bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Multispecies Coalescent Model Outperforms Concatenation across Diverse Phylogenomic Data Sets

Xiaodong Jiang[1], Scott V. Edwards[2*], Liang Liu[1,3*]

[1]Department of Statistics, University of Georgia, Athens, GA 30602

[2]Department of Organismic and Evolutionary Biology, Harvard, Cambridge, MA

[3]Institute of Bioinformatics, University of Georgia, Athens, GA 30602

*Corresponding authors

Scott Edwards

Department of Organismic and Evolutionary Biology

Harvard, Cambridge, MA

Tel: 617-384-8082

Email: sedwards@fas.harvard.edu

Liang Liu

Department of Statistics

Institute of Bioinformatics

University of Georgia

Tel: 706-542-3309

Email: lliu@uga.edu

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

A statistical framework of model comparison and model validation is essential to resolving the debates over concatenation and coalescent models in phylogenomic data analysis. A set of statistical tests are here applied and developed to evaluate and compare the adequacy of substitution, concatenation, and multispecies coalescent (MSC) models across 47 phylogenomic data sets collected across tree of life. Tests for substitution models and the concatenation assumption of topologically concordant gene trees suggest that a poor fit of substitution models (44% of loci rejecting the substitution model) and concatenation models (38% of loci rejecting the hypothesis of topologically congruent gene trees) is widespread. Logistic regression shows that the proportions of GC content and informative sites are both negatively correlated with the fit of substitution models across loci. Moreover, a substantial violation of the concatenation assumption of congruent gene trees is consistently observed across 6 major groups (birds, mammals, fish, insects, reptiles, and others, including other invertebrates). In contrast, among those loci adequately described by a given substitution model, the proportion of loci rejecting the MSC model is 11%, significantly lower than those rejecting the substitution and concatenation models, and Bayesian model comparison strongly favors the MSC over concatenation across all data sets. Species tree inference suggests that loci rejecting the MSC have little effect on species tree estimation. Due to computational constraints, the Bayesian model validation and comparison analyses were conducted on the reduced data sets. A complete analysis of phylogenomic data requires the development of efficient algorithms for phylogenetic inference. Nevertheless, the concatenation assumption of congruent gene trees rarely holds for phylogenomic data with more than 10 loci. Thus, for large phylogenomic data sets, model comparison analyses are expected to consistently and more strongly favor the coalescent model over the concatenation model. Our analysis reveals the value of model validation and comparison in phylogenomic data analysis, as well as the

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

need for further improvements of multilocus models and computational tools for phylogenetic inference.

Keywords: congruent gene trees, Bayesian model validation, Bayes factor, posterior predictive simulation, coalescent prior, independent prior, Metazoa

Due to the increasing growth in dimensionality and complexity of multilocus sequence data, accurate phylogenetic inference for understanding the evolutionary history of species faces substantial computational and modeling challenges (RANNALA AND YANG 2008; LIU *et al.* 2015; EDWARDS *et al.* 2016). With the assumption of topologically congruent (TC) genealogies across loci, a phylogenetic tree estimated from concatenated sequences is often used as the estimate of the species tree, despite the fact that the concatenation model oversimplifies the complexity inherent in the diversification of species by ignoring many biological phenomena, such as deep coalescence, hybridization, recombination, and gene duplication and loss, that are commonly observed during the history of species (MADDISON 1997; RANNALA AND YANG 2003; BRAVO *et al.* 2019). Since the advent of multilocus sequence data, there has been ongoing effort toward building stochastic models for handling gene tree variation. Given that incomplete lineage sorting (ILS) is likely the most common biological source of gene tree variation (EDWARDS 2009), some of the earliest efforts included parsimony and Bayesian multispecies coalescent models based on a coalescence process running along the lineages of the species tree (PAGE 1998; LIU AND PEARL 2007; LIU 2008; HELED AND DRUMMOND 2010). Subsequently, multispecies coalescent (MSC) model has been updated by adding biological parameters, such as gene flow, rate variation among lineages, recombination and hybridization (KUBATKO 2009; HEY 2010; WANG *et al.* 2014). In other developments, methods have been proposed to ask whether a tree is the best model

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for a given data set, or whether reticulations in the form of gene flow or lineage merging are more appropriate (see below) (MORET *et al.* 2004; JACKSON *et al.* 2017; BURBRINK AND GEHARA 2018). Meanwhile, skeptics of the MSC and advocates for simpler models, particularly concatenation models, have raised critical questions about appropriateness of the MSC and argued that observed gene tree variation is often not caused by ILS, but instead by gene tree estimation errors (SPRINGER AND GATESY 2016; SCORNAVACCA AND GALTIER 2017; RICHARDS *et al.* 2018). Given the inconsistency of concatenation methods under some regions of tree space in which coalescent methods of tree building are still consistent (LIU *et al.* 2010), to many researchers it is the convincing evidence to choose coalescent over concatenation methods for species tree inference. However, since the mathematical proofs or simulations for inconsistency of concatenation methods assume that the MSC model is true (KUBATKO AND DEGNAN 2007; ROCH AND STEEL 2015), it is important to show empirical evidence for validating and comparing the concatenation and MSC models. . Resolution of debates over concatenation and coalescent models requires a statistical framework of model comparison and model validation on a variety of empirical and simulated data.

Model fit and model adequacy in phylogenomics

Discussions about the necessity or adequacy of MSC models versus concatenation have taken three principle forms. One form involves questioning the phylogenetic signal in data sets designed for application of the MSC. For example, several authors have suggested that gene tree error (whether deriving from alignment artifacts, low signal, or gene tree estimation error), rather than ILS, is responsible for most if not all observed gene tree variation (GATESY AND SPRINGER 2014; ARCILA *et al.* 2017). If demonstrable gene tree variation can be ruled out for a given data set, this logic goes, then concatenation is a reasonable fallback model for analysis. This logic reasonably implies that the simpler model inherent in concatenation is favored, especially when gene tree variation can be shown to be

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

low or negligible. The problem with this logic, however, is that gene tree variation is only one motivation for MSC models. The other, more fundamental, motivation for MSC models is the conditional independence of loci in the genome, wherein recombination and random drift render the topologies, but more often the branch lengths of different loci independent of one another, conditional on the species tree, which necessarily influences the shape of all gene trees in the genome (EDWARDS *et al.* 2016). This point leads to the second common argument against MSC models: that violation of MSC model assumptions, such as evidence for recombination within loci or lack of recombination between loci or violations of neutrality, render MSC models poor descriptors of actual data sets, again recommending concatenation or other approaches as more robust alternatives (GATESY AND SPRINGER 2014; SCORNAVACCA AND GALTIER 2017). Again, however, demonstration of violation of a model's assumptions does not necessarily imply that that model is not a reasonable, or even the best available, description of the data. Indeed, we know of no violation of assumptions of an MSC model that is not also a violation of the concatenation model, especially since the concatenation model is best described as a special case of the MSC model, wherein all gene trees are topologically identical (LIU *et al.* 2015).

A third approach to deciding whether MSC, concatenation or other models might best apply to a given data set is testing for model fit and model adequacy (BROWN AND THOMSON 2018). Only a few papers have explicitly tested the fit and adequacy of the MSC, and, crucially, in doing so, have usually neglected to compare the fit of the major alternative model, concatenation, to that of the MSC. Reid et al. (REID *et al.* 2014) applied posterior predictive simulation (PPS), a Bayesian modeling approach, to a series of moderately sized data sets and concluded that a "poor fit to the multispecies coalescent is widely detectable in empirical data." Although this study was a major advance in our understanding of the MSC as applied to real data sets, we wonder whether the sweeping nature of this conclusion is

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reasonable and suggest that it may have been overly pessimistic. First, they definitively rejected the fit of the MSC at the level of gene trees for only 4 out of 25 data sets, and only 7 total loci (2.9%), hardly suggesting that poor fit at the level of the coalescent is "widespread". Reid et al. (2014) also suggest that a large percentage of partitions or loci in data sets, sometimes as high as 50%, violate the MSC; but the largest data set in their analysis consisted of only 20 loci, with 15 out of 25 data sets consisting of less than 10 loci, thereby possibly exaggerating the extent of violations of the MSC. Many of these rejections were based not on a poor fit of coalescent assumptions but on deviations of the loci from assumed substitution models, which is hardly a direct rejection of the MSC itself. In several cases, Reid et al (REID *et al.* 2014) were unable to distinguish whether the MSC was a poor fit due to analyses at the level of estimating gene trees and substitution models or at the level of the coalescent model itself. Additionally, their use of a χ^2 test for comparing the observed site patterns with those expected from the assumed substitution models could not accommodate missing data, making this aspect of their model testing problematic and raising the possibility that different substitution models could improve model fit. Moreover, several of the data sets in their analysis can be reasonably thought of as phylogeographic data sets (e.g., Leache (2009) and Walstrom et al. (2012)), in which gene flow is likely or expected, rather than phylogenomic data sets, in which gene flow is less expected or unexpected. Arguably, none of the data sets they analyzed could be called robustly phylogenomic in the modern sense: only 5 of the data sets interrogated relationships above the level of genus and none of them examined relationships among families suborders or orders, where, again, gene flow among lineages is less expected. It is therefore perhaps not surprising that such data sets do not conform well to the MSC. Jackson et al. (2017) recently applied a novel model fitting algorithm, Phrapl, to a series of multilocus phylogeographic data sets, concluding that a pure isolation model, such as the MSC, is rejected in favor of models including gene flow and

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

other reticulate events. Here, again, however, the data sets analyzed are explicitly termed phylogeographic, and, although the MSC was often applied to these data sets in the original papers, it is unsurprising that this model is a poor fit compared to models that include gene flow. Like all models, the MSC has a particular domain of application, one that we suggest is even wider than that in which concatenation is appropriate, but not so wide as to be applicable to data sets that demonstrably include gene flow or hybridization.

Most arguments in favor of concatenation, such as those summarized above, assert the superiority of concatenation by noting widespread perceived violations of the MSC or lack of demonstrable gene tree variation, and there have only been a few scattered examples of statistical comparisons of model fit and model adequacy between the MSC and concatenation. To our knowledge, only Liu and Pearl (2007) and Edwards et al. (2007) have explicitly compared the fit of MSC and concatenation models to the same data set, and asked which is a better fit (model comparison) and whether either model can account for the details of the multilocus data (model adequacy). Tests of model fit and model adequacy have been conspicuously absent from discussions about the relative merits of the MSC versus concatenation, although they have begun to appear in discussions of the relative merits of simple and more elaborate MSC models (WEN *et al.* 2016; JACKSON *et al.* 2017). Testing whether the MSC or concatenation can adequately describe multilocus data sets, and which model can describe those data better, will go a long way towards addressing concerns about the MSC and towards delimiting its appropriate domain of application.

In this paper, we will address several questions regarding the MSC and its application to empirical data sets: (1) is gene tree variation primarily caused by estimation error? (2) how well do the assumed substitution models fit multilocus sequence data and how does this fit drive the overall fit of the MSC? (3) how well does the multispecies coalescent model, divorced from shortcomings of the substitution model, fit empirical data? and (4) which

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model fits empirical multilocus data sets better, concatenation or the MSC? Question 1 can be addressed by a likelihood ratio test (LRT), where the null hypothesis is that all loci have the same tree topology (but possibly different branch lengths) versus the alternative hypothesis that allows gene tree variation across loci (testing TC gene trees; Figure 1). We addressed question 2 with the same χ^2 test used by Reid et al. (REID *et al.* 2014), comparing the observed and expected frequencies of site patterns, but this test cannot be applied to sequences with missing characters. We therefore modified this χ^2 test to handle missing characters (goodness of fit of substitution models; Figure 1). To evaluate the effect of substitution models on the fit of the MSC, model validation analysis was conducted with different substitution models for the same data sets. Question 3, the adequacy of the MSC, can be addressed using posterior predictive simulation (PPS) in a Bayesian framework (REID *et al.* 2014) implemented in a Bayesian phylogenetic program BEAST (SUCHARD *et al.* 2018). Unlike previous studies utilizing sequence data simulated under the MSC model to validate coalescent methods (KUBATKO AND DEGNAN 2007), PPS directly evaluates the fit of the MSC model by comparing the posterior gene trees and the gene trees simulated from the MSC (REID *et al.* 2014). However, the posterior gene trees generated with a coalescent prior are biased towards the MSC, especially when the alignments lack phylogenetic signal and Bayesian inference of gene trees is primarily driven by the coalescent prior. Validation analysis should instead compare the simulated gene trees with empirical gene trees estimated from multilocus sequence data without any influence of the MSC (i.e., the posterior gene trees generated independently across loci). Additionally, to reduce the influence of substitution models on the fit of the MSC, we perform the model validation only for loci that fit the substitution model (Bayesian model validation; Figure 1). Question 4, model comparison between the concatenation and MSC models, can again be addressed in a Bayesian framework, using Bayes factors (Bayesian model comparison; Figure 1) or other

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

posterior predictive approaches (LEWIS *et al.* 2014). By addressing these questions, we aim to directly compare competing models, especially as they apply to phylogenomic data sets where concatenation might plausibly be applied.

MATERIALS AND METHODS

Phylogenomic data sets

This study consists of 47 empirical data sets, including 25 data sets from Reid et al (REID *et al.* 2014) and 22 phylogenomic data sets across the tree of life. We chose phylogenomic data sets primarily based on their having been sampled from multiple species, usually more than 10 at the level of family or above, for their coverage of at least 50 loci and for their availability in already aligned nexus or phylip format on an easy accessible open access data base. We also eased our search by focusing primarily although not exclusively on data sets from the journal Molecular Phylogenetics and Evolution. The 22 phylogenomic data sets were downloaded from the data links available in the original papers (Table 1). The genetic markers of 22 phylogenomic data sets are highly diversified, including CDS matrices, exons, and UCEs (Table 1). There are 12 to 207 species in the 22 additional data sets, and the number of loci ranges from 110 to 30,636 (Table 1). Due to computational limits of Bayesian phylogenetic analyses, each of the 22 phylogenomic data sets was reduced to only include the alignments of the 10 most fully populated species across loci, and loci with missing sequences were removed from further analysis. Removing species from alignments will, if anything, increase gene tree similarity across loci as compared with data sets with the full complement of species, and therefore increase the fit of the concatenation model to the data. After data reduction, the data sets contained alignments of 36 (AITKEN *et al.* 2017) to 4709 (WU *et al.* 2018) loci, each with 10 species (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The sequence divergence (i.e., the average pairwise p-distance) of the 22 reduced phylogenomic data sets was significantly higher (p-value < 0.01) than that of 25 Reid et al. data sets (Figure A1), indicating that these new data sets are more classically 'phylogenomic' than those of Reid et al. (2014). Indeed, several of the data sets we analyze are at very high taxonomic levels, such as those of metazoans (WHELAN *et al.* 2015; SIMION *et al.* 2017) or mammals (SCORNAVACCA AND GALTIER 2017; WU *et al.* 2018). Some of these data sets, such as those from metazoans, are at such high taxonomic levels, that MSC models have to our knowledge never been applied, perhaps in the mistaken idea that ILS is extremely unlikely to occur among such deep lineages. However, as pointed out by many authors, even if ILS is harder to detect among deep lineages, it is no less likely to occur among deeply diverging than recently diverging lineages, since it is the length of internodes, rather than their depth in time, that is most relevant to MSC processes.

Goodness of fit of substitution models

Let $X = \{x_1, \ldots, x_k\}$ be the counts of *k* site patterns in a locus alignment. The random variables $X = \{x_1, \ldots, x_k\}$ have a multinomial distribution with $\sum_{i=1}^{k} x_i = l$, where *l* denotes the length of the alignment. Let *M* be the pre-selected substitution model. When there is no missing data, the χ^2 goodness of fit test can evaluate adequacy of model *M* by comparing the observed and expected counts of site patterns (REEVES 1992; GOLDMAN 1993; JHWUENG 2013). The test statistic is $\chi^2 = \sum_{i=1}^{k} \frac{(x_i - lp_i)^2}{lp_i}$, in which p_i is the probability of site pattern *i*, which can be estimated under the null hypothesis using the maximum likelihood (ML) estimates of the tree topology, branch lengths, and the parameters in model *M*. Asymptotically, the test statistic has the χ^2 distribution with (*k*-1) degrees of freedom. Since this χ^2 test statistic is based on the frequencies of sites for fully populated characters, sites with missing characters (gaps, ambiguous nucleotides, and unidentified regions) are excluded from analysis, even though only a small portion of nucleotides are missing in each site. To

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

incorporate partially missing sites, a modified χ^2 test was developed to calculate the observed and expected counts of site patterns in the presence of missing data (WADDELL 2005). We instead calculate the marginal proportion of a site with missing characters and then compare the marginal proportion of the site with its expectation. For example, there are two missing characters in a site {??AC} of 4 species (S_1, S_2, S_3, S_4), i.e., the nucleotides from species S_1 and S_2 are missing. The marginal proportion y_{AC} of {AC} in the alignment of S_3 and S_4 is given by $y_{AC} = x_{AC}/z$, in which x_{AC} is the count of {AC} and z is the number of sites without missing characters in the alignment of S_3 and S_4. Since a small z indicates a large amount of missing data in the alignment, we ignore the sites for which $z/l \leq 0.8$. The test statistic $t = \sum_{j=1}^{n} |y_j - p_j|$, in which y_j is the observed proportion of a site pattern with or without missing characters, and p_j is the probability of the site pattern under the null hypothesis, and n is the number of site patterns for which $z/l > 0.8$. The null distribution of the test statistic is estimated by parametric bootstrap samples generated from the ML estimates of the tree topology, branch lengths, and the parameters in model *M*.

Marginal tests of base compositions between pairs of taxa (TAVARE 1986) have been found to be more powerful than the χ^2 goodness of fit test of the substitution models to phylogenetic data (WADDELL *et al.* 2009). We therefore perform two marginal tests. The first test is to detect heterogeneity of base compositions across species, when the observed base frequencies of individual species deviate significantly from the overall average base frequencies across species. Let $\{x_A, x_C, x_G, x_T\}$ be the observed frequencies of nucleotides A, C, G, and T of a species. The frequency of nucleotide i $(i = A, C, G, T)$ is equal to $x_i = N_i/N$, where N_i is the number of nucleotide i and N is the total number of nucleotides excluding missing characters in the sequence. Under the null hypothesis of homogeneous substitution models, all species are expected to have the same base frequencies and thus the expected base frequencies under the null hypothesis are estimated by the overall average frequencies

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$\{p_A, p_C, p_G, p_T\}$ of A, C, G, and T across species. The χ^2 test of the observed frequencies $\{x_A, x_C, x_G, x_T\}$ against the expected frequencies $\{p_A, p_C, p_G, p_T\}$ is carried out for each species to detect those species whose base frequencies significantly deviate from the null hypothesis that all species have the same base frequencies. Similarly, the second marginal test is applied to the frequencies of double-nucleotides (doublets) between pairs of species to find pairs of species for which doublet frequencies are inconsistent with the pre-selected substitution model. The frequency of doublet ij ($i = A, C, G, T$ and $j = A, C, G, T$) is equal to $x_{ij} = N_{ij}/N$, where N_{ij} is the number of doublet ij and N is the total number of doublets excluding those with missing characters.

The bootstrap and two marginal tests were conducted to evaluate goodness of fit of the GTRGAMMA model to 47 empirical data sets (Figure 1). We chose the GTRGAMMA model because the most parameter-rich model (GTRGAMMA in RAxML) is sufficient for reliable phylogenetic inference (Abadi et al. 2019) and a more complex model tends to be a better fit to phylogenomic data (LIU *et al.* 2017). For each locus, the ML estimates of the phylogenetic tree and other parameters were obtained by RAxML v8.2.3 (STAMATAKIS 2014) using the GTRGAMMA model. Then, 100,000 base pairs of sequence were simulated from the estimated phylogenetic tree using Seq-Gen v1.3.2x (RAMBAUT AND GRASSLY 1997). The expected frequencies of site patterns were estimated by the corresponding frequencies in the simulated sequences. In addition, 100 bootstrap samples were generated by simulating DNA sequences from the concatenation tree for each locus. If taxa were missing from the loci, they were pruned from the concatenation tree and DNA sequences were simulated from the pruned concatenation tree. The test statistic was calculated for each bootstrap sample, and the observed test statistic t* was compared with the bootstrap test statistics $\{t_1, \ldots, t_{100}\}$ and p-value = (# of $t_i > t^*$) / 100. For the marginal test of base frequencies, the sequence of a species was considered significant if its p-value was less than 0.05 divided by the number of

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

297 species. A locus was significant if it had at least one significant species. Similarly, for the test
298 of doublet frequencies between pairs of species, a pair of species was significant if its p-value
299 was less than 0.05 divided by the number of pairs. A locus was significant if it has at least
300 one significant pair. The bootstrap and two marginal tests were applied to each locus of the
301 47 data sets. A locus waas significant if any of above three tests was significant for the locus.

302 *Testing for topologically congruent gene trees*

303 We ask the question whether a single gene tree topology can adequately explain a
304 given multilocus data set, using a variant of the LRT. Similar types of LRTs were developed
305 to test if alternative trees are congruent with the maximum likelihood (ML) tree for a single
306 locus (SHIMODAIRA AND HASEGAWA 1999; SHIMODAIRA 2002). McVay and Carstens (2013)
307 proposed a parametric bootstrap approach to assess the extent to which gene tree variation
308 can be attributed to phylogenetic estimation error. Here, we develop a LRT to evaluate the
309 concatenation assumption of congruent gene trees for multiple loci. Let $D = (D_1, D_2, \ldots, D_K)$
310 be the concatenated alignments of a multilocus sequences data set, in which D_i represents the
311 alignments of locus *i* and *K* is the count of loci. The tree topology of locus *i* is denoted by τ_i.
312 Under the concatenation model, all loci are assumed to have the same tree topology, i..e,
313 $\tau_1 = \cdots = \tau_K$. We develop an LRT to evaluate the null hypothesis that all gene trees have the
314 same topology, i.e., $\tau_1 = \cdots = \tau_K$ versus the alternative hypothesis that not all gene trees are
315 topologically identical. The test statistic is defined as $t = \log(l_1) - \log(l_0)$, in which l_0 and
316 l_1 are the likelihoods of the null and alternative hypotheses. Under the null hypothesis, the
317 ML tree is built from the concatenated alignments across loci using RAxML v8.2.3
318 (STAMATAKIS 2014) with the GTRGAMMA model (TAVARE 1986; YANG 1994). Using the
319 ML tree instead of the true concatenation tree for the null hypothesis may lead to a biased test.
320 However, since the concatenated sequences of phylogenomic data consist of millions of base
321 pairs, the ML tree is very similar, if not identical, to the true concatenation tree. hus, the bias

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

induced by the topologcial difference between the ML and true concatenation tree is negaligible for phylogenomic data. Let w_i be the log-likelihood of locus *i* by refitting branch lengths and substitution model parameters to the concatenation tree with missing taxa being removed. The log-likelihood under the null hypothesis is equal to the sum of the log-likelihoods of individual loci, i.e., $\log(l_0) = \sum_{i=1}^{K} w_i$. Refitting branch lengths and model parameters on a fixed tree topology was performed in RAxML using the command line "*raxml-HPC-AVX –s datafile –m GTRGAMMA –n outputfile –f e –t fixtree*". To find the log-likelihood of the alternative hypothesis, maximum likelihood trees were independently built for individual loci using RAxML. Since model parameters include tree topologies, the χ^2 distribution is not a good approximation to the null distribution of the test statistic (JHWUENG *et al.* 2014). Therefore, the null distribution of the test statistic was approximated by a parametric bootstrap. Bootstrap samples were generated under the null hypothesis by simulating DNA sequences from the concatenation tree pruned for available species at each locus. Since the original alignments include missing characters (gaps, ambiguous nucleotides, and unidentified regions), bootstrap samples should involve a similar pattern of missing characters. Thus, corresponding nucleotides in bootstrap samples were replaced by missing characters. Let t_i be the value of the test statistic *t* for the bootstrap sample *i* = 1, …, *B*. The loglikelihoods of the null and alternative hypothesis in the test statistic *t* were generated using RAxML. The values $\{t_i, i = 1, \ldots, B\}$ of the test statistic for *B* bootstrap samples were used to approximate the null distribution of the test statistic *t*. The p-value was estimated by the proportion of $\{t_i, i = 1, \ldots, B\}$ that were greater than or equal to the test statistic *t* calculated from real data. Rejection of the null hypothesis indicates that some gene trees are incongruent with the concatenation tree. Then, the LRT was further applied to each locus to identify alignments that reject the null (concatenation) tree. Let τ be the concatenation tree and τ_i is the gene tree *i*. The null and alternative hypotheses are $H_0: \tau_i = \tau$ and $H_1: \tau_i \neq \tau$. The test

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

statistic is $t = \log(l_1) - \log(l_0)$, in which $\log(l_0)$ and $\log(l_1)$ are the log-likelihoods of locus *i* under the null and alternative hypotheses. The null distribution of the test statistic was approximated by the bootstrap samples generated under the null hypothesis and the p-value was equal to the proportion of bootstrap test statistics that are greater than or equal to the observed test statistic. A locus was significant if its p-value was less than or equal to 0.05 divided by the number of loci (Bonferroni correction for multiple comparisons).

Bayesian validation of the multispecies coalescent model

Like Reid et al. (2014), the Bayesian MSC model is here validated by Bayesian predictive simulation. Reid et al. recommended comparing the posterior coalescent gene trees to the gene trees simulated from the posterior species trees. However, the posterior coalescent gene trees have been influenced by the multispecies coalescent prior, which is implemented in packages such as BEST (LIU 2008) or *BEAST (HELED AND DRUMMOND 2010). The posterior gene trees with an MSC prior are biased towards the MSC model to some extent, compared to gene trees generated without an MSC prior, making this test conservative (i.e., in favor of the MSC). Thus, here, simulated gene trees are compared to posterior gene trees generated without the MSC prior. We validate the MSC model using two tests. The first test, as in Reid et al. (2014), involves two comparisons - comparing the multispecies coalescent likelihoods of the simulated and posterior coalescent gene trees and comparing the number of deep coalescences of the simulated and posterior coalescent gene trees. The first test is rejected if either or both of two comparisons reject the MSC model. In the second test, not used by Reid et al. (2014), the posterior gene trees are first estimated independently across loci in BEAST by unlinking the substitution models, clock models, and trees. If the 99.9% posterior credible regions of the log-likelihoods of the independent model and the MSC model are not overlapped, we conclude that there is a significant difference between the posterior independent and coalescent gene trees. We infer a poor fit of the MSC

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model to the data if either or both of the two tests are rejected. Both tests can also be applied to each locus in the data to identify gene trees that significantly deviate from the MSC model.

To alleviate the impact of poorly fitting substitution models on the fit of the MSC, the model validation and comparison analyses were only performed on the loci that fit the assumed substitution model (GTRGAMMA). Since the number of loci in 25 Reid et al. data sets is insufficient for filtering out the unfit loci, selection of loci that fit the substitution model was only performed on 22 reduced phylogenomic data sets. After locus selection, 9 of 22 data sets had ≤ 3 loci that fit the substitution model and thus they were removed from further analysis. We then randomly selected 50 loci from the remaining 14 phylogenomic data sets, resulting in a total of 39 data sets (25 Reid et al. data sets + 14 phylogenomic data sets with 50 loci) for the model validation analysis (Figure 1). The xml input files of the 25 data sets from Reid et al. were available in the data package provided by the authors. Reid et al. (2014)'s *BEAST analyses of those data sets assumed HKY and TN93 (+GAMMA) substitution models for individual loci and a multispecies coalescent prior for gene trees. To evaluate the effect of substitution models on the fit of the MSC model, Bayesian model validation of 25 Reid et al. data sets was reconducted with the GTRGAMMA model. For 14 phylogenomic data sets, to reduce the effect of substitution models on the overall fit of the MSC, the validation analysis was conducted only for loci for which the GTRGAMMA model was a good fit. Consequently, rejections of the MSC imply a poor fit of coalescent assumptions rather than a poor fit of the assumed substitution model. The xml input files of 14 data sets were generated using BEAUti v1.8.4, assuming the GTRGAMMA substitution model for all loci.

Two independent runs were carried out for each analysis and convergence was checked by comparing the outputs from two runs. The first 50% of MCMC samples were discarded as burn-in. Then, 1000 samples were selected from the remaining MCMC samples

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and used as input for Bayesian model validation. The first test of Bayesian model validation was conducted using the R package starbeastPPS (2014). To perform the second test, 39 input data sets were re-analyzed by unlinking the substitution models, clock models, and trees across loci in BEAST, which produced posterior gene trees independently across loci without the MSC prior. The 99.9% credible region of the difference between the log-likelihoods of the independent and coalescent models was calculated in R. The two tests were also applied to each locus to identify gene trees that significantly deviate from the MSC model.

To evaluate the effect of loci rejecting the MSC on species tree estimation, species trees were built from all loci and only loci that did not reject the MSC for each of the 14 phylogenomic data sets using NJst (LIU AND YU 2011) implemented in an R package Phybase (LIU AND YU 2010). This analysis was not performed for 25 Reid et al. data sets due to their small numbers of loci for species tree estimation. We calculated the Robinson-Foulds (RF) distance (ROBINSON AND FOULDS 1981) of two species trees reconstructed for each data set using the function dist.topo in an R package APE (PARADIS AND SCHLIEP 2019). For the species trees with a positive distance, to evaluate statistical significance of the difference between two species trees, bootstrap support values of the incongruent branches were calculated using bootstrap gene trees estimated from alignments. Specifically, bootstrap gene trees were built for each locus and then used as input data to calculate bootstrap NJst trees. The bootstrap support value of a branch was equal to the count of bootstrap NJst trees supporting the clade indicated by the branch. Romiguier et al. (2013) suggested that high GC content may cause problems for phylogenetic inference under the MSC model, and that selecting AT-rich loci can improve the resolution of estimated phylogenies. To investigate the association between GC content and poor fit of the MSC, we calculated GC content for loci rejecting and accepting the MSC, respectively. A two-sample t-test was used to find

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significant difference in GC content between loci rejecting the MSC and those accepting the MSC.

Bayesian model comparison for MSC versus concatenation

Bayesian model comparison for concatenation versus the MSC was evaluated using Bayes factors (BF) (KASS AND RAFTERY 1995), the ratio of marginal likelihoods $BF = P(D|M_1)/P(D|M_2)$, in which D denotes data and M_1 and M_2 are two competing models. Here, M_1 is the MSC model and M_2 is the concatenation model. The Bayesian concatenation analyses assumed a partition model by unlinking substitution models and clock models across loci. The marginal log-likelihoods of the Bayesian multispecies coalescent and concatenation models were estimated using path sampling / stepping-stone sampling with 100 path steps implemented in BEAST. A value of log(BF) > 10 indicates that the Bayesian multispecies coalescent model M_1 is strongly favored by the data versus the Bayesian concatenation model M_2.). The model comparison analysis was applied to the (39) data sets used in the Bayesian model validation analysis. Since the model comparison analysis implemented in BEAST does not allow missing taxa in any locus, 8 data sets with missing data were removed, further reducing the number of input data for the model comparison analysis to 31 (Figure 1).

To demonstrate that model assumptions may influence species tree inference, we reconstructed species trees for 5 phylogenomic data sets using concatenation and a coalescent method (NJst). In addition, we subsampled 25%, 50%, and 75% of loci from the original phylogenomic data, and compared the species trees built from the subsamples with those for the full phylogenomic data. The concatenation trees were estimated by RAxML with the GTRCAT model. The NJst trees were built using the function sptree.njst() in an R package Phybase. We calculated the bootstrap support values for each estimated species tree. Two estimated species trees are said to be significantly incongruent if two trees have a conflict branch with a bootstrap support value > 70. The subsampling analysis was repeated 10 times

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and we reported the proportion of subsamples (out of 10) for which the estimated species tree was significantly incongruent with the species tree built from the full phylogenomic data.

RESULTS

Goodness of fit of substitution models

There was a total of 20,032 loci (CDS, UCEs or exons) throughout the 47 empirical data sets; 241 loci from the 25 Reid et al. data sets and 19,791 loci from the 22 phylogenomic data sets. The marginal test of base frequencies identified a total of 1,362 (7%) loci/alignments (or 9% loci per data set in Figure 2a) for which at least one sequence significantly deviates from the average base frequencies expected from the GTRGAMMA model. The doublet test indicated that 6,990 (35%) loci/alignments (or 51% loci per data set in Figure 2a) have at least a pair of species whose doublet patterns are significantly different from the patterns expected from the GTRGAMMA model. The marginal test favors complex substitution models, because a complex model has a higher likelihood than a simple model, indicating that the expected frequencies under the complex model are more consistent with the observed frequencies of site patterns in the alignments. Thus, rejection of the GTRGAMMA model suggests that a more complex substitution model should be fit to the data. In the bootstrap test for site patterns, the GTRGAMMA model fails to fit the alignments of 5,718 (24%) loci across the 47 data sets (or 34% of loci per data set in Figure 2a). The doublet test appears to be more likely than the bootstrap test for site patterns to reject the GTRGAMMA model (Figure 2a), indicating the necessity of marginal tests for goodness of fit of substitution models. Because the doublet test is more likely to reject the GTRGAMMA model than the other two tests, the intersection test (the intersection of three tests, i.e., at least one of two marginal tests and the bootstrap test reject the GTRGAMMA model) appears to be primarily driven by the doublet test (Figure 2a). Overall, nearly half of the alignments

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(8,775 loci or 44%) were found significant (p-value < 0.05) in the intersection test, and the proportion of significant loci identified by the intersection test ranges from 6% to 100% across the 47 data sets (Figure 2a). A two-sample t-test finds no significant difference in the proportion of loci rejecting the GTRGAMMA model between the 25 Reid et al. data sets and the 22 phylogenomic data sets (Figure 2b).

To understand the causes of poor substitution model fit, we investigated the relationship between GC content (and proportion of informative sites) and the rejection of the GTRGAMMA model. A two-sample t-test suggests that the proportions of GC content and informative sites of loci rejecting the GTRGAMMA model are significantly higher (p-value < 0.05) than those for loci that fit the GTRGAMMA model (Figure 2c-d). We fitted a logistic regression for all loci across the 47 data sets, where non-significance (0) or significance (1) of a locus in the intersection test is the binary response variable, and the proportions of GC content and informative sites are two explanatory variables. In the fitted logistic regression, the coefficients of two explanatory variables are significantly negative with p-value < 0.01. We further fit a logistic regression to each of the 47 data sets. The coefficient of GC content is negative/positive for 29/18 data sets, among which 6/1 negative/positive coefficients are significant at the α level of 5% (Figure 2e). Similarly, the coefficient for the number of informative sites is negative/positive for 33/14 data sets, among which 14/0 negative/positive coefficients are significant (Figure 2e). The preponderance of significantly negative coefficients indicates that GC content and the number informative sites are both negatively correlated with the probability of a locus rejecting the GTRGAMMA model, i.e., a higher GC content and/or proportion of informative sites tends to increase the chance of a poor fit of the GTRGAMMA substitution model.

Testing topologically congruent gene trees

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The LRT for TC gene trees rejected the null hypothesis of tree congruence for all 47 empirical data sets with p-value < 0.05. Thus, all empirical data sets in this study strongly favor the alternative hypothesis of incongruent gene trees, a pattern that cannot be adequately explained by gene tree estimation errors. The p-values of the data sets with 10 or more loci are very close to 0, indicating that the assumption of TC gene trees is rarely satisfied for phylogenomic data, which often involve thousands of loci.

A two-sample t-test finds no significant difference (p-value = 0.10) in the proportion of loci rejecting the null hypothesis of TC gene trees between the 22 phylogenomic data sets and the 25 Reid et al. data sets (Figure 3a). The topological congruence LRT on individual loci suggests that 38% of gene trees across 47 data sets are statistically incongruent with the concatenation tree (Figure 3b). When the 47 data sets are grouped into 6 categories – mammals (11), birds (11), insects (6), fish (5), reptiles (5), and others (9), an ANOVA analysis finds no significant difference in the proportion of loci rejecting the hypothesis of TC gene trees among 6 groups (Figure 3c). Both the two-sample t-test and ANOVA indicate that the proportion of loci rejecting the hypothesis of gene tree congruence is similar across groups and data sets. A linear regression line was fit for the log scale of the number of incongruent loci rejecting the hypothesis of TC gene trees (y) versus the log scale of the number of loci (x), i.e., $\log(y) = 0.87 \times \log(x) - 0.08$ with a significant (p-value < 0.01) positive correlation between $\log(x)$ and $\log(y)$ (Figure 3d). This result is consistent with the previous conclusion that phylogenomic data sets with more loci are more likely to reject the assumption of TC gene trees; namely, the observed gene tree variation cannot be adequately explained by gene tree estimation error. Moreover, both ANOVA and linear regression analyses suggest a constant and high proportion (38%) of loci rejecting the assumption of TC gene trees across 47 data sets, providing strong evidence for violation of the concatenation assumption of congruent gene trees in phylogenomic data across the tree of life. When the 22

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phylogenomic data sets are grouped by data types – CDS (10), EXON (4), UCE (8), the t-tests for pairwise comparisons find no significant difference (p-value = 0.1) for the proportion of loci rejecting the assumption of TC gene trees between the CDS and Exon groups, but the proportions of both groups are significantly (p-value < 0.01) higher than that of the UCE group (Figure 2e). This result indicates that the congruent gene tree assumption of the concatenation model is more likely to hold for the UCE data than for the CDS and Exon data.

To reduce the potential bias caused by an unfit substitution model, the LRT was only applied to the loci that fit the GTRGAMMA model. We filtered out 27 data sets with < 5 loci that fit the GTRGAMMA model and applied the LRT to the remaining 20 data sets. The null hypothesis of tree congruence was rejected for all 20 data sets with p-value < 0.05. The LRT on individual loci suggests that 46% of gene trees across 20 data sets are statistically incongruent with the concatenation tree, which is lower (but not significantly) than the proportion (48%) when the LRT was applied to all loci of the 20 data sets (Figure A2).

Bayesian model validation

The coalescent methods have been widely used for estimating species trees from phylogenomic data. Due to computational constraints, however, few studies have evaluated the fit of the MSC to the multilocus sequences. Here, we validate the MSC model using two tests based on Bayesian predictive simulation. The first test (i.e., PPS proposed by Reid et al. (2014)) compares the simulated gene trees with the posterior coalescent gene trees generated with the MSC prior, whereas the second test (i.e., the independent test) compares the posterior coalescent gene trees with the posterior independent gene trees generated with the independent prior. The analysis of 25 Reid et al. data sets found that 8 (32%) data sets failed either or both of two tests (Figure 4a), among which 3 data sets were also found to poorly fit to the MSC by Reid et al (2014; *Certhiidae, Tamias, Aliatypus*). However, the xml input

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

files provided by Reid et al (2014) assumed the HKY and TN93 (+GAMMA) substitution models for all loci. When the *BEAST analyses were rerun with the GTRGAMMA model, only two data sets provided were rejected by either or both of two tests (Table 2). Thus, the choice of substitution models has major effects on the fit of the MSC. In addition, Bayesian model validation for the 14 phylogenomic data sets for which the GTRGAMMA model was a good fit found that 12 data sets failed the first test and all 14 data sets failed the second test. Thus, the MSC model failed to fit all 14 phylogenomic data sets (Figure 4a). Since these 14 data sets have more loci than the Reid et al. data sets, it implies that phylogenomic data with more loci are more likely to reject the MSC. In addition, Bayesian model validation for individual loci of the 14 phylogenomic data sets found that 10.88% of loci rejected the MSC (Figure 4b), significantly higher (p-value < 0.01) than the proportion (4.17%) for the 25 Reid et al. data sets, indicating that the probability of a locus rejecting the MSC increases as the number of loci grows.

To evaluate the effect of loci rejecting the MSC on species tree estimation, two species trees were reconstructed from all loci and only loci that fit to the MSC for each of 14 phylogenomic data sets. A majority (9) of 14 data sets produced two identical species trees (i.e., RF = 0), whereas RF = 2 for 4 data sets (Ant, Cracid, Mammal2017, Squirrel) and RF = 4 for 1 data set (Figure 4c). Note that RF = 2 or 4 indicates only 1 or 2 conflicting branches in two species trees. Since the incongruent branches are not strongly supported (bootstrap support values < 60), the conflict between two different species trees is not significant. This analysis suggests that including loci that fail to fit the MSC has little impact on species tree estimation, when a small portion (10.88%) of loci rejects the MSC.

To investigate the association between GC content and poor fit of the MSC, we calculated the average GC content of loci rejecting or accepting the MSC for two phylogenomic data sets for which the number of loci rejecting the MSC is large (Birds and

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mammal2017, in which the number of loci rejecting the MSC is 10 and 17, respectively; Figure 4d). Other phylogenomic data sets contain insufficient number of loci rejecting the MSC for the analysis. A two-sample t-test for the overall average GC content (combining two data sets) and the average GC content of each of two data sets unanimously found that the difference in GC content between loci rejecting the MSC and loci accepting the MSC was not significant (Figure 4d), showing little evidence for a positive association between high GC content and poor fit of the MSC.

Bayesian model comparison for MSC versus concatenation

Bayesian model comparison was applied to 31 data sets for which there was no missing data, including 17 Reid et al. data sets and 14 phylogenomic data sets. The BFs (on logarithmic scale) of 22 data sets are greater than 100 and the BFs of the remaining 9 data sets are between 10 and 100. Overall, the high BFs imply that all 31 data sets strongly favor the MSC rather than the concatenation model (Figure 5). This Bayesian model comparison analysis is consistent with the LRT results for congruent gene trees, which reject the concatenation assumption of congruent gene trees and thus favor the MSC for all 47 data sets.

To demonstrate the impact of model assumptions on species tree inference, species trees were estimated for 4 phylogenomic data sets (Cracids 23 species and 430 loci, Gallophesants 18 species and 1479 loci, Lizard 29 species and 1852 loci, Shrews 19 species and 1112 loci, Table 1) using concatenation and a coalescent method NJst. We found 3 data sets (Cracids, Gallophesants, Lizard) for which the concatenation trees were significantly incongruent with the corresponding NJst trees (Figure A3), indicating that different models may yield conflict species trees. In the subsampling analysis, the average proportion of significantly incongruent concatenation trees across 4 data sets is 0.19, much higher than the average proportion (0.008) of significantly incongruent NJst trees (Figure A4), suggesting

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that concatenation is more likely than coalescent methods in producing wrong relationships with high bootstrap support values.

DISCUSSION AND CONCLUSIONS

Model validation and comparison are essential to accurate phylogenetic inference for genome-scale sequence data. Many recent disputes about the utility of the MSC model in phylogenomics have rested on perceived model violations of the MSC, rather than direct tests of the explanatory power of the MSC versus concatenation. Here we developed and implemented a set of statistical tests to evaluate the adequacy of substitution models, the concatenation model, and the MSC model. In particular, we tried to distinguish two possible sources of rejection of the MSC in empirical data sets: rejection due to violation of the substitution model and rejection due to violation of the MSC. The LRT results for congruent gene trees reveals strong evidence in 47 data sets against the concatenation assumption of congruent gene trees across loci. Crucially, this test suggests that the gene tree variation is real and cannot be explained simply by gene tree estimation error, a point of increasing concern among skeptics of the MSC model (RICHARDS *et al.* 2018). This result is consistent with the subsequent Bayesian model comparison analysis, which unanimously favored the MSC over concatenation for all phylogenomic data sets under consideration. Moreover, the proportion of gene trees significantly deviating from the concatenation tree (38%) is consistently high across taxonomic groups (bird, fish, mammal, insect, reptile, and others), and our linear regressions suggest that the concatenation assumption of congruent gene trees is more seriously violated as the number of loci continues to grow in phylogenomic data across the tree of life (BRAVO *et al.* 2019).

The fact that Bayesian model comparison strongly favors the coalescent over concatenation does not necessarily validate the use of the MSC model for analyzing phylogenomic data. In our Bayesian model validation, the MSC model is a good fit for the

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

majority data sets in Reid et al. (REID *et al.* 2014), but the choice of substitution models has a strong influence on the fit of the MSC model. To alleviate the effect of substitution models, we applied Bayesian model validation to the loci of 14 phylogenomic data sets that fit the assumed substitution model. The MSC failed to fit all 14 phylogenomic data sets, but the proportion of loci rejecting the MSC was only 11%, significantly smaller than those for substitution models (44%) and concatenation models (38%). Thus, deficiencies in the fit of data to substitution models and to concatenation models appear to be much more severe than fit to the MSC model, suggesting that more attention should be given to appropriately modeling the evolution of single nucleotides (i.e., substitution models) and gene tree variation, though continuous efforts for improving models at the level of both sequences and gene trees are ultimately desirable.

An empirical study of placental mammals (ROMIGUIER *et al.* 2013) suggested that GC-rich regions are associated with high recombination rates which may be problematic for species tree inference under the MSC model. Our analysis, however, finds no convincing evidence for the positive association between high GC content and poor fit of the MSC. Instead, we find that high GC content is strongly associated with poor fit of substitution models. Thus, the shifting phylogenetic relationships of placental mammals for GC-rich regions found by Romiguier et al. (2013) may be caused not by poor fit of the MSC, but the conflicts among gene trees due to poorly fitted substitution models.

Stochastic models for phylogenomic data should consider the cumulative effect of the mutation process of nucleotides (molecular evolution) and biological processes rooted in population genetics that have played important roles in the evolution of species. Some have argued (Edwards 2009; Liu et al. 2015) that, among the relevant biological processes, the coalescence process, which assumes random drift, should serve as the null model, and other biological factors, such as gene flow and hybridization, can be added to the null model if the

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

null model cannot adequately explain the observed gene tree variation. As the number of loci continues to increase, some loci are bound to reject the MSC model. Brown and Thomson (2017) suggests that a small number of extremely influential loci can significantly change the estimates of phylogenetic trees. Our analysis indicates that if the loci rejecting the MSC (or concatenation) only account for a small proportion of the empirical data (e.g. 11% in this study), the MSC model (or concatenation) can still be applied to entire data sets or to data sets purged of the loci that violate the MSC. On the other hand, a large number of loci rejecting the MSC (or concatenation) suggests that additional biological phenomena may have occurred and must be added to the stochastic model when analyzing such data sets.

Mathematical models are variably robust to assumption violations. Several authors have identified numerous putative biological violations of the MSC in empirical data sets, including recombination within loci, pseudo-concatenation of loci such as occurs in transcriptome data as well as natural selection (GATESY AND SPRINGER 2014; SCORNAVACCA AND GALTIER 2017). However, even in these data sets, despite numerous putative violations of the MSC, the MSC is a better fit than concatenation, suggesting that violations of the MSC may not recommend falling back on concatenation as an alternative method of analysis (LIU *et al.* 2015; EDWARDS *et al.* 2016). The analysis of empirical data in this study suggests that, although 11% of loci reject the MSC, there is no significant difference between the species trees estimated from all loci and only loci that fit the MSC model. Thus, gene-tree-based coalescent methods are robust to a certain degree of violation of coalescent assumptions. In this study, model validation and species tree analyses were conducted on reduced phylogenomic data sets of 50 loci and 10 species each. Our analyses suggest that the concatenation assumption of congruent gene trees rarely holds for phylogenomic data with more than 10 loci. Thus, for large phylogenomic data sets, model comparison analyses are expected to consistently and more strongly favor the coalescent model over the concatenation

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model. As the number of loci increases, model validation analysis is more likely to reject the MSC. Adding species above the level of genus in phylogenomic data, however, will introduce additional gene tree variation caused by gene flow, gene duplication/loss, etc. Thus, unlike increasing numbers of loci, growth in the number of taxa in phylogenomic data is likely to increase the proportion of loci rejecting the MSC. When the majority of loci do not support the MSC, the coalescent methods will eventually fail to accurately reconstruct species trees from the full phylogenomic data, even though model comparison is still in favor of the MSC over the concatenation model. In such cases, MSC network models, in which gene flow and lineage merging are incorporated, may better fit phylogenomic data sets than the standard MSC model (WEN *et al.* 2016; BASTIDE *et al.* 2018).

SUPPLEMENTARY MATERIAL

The alignments of the input data sets for the substitution model, TC gene trees, model validation, and model comparison analyses are available from the Dryad Digital Repository: http://dx.doi.org/ 10.5061/dryad.7q6q3s0

FUNDING

This research was supported by NSF grant DEB 1355343 (EAR 1355292) to S.V.E, and NIH/NIAID R01AI093856 to L.L.

ACKNOWLEDGEMENTS

We thank Noah Reid for sharing 25 multilocus sequence data and the xml *BEAST input files and Bryan Carstens for the valuable comments. We thank Georgia Advanced Computing Research Center (GACRC) for their computing resource.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Abadi, S., D. Azouri, T. Pupko and I. Mayrose, 2019 Model selection may not be a mandatory step for phylogeny reconstruction. Nature Communications 10: 934.

Aitken, A. L., D. J. Clarke, D. D. McKenna, S. Shin, S. Haddad *et al.*, 2017 Phylogenomic Data Yield New and Robust Insights into the Phylogeny and Evolution of Weevils. Molecular Biology and Evolution 35: 823-836.

Arcila, D., G. Orti, R. Vari, J. W. Armbruster, M. L. J. Stiassny *et al.*, 2017 Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat Ecol Evol 1: 20.

Bastide, P., C. Solis-Lemus, R. Kriebel, K. William Sparks and C. Ane, 2018 Phylogenetic Comparative Methods on Phylogenetic Networks with Reticulations. Syst Biol 67: 800-820.

Blaimer, B. B., J. R. Mawdsley and S. G. Brady, 2018a Multiple origins of sexual dichromatism and aposematism within large carpenter bees. Evolution 72: 1874-1889.

Blaimer, B. B., P. S. Ward, T. R. Schultz, B. L. Fisher and S. G. Brady, 2018b Paleotropical Diversification Dominates the Evolution of the Hyperdiverse Ant Tribe Crematogastrini (Hymenoptera: Formicidae). Insect Systematics and Diversity 2: 3-3.

Blom, M. P. K., J. G. Bragg, S. Potter and C. Moritz, 2017 Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards. Syst Biol 66: 352-366.

Bogarín, D., O. A. Pérez-Escobar, D. Groenenberg, S. D. Holland, A. P. Karremans *et al.*, 2018 Anchored hybrid enrichment generated nuclear, plastid and mitochondrial markers resolve the Lepanthes horrida (Orchidaceae: Pleurothallidinae) species complex. Molecular Phylogenetics and Evolution 129: 27-47.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

718 Bravo, G. A., A. Antonelli, C. D. Bacon, K. Bartoszek, M. P. K. Blom *et al.*, 2019 Embracing
719 heterogeneity: coalescing the Tree of Life and the future of phylogenomics. PeerJ 7:
720 e6399.
721 Brown, J. M., and R. C. Thomson, 2017 Bayes Factors Unmask Highly Variable Information
722 Content, Bias, and Extreme Influence in Phylogenomic Analyses. Syst Biol 66: 517-
723 530.
724 Brown, J. M., and R. C. Thomson, 2018 Evaluating Model Performance in Evolutionary
725 Biology. Annual Review of Ecology, Evolution, and Systematics, Vol 49 49: 95-114.
726 Burbrink, F. T., and M. Gehara, 2018 The Biogeography of Deep Time Phylogenetic
727 Reticulation. Syst Biol 67: 743-744.
728 Cook, J. A., B. S. McLean, K. C. Bell, J. M. Allen and K. M. Helgen, 2018 Impacts of Inference
729 Method and Data set Filtering on Phylogenomic Resolution in a Rapid Radiation of
730 Ground Squirrels (Xerinae: Marmotini).
731 Cui, R., M. Schumer, K. Kruesi, R. Walter, P. Andolfatto *et al.*, 2013 PHYLOGENOMICS
732 REVEALS EXTENSIVE RETICULATE EVOLUTION IN XIPHOPHORUS FISHES. Evolution 67:
733 2166-2179.
734 Edwards, S. V., 2009 Is a New and General Theory of Molecular Systematics Emerging?
735 Evolution 63: 1-19.
736 Edwards, S. V., L. Liu and D. K. Pearl, 2007 High-resolution species trees without
737 concatenation. Proceedings of the National Academy of Sciences of the United
738 States of America 104: 5936-5941.
739 Edwards, S. V., Z. X. Xi, A. Janke, B. C. Faircloth, J. E. McCormack *et al.*, 2016 Implementing
740 and testing the multispecies coalescent model: A valuable paradigm for
741 phylogenomics. Molecular Phylogenetics and Evolution 94: 447-462.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

742 Espeland, M., J. Breinholt, K. R. Willmott, A. D. Warren, R. Vila *et al.*, 2018 A Comprehensive
743 and Dated Phylogenomic Analysis of Butterflies. Current Biology 28: 770-778.e775.
744 Gatesy, J., and M. S. Springer, 2014 Phylogenetic analysis at deep timescales: Unreliable
745 gene trees, bypassed hidden support, and the coalescence/concatalescence
746 conundrum. Molecular Phylogenetics and Evolution 80: 231-266.
747 Giarla, T. C., and J. A. Esselstyn, 2015 The Challenges of Resolving a Rapid, Recent Radiation:
748 Empirical and Simulated Phylogenomics of Philippine Shrews. Systematic Biology 64:
749 727-740.
750 Goldman, N., 1993 Statistical Tests of Models of DNA Substitution. Journal of Molecular
751 Evolution 36: 182-198.
752 Heled, J., and A. J. Drummond, 2010 Bayesian inference of species trees from multilocus
753 data. Mol Biol Evol 27: 570-580.
754 Hey, J., 2010 Isolation with Migration Models for More Than Two Populations. Molecular
755 Biology and Evolution 27: 905-920.
756 Hosner, P., E. Braun and R. Kimball, 2016 *Rapid and recent diversification of curassows,*
757 *guans, and chachalacas (Galliformes: Cracidae) out of Mesoamerica: Phylogeny*
758 *inferred from mitochondrial, intron, and ultraconserved element sequences*.
759 Jackson, N. D., A. E. Morales, B. C. Carstens and B. C. O'Meara, 2017 PHRAPL:
760 Phylogeographic Inference Using Approximate Likelihoods. Systematic Biology 66:
761 1045-1053.
762 Jhwueng, D. C., 2013 Assessing the Goodness of Fit of Phylogenetic Comparative Methods: A
763 Meta-Analysis and Simulation Study. Plos One 8.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

764 Jhwueng, D. C., S. Huzurbazar, B. C. O'Meara and L. Liu, 2014 Investigating the performance
765 of AIC in selecting phylogenetic models. Statistical Applications in Genetics and
766 Molecular Biology 13: 459-475.
767 Kass, R. E., and A. E. Raftery, 1995 Bayes Factors. Journal of the American Statistical
768 Association 90: 773-795.
769 Kubatko, L. S., 2009 Identifying Hybridization Events in the Presence of Coalescence via
770 Model Selection. Systematic Biology 58: 478-488.
771 Kubatko, L. S., and J. H. Degnan, 2007 Inconsistency of phylogenetic estimates from
772 concatenated data under coalescence. Systematic Biology 56: 17-24.
773 Lax, G., Y. Eglit, L. Eme, E. M. Bertrand, A. J. Roger *et al.*, 2018 Hemimastigophora is a novel
774 supra-kingdom-level lineage of eukaryotes. Nature 564: 410-414.
775 Leache, A. D., 2009 Species tree discordance traces to phylogeographic clade boundaries in
776 North American fence lizards (Sceloporus). Syst Biol 58: 547-559.
777 Lewis, P. O., W. G. Xie, M. H. Chen, Y. Fan and L. Kuo, 2014 Posterior Predictive Bayesian
778 Phylogenetic Model Selection. Systematic Biology 63: 309-321.
779 Liu, L., 2008 BEST: Bayesian estimation of species trees under the coalescent model.
780 Bioinformatics 24: 2542-2543.
781 Liu, L., and D. K. Pearl, 2007 Species trees from gene trees: Reconstructing Bayesian
782 posterior distributions of a species phylogeny using estimated gene tree
783 distributions. Systematic Biology 56: 504-514.
784 Liu, L., S. Y. Wu and L. L. Yu, 2015 Coalescent methods for estimating species trees from
785 phylogenomic data. Journal of Systematics and Evolution 53: 380-390.
786 Liu, L., and L. Yu, 2010 Phybase: an R package for species tree analysis. Bioinformatics 26:
787 962-963.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

788 Liu, L., and L. Yu, 2011 Estimating Species Trees from Unrooted Gene Trees. Systematic
789 Biology 60: 661-667.
790 Liu, L., L. Yu and S. V. Edwards, 2010 A maximum pseudo-likelihood approach for estimating
791 species trees under the coalescent model. BMC Evol Biol 10: 302.
792 Liu, L., J. Zhang, F. E. Rheindt, F. Lei, Y. Qu *et al.*, 2017 Genomic evidence reveals a radiation
793 of placental mammals uninterrupted by the KPg boundary. Proc Natl Acad Sci U S A
794 114: E7282-E7290.
795 Locke, S. A., A. Van Dam, M. Caffara, H. A. Pinto, D. López-Hernández *et al.*, 2018 Validity of
796 the Diplostomoidea and Diplostomida (Digenea, Platyhelminthes) upheld in
797 phylogenomic analysis. International Journal for Parasitology 48: 1043-1059.
798 MacGuigan, D. J., and T. J. Near, 2018 Phylogenomic Signatures of Ancient Introgression in a
799 Rogue Lineage of Darters (Teleostei: Percidae).
800 Maddison, W. P., 1997 Gene trees in species trees. Systematic Biology 46: 523-536.
801 McGee, M. D., B. C. Faircloth, S. R. Borstein, J. Zheng, C. Darrin Hulsey *et al.*, 2016 Replicated
802 divergence in cichlid radiations mirrors a major vertebrate innovation. Proc Biol Sci
803 283.
804 McVay, J. V., and B. Carstens, 2013 Phylogenetic model choice: justifying a species tree or
805 concatenation analysis. J Phylogen Evolution Biol 1: 114.
806 Meiklejohn, K. A., B. C. Faircloth, T. C. Glenn, R. T. Kimball and E. L. Braun, 2016 Analysis of a
807 Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in
808 Some Multispecies Coalescent Methods. Syst Biol 65: 612-627.
809 Moret, B. M. E., L. Nakhleh, T. Warnow, C. R. Linder, A. Tholse *et al.*, 2004 Phylogenetic
810 networks: Modeling, reconstructibility, and accuracy. Ieee-Acm Transactions on
811 Computational Biology and Bioinformatics 1: 13-23.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

812 O'Hara, T. D., A. F. Hugall, P. A. Cisternas, E. Boissin, G. Bribiesca-Contreras *et al.*, 2019
813 Phylogenomics, life history and morphological evolution of ophiocomid brittlestars.
814 Molecular Phylogenetics and Evolution 130: 67-80.
815 Page, R. D. M., 1998 GeneTree: comparing gene and species phylogenies using reconciled
816 trees. Bioinformatics 14: 819-820.
817 Paradis, E., and K. Schliep, 2019 ape 5.0: an environment for modern phylogenetics and
818 evolutionary analyses in R. Bioinformatics 35: 526-528.
819 Prum, R. O., J. S. Berv, A. Dornburg, D. J. Field, J. P. Townsend *et al.*, 2015 A comprehensive
820 phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature
821 526: 569.
822 Rambaut, A., and N. C. Grassly, 1997 Seq-Gen: an application for the Monte Carlo simulation
823 of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13: 235-
824 238.
825 Rannala, B., and Z. H. Yang, 2003 Bayes estimation of species divergence times and ancestral
826 population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.
827 Rannala, B., and Z. H. Yang, 2008 Phylogenetic inference using whole Genomes. Annual
828 Review of Genomics and Human Genetics 9: 217-231.
829 Reeves, J. H., 1992 Heterogeneity in the Substitution Process of Amino-Acid Sites of Proteins
830 Coded for by Mitochondrial-DNA. Journal of Molecular Evolution 35: 17-31.
831 Reid, N. M., S. M. Hird, J. M. Brown, T. A. Pelletier, J. D. McVay *et al.*, 2014 Poor Fit to the
832 Multispecies Coalescent is Widely Detectable in Empirical Data. Systematic Biology
833 63: 322-333.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

834 Richards, E. J., J. M. Brown, A. J. Barley, R. A. Chong and R. C. Thomson, 2018 Variation
835 Across Mitochondrial Gene Trees Provides Evidence for Systematic Error: How Much
836 Gene Tree Variation Is Biological? Syst Biol 67: 847-860.
837 Robinson, D. F., and L. R. Foulds, 1981 Comparison of phylogenetic trees. Math Biol 53: 131-
838 147.
839 Roch, S., and M. Steel, 2015 Likelihood-based tree reconstruction on a concatenation of
840 aligned sequence data sets can be statistically inconsistent. Theoretical Population
841 Biology 100: 56-62.
842 Romiguier, J., V. Ranwez, F. Delsuc, N. Galtier and E. J. Douzery, 2013 Less is more in
843 mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the
844 root of placental mammals. Mol Biol Evol 30: 2134-2144.
845 Sann, M., O. Niehuis, R. S. Peters, C. Mayer, A. Kozlov *et al.*, 2018 Phylogenomic analysis of
846 Apoidea sheds new light on the sister group of bees. BMC Evolutionary Biology 18:
847 71.
848 Scornavacca, C., and N. Galtier, 2017 Incomplete Lineage Sorting in Mammalian
849 Phylogenomics. Syst Biol 66: 112-120.
850 Shimodaira, H., 2002 An approximately unbiased test of phylogenetic tree selection.
851 Systematic Biology 51: 492-508.
852 Shimodaira, H., and M. Hasegawa, 1999 Multiple comparisons of log-likelihoods with
853 applications to phylogenetic inference. Molecular Biology and Evolution 16: 1114-
854 1116.
855 Simion, P., H. Philippe, D. Baurain, M. Jager, D. J. Richter *et al.*, 2017 A Large and Consistent
856 Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals.
857 Current Biology 27: 958-967.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

858 Springer, M. S., and J. Gatesy, 2016 The gene tree delusion. Molecular Phylogenetics and
859 Evolution 94: 1-33.
860 Stamatakis, A., 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of
861 large phylogenies. Bioinformatics 30: 1312-1313.
862 Straube, N., C. Li, M. Mertzen, H. Yuan and T. J. B. E. B. Moritz, 2018 A phylogenomic
863 approach to reconstruct interrelationships of main clupeocephalan lineages with a
864 critical discussion of morphological apomorphies. 18: 158.
865 Suchard, M. A., P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond *et al.*, 2018 Bayesian
866 phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4:
867 vey016.
868 Tavare, S., 1986 Some probabilistic and statistical problems in the analysis of DNA
869 sequences. Lectures on Mathematics in the Life Sciences 17: 57-86.
870 Waddell, P. J., 2005 Measuring the fit of sequence data to phylogenetic model: Allowing for
871 missing data (vol 22, pg 395, 2005). Molecular Biology and Evolution 22: 1157-1157.
872 Waddell, P. J., R. Ota and D. Penny, 2009 Measuring fit of sequence data to phylogenetic
873 model: gain of power using marginal tests. Journal of Molecular Evolution 69: 289-
874 299.
875 Walstrom, V. W., J. Klicka and G. M. Spellman, 2012 Speciation in the White-breasted
876 Nuthatch (Sitta carolinensis): a multilocus perspective. Mol Ecol 21: 907-920.
877 Wang, Y., Y. Zhou, L. F. Li, X. Chen, Y. T. Liu *et al.*, 2014 A new method for modeling
878 coalescent processes with recombination. Bmc Bioinformatics 15.
879 Wen, D., Y. Yu and L. Nakhleh, 2016 Bayesian Inference of Reticulate Phylogenies under the
880 Multispecies Network Coalescent. PLoS Genet 12: e1006006.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Whelan, N. V., K. M. Kocot, L. L. Moroz and K. M. Halanych, 2015 Error, signal, and the placement of Ctenophora sister to all other animals. Proceedings of the National Academy of Sciences of the United States of America 112: 5773-5778.

Wu, S., S. Edwards and L. Liu, 2018 Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data in Brief 18: 1972-1975.

Yang, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39: 306-314.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

906 TABLE 1: Summary of 22 phylogenomic data sets analyzed here. In the columns OTUs and
907 Loci, the bottom number is the count of taxa (or loci) in the original data set; the top number
908 is the count of taxa (or loci) after data reduction.

Data set	OTUs	Loci	Markers
Ant (BLAIMER *et al.* 2018b)	10/153	1509/1763	UCE
Bee (SANN *et al.* 2018)	10/95	193/195	CDS
Bird (PRUM *et al.* 2015)	10/198	259/259	CDS
Lizard (BLOM *et al.* 2017)	10/29	1831/1852	Exon
Brittlestar (O'HARA *et al.* 2019)	10/46	407/416	CDS
Butterfly (ESPELAND *et al.* 2018)	10/207	325/352	Exon
CarpenterBee (BLAIMER *et al.* 2018a)	10/179	597/753	UCE
Cichlid (MCGEE *et al.* 2016)	10/50	924/1043	UCE
Clupeocephalan (STRAUBE *et al.* 2018)	10/52	48/829	Exon
Cracids (HOSNER *et al.* 2016)	10/23	430/430	UCE
Darter (MACGUIGAN AND NEAR 2018)	10/112	363/30636	Exon
Diplostomoidea (LOCKE *et al.* 2018)	10/12	324/517	UCE
Gallopheasant (MEIKLEJOHN *et al.* 2016)	10/18	1479/1479	UCE
Hemimastigophora (LAX *et al.* 2018)	10/61	239/351	CDS
Lepanthe (BOGARÍN *et al.* 2018)	10/33	334/433	CDS
Mammal2018 (WU *et al.* 2018)	10/90	4709/5162	CDS
Mammal2017 (SCORNAVACCA AND GALTIER 2017)	10/97	108/110	CDS
Metazoan2015 (WHELAN *et al.* 2015)	10/76	36/115	CDS
Shrew (GIARLA AND ESSELSTYN 2015)	10/19	966/1112	UCE
Squirrel (COOK *et al.* 2018)	10/74	3209/3951	UCE
Weevil (AITKEN *et al.* 2017)	10/67	318/521	CDS
Fish (CUI *et al.* 2013)	10/27	1183/1183	CDS

909

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 2: Bayesian model validation for the 25 data sets in Reid et al. (2014). The validation analysis involves two tests - starbeastPPS and the independent test. Significance symbols *<0.05, **<0.025, ***<0.01, ****<0.001, and ns denotes nonsignificant. The last column GTRGAMMA indicates that the Bayesian model validation analyses assuming HKY and TN93 + GAMMA were reconducted with the GTRGAMMA model.

Data set	starbeastPPS	Independent test	GTRGAMMA	
			starbeastPPS	Independent test
Aliatypus	***	***	ns	ns
Certhiidae	**	***	****	***
Cheirogaleidae	ns	***	ns	ns
Liolaemus	ns	***	ns	ns
Malurus	ns	***	ns	ns
Sceloporus	ns	***	ns	ns
Sitta	ns	***	ns	ns
Tamias	****	ns	****	ns

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure legends

Figure 1: The flowchart of phylogenetic analyses in this study. The alignments of 47 data sets were used to test the hypothesis of topologically congruent (TC) gene trees and goodness of fit of substitution models. The model validation and comparison analyses were only performed for the loci that fit the GTRGAMMA model, reducing the number of input data for model validation to 39 (25 Reid et al. data sets and 14 phylogenomic data sets). Because the model comparison analysis implemented in BEAST does not allow missing taxa in any locus, 8 Reid et al. data sets were removed, further reducing the number of input data for the model comparison analysis to 31 (17 Reid et al. data sets and 14 phylogenomic data sets).

Figure 2: Goodness of fit of substitution models for 47 data sets. a) The violin boxplot of the proportion of loci rejecting the GTRGAMMA model in 47 data sets by the base frequency test, doublet test, bootstrap test, and the intersection of three tests (i.e., the intersection test), respectively. b) The violin boxplot of the proportion of loci rejecting the GTRGAMMA model for the 25 Reid et al. data sets and the 22 phylogenomic data sets. A two-sample t-test shows no significant difference between two sets of data with p-value = 0.96. c) Boxplots of GC content of loci rejecting or accepting the GTRGAMMA model. d) Boxplots of informative sites of loci rejecting or accepting the GTRGAMMA model. e) Logistic regression of whether a locus rejects the GTRGAMMA model (response variable) versus GC content and informative sites (two explanatory variables). The y-axis is the count of data sets with a negative or positive coefficient of GC content (or number of informative sites) estimated in the logistic regression. The bars in dark grey denote the count of data sets for which the coefficient (negative or positive) of GC (or informative sites) is significant (p-value < 0.05) in the logistic regression.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3: LRT for topologically congruent gene trees. a) A scatter plot of the p-value versus the number of loci rejecting the hypothesis of TC gene trees across 47 data sets. b) Boxplots of the proportion of loci rejecting the hypothesis of TC gene trees for 25 Reid et al. data sets and 22 phylogenomic data sets. A two-sample t-test shows no significant difference (p-value = 0.10) in the proportion of loci rejecting the hypothesis of TC gene trees between the 25 Reid et al. data sets and the 22 phylogenomic data sets. c) Proportion of loci rejecting the hypothesis of TC gene trees across 47 data sets. d) Proportion of loci rejecting the hypothesis of TC gene trees in 6 groups. The 47 data sets fall into 6 groups (bird, mammal, insect, fish, reptile, and others). The filled black diamond represents the average proportion of loci rejecting the hypothesis of TC gene trees in each group. ANOVA finds no significant difference (pvalue = 0.71) in the proportion of loci rejecting the hypothesis of TC gene trees among 6 groups. e) A linear regression line fitted for the log of the number of loci rejecting the hypothesis of TC gene trees (y) versus the log of the number of loci (x). The regression line is $log\ (y) = 0.87 \times log\ (x) - 0.08$, where the coefficient 0.87 is significantly greater than 0 with pvalue < 0.01.

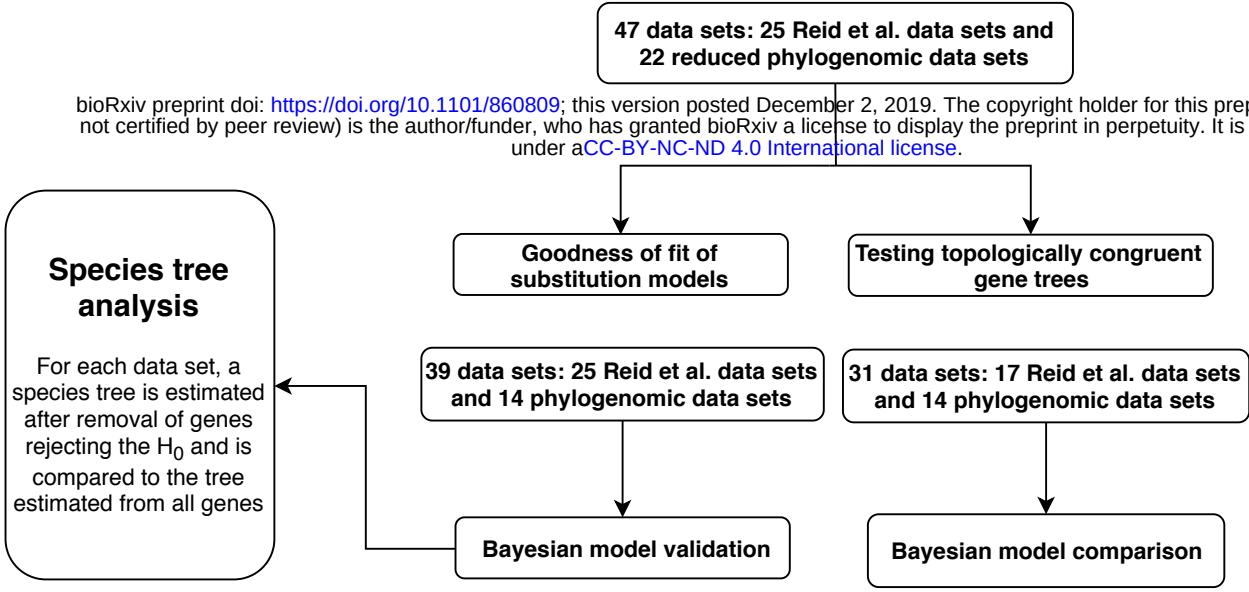
Figure 4: Bayesian model validation. The MSC model is validated by two tests. The first test proposed in Reid et al. (2014) compares the simulated and posterior coalescent gene tree. The second test compares the gene trees estimated with the coalescent prior and those estimated with independent prior. The MSC model is a poor fit to the data if either or both of the two tests are rejected. a) The proportion of data sets rejecting the MSC model. 32% of 25 Reid et al. data sets and 100% of 14 phylogenomic data sets reject the MSC model. b) The tests were applied to each locus to find the proportion of loci rejecting the MSC model. The proportion (10.88%) of loci rejecting the MSC for 22 phylogenomic data sets is significantly (p-value < 0.01) higher than the proportion (4.17%) of loci in 25 Reid et al. data sets that reject the MSC

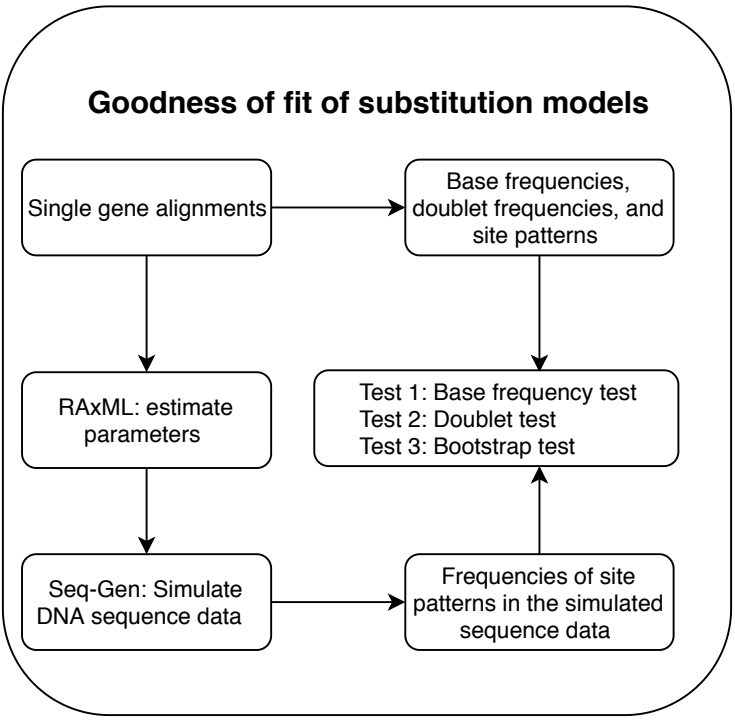
bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

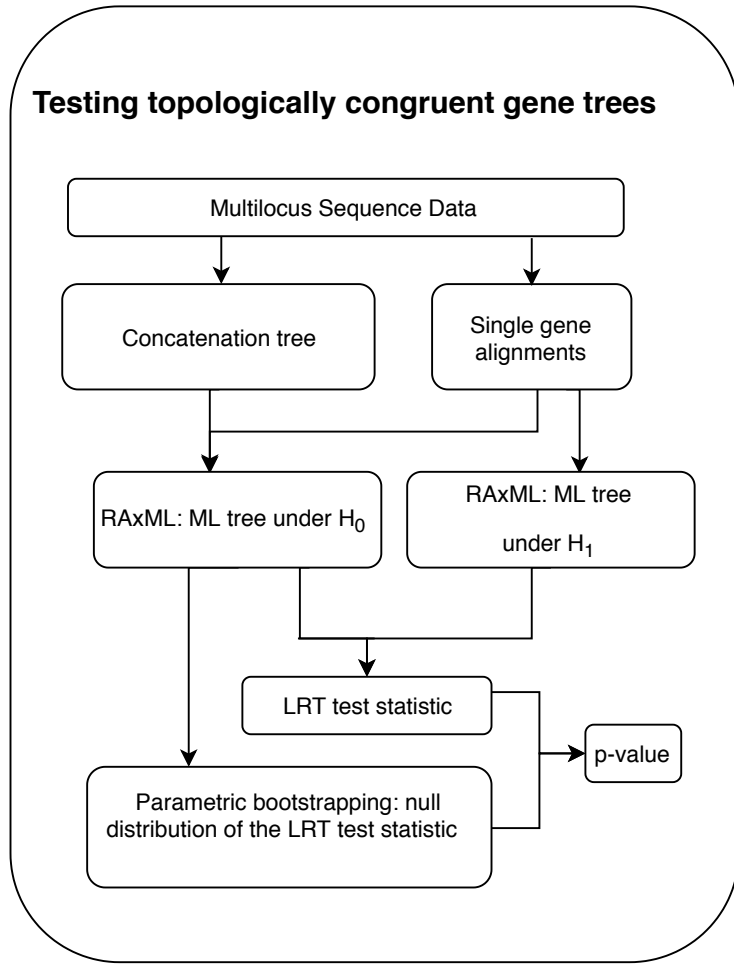
model. c) A scatter plot of the RF distance versus the proportion of loci rejecting the MSC model. The RF distance between two species trees built from all loci and only loci that fit to the MSC is calculated for 14 phylogenomic data sets; RF = 0 for 9 data sets, RF = 2 for 4, RF = 4 for 1 data set. The number below each point is the bootstrap support value on the corresponding incongruent branch in two species trees. Note that two points with RF = 2 are overlapping, and two values below the point with RF = 4 (i.e., two incongruent branches) are the bootstrap support values on two incongruent branches. d) Testing the association between GC content and poor fit of the MSC. We calculated the average GC content of loci rejecting or accepting the MSC for two phylogenomic data sets with large numbers of loci rejecting the MSC (Birds and mammal2017). We performed a two-sample t-test for the overall average GC content (combining two data sets) and the average GC content of each of two data sets, respectively. All three tests found no significant difference (p-value > 0.3) in GC content between loci rejecting the MSC and loci accepting the MSC.

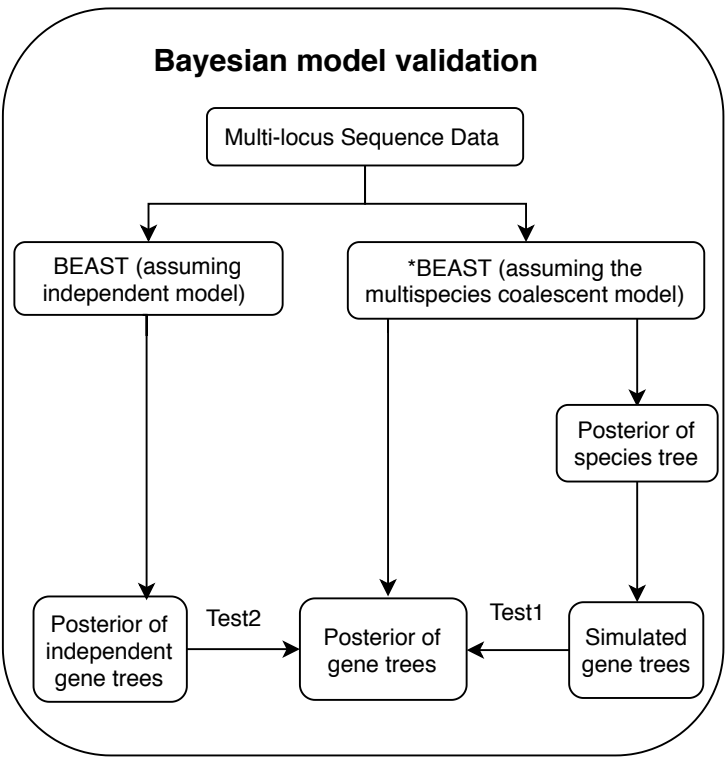
Figure 5: Bayesian model comparison. The log-scale Bayes factors of coalescent vs concatenation (unlinking substitution model parameters) for 31 data sets, including 17 Reid et al. data sets and 14 phylogenomic data sets.

bioRxiv preprint doi: https://doi.org/10.1101/860809; this version posted December 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.









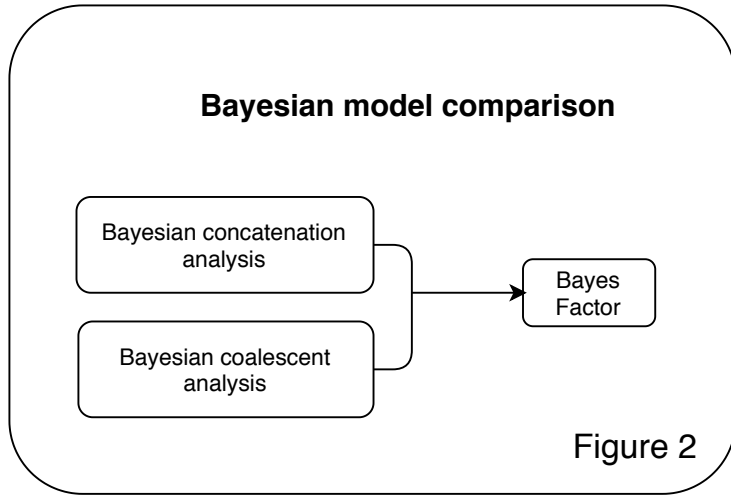
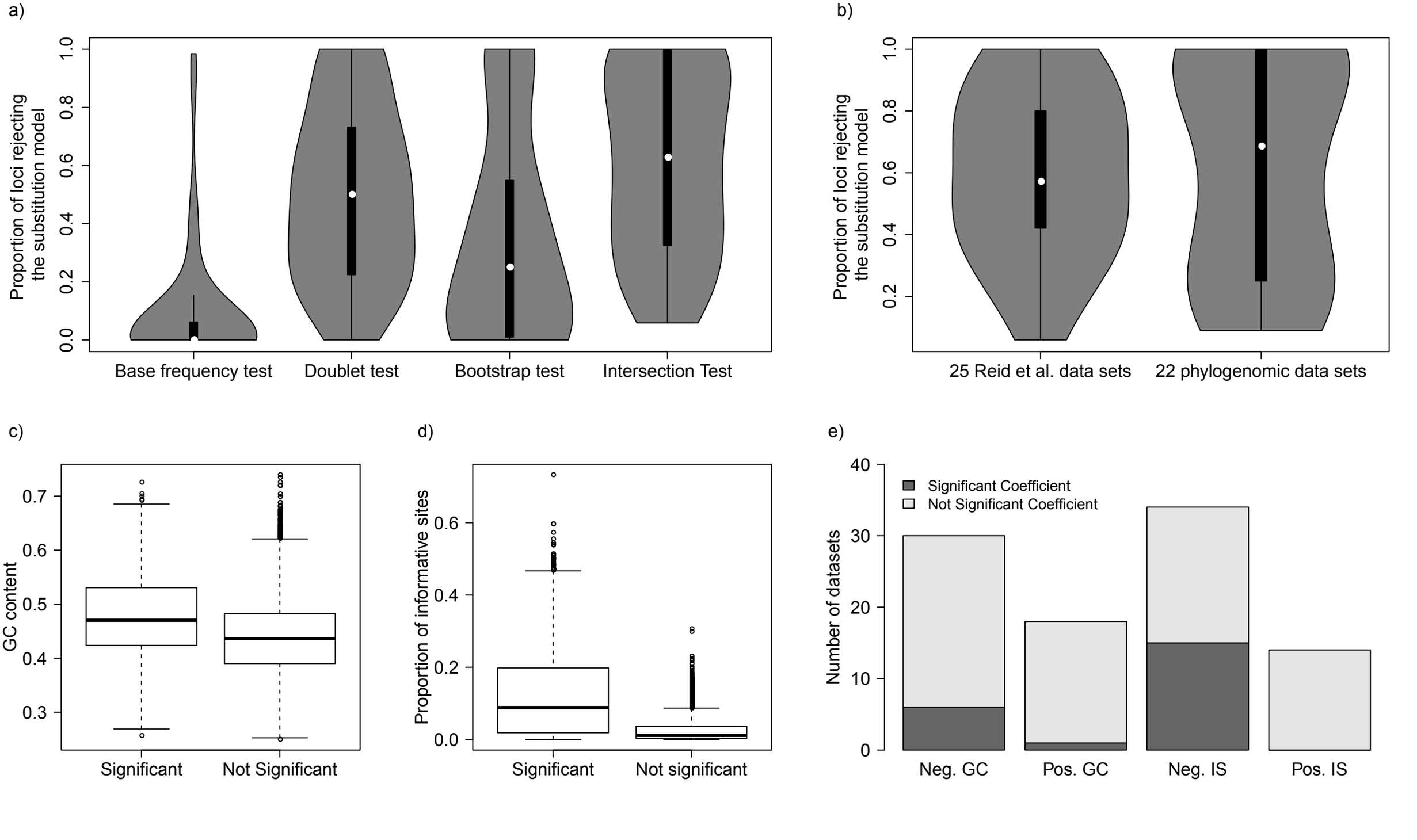
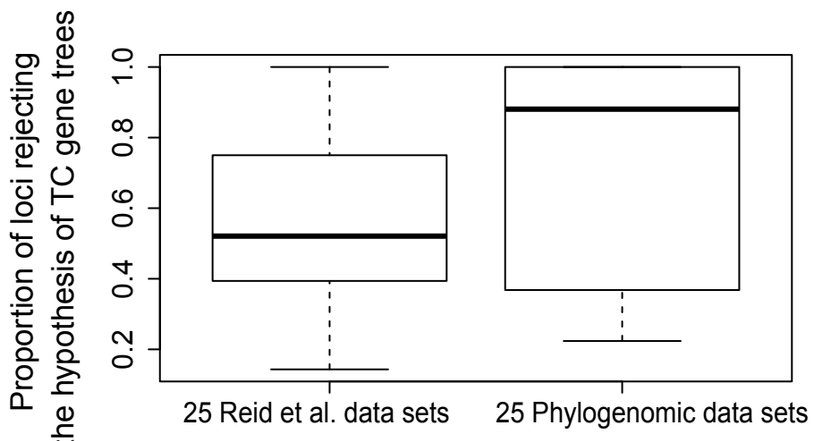


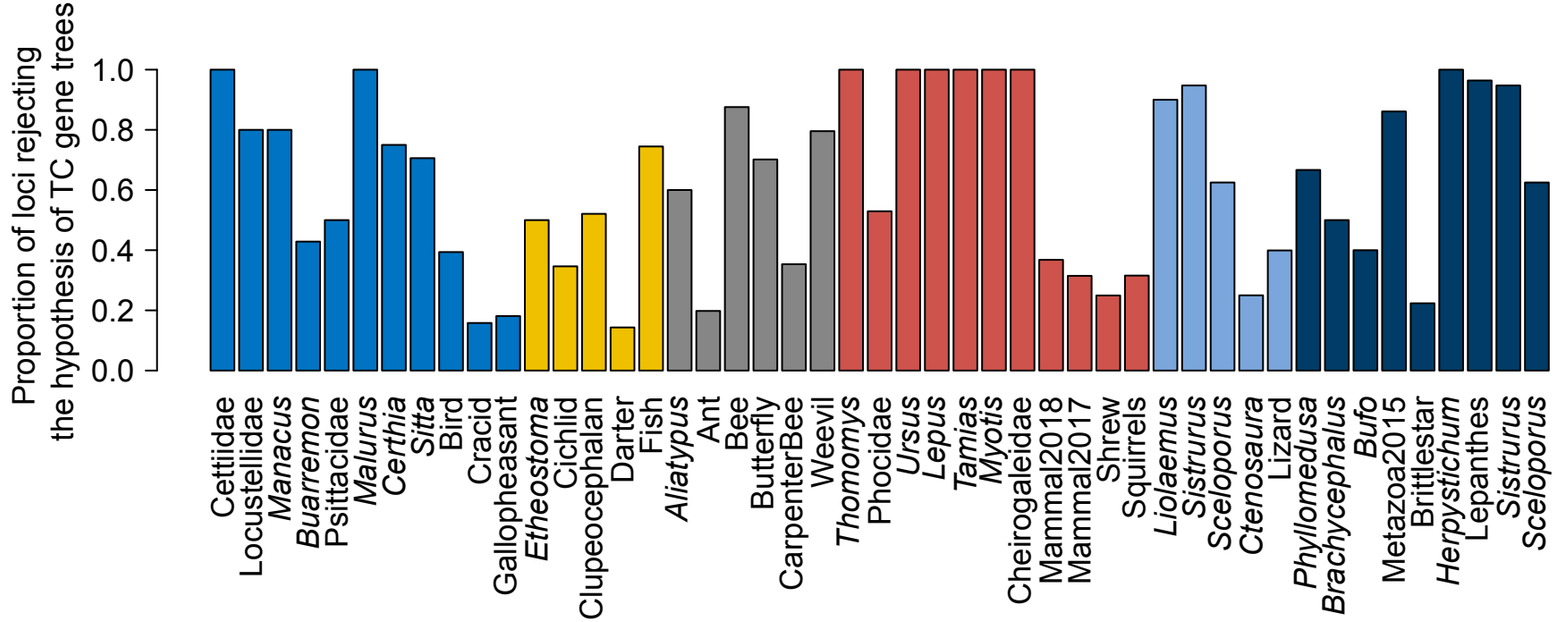
Figure 2


a)
Proportion of loci rejecting the substitution model
0.0
0.2
0.4
0.6
0.8
1.0
Base frequency test
Doublet test
Bootstrap test
Intersection Test
b)
Proportion of loci rejecting the substitution model
0.2
0.4
0.6
0.8
1.0
25 Reid et al. data sets
22 phylogenomic data sets
c)
GC content
0.3
0.4
0.5
0.6
0.7
Significant
Not Significant
d)
Proportion of informative sites
0.0
0.2
0.4
0.6
Significant
Not significant
e)
Number of datasets
0
10
20
30
40
Significant Coefficient
Not Significant Coefficient
Neg. GC
Pos. GC
Neg. IS
Pos. IS

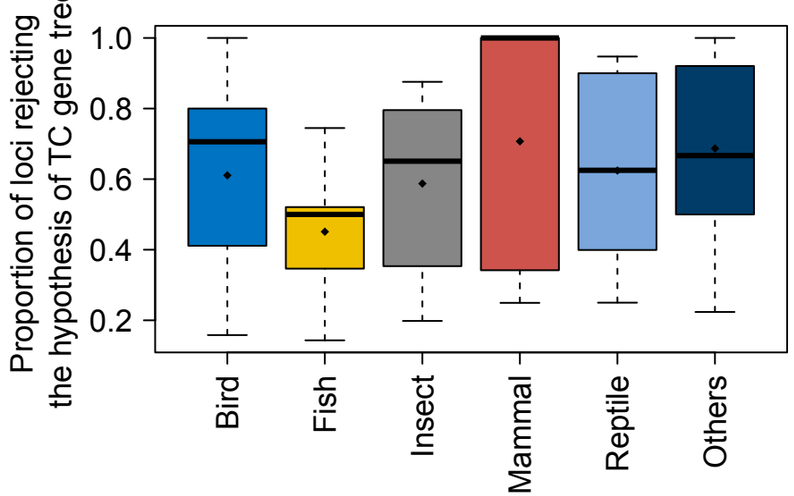
a)


Proportion of loci rejecting
the hypothesis of TC gene trees
0.2
0.4
0.6
0.8
1.0
25 Reid et al. data sets
25 Phylogenomic data sets

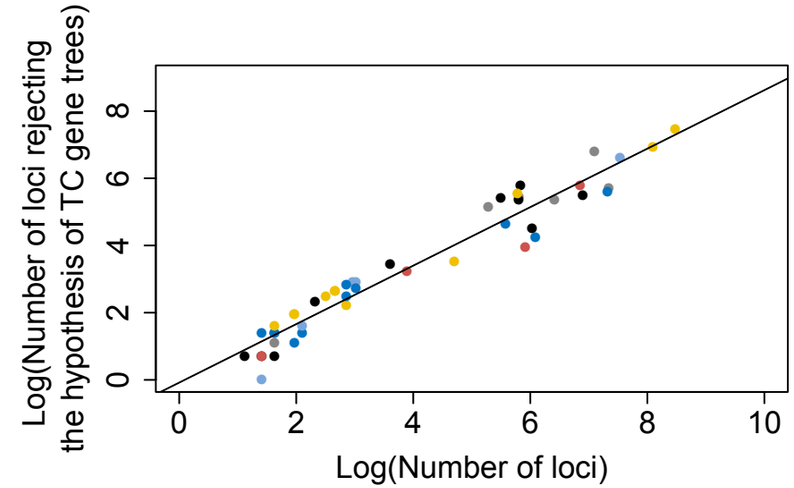
b)


Proportion of loci rejecting
the hypothesis of TC gene trees
0.0
0.2
0.4
0.6
0.8
1.0
Cettiidae
Locustellidae
Manacus
Buarremon
Psittacidae
Malurus
Certhia
Sitta
Bird
Cracid
Gallopheasant
Etheostoma
Cichlid
Clupeocephalan
Darter
Fish
Aliatypus
Ant
Bee
Butterfly
CarpenterBee
Weevil
Thomomys
Phocidae
Ursus
Lepus
Tamias
Myotis
Cheirogaleidae
Mammal2018
Mammal2017
Shrew
Squirrels
Liolaemus
Sistrurus
Sceloporus
Ctenosaura
Lizard
Phyllomedusa
Brachycephalus
Bufo
Metazoa2015
Brittlestar
Herpystichum
Lepanthes
Sistrurus
Sceloporus

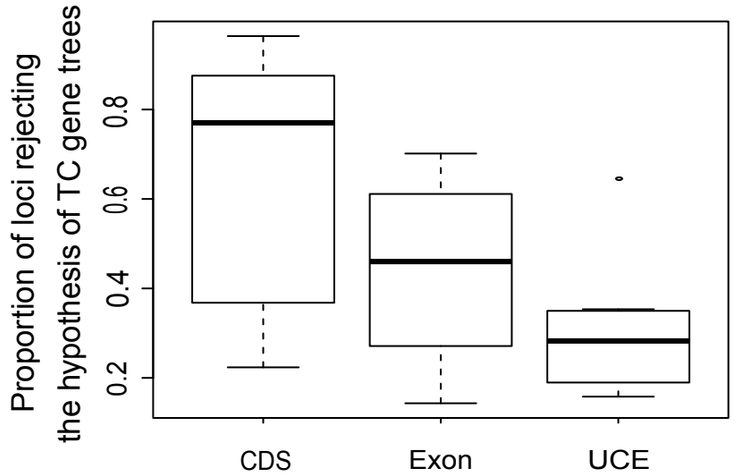
c)


Proportion of loci rejecting
the hypothesis of TC gene trees
0.2
0.4
0.6
0.8
1.0
Bird
Fish
Insect
Mammal
Reptile
Others

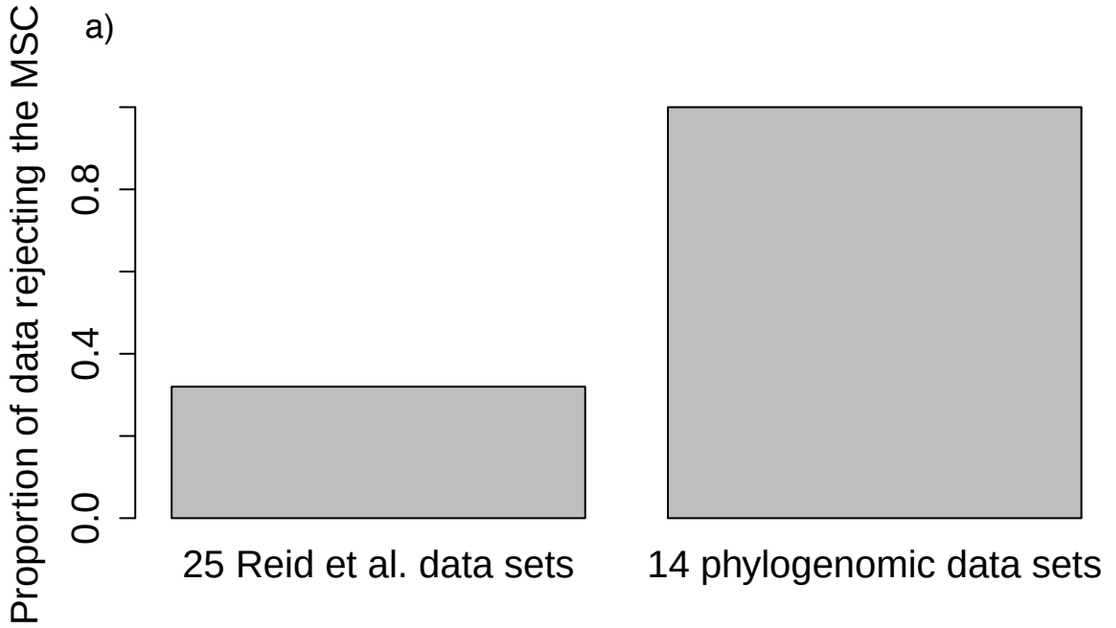
d)

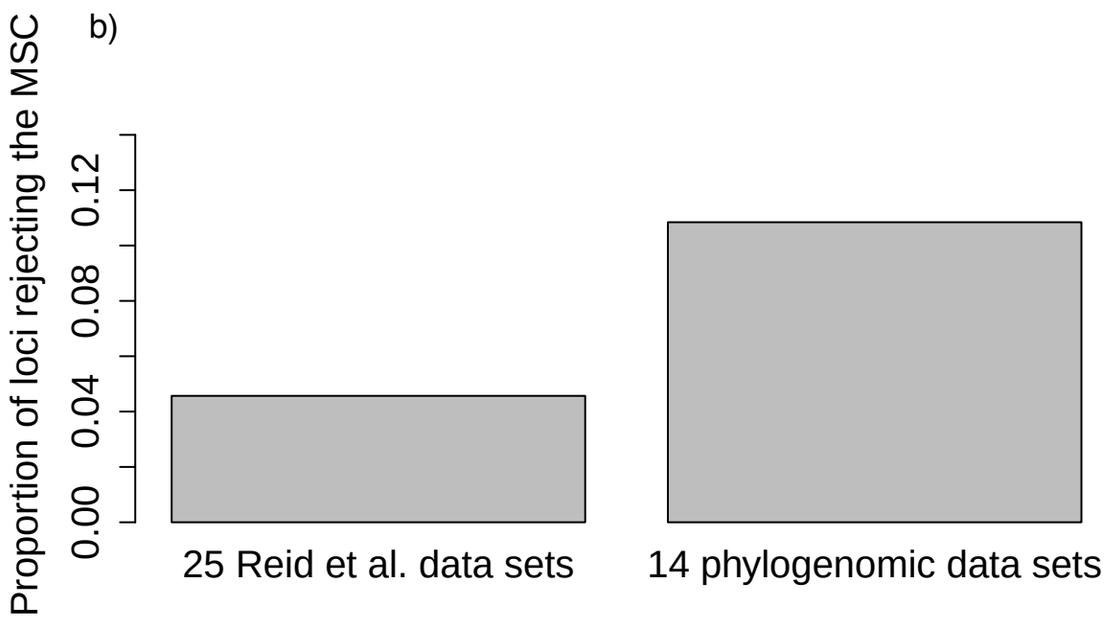

Log(Number of loci rejecting
the hypothesis of TC gene trees)
0
2
4
6
8
0
2
4
6
8
10
Log(Number of loci)

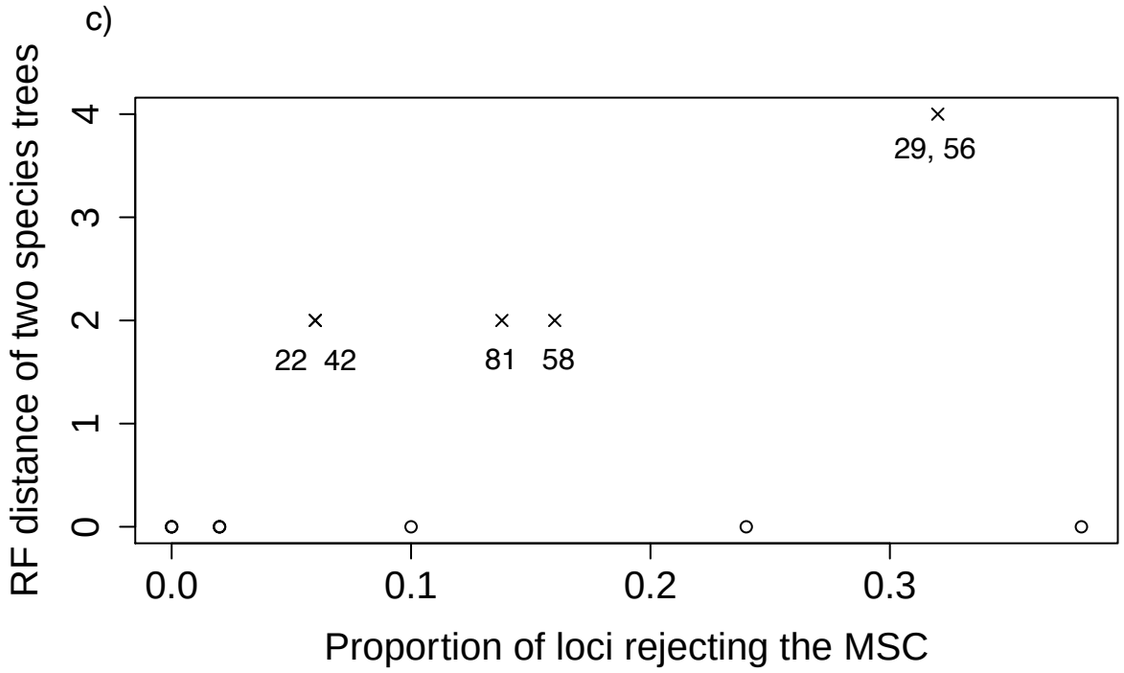
e)


Proportion of loci rejecting
the hypothesis of TC gene trees
0.2
0.4
0.6
0.8
CDS
Exon
UCE

Bird
Fish
Insect
Mammal
Reptile
Others


a)
Proportion of data rejecting the MSC
0.0
0.4
0.8
25 Reid et al. data sets
14 phylogenomic data sets


b)
Proportion of loci rejecting the MSC
0.00
0.04
0.08
0.12
25 Reid et al. data sets
14 phylogenomic data sets


c)
RF distance of two species trees
0
1
2
3
4
29, 56
22
42
81
58
0.0
0.1
0.2
0.3
Proportion of loci rejecting the MSC

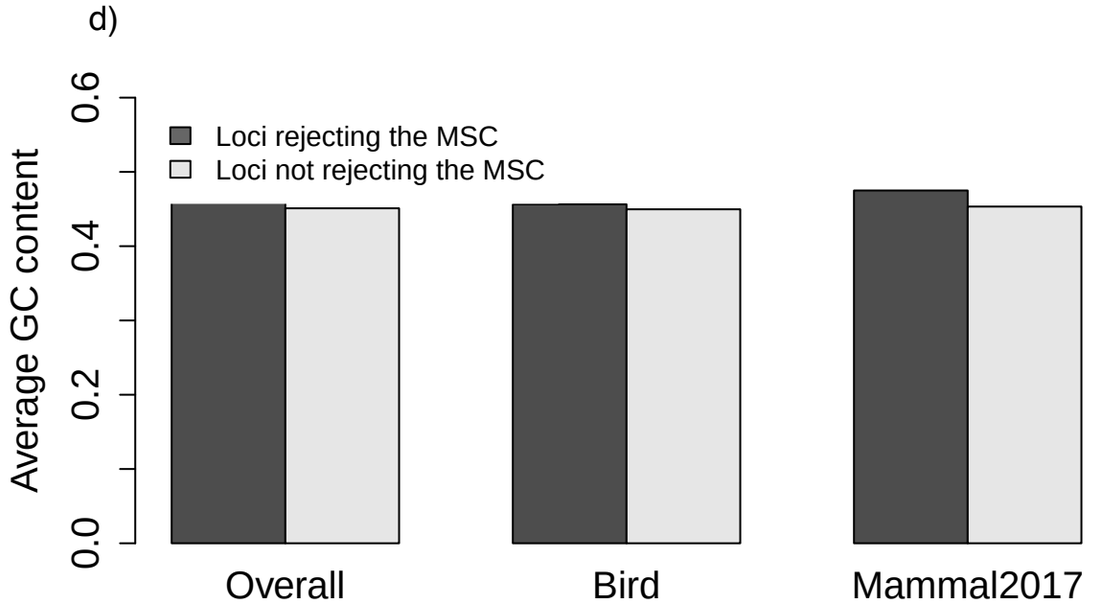

d)
Average GC content
0.0
0.2
0.4
0.6
Loci rejecting the MSC
Loci not rejecting the MSC
Overall
Bird
Mammal2017

Figure 5

Log(Bayes Factor)
4000
3000
2000
1000
0
17 Reid et al. data sets
14 phylogenomic data sets
Brachycephalus
Buarremon
Certhia
Cettiidae
Cheirogaleidae
Ctenosaura
Herpystichum
Locustellidae
Malurus
Manacus
Myotis
Psittacidae
Sceloporus
Sistrurus
Tamiae
Thomomys
Ursus
Ant
Bird
Lizard
BrittleStar
CarpenterBee
Cichlid
Cracid
Darter
Gallopheasant
Lepanthes
Mammal2017
Mammal2018
Shrew
Squirrel