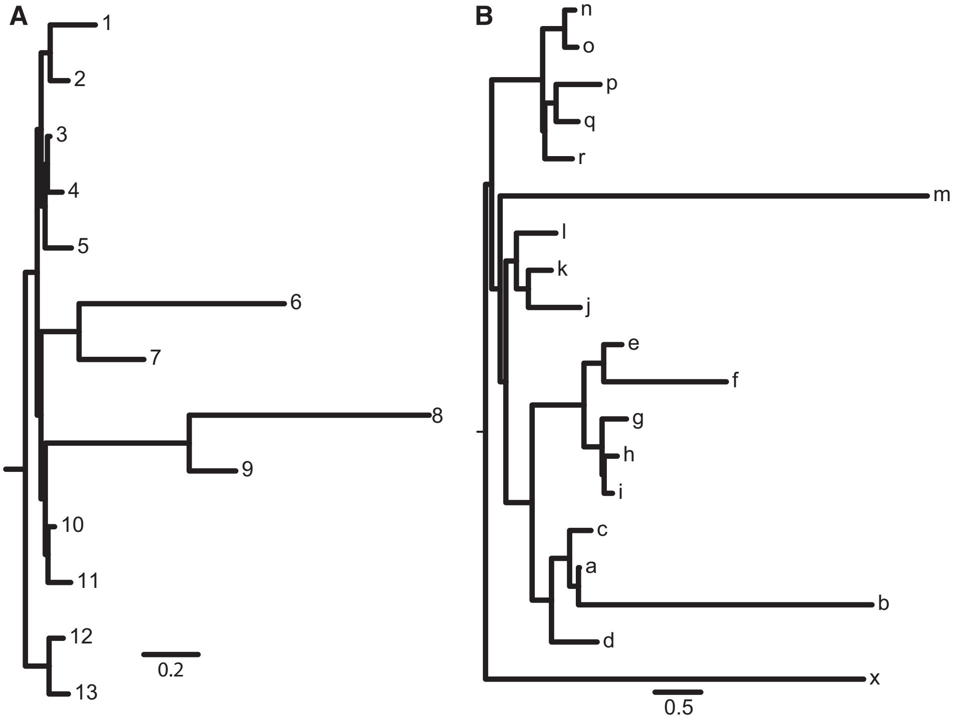

FIGURE 4. Topologies used for simulating large amino acid sequence datasets. A) Datasets L13.TA.1-4. B) Datasets L19.TB.1-2. (see Table 1).

models were used to generate amino acid sequences on known trees by randomly assigning 40 genes to be simulated under the WAG, MTREV, BLOSUM62, JTT, and DCmut substitution models, respectively. All genes were simulated with identical amino acid composition. This resulted in five different substitutional categories for each dataset. Four categories of a discrete gamma distribution ($\alpha = 1.0$) of rate heterogeneity were incorporated into each model for datasets generated with INDELible. Large datasets were designed to have characteristics of empirical datasets, so each dataset had indels, and they were simulated similarly to small datasets (see Table 1).

Like the small datasets, we simulated large datasets with only inter-gene heterogeneity as well as datasets with intra-gene heterogeneity, but for the latter, three genes were randomly combined instead of two as with the small datasets; this was done 66 times without replacement and the remaining two genes were also combined. We also attempted to simulate data under CAT-GTR. Simulating data under the CAT-GTR model is done with Bayesian posterior predictive simulation, which uses empirical data as a starting point so CAT-GTR can infer substitutional heterogeneity present in the dataset for use as a simulation parameter. A major caveat of this approach is the assumption that CAT-GTR accurately captures the real level of heterogeneity in any given dataset, but this may not always be a valid assumption (see below). Data simulated with CAT-GTR failed to produce well-supported, accurate trees, even when using CAT-GTR for tree inference. Therefore, the usefulness of data we generated under CAT-GTR in comparing model performance is limited. Additional details concerning data simulated under CAT-GTR and associated results are discussed in Supplementary Material available on Dryad.

Empirical DataSets

We reanalyzed two metazoan datasets from Philippe et al. (2009) and Nosenko et al. (2013). We retrieved the Philippe et al. (2009) dataset with all outgroups included and the Nosenko et al. (2013) dataset with both ribosomal and non-ribosomal proteins included from the original publications (details in Supplementary Material available on Dryad). These super matrices were split into individual genes with a custom R script (github.com/nathanwhelan/Split_supermatrix_into_partitions) using annotations provided in the original studies. We chose to reexamine Philippe et al. (2009) because Philippe et al. (2011b) used this dataset to purportedly demonstrate that CAT models are better than site-homogeneous models, but it was only analyzed with CAT-F81 and a single WAG model, not CAT-GTR or partitioning. Nosenko et al. (2013) was reanalyzed because it was originally analyzed with both CAT models but not with data partitioning. Furthermore, both studies recovered trees with CAT-F81 that placed sponges as the sister group to all other animals and ctenophores as the sister group to cnidarians, which are hypotheses that have recently been contradicted by numerous studies (Ryan et al. 2013;

Moroz et al. 2014; Borowiec et al. 2015; Chang et al. 2015; Whelan et al. 2015b). Unlike studies on other animal groups where CAT models and site-homogeneous models have inferred different topologies [e.g., among lophotrochozoans (Egger et al. 2015), placement of acoels and *Xenoturbella* (Brinkmann and Philippe 2008; Philippe et al. 2011a; Cannon et al. 2016; Rouse et al. 2016)] an emerging consensus has developed concerning relationships among non-bilaterian phyla (Ryan et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Chang et al. 2015; Whelan et al. 2015a,b). Even though a recent reanalysis of some small datasets disputed the finding of ctenophores as the sister group to other animals (Pisani et al. 2015, but see Halanych et al. 2016), a reanalysis of Philippe et al. (2009) and Nosenko et al. (2013) in the context of CAT models versus partitioning is warranted as no recent study has recovered ctenophores and cnidarians sister to each other as in Philippe et al. (2009) and Nosenko et al. (2013).

Model Testing and Phylogenetic Inference

Both CAT models are site-heterogeneous models, so if they perform well in modeling substitutional heterogeneity then we expect them to accurately capture the heterogeneity as well as, or better than, the best-fit partitions inferred with PartitionFinder. Furthermore, models used to simulate data can be described with a GTR model so we expect CAT-GTR to perform as well as, or better than, partitioning. Partitioning can be expected to perform worse than CAT models when using genes that had intra-gene substitutional heterogeneity because partitioning was blind to some of the heterogeneity present.

For each dataset, the potential of LBA was examined using MP tree inference as this method is more susceptible to LBA than model-based methods (Felsenstein 1978; Huelsenbeck 1997). In other words, MP served as a control to determine if datasets were susceptible to LBA, an issue of model performance we wished to explore. MP analyses were done with PAUP* (people.sc.fsu.edu/d̃swofford/paup_test/). MP trees were inferred with starting trees obtained through stepwise addition and 10 replicates of random sequence-addition. If more than one most parsimonious tree was recovered, a strict consensus of equally parsimonious trees was generated after each analysis. Nodal support was measured with 100 bootstrap (BS) replicates using the same search parameters as the full MP search.

Even though exact models (i.e., WAG, BLOSUM62, JTT, DCmut, MTREV) and number of appropriate partitions (i.e., five) that datasets were simulated under was known, we wanted to avoid biasing our analyses toward any one model or method. Therefore, PartitionFinder 1.1.1 (Lanfear et al. 2012, 2014) was used to determine inferred best-fit partitions and models for each dataset. Models tested by PartitionFinder were WAG$+\Gamma$, BLOSUM62$+\Gamma$, MTREV$+\Gamma$, MTMAM$+\Gamma$, RTREV$+\Gamma$, CPREV$+\Gamma$, DCmut$+\Gamma$, DAYHOFF$+\Gamma$, JTT$+\Gamma$, LG$+\Gamma$, VT$+\Gamma$, and corresponding models with empirically derived amino acid frequencies (i.e., + F). PartitionFinder requires starting blocks, or genes, to be specified and then tests which, if any, genes should be combined for use with the same substitution model. For analyses of datasets with only inter-gene heterogeneity, individual genes served as starting blocks, but for analyses of datasets with intra-gene heterogeneity we used combined genes (described above) as starting blocks. All PartitionFinder analyses used linked branch lengths. The greedy search algorithm in PartitionFinder was used for small datasets. Twenty percent relaxed clustering, which is a faster alternative to the greedy search in PartitionFinder designed for phylogenomic scale datasets (Lanfear et al. 2014), was used for the large simulated datasets, datasets simulated with CAT-GTR, and empirical datasets.

The AVX executable of RAxML 8.2.4 Stamatakis (2014) was used for ML analyses with best-fit partitions and models inferred by PartitionFinder, a gamma distribution with four discrete categories to model rate heterogeneity, and 100 fast BS replicates to assess nodal support. For each dataset, three ML analyses were done: 1) with a single protein GTR model applied to the entire amino acid matrix, 2) partitioning scheme resulting from PartitionFinder analysis using gene blocks lacking intra-gene heterogeneity, and 3) partitioning scheme resulting from PartitionFinder analysis using gene blocks with intra-gene heterogeneity. Using a single protein GTR model applied to the whole dataset was done to explore the practical effect of completely ignoring substitutional heterogeneity. Even though ML analyses can become stuck in local optima, we only ran RAxML once as is typically done in empirical studies (Moroz et al. 2014; Whelan et al. 2015b).

We used PhyloBayes MPI 1.5a (Lartillot et al. 2013) for Bayesian inference, and trees for each dataset were inferred with both CAT-GTR and CAT-F81 models. When CAT-F81 recovered a worse branching pattern than other model-based methods, we also analyzed those datasets in PhyloBayes with a single F81 model. This was done to explore whether failures of CAT-F81 were a result of problems with CAT or with applying a F81 model to datasets. Four discrete gamma categories were employed in each PhyloBayes analysis to model rate heterogeneity. Default priors were used because PhyloBayes MPI does not let users modify priors. For each analysis, two chains were run, but additional chains were used if one appeared stuck in a local optima based on trace plots viewed in Tracer (Rambaut and Drummond 2007). Burn-in for each BI analysis was assessed with Tracer by determining when chains had reached stationarity (Table 2). Convergence of Bayesian runs cannot be known with certainty, but apparent convergence was considered to have occurred when the post burn-in maximum difference in topologies among chains and relative difference of parameters among chains was below 0.3 and when the effective sample size for each parameter was greater than 50 as measured by bpcomp and tracecomp (Lartillot et al.

Table 2. Burn-in and total MCMC generations for CAT-GTR and CAT-G81 analyses

	CAT-F81		CAT-GTR	
Dataset	Burn-in	Generations per chain	Burn-in	Generations per chain
S10.T1.1_iSG	1,000	39,218	1,000	3,637
S10.T1.2_iSG	1,000	62,456	1,000	23,537
S10.T1.3_iSG	1,000	6,234	1,000	17,771
S10.T1.4_iSG	500	6,071	1,500	5,498
S10.T1.5_iSG	5,000	27,223	1,000	7,040
S10.T1.6_iSG	5,000	15,633	16,000	17,636
S10.T2.1_iSG	10,000	67,354	500	20,111
S10.T2.2_iSG	20,000	48,099	10,000	35,304
S10.T2.3_iSG	20,000	67,390	10,000	23,538
S10.T2.4_iSG	10,000	53,258	2,000	16,911
S10.T2.5_iSG	10,000	26,210	16,000	22,970
S10.T2.6_iSG	10,000	32,025	6,000	16,358
S50.T1.1_iSG	10,000	31,747	1,000	10,619
S50.T1.2_iSG	10,000	61,280	1,000	11,326
S50.T1.3_iSG	40,000	70,841	20,000	56,404
S50.T1.4_iSG	1,500	9,056	5,000	11,533
S50.T2.1_iSG	15,000	20,465	1,000	38,444
S50.T2.2_iSG	10,000	29,081	8,000	13,796
S50.T2.3_iSG	15,000	26,078	10,000	31,036
S50.T2.4_iSG	10,000	34,304	5,000	32,164
L13.TA.1_iSG	8,000	36,945	2,500	17,048
L13.TA.2_iSG	2,500	24,265	1,000	21,015
L13.TA.3_iSG	1,000	34,945	2,000	11,658
L13.TA.4_iSG	2,500	10,968	5,000	20,320
L19.TB.1_iSG	6,000	32,665	10,000	21,417
L19.TB.2_iSG	7,500	35,291	5,000	47,241
S10.T1.1_I	5,000	30,000	5,000	44,777
S10.T1.2_I	10,000	32,513	35,000	45,546
S10.T1.3_I	8,000	32,812	8,000	9,676
S10.T1.4_I	17,000	21,067	8,000	16,872
S10.T1.5_I	5,000	25,345	15,000	24,559
S10.T1.6_I	10,000	21,652	5,000	14,379
S10.T2.1_I	10,000	27,927	25,000	42,226
S10.T2.2_I	20,000	75,665	5,000	17,276
S10.T2.3_I	10,000	15,600	20,000	33,386
S10.T2.4_I	13,000	23,459	15,000	27,236
S10.T2.5_I	13,000	22,111	13,000	17,726
S10.T2.6_I	20,000	26,742	4,500	14,100
S50.T1.1_I	10,000	26,559	5,000	19,403
S50.T1.2_I	10,000	20,421	6,090	16,801
S50.T1.3_I	20,000	70,601	15,000	44,481
S50.T1.4_I	8,200	12,145	2,500	10,146
S50.T2.1_I	45,000	47,907	5,000	13,092
S50.T2.2_I	25,000	33,496	5,000	10,388
S50.T2.3_I	61,000	113,368	15,000	42,576
S50.T2.4_I	18,000	21,451	25,000	37,446
L13.TA.1_I	10,000	16,336	3,000	7,923
L13.TA.2_I	2,000	20,639	6,100	11,269
L13.TA.3_I	25,000	60,841	2,000	27,429
L13.TA.4_I	6,000	11,999	3,000	6,956
L19.TB.1_I	15,000	33,933	5,000	25,757
L19.TB.2_I	12,000	25,839	4,000	10,358
Nosenko et al.	1,000	4,633	7,200	14,985
Philippe et al.	1,000	12,267	1,500	14,000

2013). For each BI analysis, once convergence appeared to have been reached, a majority rule consensus of the post-burin posterior distribution of trees was generated using bpcomp and nodal support was measured by posterior probability (PP).

Our goal was not to compare PhyloBayes versus RAxML, *per se*, but limited overlap in what methods could be used in any given phylogenetic program necessitated practical considerations about which programs to use. For example, PhyloBayes is the only program that implements site-heterogeneous CAT models, but PhyloBayes does not permit data partitioning. Furthermore, PhyloBayes and RAxML are, by far, the most widely used programs for phylogenomics so conclusions made using trees inferred with these programs are broadly applicable to the field.

To determine if substitutional heterogeneity was being correctly inferred, the number of categories inferred by CAT-F81 and CAT-GTR was recorded after each PhyloBayes analysis by calculating the average number of categories across post burn-in posterior distributions. Total partitions inferred by PartitionFinder were also documented. We plotted inferred categories/partitions versus number of overall amino acids (i.e., alignment columns multiplied by number of taxa minus gaps) for each dataset and performed a linear regression using R (R Core Development Team 2015) to determine how inferred partitions/categories corresponded to the amount of data and actual number of substitutional categories. For CAT models, the number of categories for all simulated datasets was analyzed at once, but for partitioning, small and large simulated datasets were analyzed separately because two different methods were used within PartitionFinder based on dataset size (i.e., greedy search for small datasets, 20% relaxed clustering for large datasets). Datasets simulated with indel-Seq-Gen and INDELible were also explored separately to determine if inferred number of categories/partitions differed when a gamma distribution of rate heterogeneity was incorporated in data simulations.

Accuracy of phylogenetic inference with simulated data was measured with normalized Robinson–Foulds (RF) distances (Robinson and Foulds 1981; Kupczok et al. 2010) and the branch-score distance metric of Kuhner and Felsenstein (1994). Raw RF distances between inferred trees and true trees were calculated with the R package Phangorn (Schliep 2011) and normalized by dividing raw RF distances by 2*(n−3) where n is the number of taxa; only normalized RF distances are reported. Branch-score distances were directly calculated with Phangorn. An RF distance of zero indicates that the inferred tree matches the branching order of the known tree, whereas a branch-score distance of zero would indicate that the inferred tree had identical branch lengths and branching order as the known tree. Larger distances of each metric indicate less accurate trees. Branch-score distance was not calculated for MP trees because the branch length scale was not in substitutions per site as with other trees.

Computational time for each model-based phylogenetic analysis was tracked to quantify differences in computational demand. CPU time for Bayesian chains stuck in local optima was not recorded in an effort to limit bias in reported CPU time resulting from stochastic issues. Bayesian analyses were run on either the Dense Memory Cluster of the Alabama Super Computer (ASC) or on the CASIC High Performance Cluster at Auburn University. Individual Bayesian analyses were run with openMPI on either Intel Xeon E5-2660 2.20 GHz cores (CASIC), Intel Nehalem 2.26 GHz cores (ASC), or 2.30 GHz AMD Opteron Magny-Cours cores (ASC). PartitionFinder was run on the SkyNet cluster at Auburn University with Intel Xeon E7-4800 2.4GHz cores, and RAxML analyses were run on ASC with Xeon E5-4640 2.40 GHz cores. We note that required computational time using different processors is not exactly equivalent. However, we were most interested in comparing required computational time for analyses with CAT models in PhyloBayes versus partitioning with PartitionFinder and RAxML analyses. The observed differences in computational time were so great between methods that differences in processor performance are effectively negligible.

Results

Model Performance on Small DataSets

Maximum parsimony analyses of 10- and 50-taxon small datasets resulted in trees with correct and incorrect branching patterns according to RF distances (Table 3). At least some incorrect relationships inferred with MP appeared similar to what would be expected from LBA (Supplementary Figs. S1, S2A, S3A and Supplementary Material available on Dryad). Maximum likelihood phylogenetic inference with a single amino acid GTR model of small datasets generated with indel-Seq-Gen almost always performed worse than partitioning and CAT-GTR in inferring the true tree based on RF and branch-score distances (Table 3; Supplementary Material available on Dryad). Results were less clear when comparing tree inference accuracy with GTR to partitioning and CAT-GTR on datasets simulated with INDELible. A single GTR model never recovered a tree with a lower RF distance than partitioning, but GTR recovered a tree with a lower RF distance than the tree recovered with CAT-GTR on dataset S50.T1.2_I (Table 3). When branching patterns were identical among models for datasets simulated with INDELible, GTR often, but not always, performed better at inferring branch lengths than partitioning on datasets with among-gene substitutional heterogeneity based on branch-score distances (Table 3). Results with a single GTR model suggest that simulated datasets possessed at least some substitutional heterogeneity that was not adequately accounted for when analyzed with a single site-homogeneous substitution model.

PartitionFinder did not always recover the correct number of partitions (i.e., five) or correct models for small datasets, but it did generally perform better at inferring the correct number of partitions and correct models on datasets with more characters and taxa (Supplementary Table S2 and Supplementary Material available on Dryad). The number of partitions inferred by PartitionFinder for small datasets was significantly correlated with dataset size (indel-Seq-Gen: P=0.004; INDELible: P=0.005), but linear regression fit was poor (indel-Seq-Gen: $R^2 = 0.371$; INDELible: $R^2 = 0.391$; Fig. 5A, C; Supplementary Table S3 available on Dryad). Despite the correlation with increased dataset size, PartitionFinder inferred the correct number of substitutional categories for 28 of 40 small datasets (Fig. 5A, C; Supplementary Table S2 available on Dryad).

TABLE 3. Metrics comparing inferred trees to true tree (RF; branch score)

	CAT-GTR	CAT-F81	Partitioning	Partitioning, intra-gene	GTR	Parsimony	F81
S10.T1.1_iSG	0.000 ; 0.680	0.000 ; 0.497	0.000 ; 0.646	0.000 ; 0.807	0.000 ; 0.835	0.000 ; N/A	-
S10.T1.2_iSG	0.000 ; 0.847	0.000 ; 0.683	0.000 ; 0.933	0.000 ; 0.933	0.000 ; 0.918	0.000 ; N/A	-
S10.T1.3_iSG	0.000 ; 0.493	0.071 ; 1.15	0.000 ; 0.292	0.000 ; 0.292	0.000 ; 0.632	0.000 ; N/A	0.000 ; 1.26
S10.T1.4_iSG	0.000 ; 0.834	0.000 ; 1.51	0.000 ; 0.652	0.000 ; 0.953	0.000 ; 0.059	0.000 ; N/A	-
S10.T1.5_iSG	0.000 ; 0.884	0.000 ; 2.34	0.000 ; 0.674	0.000 ; 0.873	0.000 ; 0.979	0.000 ; N/A	-
S10.T1.6_iSG	0.000 ; 0.833	0.000 ; 2.30	0.000 ; 0.622	0.000 ; 0.918	0.000 ; 0.900	0.000 ; N/A	-
S10.T2.1_iSG	0.000 ; 0.164	0.000 ; 0.204	0.000 ; 0.179	0.000 ; 0.153	0.000 ; 0.190	0.000 ; N/A	-
S10.T2.2_iSG	0.000 ; 0.209	0.000 ; 0.363	0.000 ; 0.192	0.000 ; 0.206	0.000 ; 0.230	0.000 ; N/A	-
S10.T2.3_iSG	0.000 ; 0.230	0.000 ; 0.195	0.000 ; 0.218	0.000 ; 0.270	0.000 ; 0.139	0.000 ; N/A	-
S10.T2.4_iSG	0.000 ; 0.204	0.071 ; 0.167	0.000 ; 0.174	0.000 ; 0.201	0.000 ; 0.224	0.000 ; N/A	0.000 ; 0.309
S10.T2.5_iSG	0.000 ; 0.185	0.000 ; 1.95	0.000 ; 0.143	0.000 ; 0.185	0.000 ; 0.200	0.000 ; N/A	-
S10.T2.6_iSG	0.000 ; 0.190	0.000 ; 2.010	0.000 ; 0.155	0.000 ; 0.181	0.000 ; 0.206	0.000 ; N/A	-
S50.T1.1_iSG	0.021 ; 1.41	0.032 ; 1.54	0.021 ; 1.08	0.021 ; 1.25	0.021 ; 1.45	0.032 ; N/A	0.042 ; 2.13
S50.T1.2_iSG	0.011 ; 1.26	0.000 ; 1.328	0.000 ; 0 1.02	0.000 ; 1.17	0.021 ; 1.31	0.021 ; N/A	-
S50.T1.3_iSG	0.053 ; 1.50	0.053 ; 1.57	0.042 ; 1.36	0.042 ; 1.36	0.064 ; 1.53	0.085 ; N/A	0.042 ; 2.20
S50.T1.4_iSG	0.000 ; 1.33	0.021 ; 1.46	0.021 ; 1.02	0.000 ; 1.23	0.021 ; 1.95	0.021 ; N/A	0.021 ; 2.07
S50.T2.1_iSG	0.021 ; 1.90	0.053 ; 2.44	0.021 ; 1.28	0.021 ; 2.19	0.021 ; 1.95	0.021 ; N/A	0.021 ; 3.07
S50.T2.2_iSG	0.011 ; 1.91	0.032 ; 2.48	0.000 ; 1.27	0.021 ; 1.08	0.021 ; 1.95	0.032 ; N/A	0.032 ; 3.08
S50.T2.3_iSG	0.021 ; 1.99	0.021 ; 2.33	0.021 ; 1.43	0.021 ; 1.91	0.021 ; 2.04	0.042 ; N/A	0.000 ; 3.17
S50.T2.4_iSG	0.021 ; 1.86	0.032 ; 2.67	0.021 ; 1.26	0.021 ; 1.74	0.021 ; 1.95	0.032 ; N/A	0.021 ; 3.11
L13.TA.1_iSG	0.000 ; 0.260	0.000 ; 0.260	0.000 ; 0.228	0.000 ; 0.254	0.000 ; 0.275	0.100 ; N/A	-
L13.TA.2_iSG	0.000 ; 0.257	0.000 ; 0.179	0.000 ; 0.217	0.000 ; 0.239	0.000 ; 0.272	0.100 ; N/A	-
L13.TA.3_iSG	0.000 ; 0.265	0.000 ; 0.185	0.000 ; 0.214	0.000 ; 0.258	0.000 ; 0.275	0.100 ; N/A	-
L13.TA.4_iSG	0.000 ; 0.264	0.000 ; 0.182	0.000 ; 0.235	0.000 ; 0.263	0.000 ; 0.280	0.100 ; N/A	-
L19.TB.1_iSG	0.000 ; 3.25	0.000 ; 2.94	0.000 ; 0.276	0.000 ; 3.01	0.000 ; 3.37	0.125 ; N/A	-
L19.TB.2_iSG	0.000 ; 3.17	0.063 ; 2.41	0.000 ; 2.73	0.000 ; 3.00	0.000 ; 3.37	0.125 ; N/A	0.000 ; 4.26
S10.T1.1_I	0.000 ; 0.196	0.000 ; 1.01	0.000 ; 0.169	0.000 ; 0.063	0.000 ; 0.099	0.000 ; N/A	-
S10.T1.2_I	0.000 ; 0.211	0.000 ; 1.69	0.000 ; 0.285	0.000 ; 0.264	0.000 ; 0.285	0.000 ; N/A	-
S10.T1.3_I	0.000 ; 0.317	0.143 ; 1.88	0.000 ; 0.321	0.143 ; 0.606	0.143 ; 0.634	0.000 ; N/A	0.000 ; 0.757
S10.T1.4_I	0.000 ; 0.065	0.000 ; 1.69	0.000 ; 0.116	0.000 ; 0.133	0.000 ; 0.057	0.000 ; N/A	-
S10.T1.5_I	0.000 ; 0.064	0.000 ; 2.01	0.000 ; 0.072	0.000 ; 0.133	0.000 ; 0.078	0.000 ; N/A	-
S10.T1.6_I	0.000 ; 0.054	0.000 ; 2.26	0.000 ; 0.073	0.000 ; 0.177	0.000 ; 0.054	0.000 ; N/A	-
S10.T2.1_I	0.000 ; 0.084	0.000 ; 0.427	0.000 ; 0.100	0.000 ; 0.075	0.000 ; 0.080	0.000 ; N/A	-
S10.T2.2_I	0.143 ; 0.081	0.143 ; 0.274	0.143 ; 0.091	0.143; 0.119	0.143 ; 0.086	0.143 ; N/A	0.143 ; 0.280
S10.T2.3_I	0.000 ; 0.055	0.000 ; 0.245	0.000 ; 0.079	0.000 ; 0.061	0.000 ; 0.056	0.000 ; N/A	-
S10.T2.4_I	0.000 ; 0.043	0.000 ; 0.683	0.000 ; 0.054	0.000 ; 0.040	0.000 ; 0.041	0.000 ; N/A	-
S10.T2.5_I	0.000 ; 0.045	0.071 ; 0.856	0.000 ; 0.041	0.000 ; 0.090	0.000 ; 0.050	0.143 ; N/A	0.143 ; 0.260
S10.T2.6_I	0.000 ; 0.034	0.000 ; 1.05	0.000 ; 0.032	0.000 ; 0.317	0.000 ; 0.044	0.000 ; N/A	-
S50.T1.1_I	0.021 ; 0.401	0.042 ; 1.03	0.021 ; 0.288	0.021 ; 0.373	0.021 ; 0.401	0.042 ; N/A	0.042 ; 1.49
S50.T1.2_I	0.032 ; 0.235	0.021 ; 0.738	0.000 ; 0.292	0.000 ; 0.614	0.021 ; 0.234	0.021 ; N/A	0.064 ; 1.38
S50.T1.3_I	0.021 ; 0.356	0.021 ; 0.697	0.021 ; 0.327	0.021 ; 0.323	0.021 ; 0.335	0.064 ; N/A	0.042 ; 1.50
S50.T1.4_I	0.021 ; 0.151	0.021 ; 0.802	0.021 ; 0.142	0.021 ; 0.248	0.021 ; 0.165	0.021 ; N/A	0.021 ; 1.47
S50.T2.1_I	0.000 ; 0.655	0.000 ; 1.88	0.000 ; 0.547	0.000 ; 0.614	0.000 ; 0.628	0.021 ; N/A	-
S50.T2.2_I	0.000 ; 0.681	0.021 ; 1.82	0.000 ; 0.455	0.000 ; 0.481	0.000 ; 0.655	0.042 ; N/A	0.000 ; 2.47
S50.T2.3_I	0.011 ; 0.644	0.021 ; 1.23	0.021 ; 0.535	0.000 ; 0.601	0.021 ; 0.819	0.085 ; N/A	0.021 ; 2.27
S50.T2.4_I	0.000 ; 0.419	0.000 ; 1.46	0.000 ; 0.265	0.000 ; 0.465	0.000 ; 0.421	0.021 ; N/A	-
L13.TA.1_I	0.000 ; 0.013	0.000 ; 0.324	0.000 ; 0.006	0.000 ; 0.007	0.000 ; 0.031	0.100 ; N/A	-
L13.TA.2_I	0.000 ; 0.030	0.000 ; 0.288	0.000 ; 0.022	0.000 ; 0.013	0.000 ; 0.039	0.100 ; N/A	-
L13.TA.3_I	0.000 ; 0.030	0.000 ; 0.253	0.000 ; 0.022	0.000 ; 0.034	0.000 ; 0.034	0.100 ; N/A	-
L13.TA.4_I	0.000 ; 0.011	0.000 ; 0.360	0.000 ; 0.016	0.000 ; 0.025	0.000 ; 0.013	0.100 ; N/A	-
L19.TB.1_I	0.000 ; 0.105	0.000 ; 0.744	0.000 ; 0.070	0.000 ; 0.062	0.000 ; 0.062	0.375 ; N/A	-
L19.TB.2_I	0.000 ; 0.141	0.063 ; 1.72	0.000 ; 0.097	0.000 ; 0.211	0.000 ; 0.272	0.375 ; N/A	0.000 ; 2.17

Note: RF values of zero indicate identical topology to known tree, and lower branch-score values indicate more accurate trees.

In instances where combined genes were used as starting blocks to simulate intra-gene heterogeneity, PartitionFinder inferred two or three partitions for all but six datasets (Supplementary Table S2 and Supplementary Material available on Dryad); notably, since the original five genes simulated with different models were combined to form three genes (see section "Methods") the greatest number of partitions that could be inferred by PartitionFinder was three. Various models were inferred as best fit for each partition with datasets possessing intra-gene heterogeneity, including the VT model. We did not use the VT model to simulate data, but it probably offers a reasonable fit for genes that had components simulated under two different models as seen in datasets with intra-gene heterogeneity. Linear regression of the number of inferred partitions on datasets with intra-gene heterogeneity was significantly correlated to dataset size (indel-Seq-Gen: $P = 0.007$; INDELible: $P = 0.0008$), but the relationship was weak (indel-Seq-Gen: $R^2 = 0.344$; INDELible: $R^2 = 0.473$;

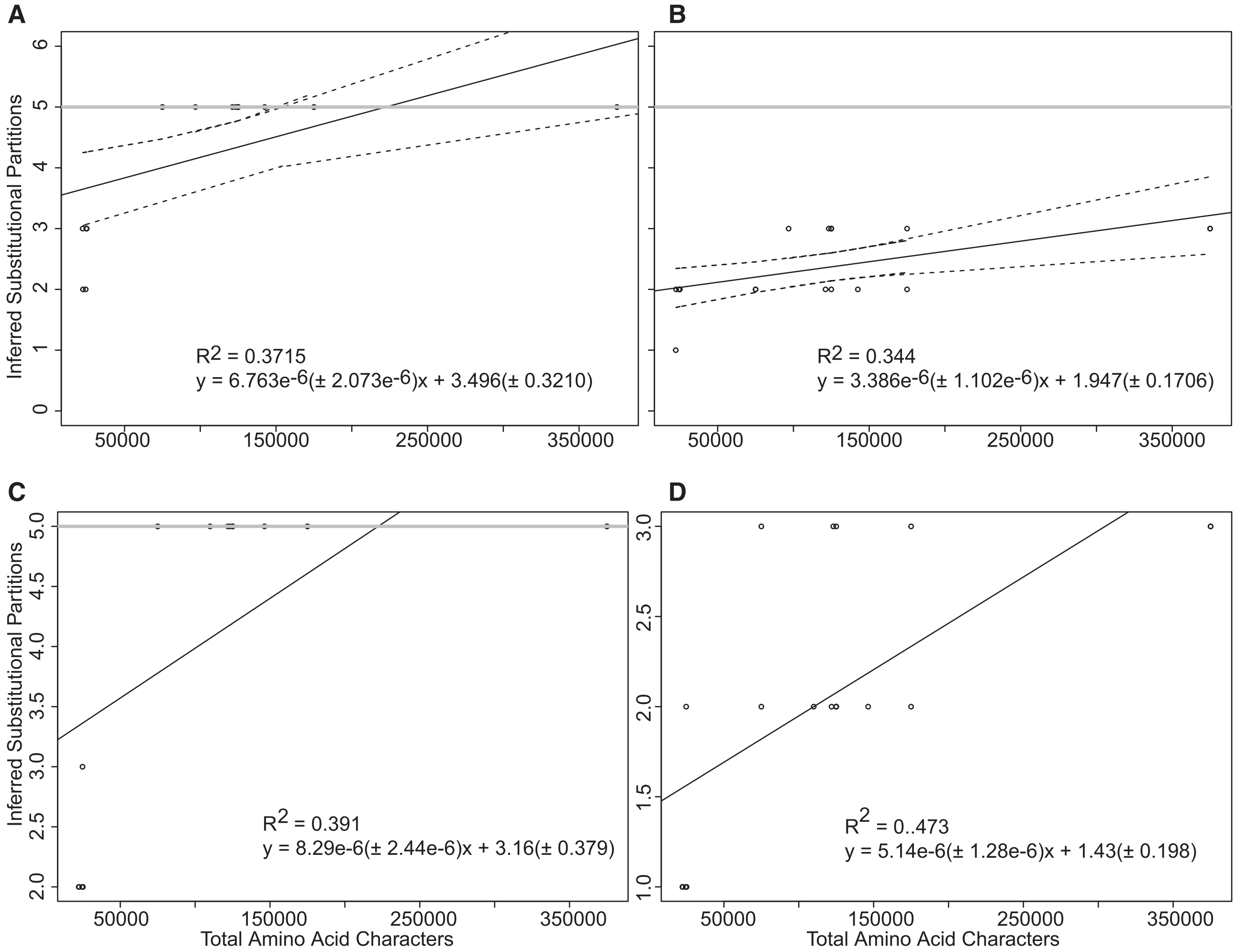

FIGURE 5. Linear regression plots of total characters versus number of substitutional partitions inferred with PartitionFinder for small datasets using the greedy search algorithm. Gray line represents correct number of substitutional partitions. A) Partitioning of genes without intra-gene heterogeniety and simulated with indel-Seq-Gen. B) Partitioning of genes that were combined to create intra-gene heterogeneity and simulated with indel-Seq-Gen. C) Partitioning of genes without intra-gene heterogeneity and simulated with INDELible. D) Partitioning of genes that were combined to create inra-gene heterogeneity and simulated with INDELible.

Fig. 5B, D; Supplementary Table S3 available on Dryad). Including data from large datasets (see below), the number of categories assigned by CAT-F81 was always much higher than the actual number of substitutional categories/partitions (Supplementary Table S2 available on Dryad), but it was only correlated with dataset size on datasets simulated with INDELible (indel-Seq-Gen: $P = 0.1436$; INDELible: $P = 3.509 \times 10^{-5}$; Fig. 6A, C; Supplementary Table S3 available on Dryad). The number of categories assigned by CAT-GTR was significantly correlated with dataset size (indel-Seq-Gen: $P = 3.357 \times 10^{-10}$; INDELible: $P = 3.34 \times 10^{-8}$) and fit of linear regression was good (indel-Seq-Gen: $R^2 = 0.8125$; INDELible: $R^2 = 0.726$; Fig. 6B, D; Supplementary Table S3 available on Dryad).

All analyses of 10-taxon datasets generated with indel-Seq-Gen recovered the correct branching order, except for CAT-F81 on datasets S10.T1.3_iSG and S10.T2.4_iSG (Table 3; Fig. 7; Supplementary Material available on Dryad). For analyses done on 10-taxon datasets simulated with INDELible, all models recovered an inaccurate branching pattern on dataset S10.T2.2_I (RF = 0.143 for all models; Table 3). Partitioning on datasets with intra-gene heterogeneity recovered an inaccurate branching pattern on one additional 10-taxon, INDELible dataset (S10.T1.3_I: RF = 0.143) and CAT-F81 recovered inaccurate branching patterns on two 10-taxon INDELible datasets (S10.T1.3: RF = 0.143; S10.T2.5_I: RF = 0.071), Furthermore, many correct relationships inferred by CAT-F81 across all 10-Taxon datasets were poorly supported (e.g., datasets S10.T1.3_iSG, S10.T1.1_I; Supplementary Figs. S4 and S5 available on Dryad). When a single F81 model was applied to 10-taxon datasets generated with indel-Seq-Gen where CAT-F81 failed to recover accurate branching patterns (i.e., S10.T1.3_iSG, S10.T2.4_iSG), F81 recovered the correct branching pattern. When a single F81 model was applied to 10-taxon datasets generated with INDELible, it

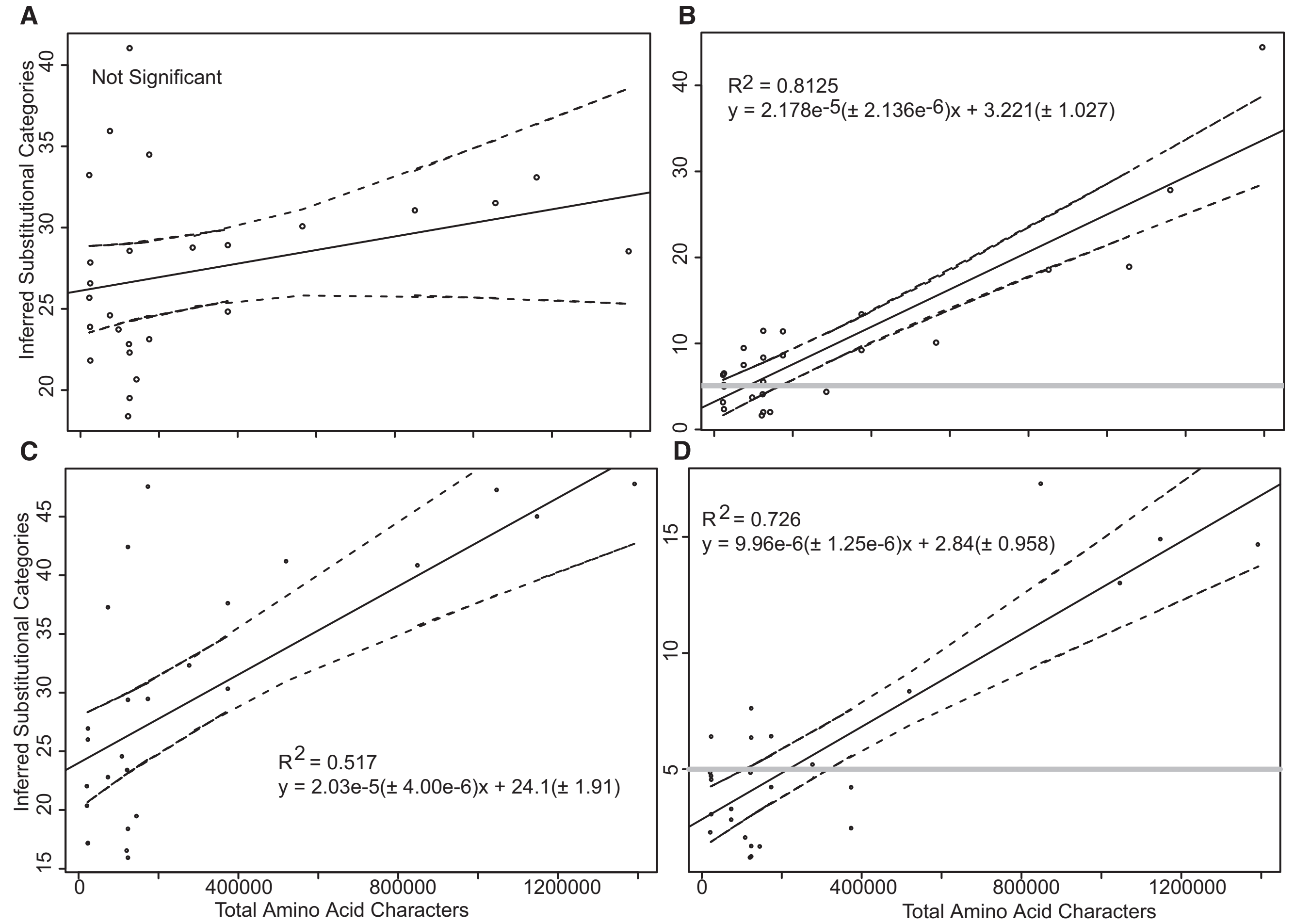

FIGURE 6. Linear regression plots of total characters versus number of substitutional categories inferred by CAT models. Gray line represents correct number of substitutional categories. A) CAT-F81 on genes simulated with indel-Seq-Gen. B) CAT-GTR on genes simulated with indel-Seq-Gen. C) CAT-F81 on genes simulated with INDELible. D) CAT-GTR on genes simulated with INDELible.

recovered a more accurate branching pattern than CAT-F81 on dataset S10.T1.3_I, an identical branching pattern on dataset S10.T2.2_I, and a less accurate branching pattern on dataset S10.T2.5_I based on RF distances (Table 3). For 10-taxon datasets that CAT-F81 failed on, both partitioning and CAT-GTR recovered correct relationships, except on dataset S10.T2.2_I where all models recovered equally inaccurate branching patterns (RF = 0.143).

Trees with correct branching patterns were much less commonly inferred with 50-taxon datasets than with 10-taxon datasets (Table 3), but incorrectly inferred relationships had low support (Fig. 8; Supplementary Figs. S2 and S6, Supplementary Material available on Dryad). In some instances, CAT models would infer a polytomy where partitioning inferred a poorly or moderately supported node (e.g., datasets S50.T1.2_iSG, S50.T1.2_I, Fig. 9). These differences were a result of RAxML producing a bifurcating tree but BI majority rule consensus trees allowing for polytomies if a node was not recovered in at least 50% of the posterior distribution. Nevertheless, similar overall patterns were observed with 50-taxon datasets as with 10-taxon datasets. For example, CAT-F81 produced trees with worse RF and branch-score distances than partitioning and CAT-GTR on seven of eight 50-taxon datasets simulated with both indel-Seq-Gen and INDELible (Table 3; Supplementary Material available on Dryad). CAT-F81 inferred a more accurate branching pattern than CAT-GTR once with data simulated with indel-Seq-Gen (dataset S50.T1.2_iSG, CAT-F81: RF = 0.000, CAT-GTR: RF = 0.011) and once with data simulated with INDELible (dataset S50.T1.2_I, CAT-F81: RF = 0.021, CAT-GTR: RF = .0.032). These were the only instances with simulated data where CAT-F81 outperformed CAT-GTR at inferring accurate relationships.

Tree inference with PhyloBayes and CAT-F81 never resulted in a more accurate branching pattern than partitioning according to RF distances (Table 3). Surprisingly, when a single F81 matrix was applied to datasets where CAT-F81 recovered incorrect branching patterns, F81 performed as well as or better than CAT-F81 at inferring accurate branching patterns on 16 out of 18 datasets (Table 3; Supplementary

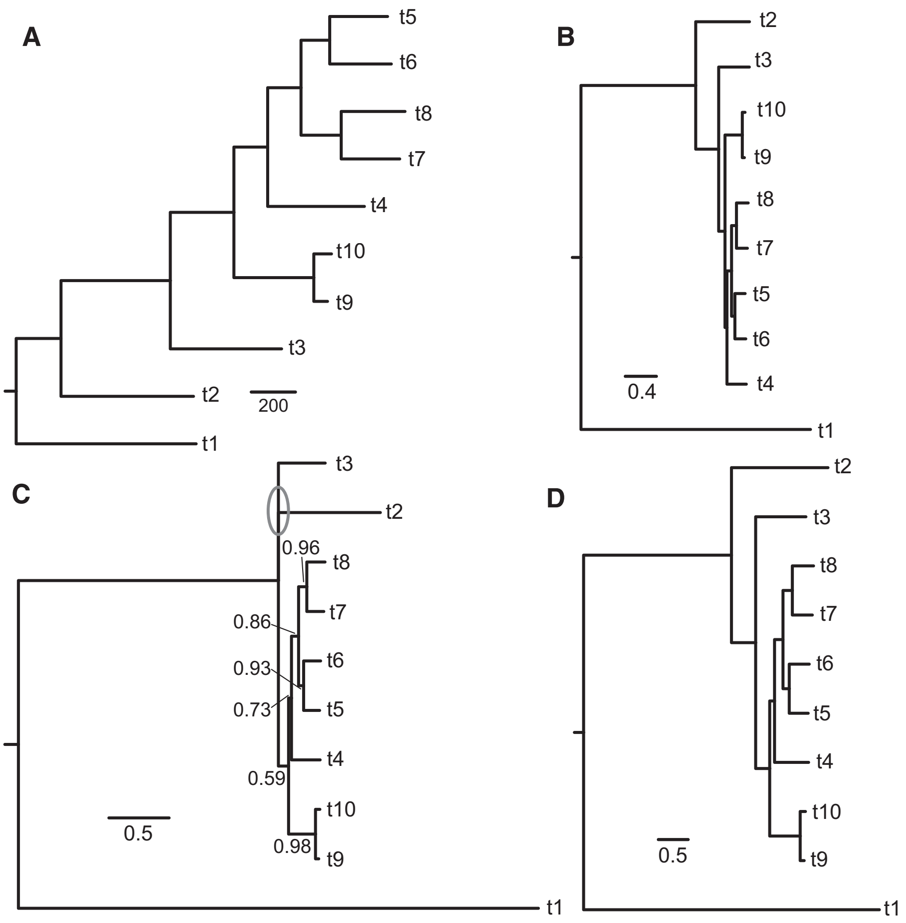

FIGURE 7. Trees inferred with dataset S10.T1.3_iSG. Gray circle indicates presence of incorrectly inferred polytomy. Nodes have 100% BS or PP unless labeled. A) Single most parsimonious tree. B) ML with data partitioning. C) BI with CAT-F81. D) BI with CAT-GTR. All models inferred correct trees with well supported, correct relationships except CAT-F81.

Material available on Dryad). CAT-F81 on 50-taxon datasets, particularly those generated with indel-Seq-Gen, also often had poor support for correctly inferred nodes, whereas other methods (i.e., CAT-GTR, partitioning, single site-homogeneous models) had strong support for the same, correctly inferred nodes (e.g., dataset S50.2.3_iSG; Fig. 8; Supplementary Figs. S2 and S6; Supplementary Material available on Dryad). CAT-GTR and partitioning recovered identical relationships with ten of the sixteen 50-taxon datasets (Table 3; Supplementary Material available on Dryad). Partitioning inferred more accurate branching patterns than CAT-GTR for four of the six datasets where they inferred different branching patterns based on RF distances (i.e., S50.T1.2_iSG, S50.T1.3_iSG, S50.T2.2_iSG, S50.T1.2_I; Table 3; Supplementary Material available on Dryad).

Analyses with CAT-GTR and CAT-F81 took considerably more CPU hours to finish than corresponding analyses with ML and partitioning (Table 4). In the most extreme instance on small datasets, CAT-GTR took over 9,558 CPU hours longer than partitioning for dataset S50.T2.4_iSG. Every BI analysis converged, but for some datasets (e.g., S50.T1.4_iSG, S50.T2.3_iSG, S50.T2.4_iSG, S10.2.3_I, S50.T1.4_I) at least one CAT-GTR chain appeared to be stuck in a local maximum, which required additional chains being run.

Model Performance on Large DataSets

All MP analyses of the simulated, large datasets resulted in incorrect branching patterns consistent with LBA (Figs. 10A, 11A; Supplementary Figs. S7–S12; Supplementary Material available on Dryad). Analyses with a single GTR model applied to entire datasets resulted in accurate branching patterns, but comparably poor branch lengths for most datasets based on branch-score distances (Table 3; Supplementary Material available on Dryad). Partitioning and CAT-GTR always recovered branching patterns identical to the known tree according to RF distances (Table 3; Figs. 10, 11; Supplementary Figs. S7–S12 available on Dryad). In contrast, CAT-F81 recovered a well-supported, incorrect placement of long-branched taxon “m” when used on dataset L19.TB.2_iSG and L19.TB.2_I (Table 3, Fig. 11C;

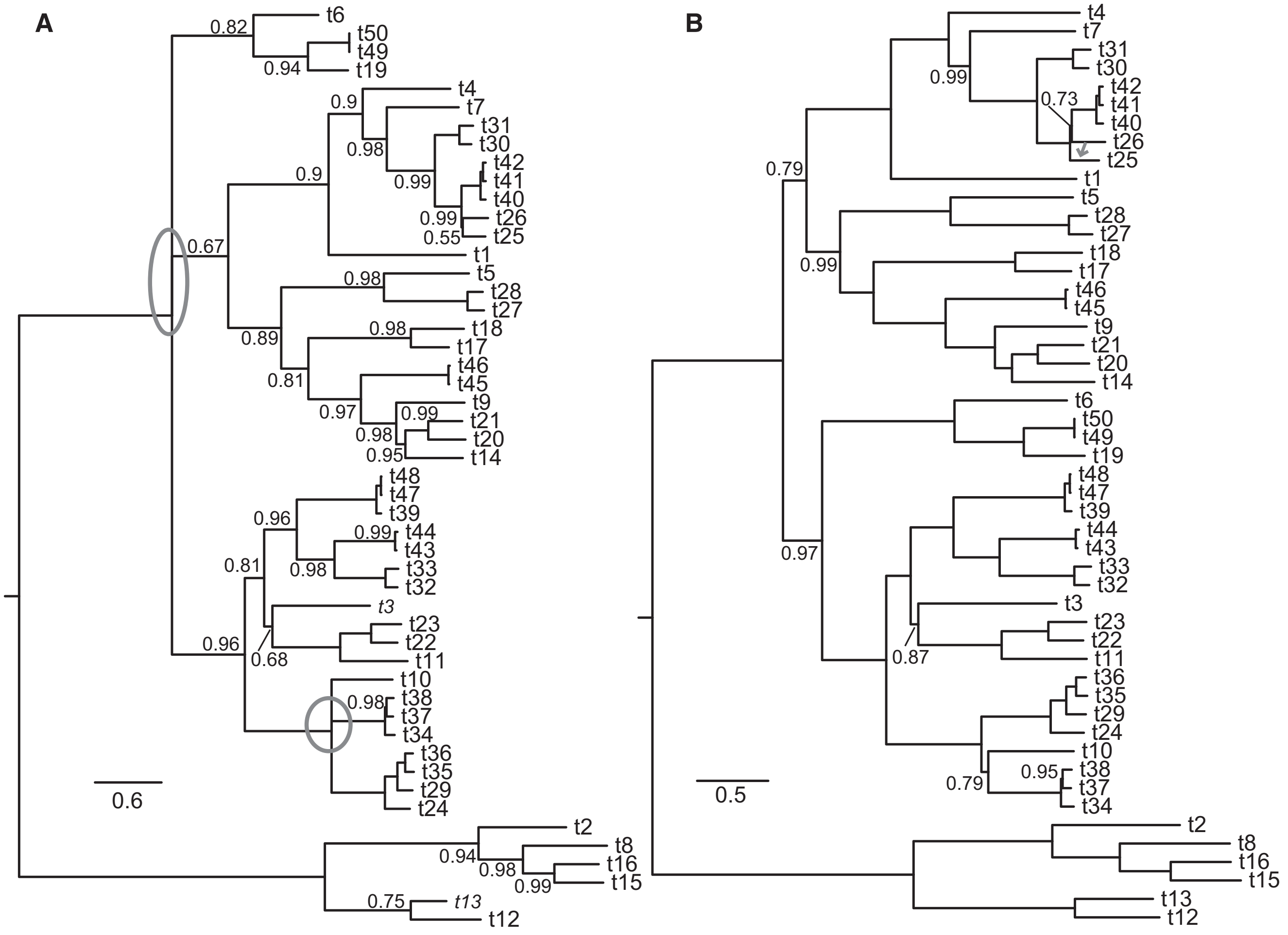

FIGURE 8. Trees inferred with CAT models and dataset S50.2.3_iSG. Nodes are supported with 100% PP unless otherwise labeled. Gray ovals represent incorrectly inferred polytomies, and red lines represent where incorrecly inferred branchs should be based on the true tree. A) BI with CAT-F81. B) BI with CAT-GTR.

Supplementary Fig. S8 available on Dryad). Instead of this branch being pulled toward the long-branched outgroup, as seen with MP, it was pulled to a more derived position inside a three-taxon clade (Fig. 11C; Supplementary Fig. S8C available on Dryad). For all other model-based analyses, there were only small differences in branch lengths, but not branching pattern, between the known tree and inferred tree (Table 3; Figs. 10, 11; Supplementary Figs.S7–S12; Supplementary Material available on Dryad). This includes when a single F81 model was applied to dataset L19.TB.2_iSG and L19.TB.2_I (Table 3; Supplementary Figs. S13 and S14 available on Dryad).

As mentioned above, there was no correlation between dataset size and the number of categories inferred by CAT-F81 for datasets inferred with indel-Seq-Gen, but the number of categories was always much higher than the real number of substitutional categories (i.e., five) both for datasets simulated with indel-Seq-Gen and INDELible (Fig. 6A, C, Supplementary Table S2 available on Dryad). The number of substitutional categories inferred by CAT-GTR was also significantly and strongly correlated with the number of characters in a dataset (Fig. 6B Supplementary Table S2 available on Dryad). PartitionFinder with non-combined gene blocks and 20% relaxed clustering also inferred a greater number of substitutional categories than what was present (Supplementary Table S2 available on Dryad), but it was not correlated with dataset size (indel-Seq-Gen: $P=$ 0.4671; INDELible: $P=$ 0.697; Supplementary Table S3 available on Dryad; Fig. 12A). PartitionFinder with combined genes (i.e., analyses with intra-gene heterogeneity) and 20% relaxed clustering had a significant positive relationship with dataset size and inferred substitutional categories (indel-Seq-Gen: $P=0.0371$, $R^2=0.7028$; INDELible: $P=0.005$, $R^2=0.892$; Supplementary Table S3 available on Dryad; Fig. 12B). As with small datasets, all BI analyses of large datasets converged, but CAT-F81 and CAT-GTR took over 418,000 more CPU hours to finish than partitioning on large datasets (Table 4).

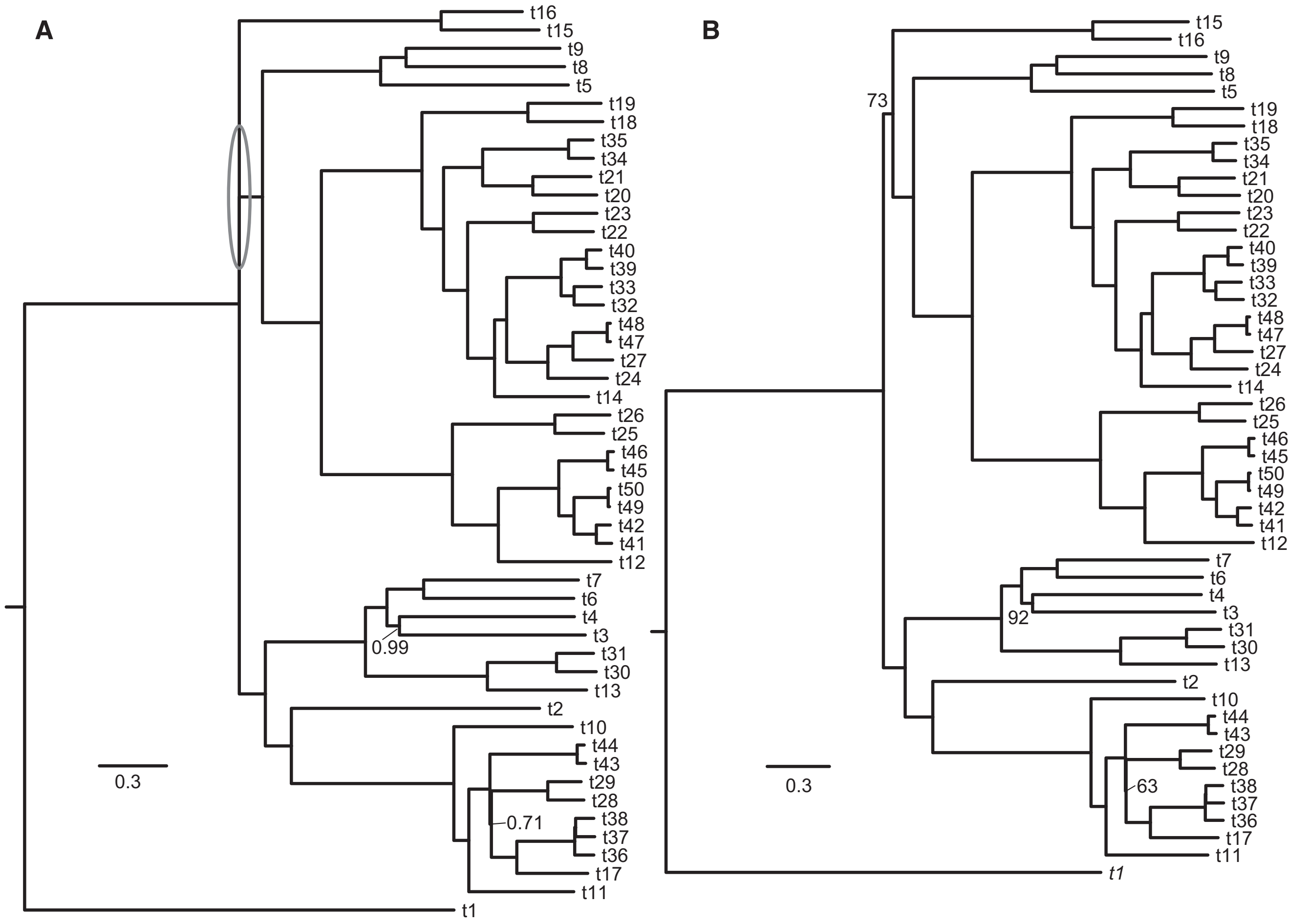

FIGURE 9. Trees inferred with dataset S50.1.2_iSG. Nodes are supported with 100% BS or PP unless otherwise noted. Gray ovals indicates incorrectly inferred polytomy (see Fig. 3). A) BI with CAT-GTR. B) ML with partitioning and RAxML. Both trees are similar except for an incorrectly inferred polytomy with CAT-GTR.

Empirical Data Analyses

When we reanalyzed the Philippe et al. (2009) dataset with MP, partitioning, CAT-F81, and CAT-GTR we recovered ctenophores as the sister group to all other animals (Supplementary Fig. S15 available on Dryad), conflicting with the reported CAT-F81 tree in Philippe et al. (2009); CAT-GTR was not used in original study. Nodal support for ctenophores as the sister group to other animals on trees inferred here with CAT-F81 and the Philippe et al. (2009) dataset was 0.93 PP. We recovered 99% BS support with partitioning and 0.97 PP with CAT-GTR for ctenophores as the sister group to all other animals. Our analyses of the Nosenko et al. (2013) dataset also conflicted with the original study as ctenophores were recovered as the sister group to all other extant metazoans with partitioning and CAT-GTR (Supplementary Fig. S16 available on Dryad; partitioning BS = 99, CAT-GTR PP = 0.98). In contrast, our CAT-F81 analysis of Nosenko et al. (2013) recovered a paraphyletic Porifera as sister groups to all other animals and ctenophores as the sister group to cnidarians (Supplementary Fig. S16 available on Dryad). CAT-F81 inferred 621.7 and 477.6 substitutional categories and CAT-GTR inferred 1049.99 and 760.76 substitutional categories for the Philippe et al. (2009) and Nosenko et al. (2013) datasets, respectively (Supplementary Table S2 available on Dryad). We assume differences between trees in the original publications and those inferred here with the same models are a result of algorithmic and/or programming improvements in more recent versions of PhyloBayes considering that Philippe et al. (2009) and Nosenko et al. (2013) used an older version of PhyloBayes (PhyloBayes 2.3 and 3.2e, respectively) than the one used here (PhyloBayes MPI 1.5a).

DISCUSSION

Our results show that CAT-F81 is less accurate in inferring correct trees than CAT-GTR and partitioning with site-homogeneous models (Table 3). CAT-F81 has occasionally been used instead of CAT-GTR when analyses with CAT-GTR were deemed too computationally demanding, even when CAT-GTR was

TABLE 4. Computational time used for each analysis

Dataset	CAT-F81 (CPU hours)	CAT-GTR (CPU hours)	Partitioning including PartitionFinder (CPU hours)	Partitioning with intra-gene heterogeneity including PartitionFinder (CPU hours)	ML with single GTR matrix (CPU hours)
S10.T1.1_I	165.24	975.38	1.39	1.48	2.27
S10.T1.2_I	281.64	992.59	1.56	1.22	2.08
S10.T1.3_I	215.18	300.26	2.15	2.06	2.90
S10.T1.4_I	590.93	1,821.89	3.72	3.29	5.43
S10.T1.5_I	546.34	3,133.15	5.63	5.50	8.66
S10.T1.6_I	1,388.38	2,061.91	7.72	7.34	12.34
S10.T2.1_I	222.44	1,048.40	1.18	1.36	2.27
S10.T2.2_I	607.65	530.54	1.55	1.37	2.37
S10.T2.3_I	103.63	974.01	1.50	1.52	2.51
S10.T2.4_I	539.15	2,779.79	3.50	3.62	5.87
S10.T2.5_I	609.34	3,810.08	6.70	5.86	10.95
S10.T2.6_I	1,538.88	3,116.28	7.16	6.88	13.22
S50.T1.1_I	803.10	3,132.52	20.88	17.59	22.32
S50.T1.2_I	805.32	3,177.38	22.58	21.39	26.39
S50.T1.3_I	3,824.19	12,880.30	33.98	32.55	42.11
S50.T1.4_I	1,413.01	5,193.53	49.72	52.22	69.94
S50.T2.1_I	1,601.42	1,494.17	19.05	17.87	24.02
S50.T2.2_I	1,455.57	2,172.33	21.11	22.13	29.43
S50.T2.3_I	7,511.40	13,874.00	46.08	40.15	55.32
S50.T2.4_I	2,611.23	17,662.00	47.14	42.55	65.98
L13.TA.1_I	19,303.03	10,433.19	162.86	61.81	62.38
L13.TA.2_I	9,642.86	18,087.57	59.86	43.32	32.59
L13.TA.3_I	17,824.85	24,925.02	19.34	13.03	17.48
L13.TA.4_I	11,003.42	18,020.56	174.78	57.58	60.34
L19.TB.1_I	9,915.97	23,632.76	204.81	187.01	252.00
L19.TB.2_I	18,076.70	14,744.94	330.33	232.53	287.36
S10.T1.1_iSG	384.81	163.75	6.01	1.57	3.43
S10.T1.2_iSG	607.42	584.89	5.69	1.51	3.04
S10.T1.3_iSG	496.13	530.24	6.98	1.84	9.74
S10.T1.4_iSG	61.89	363.67	6.59	4.66	9.99
S10.T1.5_iSG	408.99	967.43	8.65	7.19	21.48
S10.T1.6_iSG	729.60	3,801.69	35.73	10.01	18.98
S10.T2.1_iSG	842.13	930.22	1.95	1.42	2.95
S10.T2.2_iSG	487.82	627.62	2.10	1.64	7.99
S10.T2.3_iSG	540.16	695.05	1.95	1.57	1.31
S10.T2.4_iSG	878.70	933.00	4.91	4.81	10.77
S10.T2.5_iSG	384.23	1,414.17	7.80	7.15	15.09
S10.T2.6_iSG	829.55	3,272.34	10.64	10.29	29.65
S50.T1.1_iSG	959.71	995.44	21.82	20.19	29.95
S50.T1.2_iSG	1,313.43	1,458.69	22.92	21.85	30.87
S50.T1.3_iSG	2,538.68	7,014.88	36.93	33.49	45.90
S50.T1.4_iSG	916.75	6,441.94	63.69	49.98	56.08
S50.T2.1_iSG	1,677.56	7,589.39	23.88	23.05	29.71
S50.T2.2_iSG	1,027.90	1,648.33	29.53	26.36	34.89
S50.T2.3_iSG	1,068.38	6,770.47	50.82	43.16	65.09
S50.T2.4_iSG	1,437.61	9,611.40	54.57	52.96	56.85
L13.TA.1_iSG	13,000.94	18,283.30	1,277.70	269.45	94.75
L13.TA.2_iSG	4,620.36	13,480.51	286.25	62.13	53.26
L13.TA.3_iSG	3,967.73	4,666.39	72.00	24.53	26.63
L13.TA.4_iSG	2,823.51	22,539.99	1,146.27	245.67	88.52
L19.TB.1_iSG	49,140.75	37,150.44	1,408.67	393.54	267.30
L19.TB.2_iSG	17,250.33	56,176.79	1,945.49	749.34	343.65
Nosenko et al.	1,006.13	23,198.15	333.07	–	–
Philippe et al.	8,895.73	14,622.54	330.36	–	
Total	219,894.38	436,907.26	8,459.26	2,952.57	2,476.38

acknowledged to be a better-fitting model than CAT-F81 (e.g., Nosenko et al. 2013; Pisani et al. 2015). We argue that CAT-F81 should not be used in future studies, due to concerns about inaccurate inference. In contrast, both CAT-GTR and partitioning performed similarly in inferring correct branching patterns based on RF distances (Table 3). A single protein GTR model performed rather well in inferring accurate branching patterns, but when all simulated datasets analyzed here are considered in aggregate, GTR inferred slightly less accurate trees than CAT-GTR and partitioning. Broadly, we found that partitioning performed as well

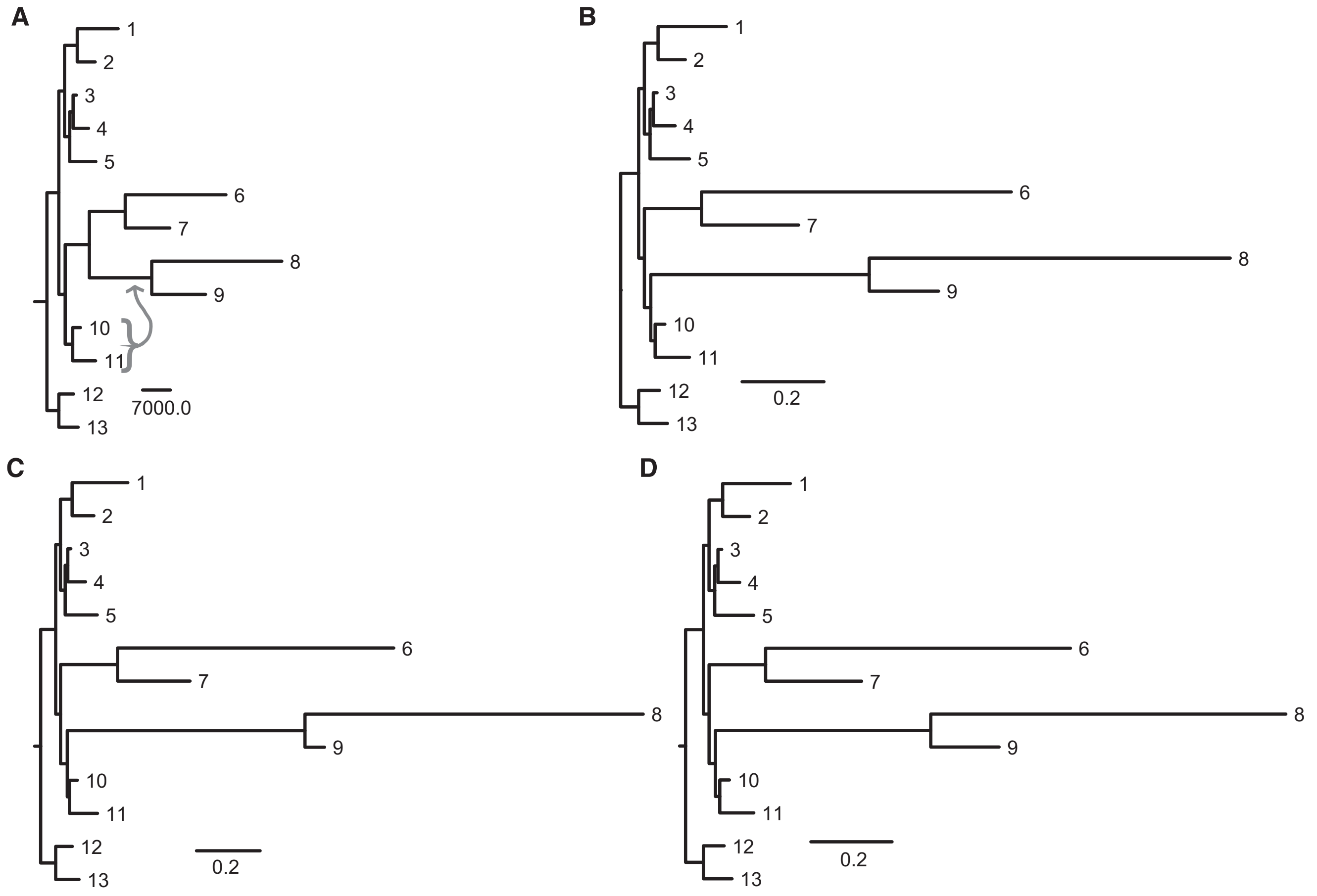

FIGURE 10. Trees inferred with dataset L13.TA.1_I. Gray arrows indicate where branches should be based on the known tree (see Fig. 4). All nodes have 100% BS or PP. A) Single most parsimonious tree. B) ML with data partitioning. C) BI with CAT-F81. D) BI with CAT-GTR.

as, or slightly better than, PhyloBayes and CAT-GTR at inferring accurate relationships, including with datasets that were susceptible to LBA.

Topological Accuracy

In instances when CAT-F81 recovered inaccurate branching orders with small datasets, the recovered pattern was often similar to errors recovered with MP and indicative of LBA, but incorrect branching patterns were poorly supported (Figs. 7 and 8; Supplementary Figs. S2 and S3; Supplementary Tree Files available on Dryad). In contrast, when applied to large dataset LB.TB.2_iSG and LB.TB.2_I, CAT-F81 recovered strong support for an incorrect relationship (Fig. 11C; Supplementary Fig. S13C available on Dryad). However, this relationship was not the same as the incorrect relationships inferred by MP on datasets LB.TB.2_iSG and LB.TB.2_I, and it may not be a LBA artifact (Fig. 11C; Supplementary Fig. S13C available on Dryad). Errors observed with CAT-F81 conflict with past claims that this model performs well on datasets susceptible to LBA (Brinkmann and Philippe 2008; Philippe et al. 2009, 2011b; Nosenko et al. 2013). Even more troubling is the observation that CAT-F81 sometimes performed worse at inferring correct branching patterns than MP and a single F81 model applied to an entire dataset (e.g., S10.T2.4_iSG, S50.T2.1_iSG, LB.TB.2_iSG, S10.T1.3_I, LB.TB.2_I; Table 3). Some problems observed with CAT-F81 may stem from applying a model that assumes equal substitution frequencies among all amino acids (i.e., F81). However, when we applied a single F81 model across the entirety of datasets CAT-F81 failed on, a more accurate branching order was recovered with 10 out of 20 datasets according to RF distances (Table 3), suggesting problems with the CAT component of CAT-F81.

Our analyses with CAT-GTR and partitioning recovered similar, or identical, branching patterns even on simulated datasets with intra-gene heterogeneity and empirical datasets (Table 3; Supplementary Material available on Dryad), but CAT-GTR recovered less accurate branching patterns than partitioning on datasets that lacked intra-gene heterogeneity more times (i.e., four) than CAT-GTR did better than partitioning (i.e., twice; Table 3). As a site-heterogeneous model, CAT-GTR should have been able to model intra-gene heterogeneity that partitioning was blind to. Yet, CAT-GTR recovered a less accurate branching pattern than

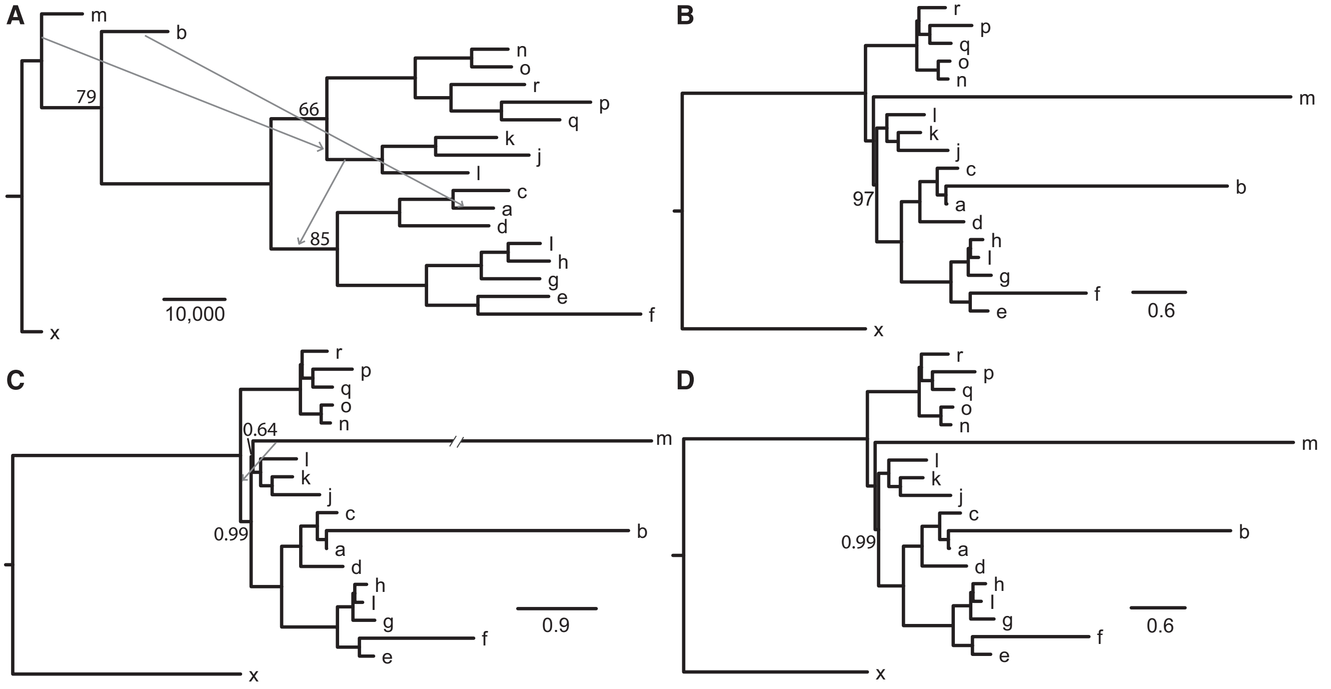

FIGURE 11. Trees inferred with dataset L19.TB.2_I. Gray arrows indicate where branches should be based on the known tree. Unless otherwise labeled, nodes have 100% BS or PP. A) Single most-parsimonious tree. B) ML with data partitioning. C) BI with CAT-F81. D) BI with CAT-GTR.

partitioning on datasets with intra-gene heterogeneity on four datasets (i.e., S50.T1.2_iSG, S50.T1.3_iSG, S50.T1.2_I, S50.T2.3_I; Table 3), and CAT-GTR only recovered a more accurate topology than partitioning with intra-gene heterogeneity twice (S50.T2.2_iSG, S10.T1.3_I; Table 3). Notably, partitioning appears robust to at least some intra-gene substitutional heterogeneity.

Accuracy of Branch Length Inference

Even though many biologists are likely more interested in branching pattern than branch lengths, accurate branch lengths are needed for many analyses such as inferences about phenotypic evolution or timing of cladogenesis. Both CAT models implemented in PhyloBayes and partitioning in RAxML report branch lengths in substitutions per site so, we consider the general tendency of CAT models to perform worse than partitioning in terms of branch lengths, according to branch-score distances (Table 3), to be notable. We were especially surprised that partitioning in RAxML inferred more accurate branch lengths, based on branch-score distances (Table 3), than CAT-GTR in PhyloBayes on some datasets when genes were combined to create intra-gene heterogeneity. Notably, potential problems with completely ignoring substitutional heterogeneity in datasets can be seen when comparing methods that account for substitutional heterogeneity in some fashion (i.e., CAT models and partitioning) and those that do not (i.e., a single GTR or F81 model). For instance, even though F81 performed better than CAT-F81 on dataset L19.TB.2_I in inferring correct relationships (F81: RF = 0.000; CAT-F81: RF = 0.021), branch lengths of the F81 tree were considerably worse than those inferred with CAT-GTR and partitioning based on branch-score distance (F81: branch-score = 2.17; CAT-GTR: branch-score = 0.141; partitioning: branch-score = 0.097). However, on datasets simulated with INDELible, a single protein GTR model surprisingly recovered trees with more accurate branch lengths than partitioning and CAT-GTR in some instances (Table 3). Nevertheless, we do not advocate for the use of a single protein GTR model on large datasets, because it did recover less accurate relationships than partitioning and CAT-GTR in two instances, whereas partitioning always recovered as accurate or more accurate branching patterns than a single protein GTR model.

Computational Time and Estimation of Substitutional Heterogeneity

One possible reason CAT models were often less accurate than partitioning in inferring branch lengths according to branch-score distances is their tendency to overestimate the true number of substitutional categories, particularly for larger datasets. We were surprised that the number of categories inferred by the two different CAT models often differed, as substitutional heterogeneity was the same regardless of the CAT model used. Of concern, as more sequence data was present, CAT-GTR performed worse in estimating substitutional heterogeneity on our simulated datasets. CAT-F81 also overestimated the number of substitutional

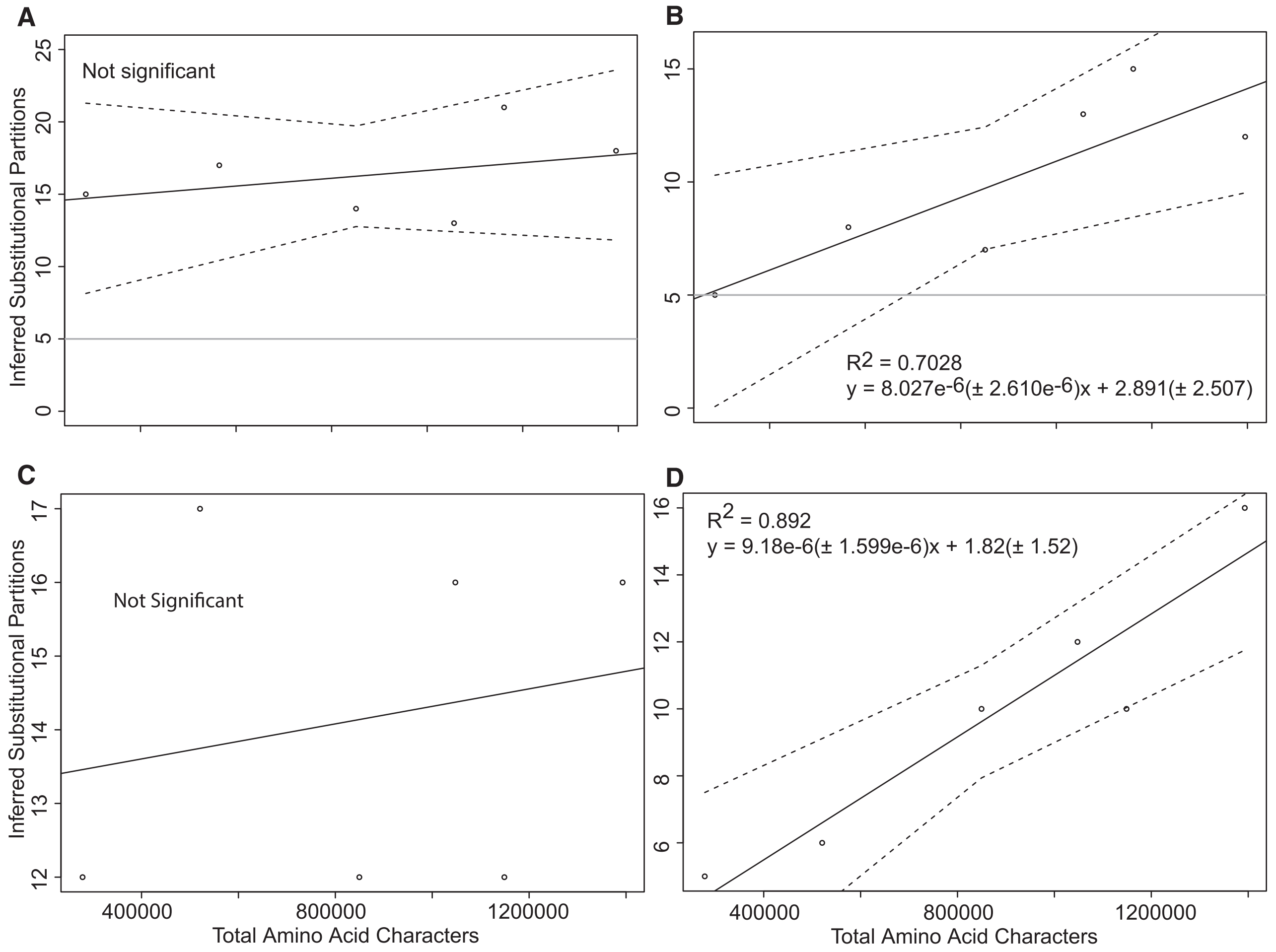

FIGURE 12. Linear regression plots of total characters versus number of substitutional partitions inferred with PartitionFinder for small datasets using the 20% relaxed clustering algorithm. Gray line represents correct number of substitutional partitions. A) Partitioning of genes without intra-gene heterogeniety simulated with indel-Seq-Gen, B) Partitioning of genes that were combined to create intra-gene heterogeneity simulated with indel-Seq-Gen. C) Partitioning of genes without intra-gene heterogeneity simulated with INDELible. D) Partitioning of genes that were combined to create intra-gene heterogeneity simulated with INDELible.

categories in our simulated datasets, but this was only correlated with dataset size when analyzing datasets simulated with INDELible. Such behavior of CAT models may have more to do with their implementation in PhyloBayes MPI than CAT models *per se*. For instance, because CAT models are implemented in a Bayesian framework, prior choice likely affects how sites are assigned to categories, but priors cannot be modified with PhyloBayes MPI. Thus, how different priors may affect CAT model performance in inferring substitutional heterogeneity and accurate trees is unclear. PartitionFinder was also not perfect in determining the number of substitutional categories, particularly on larger datasets when relaxed clustering was used and when PartitionFinder was blind to intra-gene heterogeneity. Supposed advantages of CAT models hinge on their ability to accurately estimate substitutional heterogeneity. However, we present strong evidence that how many categories are inferred in any given analysis by current implementation of CAT models, particularly CAT-GTR, has more to do with dataset size than actual substitutional heterogeneity (Fig. 6; Supplementary Tables S2 and S3 available on Dryad).

Knowing the true extent of substitutional heterogeneity in empirical datasets is impossible, but given simulation results we suspect substitutional heterogeneity in empirical datasets is overestimated by CAT models. For instance, CAT-GTR inferred an average of 760.76 substitutional categories for the Nosenko et al. (2013) dataset for only 20,242 amino acid positions, and a similar pattern was seen with Philippe et al. (2009; i.e., 1049.99 categories for 30,257 amino acid positions). Even if CAT-GTR inferred the true amount of substitutional heterogeneity present in the Nosenko et al. (2013) dataset, it means that, on average, each category will possess only 26.6 amino acids. When, or if, so few amino acids applied to each category will result in incorrect branching patterns is unclear with our data, but it may be the reason why CAT-GTR analyses sometimes infer odd relationships that are not recreated in other studies [e.g., sponges and ctenophores as sister groups to each other (Ryan et al. 2013), acoels + *Xenoturbella* as the sister group to protostomes (Rouse et al. 2016)].

In total, our simulation analyses required over 670,414 CPU hours (Table 4). Over 97% of this time was spent running analyses with CAT models, and our calculations do not include CPU hours required for restarting chains stuck in local optima. Computational demands and convergence issues associated with our relatively simple datasets support anecdotal statements that BI with CAT models is not computationally feasible for studies analyzing numerous, large, empirical datasets (Nosenko et al. 2013; Ryan et al. 2013; Struck et al. 2014; Borowiec et al. 2015; Pisani et al. 2015; Whelan et al. 2015b). Most importantly, accurate trees can be more rapidly obtained with partitioning than with CAT models.

Relevance to Empirical Studies

The most dramatic differences that have been documented between trees inferred with CAT models versus other models are arguably from studies exploring nonbilaterian animal relationships (Philippe et al. 2009, 2011b; Pick et al. 2010; Nosenko et al. 2013; Pisani et al. 2015), but our results conflict with a conclusion of these studies that only CAT models should be used for phylogenomic analyses. In particular, simulation results indicate that trees inferred with CAT-F81 from past studies that recovered sponges as the sister group to all other animals (i.e, Philippe et al. 2009, 2011b; Pick et al. 2010; Nosenko et al. 2013; Pisani et al. 2015) are inaccurate. More broadly, our simulation results suggest that rather than suppressing LBA, CAT models, particularly CAT-F81, can cause systematic error in empirical studies. For instance, ctenophores tend to be pulled away from non-metazoans and toward cnidarians and bilaterians on some trees inferred with large datasets and CAT-F81 (Fig. S16; Philippe et al. 2009, 2011b; Nosenko et al. 2013); this pattern is similar to how taxon "m" was repelled from the outgroup with CAT-F81 on datasets L19.TB.2_iSG and L19.TB.2_I (Fig. 11C; Supplementary Fig. S13 available on Dryad). Surprisingly, CAT-GTR and partitioning with the Nosenko et al. (2013) dataset and partitioning, CAT-F81, and CAT-GTR with the Philippe et al. (2009) dataset—two studies that advocate for the use of CAT models and concluded that sponges are the sister group to all other animals—recovered ctenophores as the sister group to all animals. We view this as supporting simulation results indicating that CAT-GTR and partitioning will usually perform similarly in inferring relationships. We also view the result from CAT-F81 with the Nosenko et al. (2013) datasets of ctenophores as the sister group to cnidarians, a hypothesis that has been rejected by many studies (Dunn et al. 2008; Hejnol et al. 2009; Pick et al. 2010; Nesnidal et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Chang et al. 2015; Pisani et al. 2015; Whelan et al. 2015b) to support our conclusion that CAT-F81 can be critically inaccurate. Overall, our empirical analyses, in the context of our simulation results, support three conclusions: 1) as indicated in simulation, CAT-GTR and partitioning will generally result in similar branching patterns and when they disagree nodal support is often low; 2) trees inferred with CAT-F81 will sometimes disagree with trees inferred with CAT-GTR and partitioning, and such CAT-F81 trees are likely less accurate; 3) available data suggest that ctenophores are the sister group to all other animals (see Supplementary Material available on Dryad, for additional discussion).

We have focused on past studies about non-bilaterian relationships because they have often been used to postulate about CAT model accuracy. Reanalyzing every study that produced different trees with CAT models and a single site-homogenous model or partitioning was outside the scope of this study. However, conclusions that trees not inferred with CAT models must be inaccurate (Philippe et al. 2009, 2011b; Nosenko et al. 2013; Pisani et al. 2015) should be revisited. Many studies (Delsuc et al. 2008; Liu et al. 2009; Philippe et al. 2011a; Nesnidal et al. 2013; O'Hara et al. 2014; Siu-Ting et al. 2014; Struck et al. 2014; Borowiec et al. 2015; Chang et al. 2015; Feuda and Smith 2015; Luo 2015; Whelan et al. 2015b) have shown that tree inference using site-homogeneous models, with or without data partitioning, and CAT models generally recover similar trees across a variety of taxa. When trees differ, conflicting nodes are often poorly supported or have variable support depending on gene choice (Brinkmann and Philippe 2008; Philippe et al. 2009, 2011b; Egger et al. 2015; Kenny et al. 2015). Nevertheless, results from analyses using CAT models are often emphasized (Brinkmann and Philippe 2008; Philippe et al. 2009, 2011b), and in some cases, trees from Bayesian analyses that showed no evidence of convergence have been highlighted (Egger et al. 2015; Pisani et al. 2015). Given our simulation results, preference toward relationships inferred with CAT models is not supported by available evidence. This underscores problems with using empirical data, absent of any simulation studies, to postulate about the performance of any model or method. We also find no evidence to support claims that CAT models are "realistic" or that they comprehensively mitigate systematic error.

Conclusions and Future Directions

Resolving many problematic nodes on the tree of life appears within reach, and our results suggest that despite substitutional heterogeneity being present in data, accounting for it in a coarse manner with partitioning results in reasonably accurate trees. Even if CAT-GTR, by itself, describes empirical substitution processes well, using current implementations of the model do not appear to result in more accurately inferred branching patterns than partitioning. These findings are corroborated by a recent study that suggested partitioning results in relatively accurate trees (Darriba and Posada 2015), and another paper suggesting problems with CAT-GTR in inferring accurate branching patterns when missing data is high in a small number of taxa (Li et al. forthcoming). Thus, we argue that spending computational resources on analyses with CAT-GTR is unnecessary for accurate phylogenetic inference. Additionally, practices such as removing constant characters from analyses, or worse, parsimony uninformative characters that are informative in ML and BI, when using CAT-GTR (as in Tarver et al. 2016) cannot be justified as likelihood calculations are effected. Moreover, the practice of reporting unconverged CAT-GTR analyses (as in Egger et al. 2015 and Pisani et al. 2015) is not only statistically invalid (Gelman and Rubin 1992; Huelsenbeck et al. 2002), but it can no longer be rationalized under assumptions of model performance. Finally, studies that emphasized trees inferred with CAT-F81 should be reassessed, given the inaccuracies associated with CAT-F81 that were revealed in simulation.

We do not wish to imply that complex, or computationally demanding, substitution models are inherently problematic. We also are not advocating against Bayesian phylogenetics or site-heterogeneous models as a whole. However, complexity should not be confused with accuracy or realism, and any new substitution model and/or programs designed to use such models in phylogenetic inference should be thoroughly tested before wide-ranging conclusions about evolution are made. Developing a method for quantifying substitutional heterogeneity that does not merely scale with dataset size will likely help improve both site-heterogeneous models and partitioning. Furthermore, base compositional homogeneity and similar substitution processes among lineages are both assumptions of the models explore here, but these assumptions may be more problematic than currently appreciated (Jayaswal et al. 2014; Kocot et al. forthcoming). Computationally tractable models that do not make such assumptions, in addition to better implementations of site-heterogeneous models, may enhance future studies at deep and shallow scales.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.85b2m.

Funding

This work was made possible in part by a grant of high-performance computing resources and technical support from the Alabama Supercomputer Authority and was supported by the US National Aeronautics and Space Administration (NASA-NNX13AJ31G).

Acknowledgments

We thank members of the Molette Biology Laboratory for Environmental and Climate Change Studies at Auburn University, Johanna Cannon, Kevin Kocot, and Milton Tan for helpful discussions. Thanks to Lars Jermiin, Robert Lanfear, and Alexandros Stamatakis, Tom Near, and Andy Anderson for helpful comments that improved a previous version of this work. Nicolas Lartillot graciously provided advice for simulating data under CAT-GTR and provided helpful comments on a previous version. The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the U.S. Fish and Wildlife Service. This is Molette Biology Labarotory contribution 56 and Auburn University Marine Biology Program contribution 148.

References

Adachi J., Hasegawa M. 1996. Model of amino acid substitution proteins encoded by mitochondrial DNA. J. Mol. Evol. 42:459–468.

Betts M.J., Russell R.B. 2003. Amino acid properties and consequences of substitutions. In: Barnes M.R., Gray I.C. editors. Bioinformatics for geneticists. Chichester, UK: John Wiley and Sons, Ltd. p. 289–316.

Borowiec M.L., Lee E.K., Chiu J.C., Plachetzki D.C. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome datasets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics 16:987.

Brinkmann H., Philippe H. 2008. Animal phylogeny and large-scale sequencing: progress and pitfalls. J. Syst. Evol. 46:274–286.

Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.

Cannon J.T, Kocot K.M., Waits D.S., Weese D.A., Swalla B.J., Santos S.R., Halanych K.M. 2014. Phylogenomic Resolution of the Hemichordate and Echinoderm Clade. Curr. Biol. 24:2827–2832.

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

Chang E.S., Neuhof M., Rubinstein N.D., Diamant A., Philippe H., Huchon D., Cartwright P. 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. USA 112:14912–14917.

Darriba D., Posada D. 2015. The impact of partitioning on phylogenomic accuracy. bioRxiv.

Delsuc F., Tsagkogeorga G., Lartillot N., Philippe H. 2008. Additional molecular support for the new chordate phylogeny. Genesis 46:592–604.

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.,Sorensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindal M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.

Egger B., Lapraz F., Tomiczek B., Müller S., Dessimoz C., Girstmair J., Škunca N., Rawlinson K.A., Cameron C.B., Beli E., Todaro M.A., Gammoudi M., Noreña C., Telford M.J. 2015. A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. Curr. Biol. 25:1347–1353.

Felsenstein J. 1978. Cases in which parsimony or compatability methods will be positively misleading. Syst. Zool. 27:401–410.

Feuda R., Smith A.B. 2015. Phylogenetic signal dissection indentifies the root of starfishes. PLoS One 10:e0123331.

Finet C., Timme R.E., Delwiche C.F., Marlétaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr. Biol. 20:2217–2222.

Fletcher W., Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol. 26:1879–1888.

Gan T.Y., Dlamini E.M., Biftu G.F. 1997. Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. J. Hydrol. 192:81–103.

Gelman A., Rubin D.B. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–511.

Gu X., Fu Y-X., Li W-H. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assesssing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Halanych K.M., Whelan N.V., Kocot K.M., Kohn A.B., Moroz L.L. 2016. Miscues misplace sponges. Proc. Natl. Acad. Sci. USA 113:E946–E949.

Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguñà J., Bailly X., Jondelius U., Wien M., Müller W.E.G., Seaver E., Wheeler C., Martindale M.Q., Giribet G., Dunn C.W. 2009. Assessing the root of bilaterian animals with scalable phylogenomic models. Proc. R. Soc. Lond. B Biol. Sci. 276:4261–4270.

Henikoff S., Henikoff J.G.. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89:10915–10919.

Hillis D.M., Bull J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Holder M.T., Zwickl D.J, Dessimoz C. 2008. Evaluating the robustness of phylogenetic methods to among-site rate variability in substitution processes. Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci. 363:4013–4021.

Huelsenbeck J.P. 1995a. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

Huelsenbeck J.P. 1995b. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol. Biol. Evol. 12:843–849.

Huelsenbeck J.P. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46:69–74.

Huelsenbeck J.P, Larget B., Miller R.E., Ronquist F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

Jayaswal V., Wong T.K.F, Robinson J., Poladian L., Jermiin L.S. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63:726–742.

Jones D.T., Taylor W.R., Thornton J.M. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett. 339:269–274.

Kainer D., Lanfear R. 2015. The effects of partitioning on phylogenetic inference. Mol. Biol. Evol. 32:1611–1627.

Kenny J.N., Si Y.W., Hayward A., Paps J., Chu K.H., Hui J.H.L. 2015. The phylogenetic utility and functional constraint of microRNA flanking sequences. Proc. R. Soc. Lond. B Biol. Sci. 282:20142983.

Kocot K.M., STruck T.H., Merkel J., Waits D.S., Todt C., Brannock P.M., Weese D.A., Cannon J.T., Moroz L.L., Leib B., Halanych K.M. Forthcoming. Phylogenomics of Lophotrochozoa with consideration of systematic error. Syst. Biol.

Kosiol C., Goldman N. 2005. Different versions of the Dayhoff rate matrix. Mol. Biol. Evol. 22:193–199.

Kück P., Meusemann K. 2010. FASconCAT: convienent handling of data matrices. Mol. Phylogen. Evol. 56:1115–1118.

Kuhner M.K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

Kupczok A., Schmidt H.A., von Haeseler A. 2010. Accuracy of phylogeny reconstruction methods combining overlapping gene datasets. Algorithm Mol Biol 5:37.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–1701.

Lanfear R., Calcott B., Kainer D., Mayer C., Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14:82.

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

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

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

Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Le S.Q., Lartillot N., Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci. 363:3965–3976.

Li Y., Kocot K.M., Whelan N.V., Santos S.R., Waits D.S., Thornhill D.J., Halanych K.M. Forthcoming. Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods. Zool. Scr.

Liu Y., Steenkamp E.T., Brinkmann H., Forget L., Philippe H., Lang B.F. 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support. BMC Evol. Biol. 9:272.

Ludwig D., Walters C.J. 1986. Are age-structured models appropriate for catch-effort data? Can. J. Fish. Aquat. Sci. 42:1066–1072.

Luo H. 2015. Evolutionary origin of a streamlined marine bacterioplankton lineage. The ISME Journal 9:1423–1433.

Moore M.J., Soltis P.S., Bell C.D., Burleigh J.G., Soltis D.E. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc. Natl. Acad. Sci. USA 107:4623–4628.

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., Girardo D.O., Fodor A., Gusev F., Sanford R., Bruders R., Kittler E., Mills C.E., Rast J.P., Derelle R., Solovyev V.V., Kondrashov F.A., Swalla B.J., Sweedler J.V., Rogaev E.I., Halanych K.M., Kohn A.B. 2014. The ctenophore genome and the evolutionary origins of neural systems. Nature 510:109–114.

Nesnidal M., Helmkampf M., Bruchhaus I., El-Matbouli M., Hausdorf B. 2013. Agent of whirling disease meets oprhan worm: phylogenomic analyses firmly place Myxozoa in Cnidaria. PLOS One 8:e54576.

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. 2013. Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogen. Evol. 67:223–233.

O'Hara T.D., Hugall A.F., Thuy B., Moussalli A. 2014. Phylogenomic reconstruction of the class Ophiuroidea unlocks a global microfossil record. Curr. Biol. 24:1874–1879.

Pagel M., Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571–581.

Philippe H., Brinkmann H., Copley R.R., Moroz L.L., Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011a. Acoelomorph flatworms are deutrostomes related to *Xenoturbella*. Nature 470:255–258.

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

Philippe H., Delsuc F., Brinkmann H., Lartillot N.. 2005. Phylogenomics. Annu. Rev. Ecol., Evol. Syst. 36:541–562.

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. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19:706–712.

Pick K.S., Philippe H., Schreiber F., Erpenbeck D., Jackson D.J., Wrede P., Wiens M., Alié A., Morgenstern B., Manuel M., Wörheide G. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. 27:1983–1987.

Pisani D., Feuda R., Peterson K.J., Smith A.B. 2012. Resolving phylogenetic signal from noise when divergence is rapid: a new look at the old problem of echinoderm class relationships. Mol. Phylogen. Evol. 62:27–34.

Pisani D., Pett W., Dohrmann M., Feuda R., Rota-Stabelli O., Philippe H., Lartillot N., Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. USA 112:15402–15407.

R Core Development Team. 2015. R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

Rambaut A., Drummond A.J. 2007. Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer.

Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rokas A., Carroll S.B. 2006. Bushes in the tree of life. PLoS Biol. 4:e352.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol. Biol. Evol. 30:197–214.

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

Ryan J.F., Pang K., Schnitzler C.E., Nguyen A-D., Moreland R.T., Simmons D.K., Koch B.J., Francis W.R., Havlak P., Smith S.A., Putnam N.H., Hadock S.H.D., Dunn C.W., Wolfsberg T.G., Mullikin J.C., Martindale M.Q., Baxevanis A. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342:1242592.

Schliep K.P. 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Simakov O., Kawashima T., Marlétaz F., Jenkins J., Koyanagi R., Mitros T., Hisata K., Bredeson J., Shoguchi E., Gyoja F., Yue J.-X., Chen Y.-C., Freeman R.M., Sasaki A., Hikosaka-Katayama T., Sato A., Fujie M., Baughman K.W., Levine J., Gonzalez P., Cameron C., Fritzenwanker J.H., Pani A.M., Goto H., Kanda M., Arakaki N., Yamasaki S., Qu J., Cree A., Ding Y., Dinh H.H., Dugan S., Holder M., Jhangiani S.N., Kovar C.L., Lee S.L., Lewis L.R., Morton D., Nazareth L.V., Okwuonu G., Santibanez J., Chen R., Richards S., Muzny D.M., Gillis A., Peshkin L., Wu M., Humphreys T., Su Y.-H., Putnam N.H., Schmutz J., Fujiyama A., Yu J.-K., Tagawa K., Worley K.C., Gibbs R.A., Kirschner M.W., Lowe C.J., Satoh N., Rokhsar D.S., Gerhart J. 2015. Hemichordate genomes and deuterostome origins. Nature 527:459–465.

Siu-Ting K., Gower D.J., Pisani D., Kassahun R., Gebresenbet F., Menegon M., Mengistu A.A., Saber S.A., de Sá R., Wilkinson M., Loader S.P. 2014. Evolutionary relationships of the critically endangered frog *Ericabatrachus baleensis* Largen, 1991 with notes on incorporating unsampled taxa into large-scale phylogenetic analyses. BMC Evol. Biol. 14:44.

Stamatakis A. 2006. Phylogenetic models of rate heterogeneity: a high performance computing perspective. 20th International Parallel and Distributed Processing Symposium. Rhodes Island, IEEE.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stamatakis A., Meier H., Ludwig T. 2004. New fast and accurate heuristics for inference of large phylogenetic trees. Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International. IEEE. p. 193.

Strope C.L., Abel K., Scott S.D., Moriyama E.N. 2009. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. Mol. Biol. Evol. 26:2581–2593.

Struck T.H., Wey-Fabrizius A.R., Golombek A., Hering L., Weigert A., Bleidorn C., Klebow S., Iakovenko N., Hausdorf B., Petersen M., Kück P., Herlyn H., Hankeln T. 2014. Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. Mol. Biol. Evol. 31:1833–1849.

Tarver J.E., dos Reis M., Mirarab S., Moran R.J., Parker S., O'Reilly J.E., King B.L., O'Connell M.J., Asher R.J., Warnow T., Peterson K.J., Donoghue P.C.J., Pisani D. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. Genome Biol. Evol. 8:330–344.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17:57–86.

Telford M.J., Budd G.E., Philippe H. 2015. Phylogenomic insights into animal evolution. Curr. Biol. 25:R876–R887.

Timme R.E., Bachvaroff T.R., Delwiche C.F. 2012. Broad phylogenomic sampling and the sister lineage of land plants. PLoS One 7:e29696.

Tsagkogeorga G., Turon X., Hopcraft R.R., Tilak M-K., Feldstein T., Shenkar N., Loya Y., Huchon D., Duouzery E.J.P., Delsuc F. 2009. An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models. BMC Evol. Biol. 9:187.

Whelan N.V., Kocot K.M., Halanych K.M. 2015a. Employing phylogenomics to resolve the relationships among cnidarians, ctenophores, sponges, placozoans and bilaterians. Integr. Comp. Biol. 55:1084–1095.

Whelan N.V., Kocot K.M., Moroz L.L, Halanych K.M. 2015b. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA 112:5773–5778.

Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Williams T.L., Moret B.M.E. 2003. An investigation of phylogenetic likelihood methods. Bioinformatics and Bioengineering, 2003. Proceedings. Proceedings Third IEEE Symposium Bioinformatics and Bioengineering. IEEE. p. 79–86.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.