1-1-2018

Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological?

Emilie J. Richards
University of Hawai'i at Mānoa

Jeremy M. Brown
Louisiana State University

Anthony J. Barley
University of Hawai'i at Mānoa

Rebecca A. Chong
University of Hawai'i at Mānoa

Robert C. Thomson
University of Hawai'i at Mānoa

Recommended Citation

Richards, E., Brown, J., Barley, A., Chong, R., & Thomson, R. (2018). Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological?. *Systematic Biology, 67* (5), 847-860. https://doi.org/10.1093/sysbio/syy013

Syst. Biol. 67(5):847–860, 2018

DOI:10.1093/sysbio/syy013
Advance Access publication February 19, 2018

Variation Across Mitochondrial Gene Trees Provides Evidence for Systematic Error: How Much Gene Tree Variation Is Biological?

EMILIE J. RICHARDS[1,2,*], JEREMY M. BROWN[3], ANTHONY J. BARLEY[1], REBECCA A. CHONG[1], AND ROBERT C. THOMSON[1]

[1]*Department of Biology, University of Hawai'i, 2538 McCarthy Mall, Edmondson Hall 2016, Honolulu, HI 96822, USA;* [2]*Department of Biology, University of North Carolina, 120 South Road, Coker Hall CB 3280 Chapel Hill, NC 27599, USA; and* [3]*Department of Biological Sciences and Museum of Natural Science, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA*

**Correspondence to be sent to: Department of Biology, University of North Carolina, 120 South Road, Coker Hall CB 3280 Chapel Hill, NC 27599, USA. E-mail: ejr@live.unc.edu.*

Received 01 August 2017; reviews returned 19 January 2018; accepted 15 February 2018
Associate Editor: Rachel Mueller

Abstract.—The use of large genomic data sets in phylogenetics has highlighted extensive topological variation across genes. Much of this discordance is assumed to result from biological processes. However, variation among gene trees can also be a consequence of systematic error driven by poor model fit, and the relative importance of biological vs. methodological factors in explaining gene tree variation is a major unresolved question. Using mitochondrial genomes to control for biological causes of gene tree variation, we estimate the extent of gene tree discordance driven by systematic error and employ posterior prediction to highlight the role of model fit in producing this discordance. We find that the amount of discordance among mitochondrial gene trees is similar to the amount of discordance found in other studies that assume only biological causes of variation. This similarity suggests that the role of systematic error in generating gene tree variation is underappreciated and critical evaluation of fit between assumed models and the data used for inference is important for the resolution of unresolved phylogenetic questions. [Gene tree discordance; phylogenomics; posterior prediction; model adequacy.]

Large genomic data sets are increasingly being used for phylogenetic inference because they increase statistical power and reduce stochastic error, which can lead to greater phylogenetic resolution (Gee 2003; Rokas et al. 2003; Rokas and Carroll 2005). The use of these large data sets has highlighted the extensive topological variation that can be found across genes. For example, phylogenomic analysis of 1070 genes from 23 yeast genomes resulted in 1070 distinct gene trees (Salichos and Rokas 2013). This discordance is frequently viewed as the outcome of one of several biological sources of gene tree variation: incomplete coalescence, horizontal gene transfer, or gene duplication/loss events (reviewed by Maddison 1997; Nakhleh et al. 2013). Explicit modeling of these processes, when possible, can accommodate this variation during the inference of a species tree (Degnan and Rosenberg 2009; Edwards 2009; Boussau et al. 2013; Mirarab et al. 2016; Szöllősi et al. 2015; Edwards et al. 2016). However, variation among gene trees can also be a consequence of systematic error that arises when the model used for estimating the gene tree fits the data poorly. The relative importance of biological vs. methodological factors in explaining gene tree variation is a major unresolved question in phylogenetics.

When the model fails to account for important features of the data, inferences and measures of confidence can be inaccurate (Huelsenbeck and Hillis 1993; Yang 1994; Swofford et al. 2001; Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004; Brown and Lemmon 2007; Brown and Thomson 2017). Because the complexity of data sets grows with size, the potential for poor model fit to bias inferences also grows. Increasing data set size may reduce stochastic error, but it can also exacerbate systematic error and lead to high confidence in erroneous phylogenies (Phillips et al. 2004; Delsuc et al. 2005; Jeffroy et al. 2006; Philippe et al. 2011; Kumar et al. 2012). Several cases are now known where different genomic data sets support conflicting phylogenetic hypotheses with very high statistical support (e.g., Dunn et al. 2008; Philippe et al. 2009; Schierwater et al. 2009; Whelan et al. 2015), sometimes implying very different scenarios for the evolution of important traits (e.g., the origin of nervous systems). The relative roles of biological processes and systematic error in causing this conflict is not yet well understood.

One challenge with evaluating the contributions of systematic error to gene tree discordance is that biased inferences are difficult to detect reliably given that the true evolutionary history among most taxa is unknown. However, we can greatly reduce the confounding effects of biological processes on our ability to identify systematic error by making use of the mitochondrial genome as a model system. The entire mitochondrial genome is expected to have the same evolutionary history because it is haploid and uniparentally inherited, so recombination does not typically occur. While recombination and biparental inheritance have been documented in animal mitochondrial genomes, these occurrences appear to be rare relative to the ubiquity of such events in nuclear genomes (reviewed in White et al. 2008). Therefore, analyses using individual mitochondrial genes should result in concordant gene trees. Strong conflict among topologies inferred from different mitochondrial genes would therefore most easily be explained by systematic error during inference of gene trees.

While biased inferences are often difficult to identify directly, several approaches have been proposed to detect poor fit between models and data (e.g., Goldman 1993; Huelsenbeck et al. 2001; Bollback 2002; Nielsen

2002; Foster 2004; Rodrigue et al. 2009; Ripplinger and Sullivan 2010; Brown 2014; Reid et al. 2014; Slater and Pennell 2014; Doyle et al. 2015; Duchêne et al. 2015, 2017; Barley and Thomson 2016; Gruenstaeudl et al. 2016). When fit is poor, the potential exists for inferences to be biased. However, not all instances of poor fit will result in erroneous phylogenetic estimates. Comparison of inferred gene trees and measures of model fit across tightly linked mitochondrial genes offers a unique opportunity to understand how the outcome of model fit tests relate to gene tree variation driven by systematic error. One natural approach to conducting such tests in a Bayesian framework is known as posterior prediction, wherein samples are drawn from a posterior distribution and used to simulate many replicated "predictive" data sets. By comparing the predictive to the empirical data sets in various ways, the extent to which the model captures salient features of the data can be studied.

Here we analyze mitochondrial genomes for several sets of tetrapod species to characterize the extent of gene tree discordance and, using posterior prediction, begin to explore how model fit contributes to this discord. We find that the discordance among mitochondrial gene trees is extensive and similar to the amount of discordance found in studies of nuclear gene tree variation, where such discordance is assumed to result from biological factors (also see Meiklejohn et al. 2014). Additionally, this discordance is often strong and not driven by a lack of information in individual genes (i.e., stochastic error). Posterior predictive assessments provide additional evidence for the influence of systematic error in driving discordance among the gene trees in this study. However, more work is needed to determine specific causes of poor model fit and how these drive systematic error.

Methods

Data sets

We obtained all available (as of 31 July 2014) complete tetrapod mitochondrial genome sequences from GenBank, which we organized into six data sets comprising the major lineages within the clade: Crocodilians ($n=20$), Turtles ($n=53$), Squamates ($n=120$), Amphibians ($n=157$), Birds ($n=253$), and Mammals ($n=575$). We extracted all 13 protein-coding genes from each mitochondrial genome based on GenBank genome annotations. Multiple sequence alignments were then constructed based on translated codons for each mitochondrial protein-coding gene in each data set using the MUSCLE algorithm implemented in Geneious v 8 (Edgar 2004; Kearse et al. 2012). Alignment files are provided in the Supplementary Materials available on Dryad at http://dx.doi.org/10.5061/dryad.hj07m.

Initial Phylogenetic Analyses

For the initial phylogenetic analysis of each of the 78 gene alignments (i.e., 6 clades × 13 genes), we selected a best-fitting substitution model according to the Akaike information criterion (Akaike 1974) corrected for small sample size (AICc) implemented in jModelTest v 2.2 (Darriba et al. 2012). Details on the specific model chosen for each gene alignment and alignment lengths are provided in Supplementary Table S1 available on Dryad. We first obtained posterior distributions of trees and other parameters for each alignment using Markov chain Monte Carlo (MCMC) as implemented in MrBayes v 3.2.5, with the selected model and default prior settings (Ronquist et al. 2012). For each analysis, we used two independent runs (each with four Metropolis-coupled chains) and saved the state of the chains every 1000 generations. The MCMC was run until the postburn-in posterior distributions for each analysis contained 10,000 converged samples. We checked for convergence of the continuous parameters using Tracer v 1.6 (Rambaut et al. 2014) and considered a run converged when traces for all parameters appeared to be sampling from a stationary distribution and had ESS values above 1000. We assessed convergence of the tree topology using the R package rWTY v 0.1 (Warren et al. 2017). Runs were considered converged when the bipartition posterior probabilities in the MCMC chain reached a stationary frequency in the cumulative plots and showed strong correlations (Pearson's $r > 0.95$) across runs.

Characterization of Gene Tree Heterogeneity

To characterize the extent of gene tree heterogeneity among the 13 genes for a given clade, we calculated three different types of summary trees (majority-rule consensus trees, 95% consensus trees, and maximum clade credibility trees) from the posterior distribution for each gene. All summary trees are available in the Supplementary Materials. We then calculated the pairwise average number of bipartitions in a gene tree that conflict with another gene tree, hereby referred to as incompatible splits, between all of gene trees of the same type of summary tree for a given clade (Doyle et al. 2015; available from https://github.com/vinsondoyle/treeProcessing). This measure is related to the more widely used Robinson–Foulds (RF) distance (Robinson and Foulds 1981), but focuses on the number of bipartition-specific conflicts rather than bipartitions that are present in one tree but not the other. The practical effect of this change is that polytomies do not contribute to the distance. Because we are primarily interested in identifying strongly supported differences among gene trees, this was a useful property for our study. We then visualized the distributions of pairwise tree-to-tree distances among genes with violin plots using the R package ggplot2 v 2.1.0 (Wickham 2009). Since we were interested in distinguishing differences among gene trees that were strongly supported (and are more likely to be driven by systematic error) from those that had little statistical support (and may simply arise from stochastic error), we focused on discordance between 95% consensus trees (calculated using Dendropy v 4.0.3; Sukuraman and Holder 2010) for the rest of the analyses in this study.

TABLE 1. Descriptions of the model performance test statistics employed in this study

Test statistics	Type	Description	Source
Multinomial likelihood	Data	Related to the frequency of site patterns in an alignment	(Goldman 1993; Bollback 2002)
χ^2	Data	Captures variation in nucleotide frequencies	(Huelsenbeck et al. 2001; Foster 2004)
Tree length mean	Inference	The mean of marginal distributions of tree length	(Brown 2014)
Tree length variance	Inference	The variance of marginal distributions of tree length	(Brown 2014)
Entropy	Inference	The unevenness of support in the posterior distribution of trees	(Brown 2014)
Quantile-based test statistics	Inference	The overall similarity in the posterior distributions of trees based on the dispersion of trees in the posterior. Can be assessed at different positions along the distribution (see below).	(Brown 2014)
Interquartile range	Inference	The interquartile range of tree-to-tree distances	(Brown 2014)
First quartile	Inference	The first quartile of tree-to-tree distances	(Brown 2014)
Median	Inference	The median of tree-to-tree distances	(Brown 2014)
Third quartile	Inference	The third quartile tree-to-tree distances	(Brown 2014)
99th percentile	Inference	The 99th percentile of tree-to-tree distances	(Brown 2014)
999th quantile	Inference	The 999–1000th quantile of tree-to-tree distances	(Brown 2014)
9999th quantile	Inference	The 9999–10,000th quantile of tree-to-tree distances	(Brown 2014)

The type of test statistic refers to whether they are values based on the data themselves or the resulting inferences.

We also visually assessed gene tree heterogeneity by looking for non-overlapping sets of topologies among the thirteen genes in a low-dimensional projection of tree space created with non-linear dimensionality reduction (NLDR) using Treescaper v 1.0.0 (Huang et al. 2016; Wilgenbusch et al. 2017). Two-dimensional projections were created for each clade based on pairwise RF tree-to-tree distances of 3250 trees taken from the posterior distributions of all genes (250 trees per gene) using the curvilinear component analysis and stochastic gradient decent optimization recommended in Wilgenbusch et al. (2017).

Model Performance Assessment

We assessed the absolute fit of the selected models to their respective gene alignments by performing posterior predictive assessments with both data- and inference-based test statistics. Data-based test statistics measure some characteristic of the data itself (e.g., the frequency distribution of site patterns in the alignment or variation in base composition across taxa; Goldman 1993, Huelsenbeck et al. 2001) and inference-based test statistics measure some characteristic of the resulting inference (e.g., width of the posterior distribution of trees; Brown 2014). A list of the test statistics used in this study and brief descriptions of what they measure is provided in Table 1.

For the data-based assessments, posterior predictive simulation of data sets for each gene was performed using PuMA v 0.909 (Brown and ElDabaje 2009) and SeqGen v 1.3.2 (Rambaut and Grassly 1997) with 1000 parameter values and trees drawn uniformly from postburn-in MCMC samples. The data-based test statistics require that missing data be excluded from the alignments, so we removed missing data from sequences prior to simulation using PAUP* v 4.0b10 (Swofford 2003). Using each set of 1000 posterior predictive data sets and the corresponding empirical data set, we conducted two data-based assessments of model performance to characterize the ability of the model to replicate features of the empirical data set. We calculated the multinomial likelihood test statistic (Goldman 1993; Bollback 2002; Table 1) using PuMA (Brown and ElDabaje 2009) and the χ^2 statistic (Table 1) using the P4 python phylogenetics package (Foster 2004). We also checked whether the poor model fit detected in the data-based model performance tests might be driven by the presence of missing data in the original alignments. To do this, we excluded alignment columns that contained missing data from the empirical alignments and redid the model selection, generated new posterior distributions, and redid the data-based model adequacy assessments. We found that the presence of missing data leads to relatively small changes in PPES that cannot explain poor model fit (Supplementary Figs S1 and S2 available on Dryad).

For the inference-based assessments, we repeated the posterior predictive simulation of data sets for each gene alignment with 100 parameter values and trees drawn uniformly from postburn-in MCMC samples. Only 100 posterior predictive data sets were used for these tests due to the much higher computational demands involved in the inference-based assessments. We used a custom python script (available from https://github.com/jembrown/repMissPatterns) to substitute missing data for nucleotides in each simulated data set to match the patterns observed in the empirical data set. For each posterior predictive data set, we obtained a posterior distribution of trees and other parameters using MrBayes v 3.2.5 with the model and priors assumed during analysis of the empirical data. To assess convergence, we chose five replicates at random from each gene and performed the same convergence analysis used in the initial phylogenetic analyses. When all five replicates met the convergence criteria described above, the remaining 95 predictive phylogenetic analyses were considered to have converged if the average standard deviation of split frequencies also

fell below 0.01. All inference-based test statistics that were proposed in Brown (2014) were calculated in this study (Table 1) using AMP (Brown 2014, available from https://www.github.com/jembrown/amp) on 10,000 trees uniformly sampled from the postburn-in posterior distribution generated for a given data set.

After test statistic values were calculated, we quantified the position of the empirical value relative to the posterior predictive distribution for each test using effect sizes (Doyle et al. 2015). Effect sizes for each test statistic were calculated as the absolute value of the difference between the empirical and the median posterior predictive value divided by the standard deviation of the posterior predictive distribution. These effect sizes are hereafter referred to as posterior predictive effect sizes (PPES).

Correlation Among Measures of Model Performance

For each data set, we ranked genes according to the model performance results and then tested for correlations among the rankings. This allowed us to assess whether the test statistics generally agreed on model performance for each gene. To do so, we calculated the rank for each gene for each test statistic based on PPES and then calculated pairwise Spearman's rank correlation coefficients between test statistics using the R package "stats" v 3.2.2 (R Core Team 2015). For all pairwise combinations, we then selected one of the pair of test statistics at random and randomly shuffled its ranking of genes, recalculating the correlation coefficient. We repeated this procedure 1000 times in order to create a null distribution of correlation coefficients and assess the significance of the observed correlation. Correlations among test statistics were considered significant if less than 5% of the coefficients from the randomized rankings were greater than or equal to the correlation coefficient from the observed rankings.

Relationship Between Model Fit and Gene Tree Variation

As a rough measure of accuracy in the gene tree estimates, we were interested in quantifying how different the gene trees for each clade were from widely accepted estimates of phylogeny from the literature, as well as how this might relate to measures of model performance. To do so, we selected a "reference tree" from the literature for each clade that we could use as the current best estimate for that clade (Crocodilians: Oaks and Dudley 2011; Turtles: Thomson and Shaffer 2010; Squamates: Wiens et al. 2012; Amphibians: Pyron and Wiens 2011; Birds: Prum et al. 2015; Mammals: Meredith et al. 2011). Each reference tree was selected based on the availability of its posterior distributions/summary trees for analysis and similarity in taxa to those used in this study. Because we are primarily interested in strongly supported differences, we calculated the number of incompatibilities between the 95% consensus tree for each gene to the reference tree, trimming taxa as necessary so that taxon sampling matched between the two trees. We then carried out linear regression analysis between the tree distance and the PPES for each gene and model performance test. We tested for variation in mean PPES across genes among the six data sets for all 12 test statistics using the ANOVA function in the R package "stats" v 3.5.0 (R Core Team 2017). We also performed linear regression analysis between PPES and gene alignment length.

Impacts on Species Tree Inference

In order to measure the effects that the observed gene tree heterogeneity may have on analyses that attribute such variation to biological causes, we used the gene trees to infer a coalescent species tree using gene tree summary methods and duplication, transfer, and loss (DTL) events from gene tree-species tree reconciliation methods. We inferred species trees from each data set using the gene tree summary method ASTRAL (v. 5.5.7; Mirarab and Warnow 2015; Zhang et al. 2017) with default parameter settings. We then calculated the number of incompatibilities between the inferred species trees and a tree based on an unpartitioned concatenated data set of all 13 genes generated using RAxML (v. 8.2.10; Stamatakis 2014) to determine the extent of topological discord generated from these two different types of analyses. Such differences are typically attributed to variation in coalescence among gene trees rather than errors in gene tree estimation. We also looked at the impacts of gene tree estimation error on analyses that attribute gene tree variation to another biological source of variation: gene DTL events. We estimated DTL events for each gene using ALE (Szöllősi et al. 2013), an approach that reconciles a sample of gene trees with a putative species tree under a model that allows for phylogenetic discord in the form of DTL events, using a uniform sample of 10,000 topologies from the posterior distribution for the gene and the species tree we inferred above using ASTRAL for its respective data set.

Results

Agreement Among Gene Trees

Extensive gene tree heterogeneity was present across all data sets (Fig. 1). Across all data sets and consensus methods, the number of incompatibilities between genes was much greater than zero, with the exception of the Crocodilian data set, where most genes had identical 95% consensus gene trees. The amount of disagreement varied across the types of summary tree in a way that would be expected. Maximum clade credibility trees are the most highly resolved of the summary trees, but can contain many weakly supported nodes. Thus, stochastic error in the tree estimate will increase tree-to-tree distances relative to other types of summary trees. Conversely, the 95% consensus contain fewer

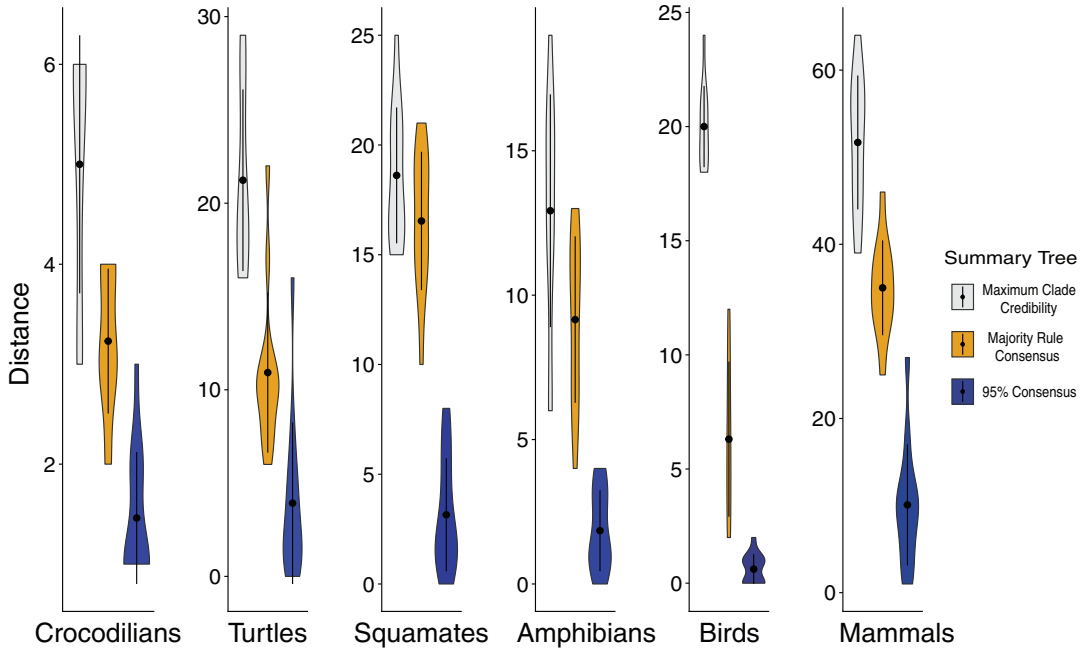

FIGURE 1. The total number of pairwise incompatibilities among all gene trees for the 6 data sets. Distances are shown between maximum clade credibility trees, majority-rule consensus trees, and 95% consensus (95C) trees. The circle represents the mean number of incompatibilities and the black bars around it represent one SD around the mean. The width of the violin plot indicates the density of gene trees with a particular tree-to-tree distance to another gene tree in the data set. There is extensive variation in topology among gene trees in each clade and across summary tree types, with the exception of some 95C trees in the Crocodilian, Turtle, and Squamate data sets.

nodes, although all have strong support, leading to comparatively smaller tree-to-tree distances. In this latter case, the tree-to-tree distance is more likely to highlight differences that can only be explained by systematic error. Among the 95% consensus trees, tree-to-tree distances were also substantially greater than zero, indicating the presence of strongly supported yet conflicting topologies among genes. In the Crocodilian data set containing 20 species, the majority of gene trees were well resolved and largely congruent. The conflicts among the Crocodilian gene trees occurred only among species-level relationships at the tips of the phylogenies. Gene trees for the larger data sets were less well resolved and conflicts among gene trees in the resolution of deeper relationships were more frequent.

We find similar patterns of gene tree heterogeneity in our low-dimensional projections of tree space across genes for each data set (Fig. 2). In all data sets except Crocodilians, we observe 13 distinct clusters of trees sampled from the posterior distributions of different genes. Some of these clusters show more substantial separation from other clusters (e.g., the cluster representing ND5 gene trees in the Turtle data set), suggesting stronger incongruence with other sets of gene trees.

The observed level of gene tree heterogeneity across tightly linked mitochondrial genes is qualitatively similar to that found in other studies of nuclear gene tree heterogeneity (Table 2). Some of these studies (e.g., Salichos and Rokas 2013) state the observed heterogeneity could have been caused by either biological or methodological sources, and that it is nearly impossible to determine their relative contributions. Other studies (e.g., Song et al. 2012; Zhong et al. 2013; Pease et al. 2016) attribute the heterogeneity to biological factors, mainly incomplete lineage sorting, and either rule out or do not consider systematic bias as a contributing factor. Most of the above-mentioned studies characterized the extent of gene tree heterogeneity by calculating pairwise RF distances among majority rule consensus trees of each locus in their data set. We also find high levels of gene tree discordance in our mitochondrial data sets when we use similar methods for characterizing gene tree heterogeneity (Table 2), indicating that systematic bias can cause levels of gene tree variation that are typically attributed to biological sources of variation.

Model Performance Assessments

The PPES resulting from the 12 model performance tests varied across genes and data sets, ranging from 0 to 1.12×10^{12} (Table 3, Supplementary Tables S2–S7 available on Dryad). This wide range is heavily influenced by Entropy, one of the inference-based test statistics, which exhibited little to no variance between posterior predictive simulations. Small differences between the empirical and median of the posterior predictive distributions lead to extremely large PPES values for some genes in all but the Crocodilians data set. This behavior of the test statistic stems from sensitivity to data set size and the complexity of sampling very large tree spaces, where the coarseness of MCMC sampling makes

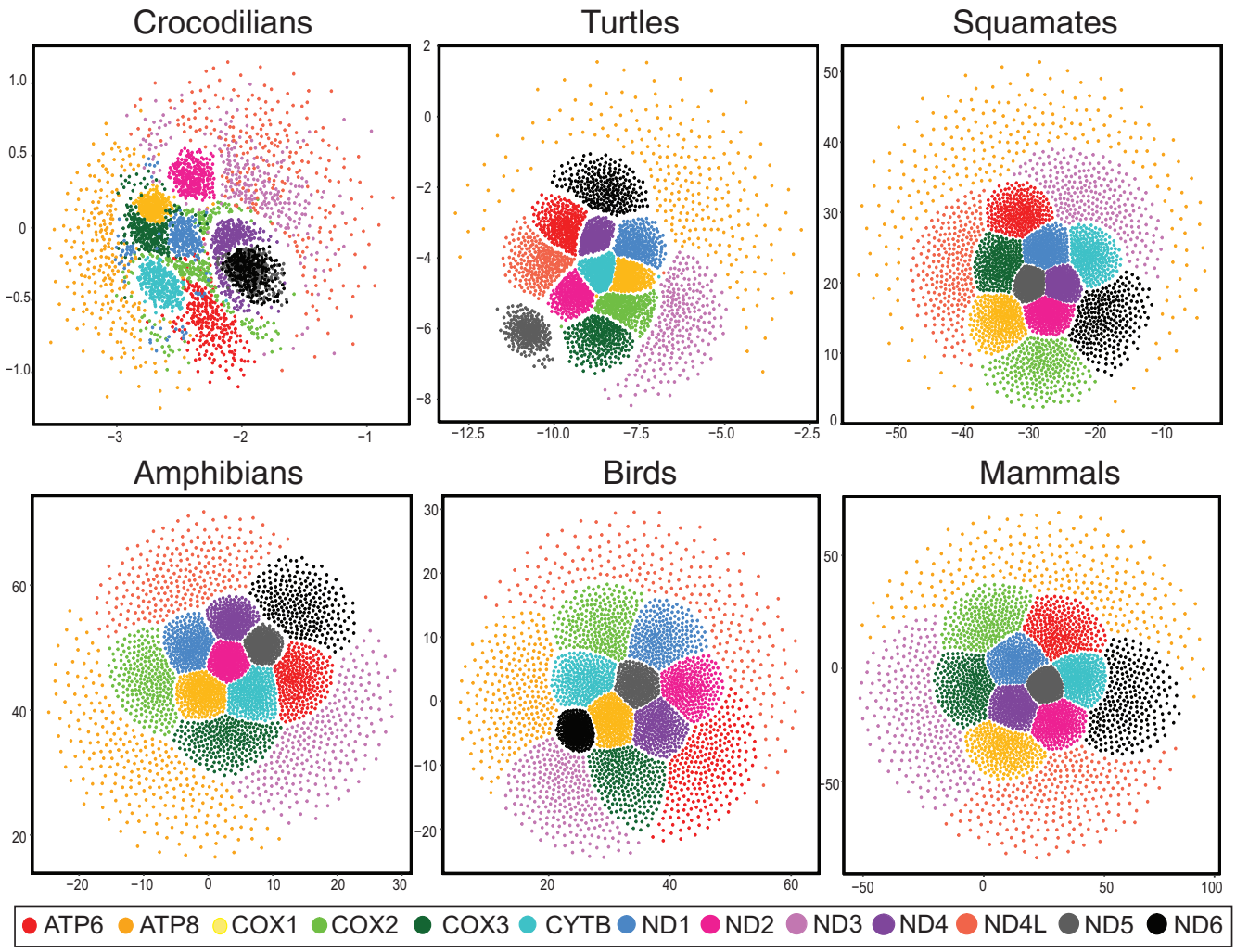

FIGURE 2. Two-dimensional NLDR representations of treespace for 13 mitochondrial genes based on RF distances between trees. Each point represents a tree taken from the posterior distribution of a given gene.

TABLE 2. Gene tree variation found in this study when compared with several other studies that focused on gene tree heterogeneity using multiple nuclear loci

Data set	Taxa	Genes	Distinct trees	Percent of possible trees found	Source
Crocodilians	20	13	12	92	This study
Turtles	53	13	13	100	This study
Squamates	120	13	13	100	This study
Amphibians	157	13	13	100	This study
Birds	253	13	13	100	This study
Mammals	575	13	13	100	This study
Yeast	23	1070	1070	100	Salichos and Rokas 2013
Vertebrates	18	1086	299	28	Salichos and Rokas 2013
Metazoans	21	225	224	99.5	Salichos and Rokas 2013
Eutherian mammals	37	447	440	98.3	Song et al. 2012
Land plants	32	184	182	98.9	Zhong et al. 2013
Tomatoes	29	2745	2743	99.9	Pease et al. 2016

it improbable to sample any individual topology more than once. In conventional phylogenetic analyses, where node probabilities are of primary interest, this issue is solved simply by summing up how frequently different bipartitions are sampled, rather than whole topologies. However, it becomes problematic when focusing on the frequencies of unique topologies, as we do here with the entropy test statistic. While large PPES for entropy might be meaningful for smaller data sets, it is unlikely that they represent extremely poor fit between

TABLE 3. The distribution of posterior PPES for each of the 12 model performance test statistics used in this study (Table 1) summarized across all six data sets

Test Statistic	Mean	SD	Median	Min	Max
Multinomial likelihood	1.65	1.64	1.42	0.002	11.4
χ^2	19.61	23.48	11.7	0.04	110.68
Tree length mean	1.91	1.85	1.35	0.026	8.21
Tree length variance	5.45	6.52	3.45	0.33	32.76
Entropy	6.61×10^{10}	2.48×10^{11}	0.96	0	1.12×10^{12}
Interquartile range	4.82	3.41	4.31	0	16.24
1st quartile	4.77	3.02	4.9	0	11.73
Median	4.95	3.16	5.19	0	12.28
3rd quartile	5.19	3.09	5.37	0	12.65
99th quantile	5.58	3.29	5.93	0	13.44
999th quantile	5.73	3.36	6.12	0	13.82
9999th quantile	5.79	3.35	6.27	0	13.89

TABLE 4. The distribution of PPES for each data set across 11 of the 12 test statistics

	Data-based test statistics					Inference-based test statistics				
Data set	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max
Crocodilians	4.15	4.57	1.58	0.16	13.92	1.05	0.78	1.03	0	3.16
Turtles	3.48	4.02	1.78	0.04	15.08	2.21	2.04	1.97	0	14.25
Squamates	12.29	14.77	3.05	0.002	49.1	6.15	3.04	6.14	0.21	28.54
Amphibians	27.82	36.06	5.58	0.09	110.68	5.99	2.37	5.67	1.47	18.08
Birds	5.29	6.17	1.86	0.09	21.46	5.19	2.37	5.47	0.07	10.81
Mammals	10.77	13.39	8.63	0.13	45.89	8.63	3.99	9.62	0.87	16.24

Entropy was removed from the pool of test statistics summarized in this table because of the extreme outlier PPES of this test statistic across the majority of the data sets (see text). PPES values for the entropy test statistic are provided in Supplementary Tables S4–S9 available on Dryad.

the model and the data for many of the large trees sampled here, where almost every topology sampled in the posterior is unique. One way to combat this issue and improve model performance assessments that use posterior probabilities of trees would be to calculate conditional clade probabilities (Höhna and Drummond 2012) or conditional clade distributions (Larget 2013). These values can provide better estimates of tree probabilities when the posterior distribution of trees is particularly diffuse and MCMC sampling alone is not sufficiently precise.

When entropy was excluded, data-based test statistics appeared to reject model fit among genes more strongly than inference-based test statistics across all six data sets, with larger PPES on average (Table 4). This result makes sense, since poor model fit must manifest itself at the level of the data in order for inferences to be affected, but not all model deficiencies noticeable in the data will affect inference. PPES for data-based test statistics ranged from 0.002 to 110.78, indicating a large range of fit between models and empirical data. The range of PPES for inference-based test statistics was smaller than for data-based test statistics and varied across data sets (Table 4). For Crocodilians, PPES across inference-based test statistics were typically small, ranging from 0 to 3.16 (Table 4 and Supplementary Table S2 available on Dryad), suggesting that the selected models appear to fit the Crocodilian gene alignments better than for the other data sets, although this may be due to differences in power to detect poor model performance across data sets of different sizes. For Turtles, PPES ranged from 0 to 14.25 (Table 4 and Supplementary Table S3 available on Dryad), indicating a mixture of model fit. Similar variation in model fit across genes was also found for the larger data sets of Squamates, Amphibians, Birds, and Mammals (Table 4, Supplementary Tables S4–S7 available on Dryad). Indeed, there were significant differences in mean PPES between data sets across nearly all test statistics (with the exception of multinomial likelihood; Supplementary Table S8 available on Dryad). While this variation may stem from differences in data set size, it is not possible to rule out that other factors that may vary between data sets. PPES and gene length were not correlated across most test statistics (with the exception of the data-based χ^2 statistic; Supplementary Table S9 available on Dryad).

Correlation Among Measures of Model Performance

Across all data sets, gene rankings were significantly correlated among the quantile-based test statistics that quantify the distances between trees in posterior distributions (Fig. 3). Within the Crocodilian and Squamate data sets, the gene rankings for the mean and variance of tree length were significantly correlated with each other. Within the Crocodilian data set, gene rankings based on entropy were correlated with gene rankings among the quantile-based test statistics. We observed a few other correlations, although these were largely weak and idiosyncratic among data sets (Fig. 3).

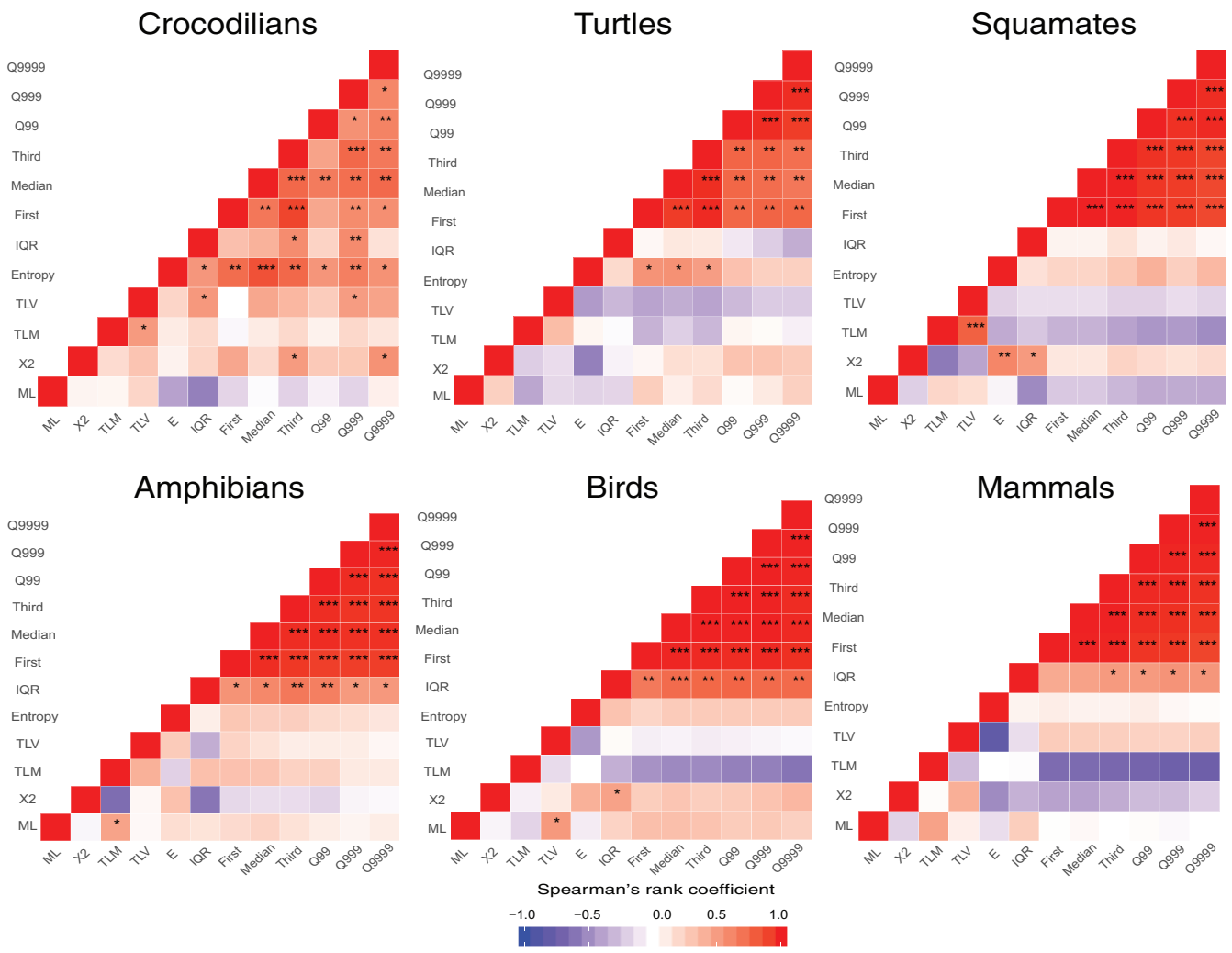

FIGURE 3. Heatmap of the Spearman's rank correlation coefficient between gene rankings among model performance tests based on PPES. Model performance tests include multinomial likelihood (ML), composition heterogeneity (X^2), tree length mean (TLM), tree length variance (TLV), statistical entropy (E), interquartile range (IQR), first quartile (First), median, third quartile (Third), 99th percentile (Q99), 999th–1000 quantile (Q999), and 9999–10000th quantile (Q9999) of tree to tree distances in posterior distributions. Stars indicate positive correlations that are significant at a significance threshold of 0.05 (*), 0.01 (**), and 0.001 (***).

Relationship Between Model Fit and Gene Tree Variation

The amount of strongly supported conflict between gene trees and reference trees varied across data sets and was low overall for Crocodilians and Birds and somewhat higher in the other clades (Table 5). There was no simple overall relationship between tree distance and PPES (Fig. 4, Supplementary Table S10 available on Dryad). Although genes did vary in their PPES, increasing PPES did not necessarily correspond to decreasing congruence between gene trees and reference trees across all data sets. However, we did observe some significant positive correlations between PPES and incongruence with the reference tree (e.g., for the 999–1000th and 9999–10,000th quantile-based test statistic in the Turtle data set; Fig. 4). We also observed some significant negative correlations in the same test statistics for the Crocodilian and Bird data sets. The negative relationships in these data sets may have to do with the combined effects of 1) a lack of strong disagreement among the gene trees and the reference tree (Table 5) and 2) an interaction between the power of a test statistic to detect poor model performance with the power of a gene to precisely estimate the phylogeny (i.e., the shortest genes often have the smallest PPES as well as the fewest incompatibilities with the reference tree, due to lack of information rather than poor fit of the model). Indeed, there was a weak but significant correlation ($r^2 = 0.09$, slope = 0.02; P-value = 0.006) between the length of a gene and the number of incompatibilities between reference trees and gene trees.

While the relationship between poor model fit and topological conflict between the gene trees and reference tree appears to be complex, we do find several cases where these methods clearly identified systematic bias or other issues in the data. While inspecting PPES results, we noted two cases where a single gene was a large

TABLE 5. The percentage of compatible bipartitions between gene trees and reference trees for each clade

Gene	Crocs (20)	Turtles (49)	Squamates (35)	Amphibians (28)	Birds (33)	Mammals (104)
ATP6	95	88	77	96	97	88
ATP8	95	98	94	96	97	99
COX1	95	98	83	86	100	93
COX2	85	98	94	86	94	94
COX3	95	92	89	93	97	91
CYTB	90	92	97	100	100	83
ND1	95	98	94	100	97	87
ND2	90	100	80	89	97	90
ND3	95	96	97	96	97	96
ND4	90	90	94	96	100	89
ND4L	95	84	100	96	100	98
ND5	90	67	86	89	97	74
ND6	95	96	97	93	100	90

The number of taxa in each data set after trimming is provided in parentheses. The percentage of bipartitions agreed upon was calculated as the number of compatible nodes divided by the total number of nodes in the tree.

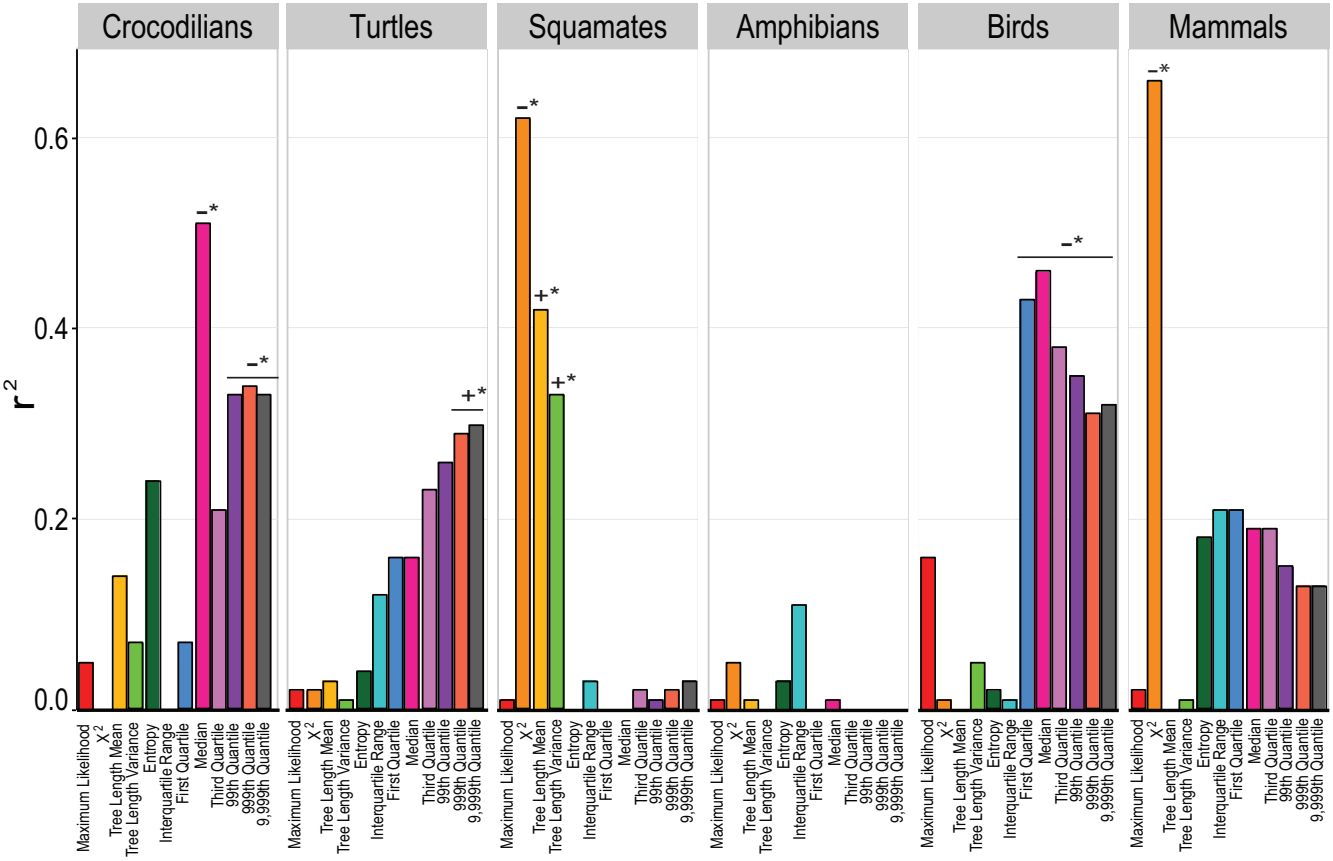

FIGURE 4. Relationship between PPES and the number of incompatibilities between 95% consensus gene tree and reference tree based on linear regression. Correlations with significantly positive or negative slopes are represented by (+*) and (−*), respectively. The values of the slope and 95% confidence intervals are provided in Supplementary Table S11 available on Dryad.

outlier for 1 or more model performance tests relative to all genes (Fig. 5). In both cases the PPES outlier was correctly signaling an issue in the analysis. Specifically, phylogenetic analysis of CYTB in the Squamate data set inadvertently included a misaligned region that affected four sequences. This misalignment increased the tree length mean and variance PPES for this gene, which were consequently much larger than these values for all other genes in the data set (Fig. 5a). The error also drove a spurious phylogenetic result that united a worm lizard with several blind snakes as a clearly erroneous clade. Once we corrected the misalignment, the tree length mean and variance PPES for CYTB were drastically reduced and the position of these taxa

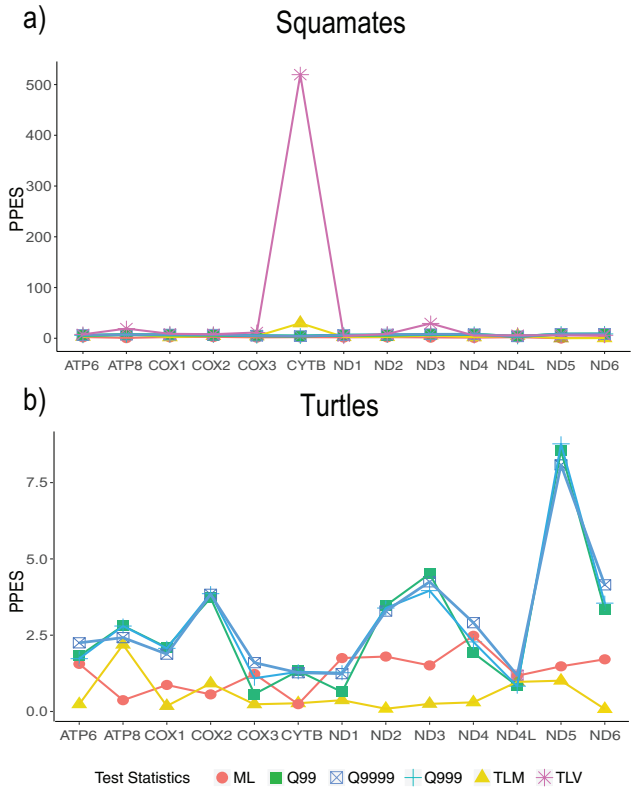

FIGURE 5. The PPES for each gene from a subset of model performance tests that highlight issues in the analysis. a) In the Squamate data set, the PPES (before the misalignment was corrected) associated with the tree length mean and variance test for the CYTB alignment are much larger than for the other genes. b) In the Turtle data set, the PPES associated with the quantile-based model performance tests of the ND5 alignment are twice as large as the PPES for ND3, the gene with the next largest PPES. Model performance tests shown here are the multinomial likelihood (ML), tree length mean (TLM), tree length variance (TLV), 99th percentile (Q99), 999th–1000 quantile (Q999), and 9999–10,000th quantile (Q9999) of tree-to-tree distances in posterior distributions.

in the gene tree returned to their more commonly accepted positions. This result also highlights that other sources of gene tree estimation error beyond those driven by poor model fit, such as alignment errors, can also contribute to non-biological sources of gene tree variation.

The quantile-based test statistics that measure the spread of the posterior distribution of trees also detected clear systematic error in the inference of the Turtle ND5 gene tree. The ND5 PPES for the 99–100th, 999–1000th, and 9999–10000th quantiles were at least twice as large as any other gene (Fig. 5b). The gene tree for ND5 supports a fundamentally different backbone of family-level relationships among turtles and contains a large number of topological conflicts with the reference tree in comparison to the rest of the gene trees in the Turtle data set (Table 5). Because the backbone relationships of turtles are well established (Barley et al. 2010; Thomson and Shaffer 2010; Crawford et al. 2015, Shaffer et al. 2017), we are confident that the ND5 gene tree is being influenced by systematic error. Supporting this, there was a significant positive correlation between the number of incompatibilities and model performance based on the quantile-based test statistics for this data set (Fig. 4, Supplementary Table S10 available on Dryad).

Impact of Non-Biological Sources of Gene Tree Variation on Species Tree Inference

As we would expect, the observed gene tree heterogeneity in this study translated into error in species tree analyses, both under a coalescent model and a model of DTL. For all data sets excluding the Crocodilians, we observed extensive topological discord between inferred species trees and concatenated trees (Supplementary Table S11 available on Dryad). The number of incompatible splits between the species and concatenated tree increased with data set size, from 10 for the Turtles data set to 88 for the Mammals data set. Similarly, when we analyze the posterior distributions for each using a method that is meant to detect DTL events, we erroneously infer transfers and losses for nearly all genes in all data sets (excluding a majority of the genes in the Crocodilian data set; Supplementary Table S12 available on Dryad), ranging from 1 to 45 events per gene.

DISCUSSION

Our analyses highlight several issues that should influence methodological choices for researchers moving forward. Most significantly, we find that the amount of gene tree variation in empirical data can be large,

irrespective of whether biological sources of gene tree variation (i.e., incomplete lineage sorting) are expected to play a significant role. Much of the conflict across genes in this study is strong and probably driven by systematic, rather than stochastic, errors. The gene tree heterogeneity we observed is qualitatively similar to other studies that attribute the variation solely to biological processes. This similarity suggests that the observation of variation among gene trees in empirical data should not necessarily be ascribed entirely to biological sources by default. Researchers should take care to check for more prosaic explanations of gene tree variation in their data (e.g., poor model fit driving systematic error) before applying a hierarchical model of gene tree variation (and assuming that it can adequately account for this variation).

"Species tree" approaches to analyzing multilocus alignments typically assume that the only source of discordance is biological (i.e., coalescent stochasticity). Consequently, non-biological sources of discordance can mislead these approaches. When we apply a coalescent based approach to estimate a species tree using the heterogeneous gene trees, we find considerable discord between species and concatenated trees. We also infer extensive DTL events across the data sets when we apply a species tree-gene tree reconciliation approach that assumes gene tree variation stems from gene-level events such as horizontal gene transfer. Only in the Crocodilian data set, where little gene tree heterogeneity is observed, is the impact of the observed gene tree variation on higher-level analyses minimal. Similar findings of large impacts of non-biological sources of gene tree variation on species tree analysis have been documented in other studies. For example, incomplete lineage sorting appears to be only a minor cause of observed phylogenetic discordance in placental mammals (Scornavacca and Galtier 2017). While several simulation-based studies have also shown that discordance related to stochastic errors in gene tree estimation can heavily influence species tree estimation (Huang et al. 2010; Molloy and Warnow 2018), the direct impact of systematic error on gene and species tree estimations is not well characterized. With increasing application of genomic data and the strong statistical power it provides for phylogenetic inference, it is important that researchers take into account both methodological and biological sources of gene tree conflict in the effort to produce accurate, highly supported trees.

The combination of pervasive gene tree variation coupled with the substantial evidence for systematic error suggests that even in genomes that have been characterized and analyzed extensively (such as the mitochondrial genome), phylogenetic analyses still have the potential to be misled. In larger data sets, such as those that sample hundreds or thousands of less well characterized loci from the nuclear genome, this potential grows further. The utility of the mitochondrial genome for this study is that we have a strong *a priori* expectation that gene trees will be concordant in the absence of poor model fit. This expectation does not hold for larger nuclear data sets, so detecting these issues is consequently both more difficult and more critical. We attempted to use variation in model fit to sort genes into those that are more or less reliable, but found that this relationship was too complex relative to the small number of genes in the mitochondrial genome to allow for such coarse characterization. Our inability to find a particularly reliable set of mitochondrial genes for phylogenetic inference does not rule out that it is possible for such an approach to do so in the nuclear data sets with hundreds of genes that are typically used today. Indeed, sorting by variation in model fit does appear to be fruitful when more loci are available (Doyle et al. 2015).

Model fit tests employing posterior predictive simulation, and related approaches, have the potential to fill an important gap in phylogenetic methodology by assessing a model's fit to a given data set. Model fit testing in a posterior predictive framework allows a great deal of flexibility to focus on different aspects of a model and their influence on inferences. These methods are being implemented in a growing number of phylogenetic software packages, making them easier to apply as a routine step in phylogenetic analyses (Lartillot et al. 2009; Höhna et al. 2017). In this study, we conducted a suite of model performance tests to explore possible sources of systematic error that may be driving extensive gene tree variation. Across most data sets, we were able to detect the presence of systematic error with some of the test statistics, particularly the upper quantile-based test statistics. However, the relationship between model performance and gene tree accuracy can be complex.

This complex relationship may stem from poor performance across all genes, leading to consistent levels of error across gene trees and difficulty in detecting a relationship with gene tree congruence. Alternatively, poor model performance in some genes may result in many subtle errors in estimated support for relationships, but not result in any one part of the tree strongly conflicting with the reference (e.g., discordance among nodes deeper in the tree that cause larger tree-to-tree distances). It is also possible that the true mitochondrial history in some of these data sets, especially those that have undergone rapid radiation, may be different than the true species history.

The specific causes of poor model fit, and their role in producing systematic error, were difficult to determine with the model performance tests used here. The implementation of more targeted data-based statistics, as well as site-specific and branch-specific inference-based test statistics could help pinpoint the specific causes of poor model fit and the regions of the tree that are most directly affected. Our difficulty with determining the sources of systematic error in this study may also stem from issues with the power of these tests to detect poor performance, as they might represent conservative measures of poor model performance (Bollback 2005, Ripplinger and Sullivan 2010, Brown 2014). The power of posterior predictive tests to detect poor model performance in a gene and the power of the gene to precisely estimate the phylogeny

are surely correlated. Precise characterization of this relationship will require simulation studies beyond the scope of this paper, but individual genes with little power to estimate their phylogeny can similarly have little power to assess model fit. However, this relationship between power and information content can be exploited by researchers. Genes with little information should be less concerning in the context of large data sets than genes with lots of power and influence on the resulting estimate. Recent studies have documented extensive variation in information content among genes in phylogenomic data sets that can have a significant impact on the inferred topology and support for clades, suggesting that the distribution of information content should be used to inform methodological choices for phylogenetic inference (Brown and Thomson 2017; Shen et al. 2017). Careful consideration of the power of a gene to accurately estimate relevant nodes of a phylogeny alongside model performance will be important for accurate, highly supported phylogenetic inference.

Conclusions

Gene tree heterogeneity in multi-locus studies is often assumed to stem from biological processes, such as incomplete lineage sorting or horizontal transfer, and several methods have been developed to model these processes. We demonstrate that systematic error can be as significant a source of variation among gene trees as biological sources, although it is not currently standard practice to check for this. Tests of absolute model fit, such as the posterior predictive framework, have the potential to fill this important gap in current phylogenetic methodology. With increasing application of genomic data and the strong statistical power it provides for phylogenetic inference, it is important that researchers better take into account the methodological sources of gene tree conflict alongside the biological in the effort to produce accurate, highly supported trees.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.hj07m.

Funding

This work was supported by the University of Hawaii Evolution, Ecology, and Conservation Biology Meredith-Carson Fellowship to E.J.R., an Arnold O. Beckman Postdoctoral Fellowship to A.J.B., and US National Science Foundation awards [DEB-1355071 and DBI-1262571 to J.M.B.; DEB-1354506 and DBI-1356796 to R.C.T.].

Acknowledgements

We utilized high-performance computing resources from the University of Hawai'i and Louisiana State University for many of the analyses conducted in this study. We thank Floyd Reed, Peter Marko, Rachel Mueller, and three anonymous reviewers for comments and advice that improved the manuscript.

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19:716–723.
Barley A.J., Spinks P.Q., Thomson R.C., Shaffer H.B. 2010. Fourteen nuclear genes provide phylogenetic resolution for difficult nodes in the turtle tree of life. Mol. Phylogenet. Evol. 55:1189–94.
Barley A.J., Thomson R.C. 2016. Assessing the performance of DNA barcoding using posterior predictive simulations. Mol. Ecol. 25:1944–1957.
Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.
Bollback J.P. 2005. Posterior mapping and posterior predictive distributions. In: Statistical methods in molecular evolution. R. Nielsen, editor. New York: Springer. p. 439–462.
Boussau B., Szöllősi G.J., Duret L., Gouy M., Tannier E., Daubin V. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23, 323–330.
Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.
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.
Brown J.M., ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25:537–538.
Brown J.M., Thomson R.C. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.
Crawford N.G., Parham J.F., Sellas A.B., Faircloth B.C., Glenn T.C., Papenfuss T.J., Henderson J.B., Hansen M.H., Simison W.B. 2015. A phylogenomic analysis of turtles. Mol. Phylogenet. Evol. 83, 250–257.
Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest2: more models, new heuristics, and parallel computing. Nat. Methods 9:772.
Delsuc F., Brinkmann H., Philippes H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Rev. Genet. 6:361–375.
Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.
Doyle V.P., Young R.E., Naylor G.J.P., Brown J.M. 2015. Can we indentify genes with increased phylogenetic reliability. Syst. Biol. 64:824–837.
Duchêne D.A., Duchêne S., Holmes E.C., Ho S.Y.W. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. Mol. Biol. Evol. 32:2986–2995.
Duchêne D.A., Duchêne S., Ho S.Y.W. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol. Biol. Evol.34:1529–1534.
Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.
Edgar R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.
Edwards S.V., Xi Z., Janke A., Faircloth B.C. McCormack J.E., Glenn T.C., Xhong B. Wu S., Lemmon E.M., Lemmon A.R., Leache A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94: 447–462.
Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.
Gee H. 2003. Evolution: ending incongruence. Nature 425:782.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.
Gruenstaeudl M., Ried N.M., Wheeler G.L., Carstens B.C. 2016. Posterior predictive checks of coalescent models: P2C2M, an R package. Mol Ecol Resour. 16:193–205.
Höhna S., Drummond A.J. 2012. Guided tree topology proposals for Bayesian phylogenetic inference. Syst. Biol. 61:1–11.
Höhna S., Coghill L.M., Mount G.G., Thomson R.C., Brown J.M. 2017. P3: Phylogenetic posterior prediction in RevBayes. Mol. Biol. Evol. doi: 10.1093/molbev/msx286.
Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error onherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different models. Syst. Biol. 59:573–583.
Huang W., Zhou G., Marchand M., Ash J.R., Morris D., Van Dooren P., Brown J.M., Gallivan K.A., Wilgenbusch J.C. 2016. Treescaper: visualizing and extracting phylogenetic signal from sets of trees. Mol. Biol. Evol. 33:3314–3316.
Huelsenbeck J.P., Rannala B. 2004.Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.
Huelsenbeck J.P., Hillis D.M. 1993. Success of phylogentic methods in the four-taxon case. Syst. Biol. 42:247–264.
Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.
Jeffroy O., Brinkmann H., Delsuc F., Philippes H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.
Kearse M., Moir R., Wilson A., Stone-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Mentjiles P., Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.
Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29: 457–472.
Larget B. 2013. The estimation of tree posterior probabilities using conditional clade probability distributions. Syst. Biol. 62:501–511.
Lartillot N., Lepage L., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.
Lemmon A.R., Moriarty E.C. 2004. The importance of proper model assumptions in Bayesian phylogenetics. Syst. Biol. 53:265–277.
Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46: 523–536.
Meiklejohn K.A., Danielson M.J., Faircloth B.C., Glenn T.C., Braun E.L., Kimball R.T. 2014. Incongruence among different mitochondrial regions: a case study using complete mitogenomes. Mol. Phylogenet. Evol. 78:314–323.
Meredith R.W., Janecka J.E., Gatesy J., Ryder O.A., Fisher C.A., Teeling E.C., Goodbla A., Eizirik E., Simao T.L.L., Stadler T., Rabosky D.L., Honeycutt R.L., Flynn J.J., Ingram C.M., Steiner C., Williams T.L., Robinson T.J., Burk-Herrick A., Westerman M., Ayoub N.A., Spinger M.S., Murphy W.J. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.
Mirarab S., Bayzid M.S., Warnow T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65:366–380.
Mirarab S., Warnow T. 2015 ASTRAL-II: coalescent-based species tree stimation with many hundreds of taxa and thousands of genes. Bioinformatics 21:44–52.
Molloy E.K., Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.
Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol. Evol. 28: 719–728.
Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Biol. 51: 729–739.
Oaks J.R., Dudley R. 2011. A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. Evolution 65: 3285–3297.
Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14:e1002379.
Philippe H., 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.
Philippe H., Brinkman H., Lavrov D.V., Littlewood D.T.J., Manuel M., Worheide W., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Phillips M.J., Delsuc F., Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1468.
Pyron R.A., Wiens J.J. 2011. A large-scale phylogeny of Amphibia including over 2800 species and a revised classification of extant frogs, salamanders, and caecilians. Mol. Phylogenet. Evol. 61: 543–583.
Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526: 569–573.
Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.
Rambaut A., Suchard M.A., Xie D., Drummond A.J. 2014. Tracer v1.6. Available from: URL http://beast.bio.ed.ac.uk/Tracer.
R Core Team. 2015. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: URL: https://www.R-project.org/.
R Core Team. 2017. A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. http://www.R-project.org
Reid N.M., Hird S.M., Brown J.M., Pelletier T.A., McVay J.D., Salter J.D., Carstens B.C. 2014. Poor fit to multispecies coalescent is widely detectable in empirical data. Syst. Biol. 63:322–333.
Ripplinger J., Sullivan J. 2010. Assessment of substitution model adequacy using frequentists and Bayesion methods. Mol. Biol. Evol. 27:2790–2803.
Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Rodrigue N., Kleinman C.L., Philippe H., Lartillot N. 2009. Computational methods for evaluating phylgoenetinc models of coding sequence evolutionwith dependence between codons. Mol. Biol. Evol. 26:1663–1676.
Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.
Rokas A., Carroll S.B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337–1344.
Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. Mr.Bayes3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.
Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–333.
Schierwater B., Eitel M., Jakob W., Osigus H.-J., Hadrys H., Dellaporta S.L.,. Kolokotronis S.-O., Desalle R. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol. 7:e20.
Scornavacca C., Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. Syst. Biol. 66:112–120.
Shaffer H.B., McCartney-Melstad E., Near T.J., Mount G., Spinks P.Q. 2017. Phylogenomic analyses of 539 highly informative loci dates a fully resolved time tree for the major clades of living turtles (Testudines). Mol. Phylogenet. Evol. 115:7–15.
Shen X.X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:126.

Slater G.J., Pennell M.W. 2014. Robust regression and posterior predictive simuation increase power to detect early bursts of trait evolution. Syst. Biol. 63:293–308.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Nat. Acad. Sci. USA 109:14942–14947.

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

Sukuraman J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics. 26:1569–1571.

Swofford D.L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Available from: URL http://paup.csit.fsu.edu.

Swofford D.L., Waddell P.J., Huelsenbeck J.P., Foster P.G., Lewis P.O., Roger J.S. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Szöllősi G.J., Rosikiewicz W., Bossau B., Tannier E., Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. Syst. Biol. 62:901–912.

Szöllősi G.J., Tannier E., Daubin V., Boussau B. 2015. The inference of gene trees with species trees. Syst. Biol. 64:e42–e62.

Thomson R.C., Shaffer H.B. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst. Biol. 59:42–58.

Warren D.L., Geneva A.J., Lanfear R. 2017. RWTY (R We There Yet): an R Package for examining convergence of Bayesian phylogenetic analyses. Mol. Biol. Evol. 34:1016–1020.

Whelan N.V., Kocot K.M., Moroz L.L., Halanych K.M. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Nat. Acad. Sci. USA doi: 10.1073/pnas.1503453112.

White D.L., Wolff J.N., Pierson M., Gemmell N.J. 2008. Reveleading the hidden complexities of mtDNA inheritance. Mol. Ecol. 17: 4925–4942.

Wickham H. 2009. Ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.

Wiens J.J., Hutter C.R., Mulcahy D.G., Noonan B.P., Townsend T.M., Sites J.W., Reeder T.W. 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biol. Lett. doi: 10.1098/rsbl.2012.0703.

Wilgenbusch J.C., Huang W., Gallivan K.A. 2017. Visualizing phylogenetic tree landscapes. BMC Bioinformatics 18:85.

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

Zhang C., Sayyari E., Mirarab S. 2017. ASTRAL-III: Increased scalability and impacts of contracting low support branches. In: Comparative Genomics: 15th International Workshop, RECOMB CG.

Zhong B., Liu L., Yan Z., Penny D. 2013. Origin of land plants using multispecies coalescent model. Trends Plant Sci. 18: 492–495.