Differences in Performance among Test Statistics for Assessing Phylogenomic Model Adequacy

David A. Duchêne[1,*], Sebastian Duchêne[2], and Simon Y.W. Ho[1]

[1]School of Life and Environmental Sciences, University of Sydney, Sydney, NSW, Australia

[2]Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Melbourne, VIC, Australia

*Corresponding author: E-mail: david.duchene@sydney.edu.au.

Accepted: May 11, 2018

Abstract

Statistical phylogenetic analyses of genomic data depend on models of nucleotide or amino acid substitution. The adequacy of these substitution models can be assessed using a number of test statistics, allowing the model to be rejected when it is found to provide a poor description of the evolutionary process. A potentially valuable use of model-adequacy test statistics is to identify when data sets are likely to produce unreliable phylogenetic estimates, but their differences in performance are rarely explored. We performed a comprehensive simulation study to identify test statistics that are sensitive to some of the most commonly cited sources of phylogenetic estimation error. Our results show that, for many test statistics, traditional thresholds for assessing model adequacy can fail to reject the model when the phylogenetic inferences are inaccurate and imprecise. This is particularly problematic when analysing loci that have few informative sites. We propose new thresholds for assessing substitution model adequacy and demonstrate their effectiveness in analyses of three phylogenomic data sets. These thresholds lead to frequent rejection of the model for loci that yield topological inferences that are imprecise and are likely to be inaccurate. We also propose the use of a summary statistic that provides a practical assessment of overall model adequacy. Our approach offers a promising means of enhancing model choice in genome-scale data sets, potentially leading to improvements in the reliability of phylogenomic inference.

Key words: model adequacy, substitution model, maximum likelihood, test statistics, birds, Laurasiatherian mammals, turtles.

Introduction

In recent years, phylogenomic analyses have provided interesting insights into the evolution of various taxonomic groups (e.g., Meredith et al. 2011; Timme et al. 2012; Misof et al. 2014). However, the evolutionary relationships in some groups of organisms have remained difficult to resolve even with large amounts of data, and different studies have sometimes yielded conflicting phylogenetic estimates (e.g., Jarvis et al. 2014; Prum et al. 2015). The sources of conflict among data sets are likely to include model misspecification, incongruence among gene trees, and the impacts of positive selection (Springer and Gatesy 2016; Reddy et al. 2017; Shen et al. 2017). Unless data from additional loci can be readily obtained, the best means of improving the reliability of the inferred species tree is to enhance the Modelling of the evolutionary process that produced the data at hand. This can be done either by improving the methods of inference (e.g., Galtier 2001; Foster 2004; Lartillot et al. 2007; Jayaswal et al. 2014) or by filtering the data according to appropriate criteria (Liu et al. 2015). Both of these measures can help to enhance the phylogenetic signal in the available data.

One potential method of improving evolutionary Modelling in phylogenetics is to perform an absolute assessment of the adequacy or plausibility of the substitution model for each locus in the data set (Doyle et al. 2015). This is different from the comparisons of relative model fit that are routinely performed in phylogenetics (Posada and Crandall 2001), and involves testing whether the chosen substitution model provides an accurate description of the evolutionary process that generated the data (Goldman 1993). Although the available substitution models are sometimes found to be good descriptions of the molecular evolutionary process (Ripplinger and Sullivan 2010), they are unlikely to be adequate for all of the loci in genome-scale data sets (Goldman 1993; Doyle et al. 2015; Duchêne et al. 2016). Methods of assessing model adequacy could, in principle, identify the

portion of the data set for which the model is realistic. This information could then be used to reject a given model or to exclude the subset of the data that is not well described by the available models (Doyle et al. 2015), leading to enhancement of the phylogenetic signal in the data.

Methods of assessing model adequacy by comparing empirical data to those simulated under the model can test whether some aspects of the model are realistic, but they might not necessarily indicate when estimates of parameters of interest are unreliable. This is because sequence data might meet the particular assumption being assessed but still violate other assumptions that are critical for estimating parameters of interest (e.g., Ho and Jermiin 2004; Brown 2014; Doyle et al. 2015). Data might also fail to meet a particular assumption of the model, yet this might not have a negative impact on the estimates of parameters of interest (Lemmon and Moriarty 2004; Brown 2014; Duchêne et al. 2017). For example, estimates of tree topology and branch lengths can be accurate even when there are substantial departures from stationary base composition (Duchêne et al. 2017). To ensure the usefulness of methods for model assessment, we must evaluate whether measures of model adequacy can predict the accuracy and precision of the inferences made using these models. Similarly, approaches that can summarize tests from multiple methods, and across hundreds of loci, are needed to make assessment of model adequacy practical in the genomic age.

Methods to assess model adequacy are based on comparing the empirical data with a distribution of data generated under the candidate model, also known as the predictive distribution. The data generated under the candidate model are also known as predictive data and can be generated using the maximum-likelihood estimates of the model parameters (Goldman 1993). Alternatively, parameter estimates can be taken from samples from the posterior distribution in a Bayesian analysis, such that predictive data account for the uncertainty in estimates of model parameters (Bollback 2002). In a Bayesian framework, simulations performed using the model are also known as posterior predictive simulations. The benefit of accounting for uncertainty in parameter estimates comes at the cost of computational demand. In this study, we focus on methods to assess model adequacy for genome-scale data sets, so we use a maximum-likelihood method of assessment. Several types of models in phylogenetics can be assessed using predictive simulations, including substitution models (Goldman 1993; Bollback 2002; Foster 2004; Brown 2014), large hierarchical models (Reid et al. 2014; Duchêne et al. 2015), models of trait evolution (Rabosky and Glor 2010; Slater and Pennell 2014), and models of diversification rates through time (Höhna et al. 2016). Predictive approaches can also be used for relative model comparison (Lewis et al. 2014), providing an alternative to commonly used information criteria or Bayes factors.

The procedure of comparing the empirical and predictive data requires the choice of a test statistic, a metric that describes some aspect of the data set or of the inferences drawn from the data set. If empirical data are similar to, or produce similar estimates to, the predictive distribution, then the model can be considered adequate. Test statistics that can be calculated directly from the data are known as data-based statistics, whereas those that describe the inferences drawn from the data are known as inference-based statistics. Some test statistics require calculations from the data as well as inferences from the data, such that they are hybrid test statistics. Critically, test statistics can differ in their sensitivity to biased inferences. Identifying which test statistics are sensitive to biased phylogenetic inferences is fundamental before we can develop a practical framework of model assessment.

The extent to which empirical data must be similar to the predictive distribution for inferences to be inaccurate remains poorly understood. The traditional approach to determining model adequacy is to reject the model if the test statistic for the original data falls above 95% or 99% of the values from simulated data. But assessment using these thresholds can lead to the model being rejected even when inferences are not necessarily unreliable, or the failure to reject a model that leads to biased inferences (Brown 2014; Duchêne et al. 2017).

In this study, we aim to characterize the performance of a wide range of test statistics for model assessment in phylogenomic data sets, and explore a method for summarizing their results. Using simulations under a range of conditions known to occur in empirical data, we characterize the ability of nine test statistics to detect poor performance of the substitution model. We define performance as the accuracy and precision of estimates of tree topology and branch lengths, which are often the parameters of interest in phylogenomic studies. We recommend using an efficient maximum-likelihood framework for model assessment that makes assessment feasible for genome-scale data sets, and focusing on the test statistics that perform well in our simulation study. Based on our simulation study, we also propose thresholds that are meaningful for identifying misleading phylogenetic inferences. We also describe the performance of a test statistic that summarizes multiple tests, which can be used to provide an overall assessment of model adequacy. We demonstrate our approach and explore the relationship between substitution model adequacy and performance in phylogenomic data sets from turtles, birds, and Laurasiatherian mammals.

Materials and Methods

Fast Assessment of Model Adequacy

Framework for Model Assessment

In a basic assessment of substitution model adequacy, phylogenetic analysis of the empirical data is performed using the model that is to be assessed. Then, a large number of data

sets of the same size as the empirical data set are generated by simulating sequence evolution under the chosen model, using the maximum-likelihood estimates of the model parameters. A chosen test statistic is calculated for each of these simulated data sets, thereby producing a distribution of values derived from the model (Goldman 1993; Foster 2004). The test statistic from the empirical data can be compared against this predictive distribution.

An ideal test statistic is able to describe the differences between the empirical data and the simulated data, especially with regard to inferences of interest (such as the tree topology and branch lengths). Therefore, selecting appropriate test statistics is critical to assessment of model adequacy. Here, we describe our framework for fast assessment of model adequacy using test statistics that can detect instances when phylogenetic inferences are inaccurate or imprecise. We assume that we have an empirical data set comprising a large number of unlinked loci from the taxa of interest.

In our framework, we obtain maximum-likelihood estimates of model parameters and the phylogeny for each gene alignment using the software PhyML 3.0 (Guindon et al. 2010). Branch support is estimated using a highly efficient, nonparametric measure of branch support with behaviour similar to the nonparametric bootstrap (Guindon et al. 2010; Anisimova et al. 2011). We take advantage of the speed of this method for assessing the performance of test statistics that are based on inferences from the data (Brown 2014).

For each locus alignment, we generate 100 data sets by simulating sequence evolution using the parameter estimates from the empirical data. These simulated data sets have the same number of taxa and sites as the empirical data. Since these simulated data are derived from the model, they can be considered a null distribution (Goldman 1993). To assess model adequacy, we calculate a number of test statistics for the empirical data set and for each simulated data set. It is common to consider the model to be inadequate when the test statistic calculated from the empirical data falls outside the central 95 or 99 percentile range of test statistics calculated from the simulated data. However, these thresholds do not necessarily reflect the points at which inferences become inaccurate, so in this study we do not use the *P*-values to test model adequacy. Instead, we used a measure of effect size based on the number of standard deviations of the predictive distribution (SDPD) between the mean and the statistic calculated from the original data (Brown 2014; Duchêne et al. 2017).

Test Statistics Considered

Almost any variable or model parameter that can be estimated from the data can be used as a test statistic in the approach described here. Instead of carrying out an exhaustive examination of possible statistics, we consider nine that have previously been proposed in the context of substitution model adequacy (table 1). The selected statistics consider several aspects of the model and potentially provide an overall examination of adequacy. These test statistics include the X_m^2 statistic for assessing stationarity of base composition (Foster 2004); *multinomial* and δ statistics for assessing overall model fit (Goldman 1993; Bollback 2002); *biochemical diversity* for assessing the diversity in base composition across sites (Lartillot et al. 2007); and *consistency index* for assessing the consistency of phylogenetic information in the data when compared with the most parsimonious possible scenario (Kluge and Farris 1969). We also use three test statistics based on inferences from the data, including *mean branch support*, *95% confidence interval in branch support*, and *sum of branch lengths* (Brown 2014). In addition, we use the *Mahalanobis* statistic, which has been used previously for assessing phylodynamic models, and provides a summary of a chosen set of test statistics (O'Hagan 2003; Drummond and Suchard 2008). This approach treats groups of test statistics as a multivariate distribution, to which the empirical data are compared using the *Mahalanobis* distance (see table 1). We have implemented our approach and the test statistics outlined here in the software PhyloMAd (github.com/duchene/phylomad).

Validation of Test Statistics Using a Simulation Study

Simulation Scenarios

To understand the performance of test statistics for assessing model adequacy, we simulated the evolution of nucleotide sequences in a number of scenarios that involve misspecified models. Under these conditions, the accuracy and precision of phylogenetic inference might be adversely affected. We performed a range of simulations in six scenarios involving realistic variation across loci (fig. 1). Our study does not include other scenarios that have the potential to create difficulties for phylogenetic inference, such as tree imbalance or extreme model parameter values. Instead, we focus on scenarios that lead to model misspecification. Our simulations were performed on fully symmetric trees with 32 tips. To derive adverse simulation scenarios, we began with a baseline phylogenetic tree with branch lengths drawn from an exponential distribution with rate of 5, such that the mean distance from the root to each tip is 1 substitution per site. We simulated the evolution of sequences along these tree to produce data sets with 200, 1,000, or 5,000 nucleotide sites.

Simulations were performed under a model in which the rates of each of the six substitution types were different. This model is the GTR + Γ (Tavaré 1986; Yang 1993) in scenarios where base composition is stationary and sites do not contain covarion-like patterns of substitution. Simulations were done using the R package PHANGORN (Schliep 2011), except for those with nonstationary base composition, which were done

Table 1
Details of Nine Test Statistics Used to Assess the Adequacy of Nucleotide Substitution Models

Test Statistic	Calculation	Model Component Assessed	Type of Statistic	Reference
X^2_m	Tree- and model-based chi-squared statistic of base frequencies across taxa. It is calculated using matrices of base composition with a row for each taxon and a column for each nucleotide. One matrix corresponds to the values expected under the tree and model of base composition, and the other corresponds to the observed base composition. The statistic is calculated using the following formula: $$X^2_m = \Sigma[(\text{obs} - \text{exp}_m)^2/\text{exp}_m]$$ where at each cell obs is the observed base frequency and exp_m is the expected frequency under the tree and phylogenetic model.	Stationarity of base composition	Data-based	Foster (2004)
Multinomial (or unconstrained) likelihood	Product of the unique site frequencies ($\hat{L}_m$), calculated as: $$\hat{L}_m = \prod_{\ell \in \mathcal{B}} (N_\ell/N)^{N_\ell}$$ where ℓ is a site pattern in the set $\mathcal{B}$, N_ℓ is the number of occurrences of pattern ℓ, and N is the total number of sites in the data.	Overall fit	Data-based	Goldman (1993a), Bollback (2002)
δ	Likelihood of the data using the unconstrained model ($\hat{L}_m$), minus the likelihood under the evolutionary model	Overall fit	Data-inference hybrid	Goldman (1993a)
Biochemical diversity	Calculated as the number of different bases occurring at each site, and the mean value taken across the alignment	Diversity in base composition across sites	Data-based	Lartillot et al. (2007)
Consistency index	Minimum possible number of substitutions in the data divided by the minimum number required to describe a given tree using parsimony. It has been considered as a measure of homoplasy, and is expected to take a value of 1 in the absence of homoplasy	Consistency of phylogenetic information in the data compared with the most parsimonious scenario	Data-inference hybrid	Kluge and Farris (1969)
Branch support	Mean of branch-support values across the maximum-likelihood tree. Branch support can be calculated in a variety of ways, including nonparametric bootstrap or approximate likelihood-ratio test (Anisimova and Gascuel 2006).	Overall fit	Inference-based	Brown (2014b)
95% CI in branch-support statistic	95% range in branch-support values across the maximum-likelihood tree	Overall fit	Inference-based	Brown (2014b)

Tree length	Sum of the branch lengths in the maximum-likelihood tree	Overall fit	Inference-based	Brown (2014b)
Mahalanobis distance	A test of model adequacy can be placed in a multivariate setting that simultaneously considers multiple test statistics by using Mahalanobis distances. The aim of this approach is to estimate a distance between the empirical test statistics and the multivariate predictive distribution from several test statistics. Individual test statistics are first standardized so that they appear on the same scale. Then it is possible to calculate the mean ($\hat{m}$) and variance covariance matrix ($\hat{V}$) of the multivariate distribution: $$\hat{m} = \frac{1}{P}\sum_{P=1}^{P} T(\text{rep}_p)$$ $$\hat{V} = \frac{1}{P-1}\sum_{P=1}^{P}\left[T(\text{rep}_p) - \hat{m}\right]^t - \left[T(\text{rep}_p) - \hat{m}\right]$$ where P is the number of predictive simulations, T is the multivariate distribution of test statistics, and rep_p is each of the predictive data sets. The empirical and replicate data sets each has a distance from the multivariate distribution that can be defined as the following: $$M(x) = (x - m)^t V^{-1}(x - m)$$ For assessing substitution model adequacy, we used two combinations of test statistics to define the multivariate distribution. One included the other eight test statistics considered in this study, and the other included the four statistics that were the most sensitive to biased phylogenetic inferences.	Summary assessment from multiple test statistics; Overall fit	Based on other test statistics	Mahalanobis (1936); O'Hagan (2003); Drummond and Suchard (2008)

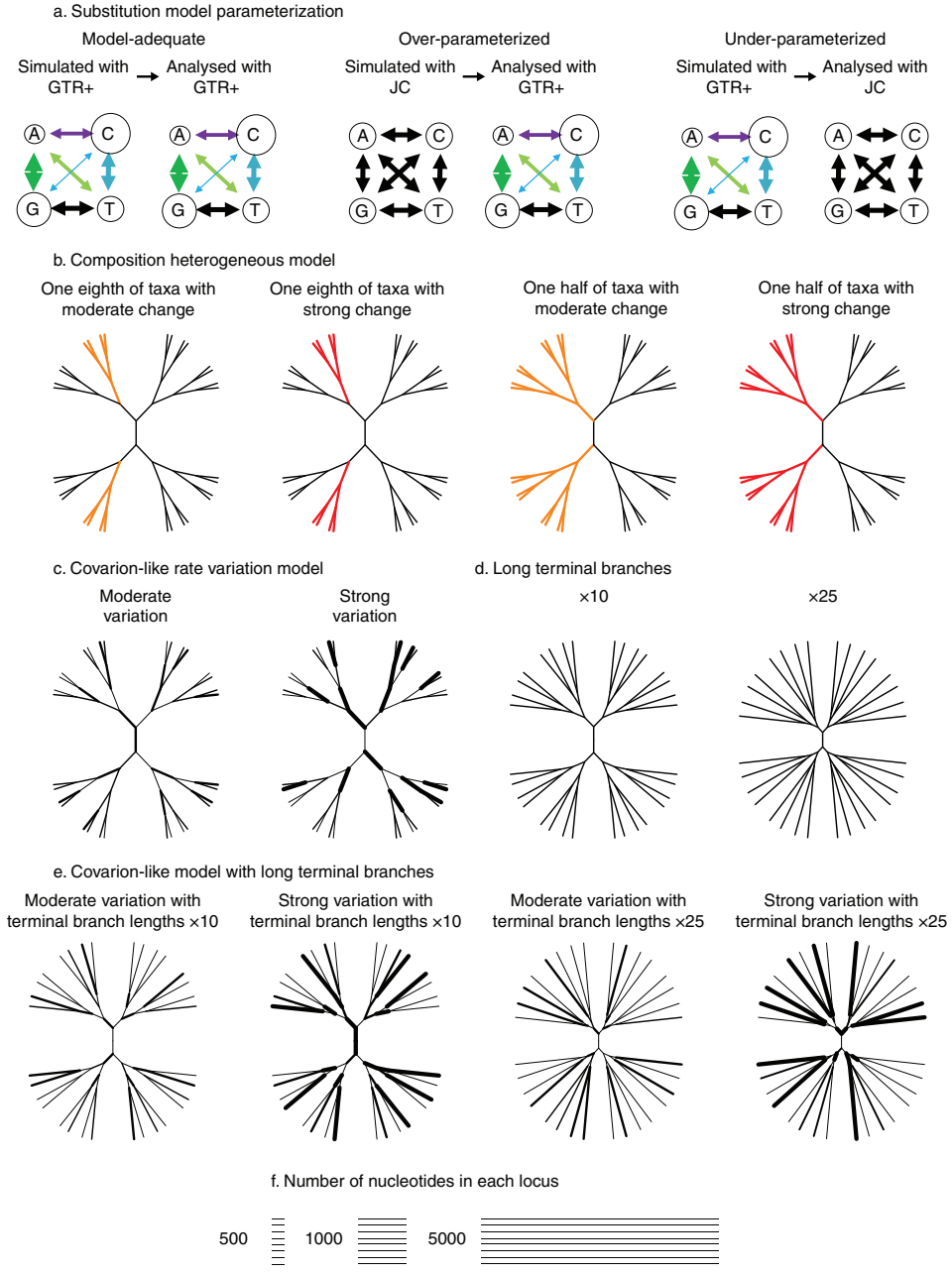

Fig. 1.—The six characteristics that were varied in simulations of sequence evolution to investigate the performance and adequacy of the candidate substitution model (GTR + Γ): (*a*) substitution model parameterization; (*b*) compositional heterogeneity; (*c*) covarion-like rate variation; (*d*) terminal branch lengths; (*e*) covarion-like rate variation and terminal branch lengths; and (*f*) sequence length for each locus. One hundred replicates were performed under each scenario from (*a*) to (*e*), under each of the sequence lengths shown in (*f*). Colors in (*a*) indicate different rate parameters, whereas in (*b*) they indicate the magnitude and proportion of taxa undergoing a change in base composition. Branch thickness corresponds to evolutionary rate in (*c*), (*d*), and (*e*).

using the software p4 (Foster 2004). The R matrix used for simulation had parameters set to 1.3472, 4.8145, 0.9304, 1.2491, 5.5587, and 1.0000, based on a previous study of placental mammals (Murphy et al. 2001). We performed an additional set of simulations under the Jukes–Cantor model (Jukes and Cantor 1969). For each simulation scenario, we performed 100 replicates for each locus length.

We first performed phylogenetic and model-adequacy analyses using substitution models that were overparameterized, underparameterized, and adequate (fig. 1*a*). This was done by evaluating three scenarios that represented different combinations of simulation model and estimation model: sequences that evolved under the JC model and were analysed using the GTR + Γ model (overparameterized

model); sequences that evolved under the GTR + Γ model and were analysed using the JC model (underparameterized model); and sequences that evolved under the GTR + Γ model and were analysed using the GTR + Γ model (adequate model).

We simulated other processes that can occur in empirical data and which can cause poor accuracy and precision in phylogenetic inference. Heterogeneity in base frequencies across lineages, which violates the assumption of compositional stationarity, can lead to the spurious joining of taxa with convergent base frequencies (Lockhart et al. 1992; Jermiin et al. 2004). We performed simulations that included two convergent changes in base composition across the tree, with the maximum topological distance from each other (fig. 1*b*). We varied the proportion of tips that underwent a change in base composition between none, one-eighth, and half of the total number of tips. We also varied the magnitude of changes in base composition across two scenarios: 1) approximately the maximum difference in base composition observed across taxa in an avian phylogenomic data set (Prum et al. 2015; from {0.25, 0.25, 0.25, 0.25} to {0.05, 0.45, 0.45, 0.05}); and 2) extreme differences in base composition (from {0.25, 0.25, 0.25, 0.25} to {0.01, 0.49, 0.49, 0.01}). The smallest base frequency in the first scenario was set to approximately the smallest base frequency observed in avian phylogenomic data. These parameters allowed data sets to have a high realistic probability of leading to biased phylogenetic inferences. The smallest base frequency in the second scenario was set to a small but nonzero value of 0.01, in order to maintain the presence of all four bases in the data.

Simulations of sequence evolution were also done under a covarion-like process (Galtier 2001), where substitution rates vary across sites and across lineages (Fitch and Markowitz 1970; Fitch 1971). This scenario has been found to cause poor estimates of phylogenetic tree lengths, but can benefit topological inference by causing the overestimation of internal branch lengths (Penny et al. 2001; Phillips 2009). For this reason, it can be considered an unusual form of model misspecification that is likely to be difficult to detect, yet might mislead the inferences. Under a covarion-like process (Galtier 2001), relative rates across lineages are described by a discretized gamma distribution. The data share the categories of the distribution, but each site has its own realization of rates across lineages (Galtier 2001). In this way, sites might share the rate along a given branch, but might not share that rate along other branches. We simulated this covarion-like process using a gamma distribution with an α parameter of 1. To vary the strength of variation across sites, rates across lineages were drawn from distributions with 1 (no covarion), 5, or 10 categories across sites, allowing for an increasing number of extreme rates (fig. 1*c*). The consequence of using different numbers of categories is that some sites will have more extreme (low and high) rates, although the rate distribution and mean rate remain unchanged.

To increase the sources of error in the data, we introduced simulation scenarios in which the lengths of the terminal branches were 10 and 25 times those in the original simulations (fig. 1*d*). These simulations led to data that resemble those across highly divergent taxa, such as those used to examine the relationships among metazoan taxa (e.g., Pisani et al. 2015). To select the parameters for these simulations, we took the original simulations involving a root-to-tip distance of 1 substitution per site to represent a period of 50 My. Under this assumption, the simulation scenarios with longer terminal branches reflect data sets across timescales of 1.25 and 2.5 billion years, resembling the scenarios simulated by Penny et al. (2001). We also performed simulations on trees with long terminal branches for each of the three covarion scenarios (fig. 1*e*), which also resembles previous investigations of these conditions (Penny et al. 2001).

Analyses of Simulated Data

For each simulated data set, phylogenetic inference was performed using a common candidate substitution model (GTR + Γ) using the software PhyML (Guindon et al. 2010), and then model adequacy was assessed as described above. We used four metrics to describe the accuracy and precision of our analyses of simulated data. The lengths of branches in estimated trees were summed, and then the sum was subtracted from the sum of branch lengths of simulated trees. This value was then divided by the sum of simulated branch lengths to describe the inaccuracy in estimates. We made the same calculation using the stemminess of each tree, which is the proportion of the inferred tree length represented by internal branches (Fiala and Sokal 1985). Stemminess was used to summarize the bias in inferences across branch lengths within each tree estimate. As a measure of error in topological inference, we calculated the unweighted Robinson–Foulds distance between the estimated and simulated trees (Robinson and Foulds 1981; Penny and Hendy 1985). Lastly, we used the support for nodes in the estimated tree as a measure of precision in the inferred topology, calculated using the Shimodaira–Hasegawa approximate likelihood-ratio test (aLRT) nonparametric measure of branch support (Guindon et al. 2010; Anisimova et al. 2011).

For each simulated data set, we assessed model adequacy using each of the nine test statistics considered in our study. In statistics, sensitivity is generally defined as the ability of a test to correctly identify instances in which a result is significant. Accordingly, we consider test statistics to be consistently sensitive to particular simulation conditions if the distribution of values from replicate simulations does not overlap with the mean of the values from the predictive distributions. For test statistics that were sensitive to inferences with poor accuracy and precision, we aimed to determine meaningful thresholds for model assessment. This is because existing tests of model adequacy are generally conservative, frequently rejecting the

model when inferences from the data are unlikely to be misleading (e.g., Duchêne et al. 2017).

We identified thresholds with a tendency to be lenient, with low risk of rejecting the model when the inferences are not misleading (low Type I error rate) at the expense of sometimes failing to reject the model when inferences are misleading (moderate Type II error rate). This approach maximizes the usage of phylogenomic data, because data can still contain useful information even when the model is rejected by a given test statistic.

The interpretation of test statistics can be sensitive to sequence length (Duchêne et al. 2017), so our thresholds take this into account. We determined the threshold for model assessment for each sensitive test statistic. Thresholds were defined as a fitted function between the sequence lengths used for simulation and the median test statistic under a simulation scenario to which the statistic is most sensitive.

Analyses of Empirical Data

The framework that we have used here for model assessment is highly computationally efficient, and therefore well suited for examining genome-scale data sets. We analysed three phylogenomic data sets to investigate the impact of assessing substitution model adequacy on the inferred tree topology. These data sets comprised 2,363 loci from 63 turtle taxa (Crawford et al. 2015), 222 loci from 200 bird taxa (Prum et al. 2015), and 96 loci from 15 Laurasiatherian mammal taxa (Zhou et al. 2012). In order to allow calculation of the multinomial likelihood, sites with gaps or unknown nucleotides were excluded. For each of the three data sets, we performed maximum-likelihood analyses using a GTR + Γ model of nucleotide substitution, and assessed model adequacy using the nine test statistics that we investigated in our simulation study (table 1).

We described the relative distances between gene trees by approximating these distances in two-dimensional space. This approach has been shown to provide an accurate description of the differences in phylogenetic signal across loci (Duchêne et al. 2018). This representation of tree space was made using multidimensional scaling (MDS), based on unweighted Robinson–Foulds distances (Robinson and Foulds 1981; Penny and Hendy 1985) for describing the distances between trees in Euclidean space (Hillis et al. 2005; Matsen 2006; Höhna and Drummond 2012). MDS finds the Euclidean positions of gene trees that minimize the sum of the distances between them (Mardia et al. 1979).

We used the MDS visualization to explore an association between tree space and the Mahalanobis distance, mean node support, tree length, and number of variable sites. This association might occur, for example, if loci that yield trees with highly supported nodes and that have high information content lead to congruent estimates of topology and have high model adequacy (Doyle et al. 2015). Alternatively, loci that lead to well-supported topologies can have some of the largest numbers of variable sites, and therefore can be the loci that show the greatest differences from predictive data (i.e., have the largest Mahalanobis distance).

To investigate whether information content was associated with distance from the model or with tree space, we used Spearman's ρ to test whether the Mahalanobis distance was correlated with mean node support, tree length, number of variable sites, and the MDS dimensions. We also used the MDS visualization to assess the performance of thresholds for assessing model adequacy. For each sensitive test statistic, we investigated the power of our threshold to reject the model in regions of tree space that are the most likely to contain misleading inferences.

Results

Accuracy and Precision under Simulation Conditions

Phylogenetic inferences are accurate and precise when there is a match between the substitution models used for simulation and analysis (fig. 2). When the model used for analysis is overparameterized, the accuracy and precision of phylogenetic estimates is similar to those obtained when the data are analysed using the correct model (supplementary figs. S1–S3, Supplementary Material online). When the model used for analysis is underparameterized, tree length is consistently underestimated (fig. 2*a*), terminal branches have a disproportionate contribution to the tree length (fig. 2*b*), and the topology is estimated with higher error than when the correct model is used (fig. 2*c*). Using an underparameterized model also leads to estimates with greater branch support (higher precision) than when the model is adequate (fig. 2*d*), in agreement with the bias-variance tradeoff found in statistical models (Burnham and Anderson 2002; Lemmon and Moriarty 2004; Wertheim et al. 2010; Liu et al. 2015).

As simulated in this study, compositional heterogeneity leads to a mild tendency to overestimate tree length, to terminal branches having a greater contribution to the tree length, and to greater error in the estimate of the tree topology than when the correct model is used (fig. 2). Branch lengths can be overestimated under conditions of compositional heterogeneity because a change in base composition can inflate the inferred numbers of the types of substitutions involved in that change (Ho and Jermiin 2004; Duchêne et al. 2017).

Analyses of data generated under a covarion-like process lead to underestimates of tree length (fig. 2*a*), but otherwise produce estimates with similar accuracy and precision to those in which the correct model was used. The lowest accuracy and precision in phylogenetic inferences was found in simulation scenarios that involved long terminal branches (fig. 2*c* and *d*). In these cases, tree length was severely underestimated and terminal branches had a small contribution to

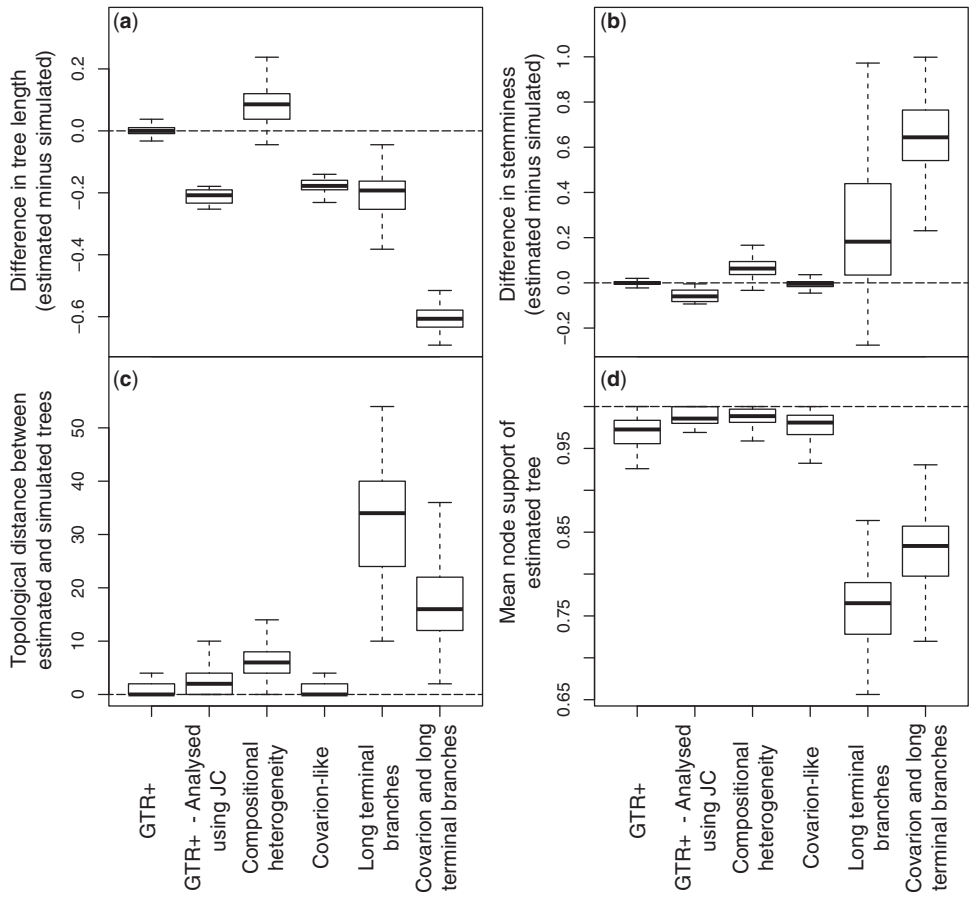

Fig. 2.—The performance of phylogenetic inference using the GTR + Γ substitution model in simulations with 5,000 nucleotides under six representative simulation conditions (for results from every simulation scenario, see supplementary figs. S1–S3, Supplementary Material online). Each box represents the results of 100 replicate analyses. Performance is described by (*a*) the length of the estimated tree minus that of the simulated tree, divided by that of the simulated tree, (*b*) the difference in stemminess, defined as the proportion of the inferred tree length represented by internal branches, (*c*) the unweighted Robinson–Foulds topological distance between estimated and simulated trees, and (*d*) the mean node support in the estimated tree, which is a measure of precision in estimates.

tree length (fig. 2*a* and *b*). The effects of the covarion process combined with long terminal branches have been studied previously, with similar outcomes (Galtier 2001; Penny et al. 2001). We also find that inferences from short sequence alignments have greater variance in accuracy and precision (supplementary figs. S2 and S3, Supplementary Material online).

Sensitivity of Model-Adequacy Test Statistics

None of the test statistics is consistently sensitive to scenarios in which the model is adequate (fig. 3*a*) or to model overparameterization (supplementary figs. S4*b*–S6*b*, Supplementary Material online). This is a desirable property and is consistent with the results of previous research (Brown 2014; Duchêne et al. 2017). Four test statistics are consistently sensitive to model underparameterization, including the *multinomial likelihood*, δ statistic, *biochemical diversity*, and the *consistency index* (fig. 3*b*). However, the δ statistic has negligible sensitivity compared with the other three test statistics, with SDPD values lower than 0.5 in data sets with 1,000 nucleotides. The same four test statistics, along with X^2_m, are consistently sensitive to compositional heterogeneity. The X^2_m statistic is overwhelmingly the most sensitive statistic to this scenario, dwarfing the signal of the other statistics (fig. 3*c*). These results suggest that most test statistics are lenient under conditions of compositional heterogeneity. However, the X^2_m statistic can be highly conservative, because some of the inferences under conditions of compositional heterogeneity have similar accuracy and precision to those obtained when using the correct model.

The *biochemical diversity* and *consistency index* statistics are consistently sensitive to the covarion-like process, in particular when the simulated sequences had a length of 5,000 nucleotides (fig. 3*d* and *f*; see supplementary figs. S5*i* and S6*i*, Supplementary Material online). Only the *biochemical*

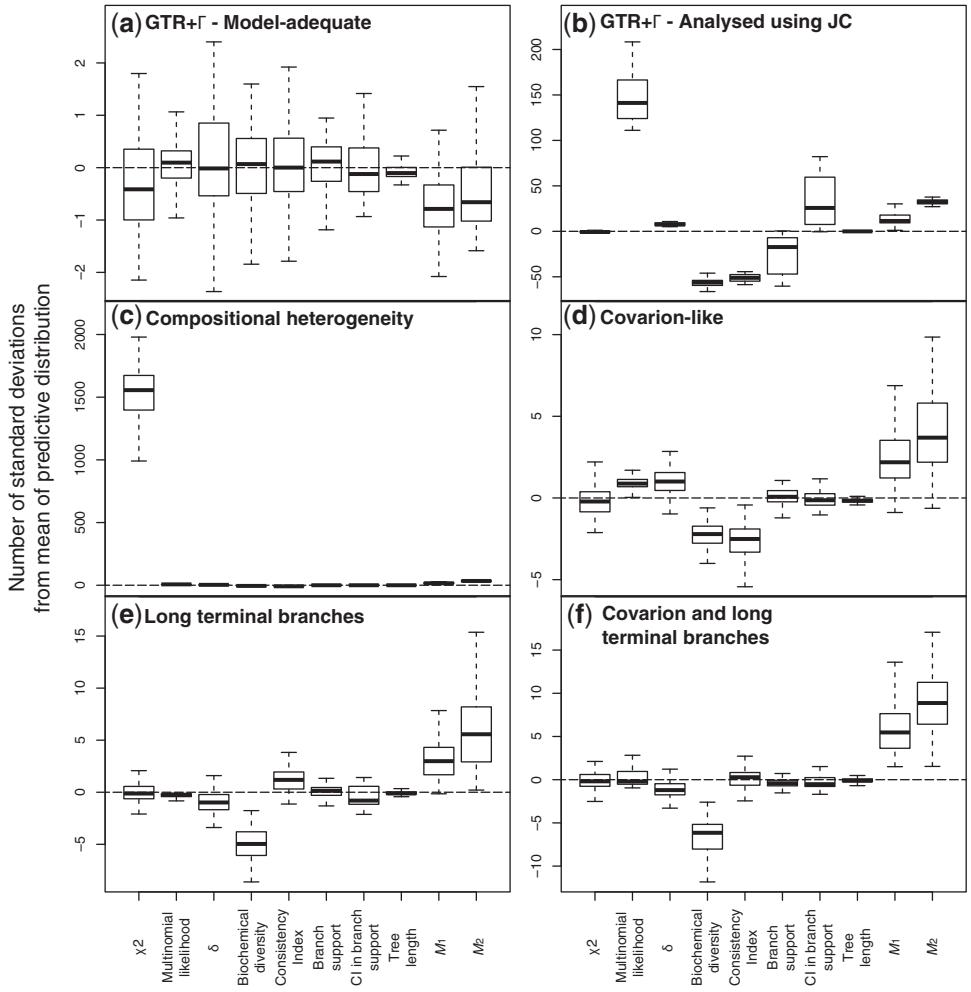

Fig. 3.—The sensitivity of nine test statistics for assessing the adequacy of the GTR + Γ substitution model in simulations with 5,000 nucleotides under six representative simulation conditions (for results from every simulation scenario, see supplementary figs. S4–S6, Supplementary Material online). The Mahalanobis test statistic was calculated to summarize all test statistics (M_1), or the four sensitive test statistics (M_2). Each box represents the results of 100 replicate analyses.

diversity statistic is consistently sensitive to long terminal branches, particularly in simulation scenarios involving sequences with lengths of 5,000 nucleotides (fig. 3e). In data sets with shorter sequences (200 and 1,000 nucleotides), two test statistics are consistently sensitive to long terminal branches, but with low sensitivity (SDPD between −1 and 1; supplementary figs. S5*j*–S5*o* and S6*j*–S6*o*, Supplementary Material online). These were the *multinomial likelihood* and the *confidence interval in branch support*. Interestingly, we find that inference-based statistics are generally very lenient. This is possibly because inferences from predictive data sets are similar to those from the original data (but see Brown 2014).

We propose an approach to model assessment that considers the test statistics that have measurable sensitivity to phylogenetic inferences with low accuracy and precision: the X^2_m, *multinomial likelihood*, *biochemical diversity*, and *consistency index*. The result of focusing on these statistics can be observed when comparing the summary *Mahalanobis* distance including all eight of the other test statistics examined here (M_1) with that including only the four most sensitive statistics (M_2). In most simulation scenarios, M_1 and M_2 are two of the most sensitive statistics, and M_2 is always more sensitive than M_1 (with the exception of scenarios in which the correct model is used). This shows that statistics with low sensitivity should be excluded. Meanwhile, examining M_2 and the four informative test statistics can provide a useful general method for examining model performance.

New Thresholds for Assessment

Analyses of phylogenomic data sets from turtles, birds, and mammals show the critical importance of using thresholds for

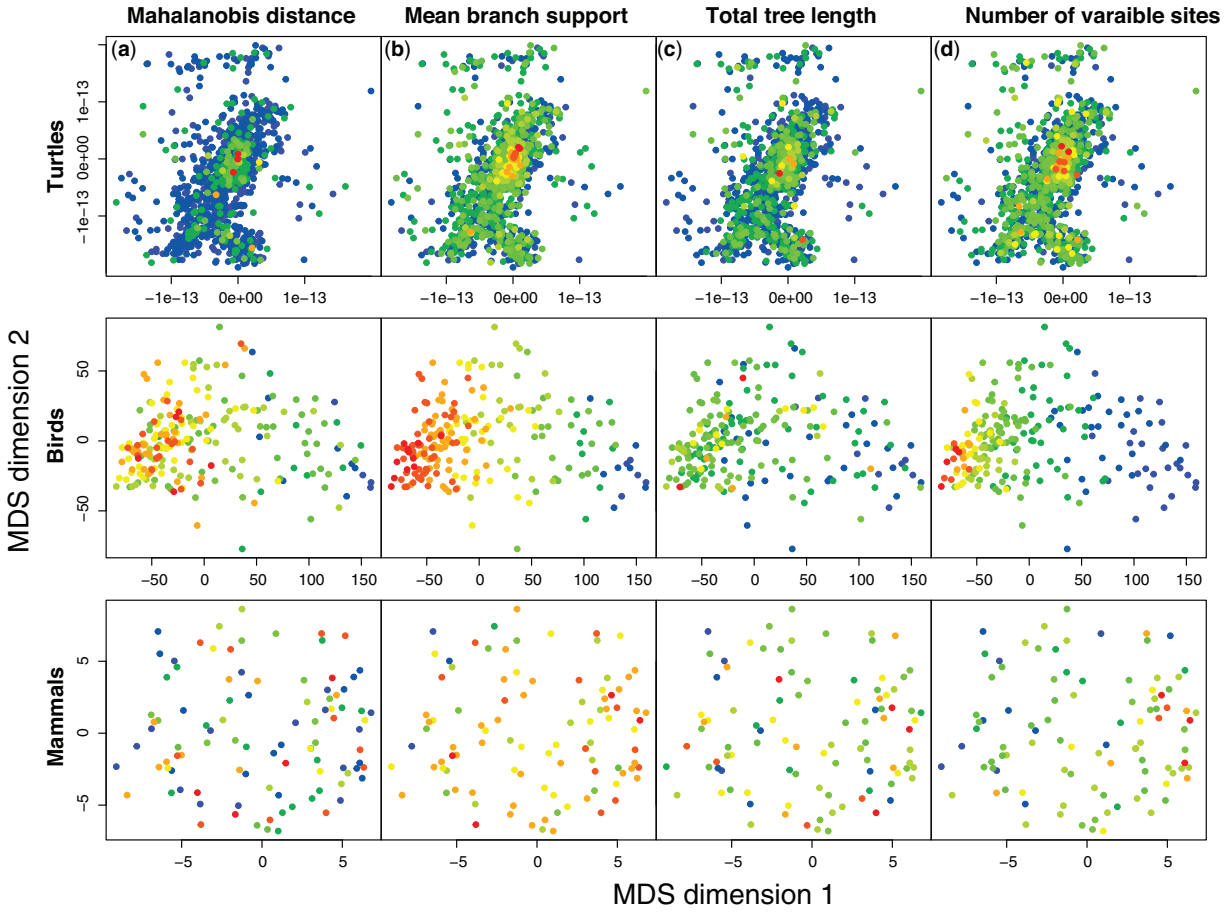

Fig. 4.—Estimated two-dimensional representation of tree-space for samples of loci from turtles, birds, and mammals. Data are colored such that warmer colors indicate higher values of (*a*) Mahalanobis distance (M_2), (*b*) mean branch support, (*c*) tree length, and (*d*) number of variable sites. High values of these variables occur in similar locations of tree-space. In the sequence data from turtles, loci with high values of M_2 are not necessarily the same as those that have high values for the other variables (see supplementary fig. S7, Supplementary Material online).

assessment that consider sequence lengths. We find striking evidence in these data that highly informative alignments have poorer perceived model adequacy. Loci with poor perceived model adequacy (the highest SDPD for M_2) yield gene-tree estimates that are similar to those of loci that yield high branch support and long branches, and that contain a large number of variable sites (fig. 4). In simulations and empirical data sets, we find that loci with gene-tree estimates that are likely to be inaccurate and imprecise have consistently good perceived model adequacy (low M_2), low mean branch support, short trees, and few variable sites (supplementary figs. S7–S8; Supplementary Material online).

We propose thresholds of model assessment based on the median value of test statistics under the scenarios to which test statistics are sensitive (supplementary fig. S9, Supplementary Material online). We use the scenarios of strong compositional heterogeneity, model underparameterization, covarion-like process with long terminal branches, and covarion-like, for determining the thresholds of X^2_m, *multinomial likelihood*, *biochemical diversity*, and *consistency index*, respectively. For deriving the threshold for M_2, we used the median values of simulations of a covarion-like process with long terminal branches.

In analyses of phylogenomic data from birds and mammals, the new thresholds have an association with estimates of the tree topology, in particular when model assessment is made using X^2_m and *biochemical diversity* (fig. 5). In these data, the model is rejected for loci in regions of tree-space with lower branch support and shorter trees. These results suggest that assessment of model adequacy might be more meaningful for data sets with longer sequences, since the mean number of sites is greater in the data from birds (741) than in the data from mammals (480) and turtles (200). Similarly, these results show that X^2_m and *biochemical diversity* statistics might provide the most informative tests out of those that we have explored.

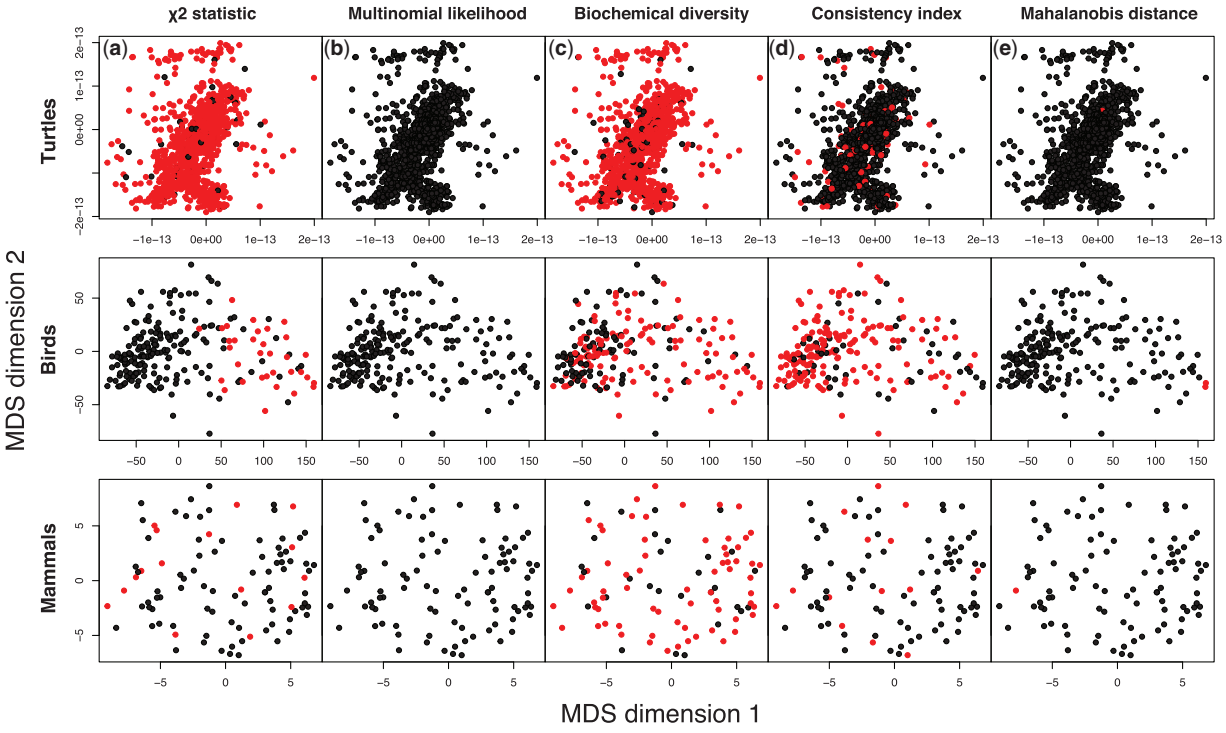

Fig. 5.—Estimated two-dimensional representation of tree-space for loci from turtles, birds, and mammals. Data are colored according to whether each locus passes (black) or fails (red) each of the five tests of model adequacy using our new thresholds for assessment.

Discussion

Assessing model adequacy can provide important insights alongside commonly used methods of model selection based on information criteria or Bayes factors. By assessing model adequacy, it is possible to consider that even the best-fitting model can provide a poor description of the process that generated the data (Gelman et al. 2014). This can be particularly useful when examining phylogenomic data sets, because a subset of the data can be selected to maximize the reliability of the model and inferences. Here, we compared the sensitivity of methods of assessing model adequacy to a diverse range of scenarios that can lead to biased inferences of tree topology and branch lengths. We show that the various test statistics for assessment differ in their sensitivity to biased inferences. We find that some test statistics can be highly lenient, whereas others can be conservative. Our results also support previous evidence that sequence length has a critical bearing on the perceived adequacy of a model (Duchêne et al. 2017). This phenomenon occurs because when longer sequences are analysed, the distribution of test statistics calculated for predictive data is narrower. As a consequence, long sequences can appear to have a greater discrepancy from predictive data compared with short sequences (Goldman 1993).

Using these insights, we have proposed thresholds for assessment that are specific to the most sensitive test statistics, implemented using a fast maximum-likelihood optimization framework. These sensitive statistics include X^2_m, *multinomial likelihood*, *biochemical diversity*, and *consistency index*. We have also shown the performance of the assessment using these four statistics simultaneously in a multivariate setting, using a statistic that we denote M_2. We focus on these five test statistics for deriving meaningful thresholds for assessment, which we hope can provide information about whether the model is likely to yield misleading inferences of tree topology or branch lengths. Exploring the power of novel test statistics for identifying misleading inferences will be an important avenue of research. For example, entropy metrics have been implemented for model assessment in a Bayesian framework (Brown 2014), and could be adapted for use in a maximum-likelihood setting by metrics that compare the empirical and predictive data to another null reference data set (Lewis et al. 2014). Similarly, more extensive simulation frameworks could lead to additional insights into the power of model assessment for detecting other potential sources of bias, such as tree imbalance or long-branch attraction.

Our results emphasize the importance of defining thresholds based on sequence length. In our simulation study and our analyses of three phylogenomic data sets, we find that the loci that yield estimates with high accuracy are usually long, such that they would be rejected using traditional thresholds for assessing model adequacy (supplementary figs. S1–S6, S8; Supplementary Material online). Critically,

the poor perceived model adequacy in analyses of long sequences might be exacerbated in empirical data, since the model is a greater simplification of the evolutionary process in these data compared with our simulations. Nonetheless, it seems unreasonable to reject the model for these data, since a simple model can be sufficient for estimating the parameters of interest accurately and precisely (Steel 2005). This leads us to propose thresholds for model assessment that are relatively lenient when applied to data generated by simulation, yet can reject the model in most cases when inferences are misleading. In analyses of phylogenomic data sets from birds and mammals, we find that our methods of assessment can reject a commonly used substitution model for loci that lead to inferences with anomalous tree topologies and low statistical support.

We also find that assessing model adequacy can be difficult for short loci. In the phylogenomic data from bird families, loci that were model-adequate according to our thresholds yielded congruent estimates of the tree topology, with strong statistical support. The data from birds comprised the longest loci of the three phylogenomic data sets that we investigated. These data also produced the most intuitive results, consistent with a previous study that also identified that model-adequate loci lead to congruent phylogenetic inferences (Doyle et al. 2015). Strikingly, when using the phylogenomic data set with the shortest sequences, based on ultraconserved elements from families of turtles, model assessment was less meaningful. In these data, model-adequate loci yielded phylogenetic estimates that were similar to those from loci for which the model was rejected. These results are congruent with those of a previous study of the X_m^2 statistic in phylogenomic data from birds, which showed that only some of the shortest loci were deemed model-inadequate and risked causing phylogenetic bias due to compositional heterogeneity (Duchêne et al. 2017).

Tests of model adequacy are frequently reliant on evaluating multiple test statistics and can be too lenient or conservative, such that it is difficult to interpret the phylogenetic performance of the model. The thresholds for assessment provided here should lead to more effective identification of models that are potentially misspecified. Nevertheless, assessing model performance remains difficult for data sets comprising short sequences. Development of novel test statistics should be accompanied by simulation studies that provide intuitive thresholds, based on the impact of model violation on inferences of interest. Other potential avenues of research include developing more summary metrics and graphics of model adequacy across the genome; assessing genome-wide models of inference, such as those developed for assessing the multispecies coalescent (Reid et al. 2014); or assessment for quantifying sequence information under the model (e.g., Klopfstein et al. 2017). Together, these advances will improve the reliability of phylogenomic inferences from sequence data.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by funding from the Australian Research Council to D.A.D. and S.Y.W.H. (grant DP160104173). S.D. was supported by a McKenzie Fellowship from the University of Melbourne. We acknowledge the University of Sydney for providing high-performance computing resources that have contributed to the research results reported within this paper.

Literature Cited

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol. 55(4):539–552.

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol. 60(5):685–699.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol. 19(7):1171–1180.

Brown JM. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst Biol. 63(3):334–348.

Burnham KP, Anderson DR. 2002. Model selection and multimodel inference: a practical information-theoretic approach. 2nd edn. New York: Springer

Crawford NG, et al. 2015. A phylogenomic analysis of turtles. Mol Phylogenet Evol. 83:250–257.

Doyle VP, Young RE, Naylor GJP, Brown JM. 2015. Can we identify genes with increased phylogenetic reliability? Syst Biol. 64(5):824–837.

Drummond AJ, Suchard MA. 2008. Fully Bayesian tests of neutrality using genealogical summary statistics. BMC Genet. 9:68.

Duchêne DA, et al. 2018. Analysis of phylogenomic tree space resolves relationships among marsupial Families. Syst Biol. 67:400–412. doi: 10.1093/sysbio/syx076.

Duchêne DA, Duchêne S, Ho SYW. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol Biol Evol. 34(6):1529–1534.

Duchêne DA, Duchêne S, Holmes EC, Ho SYW. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. Mol Biol Evol. 32(11):2986–2995.

Duchêne S, Di Giallonardo F, Holmes EC. 2016. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. Mol Biol Evol. 33(1):255–267.

Fiala KL, Sokal RR. 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution 39(3):609–622.

Fitch WM. 1971. Rate of change of concomitantly variable codons. J Mol Evol. 1(1):84–96.

Fitch WM, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet. 4:579–593.

Foster PG. 2004. Modeling compositional heterogeneity. Syst Biol. 53(3):485–495.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol. 18(5):866–873.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2014. Bayesian data analysis. Boca Raton (FL): Taylor & Francis.
Goldman N. 1993. Statistical tests of models of DNA substitution. J Mol Evol. 36(2):182–198.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.
Hillis DM, Heath TA, St John K. 2005. Analysis and visualization of tree space. Syst Biol. 54(3):471–482.
Ho SYW, Jermiin L. 2004. Tracing the decay of the historical signal in biological sequence data. Syst Biol. 53(4):623–637.
Höhna S, Drummond AJ. 2012. Guided tree topology proposals for Bayesian phylogenetic inference. Syst Biol. 61(1):1–11.
Höhna S, May MR, Moore BR. 2016. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. Bioinformatics 32(5):789–791.
Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346(6215): 1320–1331.
Jayaswal V, Wong TKF, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst Biol. 63(5):726–742.
Jermiin L, Ho SYW, Ababneh F, Robinson J, Larkum AW. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol. 53(4):638–643.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro H, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Klopfstein S, Massingham T, Goldman N. 2017. More on the best evolutionary rate for phylogenetic analysis. Syst Biol. 66:769–785. doi: 10.1093/sysbio/syx051.
Kluge AG, Farris JS. 1969. Quantitative phyletics and the evolution of anurans. Syst Biol. 18(1):1–32.
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(suppl 1):S4.
Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst Biol. 53(2):265–277.
Lewis PO, Xie W, Chen M-H, Fan Y, Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. Syst Biol. 63(3):309–321.
Liu L, Xi Z, Wu S, Davis CC, Edwards SV. 2015. Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci. 1360: 36–53.
Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AWD. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol. 34:153–162.
Mahalanobis PC. 1936. On the generalised distance in statistics. Proc Natl Inst Sci India 12:49–55.
Mardia KV, Kent JT, Bibby JM. 1979. Multivariate analysis. New York: Academic Press.
Matsen FA. 2006. A geometric approach to tree shape statistics. Syst Biol. 55:652–661.
Meredith RW, et al. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334(6055):521–524.
Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346(6210):763–767.
Murphy WJ, et al. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409(6820):614–618.
O'Hagan A. 2003. HSSS model criticism (with discussion). In: Green P, Hjiort N, Richardson S, editors. Highly structured stochastic systems. Oxford (UK): Oxford University Press. p. 423–453.
Penny D, Hendy MD. 1985. The use of tree comparison metrics. Syst Zool. 34(1):75–82.
Penny D, McComish BJ, Charleston MA, Hendy MD. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol. 53(6):711–723.
Phillips MJ. 2009. Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. Gene 441(1–2):132–140.
Pisani D, et al. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci U S A. 112(50):15402–15407.
Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. Syst Biol. [Internet] 50:580–601.
Prum RO, et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526(7574):569–573.
Rabosky DL, Glor RE. 2010. Equilibrium speciation dynamics in a model adaptive radiation of island lizards. Proc Natl Acad Sci U S A. 107(51):22178–22183.
Reddy S, et al. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian Tree of Life more than taxon sampling. Syst Biol. [Internet]. 66(5):857–879.
Reid NM, et al. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst Biol. 63(3):322–333.
Ripplinger J, Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. Mol Biol Evol. 27(12):2790–2803.
Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53(1-2):131–147.
Schliep KP. 2011. PHANGORN: phylogenetic analysis in R. Bioinformatics 27(4):592–593.
Shen X-X, Hittinger CT, Rokas A, Minh BQ, Braun EL. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol. 1(5):126.
Slater GJ, Pennell MW. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Syst Biol. 63(3):293–308.
Springer MS, Gatesy J. 2016. The gene tree delusion. Mol Phylogenet Evol. 94(Pt A):1–33.
Steel M. 2005. Should phylogenetic models be trying to "fit an elephant"? Trends Genet. 21(6):307–309.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci. 17:57–86.
Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenomic sampling and the sister lineage of land plants. PLoS One 7(1):e29696.
Wertheim JO, Sanderson MJ, Worobey M, Bjork A. 2010. Relaxed molecular clocks, the bias-variance trade-off, and the quality of phylogenetic inference. Syst Biol. 59(1):1–8.
Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10(6):1396–1401.
Zhou X, et al. 2012. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the laurasiatherian mammals. Syst Biol. 61(1):150–164.

Associate editor: David Bryant