OXFORD UNIVERSITY PRESS Systematic Biology

Poor Fit to the Multispecies Coalescent is Widely Detectable in Empirical Data

Journal:	*Systematic Biology*
Manuscript ID:	Draft
Manuscript Type:	Regular Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Reid, Noah; Louisiana State University, Biological Sciences Hird, Sarah; Louisiana State University, Biological Sciences Brown, Jeremy; Louisiana State University, Biological Sciences McVay, John; Louisiana State University, Biological Sciences Pelletier, Tara; The Ohio State University, Evolution Ecology and Organismal Biology Satler, Jordan; The Ohio State University, Evolution Ecology and Organismal Biology Carstens, Bryan; The Ohio State University, Evolution Ecology and Organismal Biology
Keywords:	model fit, posterior predictive simulation, multispecies coalescent, next-generation sequencing, species tree, gene tree, hybridization, gene duplication and extinction, species delimitation

RUNNING HEAD: POOR FIT OF EMPIRICAL DATA TO MULTISPECIES COALESCENT

Poor Fit to the Multispecies Coalescent is Widely Detectable in Empirical Data

NOAH M. REID[1*], SARAH M HIRD[1], JEREMY M BROWN[1], TARA A PELLETIER[2], JOHN D MCVAY[1], JORDAN D SATLER[2] AND BRYAN C CARSTENS[2]

[1] *Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, 70803, USA*

[2] *Department of Evolution, Ecology & Organismal Biology, Ohio State University, Columbus, Ohio, 43210, USA*

*Corresponding author, address: 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70803; phone: 225.578.4918; email: nreid1@tigers.lsu.edu

ABSTRACT

Model checking is a critical part of Bayesian data analysis, yet it remains largely unused in systematic studies. Phylogeny estimation has recently moved into an era of increasingly complex models that simultaneously account for multiple evolutionary processes, the statistical fit of these models to the data has rarely been tested. Here we develop a posterior predictive simulation (PPS)-based model check for a commonly used multispecies coalescent model, implemented in *BEAST, and apply it to 25 published datasets. We show that poor model fit is detectable in the majority of datasets; that this poor fit can mislead phylogenetic estimation; and that in some cases it stems from processes of inherent interest to systematists. We suggest that as systematists scale up to phylogenomic datasets, which will be subject to a heterogeneous array of evolutionary processes, critically evaluating the fit of models to data is an analytical step that can no longer be ignored.

Keywords: model fit, posterior predictive simulation, multispecies coalescent, next-generation sequencing, species tree, gene tree, hybridization, gene duplication and extinction, species delimitation

The introduction of multispecies coalescent models to phylogenetic inference marked a fundamental advance in systematic biology (Degnan and Rosenberg 2009, Edwards 2009). These models treat populations, rather than alleles sampled from a single individual, as the focal units in phylogenetic trees. The multispecies coalescent model connects traditional phylogenetic inference, which seeks primarily to infer patterns of divergence between species, and population genetic inference, which has typically focused on intraspecific evolutionary processes. The development of these models was motivated by the common empirical observation that genealogies estimated from different genes are often discordant (e.g. Rokas et al. (2003), Jennings and Edwards (2005)) and the discovery that, if ignored, this discordance can bias parameters of direct interest to systematists, such as the relationships and divergence times among species (Degnan and Rosenberg 2006, Kubatko and Degnan 2007, McCormack et al. 2011).

In order to reconcile discordance among gene trees and uncover true species relationships, the first gene tree/species tree models assumed that discordance is solely the result of stochastic coalescence of gene lineages within a species phylogeny (Rannala and Yang 2003, Edwards et al. 2007, Kubatko et al. 2009, Heled and Drummond 2010). These approaches estimate topology, divergence times and effective population sizes (except Kubatko et al. (2009)) of the species tree using a model where the probability of a gene tree being discordant with a species tree increases with the ratio of effective population size along a branch to the length of the branch (Takahata 1989, Rosenberg 2002). When their assumptions are met, these models are consistent (Liu and Edwards 2009) across a wide range of divergence histories (Hird et al. 2010, Leache and Rannala

2011). However, the number of independently segregating loci needed to accurately infer the species tree increases with the above ratio (Hird et al. 2010, Huang et al. 2010).

Coalescent stochasticity, however, is not the only source of gene tree discordance (Maddison 1997). Selection, hybridization, horizontal gene transfer, gene duplication/extinction, recombination and phylogenetic estimation error can also result in discordance. Maddison (1997) described the product of these disparate genealogical processes as a "cloud" of gene trees. Given that these processes are common (Zhang 2003, Mallet 2005, Charlesworth 2006), we can expect that they will be ubiquitous in new phylogenomic datasets. Unfortunately, beyond a few studies on recombination (Lanier and Knowles 2012), migration (Eckert and Carstens 2008) and horizontal gene transfer (Chung and Ane 2011), we have little idea of the extent to which these factors, unaccounted for, may bias the inference of topology and divergence times in species tree inference. Currently, no method can account for all of these factors. For example, some methods estimate species trees while accounting for gene duplication and extinction (Rasmussen and Kellis 2012), some incorporate gene flow (Gerard et al. 2011, Pickrell and Pritchard 2012), others conduct species delimitation (O'Meara 2010, Yang and Rannala 2010), and at least one models discordance without reference to a specific biological process (Ané et al. 2007). Other than a recent model restricted to a three-taxon case (Choi and Hey 2011), no model accounts for more than two factors, and none accounts for natural selection. Discordance can also be caused by methodological problems, such as errors in species delimitation and mis-specified models of DNA sequence evolution. While the potential for methodological error is well-established, very little work has been to quantitatively estimate its prevalence in empirical studies.

Given that all models must make some simplifying assumptions, and available models make assumptions that are known to be frequently violated, it is imperative to assess the statistical fit of the models to the data. Evidence of poor model fit should encourage researchers to treat phylogenetic estimates with caution and to explore important biological processes that they might not have previously considered. Model checking in this sense seeks to evaluate the absolute fit of models to the data, in order to determine whether any of the models under consideration sufficiently describe the data. This approach complements more widely applied methods of model selection that choose among a set of available models (Goldman 1993).

Posterior predictive simulation (PPS) is a commonly used method for model checking in a Bayesian framework (Gelman et al. 2009). Although the use of PPS has been advocated for phylogenetic inference (Huelsenbeck et al. 2001, Bollback 2002, Nielsen 2002, Nielsen and Bollback 2005, Brown and ElDabaje 2009), it has yet to be widely adopted. PPS has been used to show that some common macroevolutionary models are a poor fit to the true process of diversification (Rabosky et al. 2012); that two common population genetic models perform poorly in describing the history of the duck, *Anas strepsera* (Peters et al. 2012) and recommended for use in evaluating models of DNA sequence evolution in the inference of gene trees (Bollback 2002). However, aside from a recent paper suggesting its use in identifying instances of introgressive hybridization (Joly 2012) it has not been used to check the fit of multispecies coalescent models.

Here we develop a model checking method in the PPS framework to test the fit of a commonly used Bayesian multispecies coalescent model implemented in *BEAST (Drummond and Rambaut 2007, Heled and Drummond 2010) to 25 published datasets. We

then hypothesize about the sources of identified model misspecification and discuss the consequences for inferences based on the multispecies coalescent.

MATERIALS AND METHODS

*The multispecies coalescent model in *BEAST*

*BEAST implements a Bayesian hierarchical model to estimate a species tree with divergence times and effective population sizes from multilocus DNA sequence data. The model hierarchy has three levels (Fig. 1). The bottom level connects the data, a series of DNA sequence alignments for n independently segregating loci ($D = d_1, d_2, \ldots, d_n$) to their respective gene trees ($G = g_1, g_2, \ldots, g_n$) through the standard phylogenetic likelihood (Felsenstein 1981):

$$L(g_i) = P(d_i \mid g_i)$$

Gene tree likelihoods are conditioned on a chosen model of sequence evolution. Gene tree branch lengths are measured in substitutions per site, the product of mutation rate and time. The second level connects the gene trees (G) to ultrametric coalescent genealogies ($U = u_1, u_2, \ldots u_3$), whose branch lengths are proportional to time, through a molecular clock model:

$$L(u_i) = P(g_i \mid u_i)$$

Several molecular clock models, including relaxed clocks (Drummond et al. 2006) and a random local clock (Drummond and Suchard 2010) are available in *BEAST, each of which make differing assumptions about the distribution of mutation rates among branches in the gene trees (*BEAST incorporates the molecular clock model into the gene tree likelihood, but we depict it separately here for clarity). The third level connects the

ultrametric coalescent genealogies (U) with the species tree, including divergence times and effective population sizes (S), through the multispecies coalescent model:

$$L(S) = P(u_i \mid S_i)$$

The likelihood of the multispecies coalescent ($P(u_i|S)$) is calculated as the product, across all branches in the species tree, of the probabilities of the coalescent processes within each branch (Rannala and Yang 2003). The most general form of the model in *BEAST allows the population size on each branch to change linearly, with the constraint that the sum of the population sizes of daughter branches must always equal the population size of their parent (piecewise linear). The marginal posterior probability of the species tree given the data for the full model is then

$$P(S \mid D) \propto \prod_{i=1}^{n} \int_{g_i} \int_{u_i} P(d_i \mid g_i) P(g_i \mid u_i) P(u_i \mid S) P(S) du_i dg_i$$

$P(S)$ is the joint prior probability distribution on the species tree topology, branch lengths and effective population sizes. *BEAST estimates the posterior distribution of the model using a Markov chain Monte Carlo (MCMC) algorithm.

PPS approach to checking the fit of the multispecies coalescent

To conduct PPS, one first obtains the joint posterior distribution of parameters for a model (in this case, the *BEAST model, sampled via MCMC), draws sets of parameters from that joint distribution and uses them to simulate data (Fig. 1). The simulated datasets form a posterior predictive distribution representing reasonable outcomes of the model conditioned on the observed data. One can then compare the empirical data with the posterior predictive distribution using well-chosen test statistics. The ways in which the

empirical data do not match the predictive distribution can identify failures of a model to capture important biological processes.

Because *BEAST implements a hierarchical model, and we are most interested in whether the multispecies coalescent component of the model is an appropriate fit the data, our approach to using PPS isolates and checks two levels of the model independently: the multispecies coalescent and the phylogenetic likelihood. We do not try to isolate and check the fit of molecular clock models here.

We check the multispecies coalescent by comparing coalescent genealogies simulated from the posterior distribution of species trees with those estimated in the empirical analysis. We used two test quantities for the comparison, both of which directly assess the fit of the simulated and estimated coalescent genealogies to the estimated species tree: the multispecies coalescent likelihood (i.e., the probability of a coalescent genealogy given the species tree ($P(u_i|S)$), and the number of deep coalescences (Maddison 1997, Rannala and Yang 2003). We predicted that for the coalescent likelihood, poor fit would be reflected by individual loci with extremely low probabilities, a low product of probabilities across loci, or an unexpectedly high coefficient of variation of probabilities across loci. For the number of deep coalescences, we expected that poor fit would manifest itself either in individual loci with unexpectedly high or low numbers of deep coalescences, excessively high or low sums of deep coalescences across loci, or a high coefficient of variation across loci. Each of these values measures the degree of discrepancy between gene trees and species trees or across gene trees. In order to generate posterior predictive distributions with expectations of 0, we use test quantities (sensu Gelman et al. (2009)) that are conditioned on particular parameter values sampled from the posterior

distribution. To do so, we simulate one set of coalescent genealogies for each draw from the posterior (sampled from the MCMC), calculate test statistics for the coalescent genealogies from that draw as well as the coalescent genealogies simulated from the species tree in that draw, and take their difference. A 95% highest posterior predictive density interval that does not contain 0 indicates poor fit of the model to the data with respect to that test quantity. For clarity, we refer to all ultrametric gene genealogies estimated or simulated under the model as coalescent genealogies, even if there is evidence that non-coalescent processes influenced them.

Although our primary interest in this study is assessing the fit of the multispecies coalescent, there are two reasons to simultaneously assess the fit of the phylogenetic likelihood. First, poor fit of the multispecies coalescent could be due to poorly fitting models of sequence evolution that result in inaccurate estimates of coalescent genealogies. Second, we speculated that the prior distribution on gene trees induced by the coalescent model may put very low probability on gene tree topologies generated by processes other than stochastic coalescence. For such gene trees, this informative prior could result in estimates that fit the multispecies coalescent model well but strongly disagree with the underlying sequence data.

We check the fit of the phylogenetic likelihood by comparing DNA sequence data simulated from the estimated gene trees (G) with the empirical data. We use four test quantities: (1) the number of variable sites, the (2) multinomial and (3) phylogenetic likelihoods, and (4) the Goldman-Cox (GC) statistic, which is the difference between the multinomial and phylogenetic likelihoods. We expect the number of variable sites to be roughly related to total tree length. We predicted that in some cases of poor fit, the

empirical phylogenetic likelihood would be lower than expected based on the posterior predictive distribution of phylogenetic likelihoods since posterior predictive sequence datasets would not conflict with their corresponding gene trees. The multinomial likelihood is the product, across all sites in an alignment, of the frequency of their respective site patterns, and has been frequently used as a test statistic in phylogenetic PPS (Bollback 2002, Brown and ElDabaje 2009). The GC statistic has been used in assessing the fit of models of DNA sequence evolution in a likelihood framework (Goldman 1993, Ripplinger and Sullivan 2010). The multinomial and phylogenetic likelihoods are expected to converge with very large amounts of data, so when the difference between them is larger than expected, it is a sign that some part (sequence model or tree) of the evolutionary model is a poor fit to the data. It is also worth noting that the phylogenetic and multinomial likelihoods and the GC statistic are all strongly correlated with the number of variable sites in an alignment. Therefore, inaccurate estimates of tree length, even absent other reasons for poor fit, can lead to deviation in these statistics.

Empirical data

We obtained 25 datasets from Genbank or directly from the authors (Table 1). With a few exceptions, we avoided publications that dealt with hybridization directly, or those that excluded known introgressed loci. We also avoided publications whose express goal was species delimitation, because errors in species assignment are a clear violation of the multispecies coalescent. Each dataset was analyzed using the models of sequence evolution provided in the original manuscript with the exception of models requiring both a proportion of invariable sites and gamma-distributed rates across sites (RAS), because we

occasionally observed problems with convergence when using them together. In those cases we used only gamma-distributed RAS. We retained any intra-locus partitioning schemes, and for the phylogenetic likelihood model checks treated each locus subset individually. We ran each dataset twice for at least as long as in the original publication. To conduct PPS we excised the first 10% of MCMC steps as burn-in, combined both runs, and thinned them to ~2000 MCMC samples. All analyses were conducted using custom scripts in the statistical language R (R Development Core Team 2011), available on NMR's website (https://sites.google.com/site/noahmreid/) in tandem with ms (Hudson 2002), Seq-Gen (Rambaut and Grass 1997), ape (Paradis et al. 2004), and phangorn (Schliep 2011). We selected four datasets that fit the model poorly in some way and subjected them to further analysis. For two (*Tamias* and Cheirogaleidae) we removed single loci that showed poor fit, re-analyzed the rest of the data and compared the model estimates. For the other two (*Ursus* and *Sistrurus*) we eliminated the multispecies coalescent level of the model hierarchy, fit completely independent trees and clock models, and compared the gene tree estimates.

RESULTS

Datasets and analyses

We analyzed 25 datasets from papers spanning 12 orders of Eukaryotes (Table 1). The average number of operational taxonomic units (OTUs) was 13.7, the average number of alleles per dataset was 67.2, and the average number of independently segregating loci was 9.6. Seventeen datasets utilized organellar as well as nuclear DNA. For *BEAST analyses, nearly all datasets had ESS values of over 200 for all parameters. Markov chains

for a few datasets mixed poorly, but if parameter estimates from 2 independent runs were very similar after 200 million generations, we included them anyway. Results of our *BEAST analyses were consistent with published results for each dataset.

Overview of results

At the level of the coalescent genealogies, we found evidence of poor model fit in 4 datasets (16%) considering all test quantities. Seven total loci across data sets deviated from expectations (2.9%; Table 2). Two of those datasets showed poor fit at only one locus (*Certhia* and Cheirogaleidae). Three of these datasets also had poor fit at the DNA sequence level (*Aliatypus, Certhia* and *Tamias*), although not always for the same loci (50%). We were unable to identify systematic trends in the observed deviations. *Tamias* had one locus with an excess of deep coalescences and low probability (mitochondrial Cyt *b*) and one with a deficit of deep coalescences (ACR; Fig 2). *Aliatypus* had three nuclear loci with deficits of deep coalescences and low probabilities (partial results in Fig. 3). *Certhia* and Cheirogaleidae each had one locus with an excess of deep coalescences, but no deviation in probability of coalescent genealogy.

At the level of the sequence data we found evidence of poor model fit in 20 datasets (80%) with 45 partitions and 44 loci (16.9% and 18.3% respectively) deviating from expectations (Table 3). Deviations were apparent using all test statistics, but the GC statistic was the most frequent indicator (33/45 partitions). Again, there were no obvious systematic trends in the deviations, except that the empirical GC statistics tended to be smaller than expected (60% were smaller), although this was not significant under a binomial test.

While we have low sample size, mitochondrial genes (mtDNA) do not appear to be overrepresented among the loci that poorly fit in the coalescent genealogy tests. Two of the seven poorly fitting loci were mitochondrial (binomial test: successes=2, trials=7, probability=17/240, p=0.08). However, eleven of 44 of the loci that poorly fit in the sequence data tests were mitochondrial, a significant excess (binomial test: successes=11, trials=44, probability=17/240), p= 0.0002).

Case studies—removal of genes poorly fitting coalescent assumptions

The *Tamias* and Cheirogaleidae datasets each contained one locus that fit poorly at the coalescent genealogy level. In order to determine whether data that do not fit the model affect the outcome of the analysis we elected to remove these loci (the mtDNA locus Cyt *b* from *Tamias* and the nuclear locus ABCA1 from Cheirogaleidae) from their respective datasets and reanalyze them. Both loci had an excess of deep coalescences, and the *Tamias* Cyt *b* locus also showed low probability given the species tree.

After removal of Cyt *b* from *Tamias*, we did not observe substantial change in the posterior means of the relative mutation rate parameters. The species tree root height was 11% lower when Cyt *b* was included and the gene trees were on average 3% higher, although the 95% HPDs overlapped substantially (the result was the same if the gene trees were scaled to the same mutation rate or unscaled). By contrast, the scale parameter of a gamma-distributed prior on effective population sizes across branches in the species tree (the species.popMean hyperparameter), was 3.6 times larger in the dataset including Cyt *b* and the 95% HPDs for the two analyses did not overlap. Most notably, the tree topologies

between the two runs changed drastically, returning incompatible, highly supported nodes (Fig. S1). There was no evidence of poor fit using our PPS analysis after Cyt *b* was removed.

After removal of ABCA1 from the Cheirogaleidae dataset, we similarly observed few changes in the relative mutation rate parameters or root heights. The single exception was the Adora locus, whose mean relative mutation rate parameter was 1.5 times higher and whose unscaled root height was 1.4 times larger when ABCA1 was excluded. These changes apparently made the Adora tree a poorer fit to the sequence data, as all test statistics became more extreme, but none crossed the p=0.05 threshold. For all other loci, the mean root heights were 5% higher, the species tree root height was 10% higher and the popMean parameter was 10% larger when ABCA1 was included. The species tree relative branch lengths, topology and posterior probabilities were unchanged when ABCA1 was removed.

Case studies—reanalysis with independent gene trees

The *Sistrurus* and *Ursus* datasets did not show poor fit at the genealogy level, but each had several loci (3/20 and 4/14, respectively) that fit poorly at the sequence level. All poorly fitting loci had greater GC statistics than predicted. One locus from each dataset fit poorly using all test quantities. In order to test the hypothesis that poor fit at these loci stems from the multispecies coalescent-induced prior on the gene trees, we reanalyzed the data with no species tree and unlinked all parameters. As a result, all signs of poor fit vanished.

When comparing gene genealogy estimates between the analyses, results differed between the two datasets. There was no obvious reason why the independent-gene-trees

model should be a better fit to two of the three *Sistrurus* loci, as they were primarily unresolved in both analyses. The third gene, Fibrinogen beta chain (FGB), however, had one strongly supported conflict among analyses: in the independent-gene-trees model, the placement of alleles from the outgroup taxon, *Agkistrodon contortrix* were polyphyletic with respect to the ingroup, instead of as sister to the other outgroup taxon, *A. piscivorous.* Trees sampled in the MCMC for *Sistrurus* loci that fit the phylogenetic likelihood poorly tended to have very similar likelihoods whether the species tree was enforced or not.

Poorly fitting *Ursus* loci, by contrast, often had obvious topological differences across analyses. The locus nr11080, for example, yielded a paraphyletic *Ursus arctos*, with the bears from Admiralty, Baranof and Chichagof (ABC) islands being more closely related to *U. maritimus*. Under the species tree model, the ABC haplotypes were the sister group to *U. maritimus,* but under the independent-gene-trees model, they were scattered within the *U. maritimus* clade. Also in contrast to *Sistrurus*, each of the 4 poorly fitting loci showed average improvements of ~30 log-likelihood units for trees sampled during the MCMC under the independent-gene-trees model, while the other 10 loci improved ~5 log-likelihood units.

DISCUSSION

Gelman et al. (2009) argue that model checking is an essential part of Bayesian data analysis, on par with the initial formulation of models and fitting those models to data. Here we have developed the first general model-checking method for a commonly used multispecies coalescent phylogenetic inference model, and our results show that poor fit between model and data is detectable in a majority of sampled datasets. At the level of

coalescent genealogies, it is relatively straightforward to suggest biological explanations for poor fit between processes generating gene tree topologies and the specified coalescent model. By contrast, poor fit at the level of the DNA sequence data could plausibly be explained by a variety of forms of model misfit. While the test quantities used here do not uniquely identify the biological processes that violate the multispecies coalescent model, the identification of loci that fit poorly in combination with relevant biological and geographical context can suggest of directions for future analyses. Below, we discuss empirical examples of poor fit, their observed consequences, and possible approaches to take when poor fit is detected.

Empirical examples of poor fit to the multispecies coalescent model

Recent hybridization is a likely explanation for poor model fit when there is an excess of deep coalescences and closely related haplotypes are shared among species. The *Tamias* data provide a good example. Previous studies have detailed extensive mtDNA introgression between non-sister species in the genus (Good et al. 2008, Reid et al. 2010, Reid et al. 2012). In these analyses, the mtDNA is identified as having an excess of deep coalescences and low coalescent genealogy probability (Fig. 2). When the *Tamias* genealogies themselves are examined, statistically supported discordance among gene trees and species non-monophyly in the species tree are evident. When the mtDNA is removed and the data re-analyzed, all signs of poor fit, including at one nuclear locus that showed a deficit of deep coalescences, disappear. Additionally, the species tree topologies change drastically. Two species appearing as sister taxa in the tree including mtDNA are distantly related in the nuclear-only tree (Fig. S1). We expect recent hybridization to

impact analyses because without modeling it, species divergences must always post-date the divergence of shared gene lineages.

Unmodeled population structure can also presumably cause poor model fit, as exemplified by the *Aliatypus* data. There is no supported discordance among loci in relationships between populations in this group, and yet poor fit is evident in 3 out of 5 loci, and also in the summaries across loci (Fig. 3). These statistics indicate a deficit of deep coalescences (1 locus also has low probability). We hypothesize that high genetic diversity within OTUs, as a result of unmodeled population structure, leads to overestimation of effective population size, and thus an overprediction of the amount of stochastic lineage sorting. This is consistent with what is known about *Aliatypus* biology: these are terrestrial spiders that live in subterranean burrows with limited dispersal ability (Coyle and Icenogle 1994). Geographically close populations often have highly divergent mtDNA haplotypes, and those haplotypes tend to have highly restricted distributions (Satler et al. 2011).

The remaining two datasets that showed poor fit at the coalescent genealogy level, Cheirogaleidae and *Certhia*, each had one locus with an excess of deep coalescences, but did not have low probability. These issues are harder to resolve. An examination of the Cheirogaleidae locus, ABCA1, yielded a fairly well resolved genealogy that contained some species that were non-monophyletic, each with unique, highly divergent haplotypes not shared with other species. This could be a result of ancient hybridization, balancing selection, or gene duplication and extinction. Gene duplication and extinction seems unlikely, as the issue might have been expected to manifest itself in patterns of heterozygosity and been resolved through cloning. Distinguishing balancing selection from

ancient hybridization would require analyses of Dn/Ds ratios and more detailed studies of population structure.

It is also possible for systematic error in coalescent genealogy estimation to cause poor model fit at this level, even if the model is correct. For this to be the case, misspecified models of sequence evolution and/or molecular clock models would have to prefer incorrect trees strongly enough to overcome the prior probability distribution on coalescent genealogies induced by the multispecies coalescent. This may be most likely in cases when misspecification is very serious (e.g. sequences with secondary structure where sites are non-independent) or when there is a lot of information and high complexity (e.g. mtDNA). Both *Aliatypus* and *Tamias* mtDNA are found to fit the model poorly at both levels, but we believe the genealogical patterns causing poor fit in those systems are clear enough to adequately explain observed patterns of misfit.

At the level of DNA sequence data, sources of poor fit are harder to distinguish. There are two main possibilities. First, models of sequence evolution and molecular clock models could be misspecified. Second, coalescent genealogies could be a poor fit to the multispecies coalescent, but the prior distribution on coalescent genealogies induced by the model might overwhelm the signal in the data. This could strongly favor topologies and branch lengths that fit the multispecies coalescent, but are a poor representation of the sequences. We speculate that both effects are evident in our analyses. The overrepresentation of mtDNA loci among those that poorly fit at this level may result from the first effect. In particular, mtDNA contains a large number of variable sites and thus more phylogenetic signal than most autosomal loci. It seems less likely that prior probability distributions on gene genealogies could push poorly fitting loci away from their

preferred topologies. In support of this idea, of two datasets included here that partitioned their mtDNA, both had some partitions that fit the model and some that did not (e.g. Fig. 3). If the tree was the problem, we might expect all partitions of the same locus to fit poorly.

By contrast, we expect much of the poor fit we observe at nuclear loci to be a result of the second factor. Our analyses of *Sistrurus* and *Ursus* support this idea. We would not expect to see improved fit when the species tree portion of the model was eliminated if the models of sequence evolution and mutation rate variation were causing the problem. Additionally, in *Ursus* we see changes in topology and posterior probabilities that are consistent with strong prior sensitivity. Interestingly, the pattern observed at the locus mentioned above, nr11080, is likely to be a previously unacknowledged signal of hybridization at nuclear loci in this dataset. *U. maritimus* mtDNA is thought to be a result of a fixed introgression from bears related to those from the ABC islands, so it is unsurprising to discover that the ABC bears may harbor DNA that has moved in the other direction (Edwards et al. 2011, Hailer et al. 2012, Miller et al. 2012).

It is important to note that there are potentially other non-examined sources of discordance, including inaccurate taxonomic knowledge of species limits. If individuals are inaccurately assigned to OTUs in a species tree analysis, one would expect that PPS would indicate that the species tree is a poor fit to the data, although it would likely be difficult to identify the cause of this poor fit.

Consequences of poor fit

Our analyses suggest that poor fit to the multispecies coalescent model can mislead inference in empirical studies. In the case of recent hybridization, the consequences may be

severe, as species divergences are forced to post-date gene divergences. For example, when the mitochondrial DNA were removed from *Tamias*, the species tree topology changed drastically. The topologies from both analyses conflicted at strongly supported nodes and two recently hybridizing species, *T. amoenus* and *T. ruficaudus* went from sister taxa, to being only distantly related (Fig. S1). Unexpectedly, the nuclear DNA did not fit either model poorly, which may be a sign that the data are insufficiently informative or that our approach does not have high power.

When topological conflict among coalescent genealogies is the result of ancient hybridization, balancing selection, or gene duplication and extinction, the consequences may be less severe. It seems possible that such conflicts may be resolvable by invoking deep coalescence. For example, when we removed the ABCA1 locus from the Cheirogaleidae, the changes to the topology, branch lengths, and posterior support were minimal. This flexibility of the multispecies coalescent may also make such processes difficult to detect using our framework. If detecting such processes were our primary goal, rather than identifying instances of poor model fit, development of new test quantities might be necessary.

If the coalescent genealogies themselves are of interest, our results suggest that the prior probability distribution induced by the multispecies coalescent can be quite informative. This is of concern because some recent studies have used species tree-based models to improve estimates of gene genealogies (Åkerborg et al. 2009, Wu et al. 2012). This may be useful when the prior distribution is appropriate, but if important processes are unmodeled, our results suggest analyses may be misled.

Strategies for dealing with poor fit

When poor fit is detected, there are two main strategies that can be used to ameliorate it. First, remove data that violate the multispecies coalescent model. Many of the processes causing poor fit may be heterogeneous across the genome. Not every gene family is expected to be the focus of bouts of duplication and extinction, or intense selection, and not every region of the genome will introgress with equal ease. If a relatively small number of loci appear to fit poorly, it is easy to remove them and re-analyze the data. Second, the biological processes that generate variation in gene tree topologies should be explicitly modeled, as should relevant dynamics of molecular evolution. Increasingly complex multispecies coalescent models are being implemented, but there are tradeoffs. Some examine gene duplication and extinction (Rasmussen and Kellis 2012) or migration (Pickrell and Pritchard 2012) but cannot estimate divergence times.

We believe our results suggest that a concatenation approach to analyzing multilocus datasets with extensive inter-locus heterogeneity in topology may be even more perilous than simulation studies have shown (Kubatko and Degnan 2007, Edwards 2009). Those studies assume that the only source of discordance among loci is coalescent stochasticity. Here we show that other factors contribute to heterogeneity among gene trees, exacerbating the issue.

CONCLUSIONS

The very act of data analysis requires researchers to make assumptions about the evolutionary processes that have shaped the data. We demonstrate that not all empirical data are consistent with the assumptions of the multispecies coalescent model. As the

number and breadth of phylogenetic methods increases, it is far better to assess the fit between models and the data to which they are being applied than it is to assume that a certain method is appropriate to a given data set. Phylogenetics is no longer a data-poor enterprise, and we can afford to be choosy with the data that are analyzed. Posterior predictive simulation is an effective method for identifying data that violate important assumptions of analytical models.

Supplementary Material

Data (i.e., the *BEAST XML files) have been deposited in the Dryad data repository under

Acknowledgements

We thank all authors who shared previously published data. We would also like to thank SSB for supporting the symposium at Evolution 2012, and attendees of the symposium for helpful comments. NMR was supported by a grant from the Society for Systematic Biologists and by the Louisiana Board of Regents Fellowship. This work was supported by a grant from the National Science Foundation [DEB – 0918212] to BCC and by the Louisiana State University College of Science and Department of Biological Science. Portions of this research were conducted with high performance computation resources provided by LSU (http://www.hpc.lsu.edu).

References

Åkerborg Ö, Sennblad B, Arvestad L, Lagergren J. 2009. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. Proc. Natl. Acad. Sci. USA, 106:5714-5719.

Alstrom P, Fregin S, Norman JA, Ericson PGP, Christidis L, Olsson U. 2011a. Multilocus analysis of a taxonomically densely sampled dataset reveal extensive non-monophyly in the avian family locustellidae. Mol. Phylogenet. Evol., 58:513-526.

Alstrom P, Hohna S, Gelang M, Ericson PG, Olsson U. 2011b. Non-monophyly and intricate morphological evolution within the avian family cettiidae revealed by multilocus analysis of a taxonomically densely sampled dataset. BMC Evol. Biol., 11:352.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol., 24:412-426.

Belfiore NM, Liu L, Moritz C. 2008. Multilocus phylogenetics of a rapid radiation in the genus thomomys (rodentia: Geomyidae). Syst. Biol., 57:294-310.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol., 19:1171-1180.

Brown JM, ElDabaje R. 2009. Puma: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics, 25:537-538.

Brumfield RT, Liu L, Lum DE, Edwards SV. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (aves: Pipridae, manacus) from multilocus sequence data. Syst. Biol., 57:719-731.

Brunes TO, Sequeira F, Haddad CFB, Alexandrino J. 2010. Gene and species trees of a neotropical group of treefrogs: Genetic diversification in the brazilian atlantic forest and the origin of a polyploid species. Mol. Phylogenet. Evol., 57:1120-1133.

Camargo A, Avila LJ, Morando M, Sites JW. 2012. Accuracy and precision of species trees: Effects of locus, individual, and base pair sampling on inference of species trees in lizards of the liolaemus darwinii group (squamata, liolaemidae). Syst. Biol., 61:272-288.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PloS Genet., 2:e64.

Choi SC, Hey J. 2011. Joint inference of population assignment and demographic history. Genetics, 189:561-577.

Chung Y, Ane C. 2011. Comparing two bayesian methods for gene tree/species tree reconstruction: Simulations with incomplete lineage sorting and horizontal gene transfer. Syst. Biol., 60:261-275.

Clemente-Carvalho RBG, Klaczko J, Perez SI, Alves ACR, Haddad CFB, dos Reis SF. 2011. Molecular phylogenetic relationships and phenotypic diversity in miniaturized toadlets, genus brachycephalus (amphibia: Anura: Brachycephalidae). Mol. Phylogenet. Evol., 61:79-89.

Coyle FA, Icenogle WR. 1994. Natural history of the californian trapdoor spider genus aliatypus (araneae, antrodiaetidae). J. Arachnology:225-255.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PloS Genet., 2:e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol., 24:332-340.

Drummond A, Ho S, Phillips M, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol., 4:e88.

Drummond A, Rambaut A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol., 7:214.

Drummond A, Suchard M. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biology, 8:114.

Eckert AJ, Carstens BC. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogenet. Evol., 49:832-842.

Edwards CJ, Suchard MA, Lemey P, Welch JJ, Barnes I, Fulton TL, Barnett R, O'Connell TC, Coxon P, Monaghan N. 2011. Ancient hybridization and an irish origin for the modern polar bear matriline. Curr. Biol., 21:1251-1258.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? Evolution, 63:1-19.

Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA, 104:5936-5941.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol., 17:368-376.

Florez-Rodriguez A, Carling MD, Cadena CD. 2011. Reconstructing the phylogeny of "buarremon" brush-finches and near relatives (aves, emberizidae) from individual gene trees. Mol. Phylogenet. Evol., 58:297-303.

Fulton TL, Strobeck C. 2010. Multiple markers and multiple individuals refine true seal phylogeny and bring molecules and morphology back in line. Proc. R. Soc. B-Biol. Sci., 277:1065-1070.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2009. Bayesian data analysis. 2nd ed. Boca Raton, FL, Chapman and Hall/CRC.

Gerard D, Gibbs HL, Kubatko L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol. Biol., 11:291.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol., 36:182-198.

Good JM, Hird S, Reid N, Demboski JR, Steppan SJ, Martin‐Nims TR, Sullivan J. 2008. Ancient hybridization and mitochondrial capture between two species of chipmunks. Mol. Ecol., 17:1313-1327.

Hailer F, Kutschera VE, Hallström BM, Klassert D, Fain SR, Leonard JA, Arnason U, Janke A. 2012. Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science, 336:344-347.

Harrington RC, Near TJ. 2012. Phylogenetic and coalescent strategies of species delimitation in snubnose darters (percidae: Etheostoma). Syst. Biol., 61:63-79.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol., 27:570-580.

Hird S, Kubatko L, Carstens B. 2010. Rapid and accurate species tree estimation for phylogeographic investigations using replicated subsampling. Mol. Phylogenet. Evol., 57:888-898.

Huang H, He Q, Kubatko LS, Knowles LL. 2010. Sources of error inherent in species-tree estimation: Impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol., 59:573-583.

Hudson RR. 2002. Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics, 18:337-338.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science, 294:2310-2314.

Jennings WB, Edwards SV. 2005. Speciational history of australian grass finches (*poephila)* inferred from thirty gene trees. . Evolution, 59:2033-2047.

Joly S. 2012. Jml: Testing hybridization from species trees. Mol. Ecol. Resour., 12:179-184.

Joseph L, Toon A, Schirtzinger EE, Wright TF. 2011. Molecular systematics of two enigmatic genera psittacella and pezoporus illuminate the ecological radiation of australo-papuan parrots (aves: Psittaciformes). Mol. Phylogenet. Evol., 59:675-684.

Kubatko LS, Carstens BC, Knowles LL. 2009. Stem: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics, 25:971-973.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol., 56:17-24.

Kubatko LS, Gibbs HL, Bloomquist EW. 2011. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in sistrurus rattlesnakes. Syst. Biol., 60:393-409.

Lanier HC, Knowles LL. 2012. Is recombination a problem for species-tree analyses? Syst. Biol., 61:691-701.

Leaché AD. 2009. Species tree discordance traces to phylogeographic clade boundaries in north american fence lizards (sceloporus). Syst. Biol., 58:547-559.

Leache AD, Rannala B. 2011. The accuracy of species tree estimation under simulation: A comparison of methods. Syst. Biol., 60:126-137.

Lee JY, Joseph L, Edwards SV. 2012. A species tree for the australo-papuan fairy-wrens and allies (aves: Maluridae). Syst. Biol., 61:253-271.

Liu L, Edwards SV. 2009. Phylogenetic analysis in the anomaly zone. Syst. Biol., 58:452-460.

Maddison WP. 1997. Gene trees in species trees. Syst. Biol., 46:523-536.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol., 20:229-237.

Manthey JD, Klicka J, Spellman GM. 2011. Isolation-driven divergence: Speciation in a widespread north american songbird (aves: Certhiidae). Mol. Ecol., 20:4371-4384.

McCormack JE, Heled J, Delaney KS, Peterson AT, Knowles LL. 2011. Calibrating divergence times on species trees versus gene trees: Implications for speciation history of *aphelocoma* jays. Evolution, 65:184-202.

Melo-Ferreira J, Boursot P, Carneiro M, Esteves PJ, Farelo L, Alves PC. 2011. Recurrent introgression of mitochondrial DNA among hares (lepus spp.) revealed by species-tree inference and coalescent simulations. Syst. Biol., 61:367-381.

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, Kim HL, Burhans RC, Drautz DI, Wittekindt NE. 2012. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc. Natl. Acad. Sci. USA, 109:E2382-E2390.

Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Biol., 51:729-739.

Nielsen R, Bollback JP. 2005. Posterior mapping and posterior predictive distributions

statistical methods in molecular evolution. Springer New York, p. 439-462.

O'Meara BC. 2010. New heuristic methods for joint species delimitation and species tree inference. Syst. Biol., 59:59-73.

Paradis E, Claude J, Strimmer K. 2004. Ape: Analyses of phylogenetics and evolution in r language. Bioinformatics, 20:289-290.

Pasachnik SA, Echternacht AC, Fitzpatrick BM. 2010. Gene trees, species and species trees in the ctenosaura palearis clade. Conserv. Genet., 11:1767-1781.

Peters JL, Roberts TE, Winker K, McCracken KG. 2012. Heterogeneity in genetic diversity among non-coding loci fails to fit neutral coalescent models of population history. PLoS ONE, 7:e31972.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. arXiv, 206.2332.

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Rabosky DL, Slater GJ, Alfaro ME. 2012. Clade age and species richness are decoupled across the eukaryotic tree of life. PLoS Biol., 10:e1001381.

Rambaut A, Grass NC. 1997. Seq-gen: An application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. Comp. App. Biosci., 13:235-238.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics, 164:1645-1656.

Rasmussen MD, Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res., 22:755-765.

Recuero E, Canestrelli D, Voros J, Szabo K, Poyarkov NA, Arntzen JW, Crnobrnja-Isailovic J, Kidov AA, Cogalniceanu D, Caputo FP, *et al.* 2012. Multilocus species tree analyses resolve the radiation of the widespread bufo bufo species group (anura, bufonidae). Mol. Phylogenet. Evol., 62:71-86.

Reid N, Demboski JR, Sullivan J. 2012. Phylogeny estimation of the radiation of western north american chipmunks (tamias) in the face of introgression using reproductive protein genes. Syst. Biol., 61:44-62.

Reid N, Hird S, Schulte-Hostedde A, Sullivan J. 2010. Examination of nuclear loci across a zone of mitochondrial introgression between tamias ruficaudus and t. Amoenus. J. Mammal., 91:1389-1400.

Ripplinger J, Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and bayesian methods. Mol. Biol. Evol., 27:2790-2803.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature, 425:798-804.

Rosenberg NA. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol., 61:225-247.

Salicini I, Ibanez C, Juste J. 2011. Multilocus phylogeny and species delimitation within the natterer's bat species complex in the western palearctic. Mol. Phylogenet. Evol., 61:888-898.

Satler JD, Starrett J, Hayashi CY, Hedin M. 2011. Inferring species trees from gene trees in a radiation of california trapdoor spiders (araneae, antrodiaetidae, aliatypus). PLoS ONE, 6: e25355.

Schliep KP. 2011. Phangorn: Phylogenetic analysis in r. Bioinformatics, 27:592-593.

Takahata N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics, 122:957-966.

Tepe EJ, Farruggia FT, Bohs L. 2011. A 10-gene phylogeny of *solanum* section *herpystichum* (solanaceae) and a comparison of phylogenetic methods. Am. J. Bot., 98:1356-1365.

Walstrom VW, Klicka J, Spellman GM. 2012. Speciation in the white-breasted nuthatch (sitta carolinensis): A multilocus perspective. Mol. Ecol., 21:907-920.

Weisrock DW, Smith SD, Chan LM, Biebouw K, Kappeler PM, Yoder AD. 2012. Concatenation and concordance in the reconstruction of mouse lemur phylogeny: An empirical demonstration of the effect of allele sampling in phylogenetics. Mol. Biol. Evol., 29:1615-1630.

Wu Y-C, Rasmussen MD, Bansal MS, Kellis M. 2012. Treefix: Statistically informed gene tree error correction using species trees. Syst. Biol.

Yang Z, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA, 107:9264-9269.

Zhang J. 2003. Evolution by gene duplication: An update. Trends Ecol. Evol., 18:292-298.

TABLES

Table 1: Summary of empirical datasets used in this analysis.

Dataset	Authors	Year	Source	Order	Family	Genus	Level	OTUs	Alleles	Loci	Organellar DNA
Thomomys	Belfiore et al.	2008	BEAST help	Rodentia	Geomyidae	*Thomomys*	population	9	26	7	-
Manacus	Brumfield et al.	2008	web	Passeriformes	Pipridae	*Manacus*	species	5	47	5	-
Sceloporus	Leaché	2009	TreeBASE	Squamata	Iguanidae	*Sceloporus*	population	9	21	8	-
Phyllomedusa	Brunes et al.	2010	author	Anura	Hylidae	*Phyllomedusa*	species	5	27	3	+
Phocidae	Fulton and Strobeck	2010	author	Pinnipedia	Phocidae		genus	24	39	16	+
Ctenosaura	Pasachnik et al.	2010	author	Squamata	Iguanidae	*Ctenosaura*	species	5	27	4	+
Cettiidae	Alstrom et al.	2011	author	Passeriformes	Cettiidae		genus	29	97	4	+
Locustellidae	Alstrom et al.	2011	author	Passeriformes	Locustellidae		genus	41	41	5	+
Brachycephalus	Clement- Carvalho et al.	2011	genbank	Anura	Brachycephalidae	*Brachycephalus*	species	15	15	4	+
Buarremon	Florez-Rodriguez et al.	2011	author	Passeriformes	Emberizidae	*Buarremon*	species	5	5	7	+
Psittacidae	Joseph et al.	2011	author	Psittaciformes	Psittacidae		genus	27	27	8	+
Sistrurus	Kubatko et al.	2011	TreeBASE	Squamata	Crotalidae	*Sistrurus*	population	8	52	19	+
Certhia	Manthey et al.	2011	author	Passeriformes	Certhiidae	*Certhia*	population	11	141	20	-
Lepus	Melo-Ferreira et al.	2011	TreeBASE	Lagomorpha	Leporidae	*Lepus*	species	13	55	14	-
Myotis	Salicini et al.	2011	author	Chiroptera	Vespertilionidae	*Myotis*	species	6	49	7	+
Aliatypus	Satler et al.	2011	author	Araneae	Antrodiaetidae	*Aliatypus*	population	13	102	5	+
Herpystichum	Tepe et al.	2011	author	Solanales	Solanaceae	*Herpystichum*	species	12	18	10	-
Liolaemus	Carmago et al.	2012	author	Squamata	Liolaemidae	*Liolaemus*	population	16	48	20	+
Ursus	Hailer et al.	2012	supplemental	Carnivora	Ursidae	*Ursus*	population	7	90	14	-
Etheostoma	Harrington and Near	2012	TreeBASE	Perciformes	Percidae	*Etheostoma*	population	5	28	4	+
Malurus	Lee et al.	2012	TreeBASE	Passeriformes	Maluridae	*Malurus*	population	25	84	17	+
Bufo	Recuero et al.	2012	author	Anura	Bufonidae	*Bufo*	species	4	232	5	+
Tamias	Reid et al.	2012	author	Rodentia	Sciuridae	*Tamias*	species	22	232	5	+
Sitta	Walstrom et al.	2012	author	Passeriformes	Sittidae	*Sitta*	population	7	112	17	+
Cheirogaleidae	Weisrock et al.	2012	author	Primates	Cheirogaleidae		genus	20	65	12	-

Table 2: Summary of the results of the tests of the fit of the coalescent genealogies multispecies coalescent.

		Coalescent likelihood across loci - $P(u_i\|S)$				**Deep coalescences across loci**				**Individual Loci**					
		Sum		Coefficient		Sum		Coefficient		Coalescent Likelihood		Deep Coalescences			
Dataset	**Loci**	+	-	+	-	+	-	+	-	+	-	+	-	**Total**	**Poorly fitting loci (%)**
Aliatypus	5	**1**	-	-	**1**	**1**	-	-	**1**	**3**	-	**3**	-	**3**	60.0
Certhia	20	-	-	-	-	-	-	-	**1**	-	-	-	**1**	**1**	5.0
Cheirogaleidae	12	-	-	-	-	-	-	-	-	-	-	-	**1**	**1**	8.3
Tamias	5	-	-	-	-	**1**	-	**1**	-	**1**	-	**1**	**1**	**2**	40.0
Total	**240**	**1**	0	0	**1**	**2**	0	**1**	**2**	**4**	0	**4**	**3**	**7**	2.9

Table 3: Summary of the results of the tests of the fit of the sequence data to the phylogenetic Likelihood.

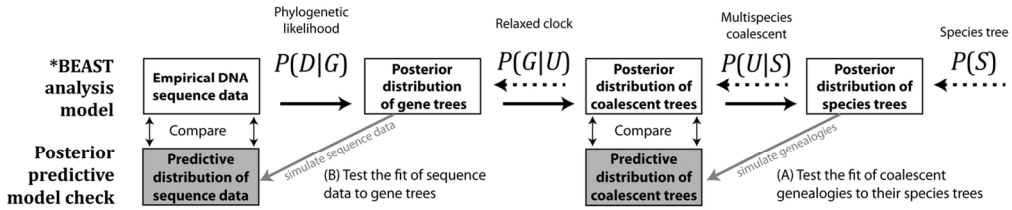
			Variable sites		Phylogenetic likelihood		Multinomial likelihood		GC statistic			% poorly fitting partitions	% poorly fitting loci
Dataset	**Loci**	**Partitions**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**Total**		
Aliatypus	5	8	**2**	-	-	**3**	-	**2**	**3**	-	**3**	37.5	40.0
Brachycephalus	4	4	-	-	-	-	-	-	-	**1**	**1**	25.0	25.0
Buarremon	7	7	-	**1**	-	-	-	-	-	-	**1**	14.3	14.3
Bufo	5	5	-	**1**	-	-	**1**	-	**1**	-	**2**	40.0	40.0
Certhia	20	20	-	**1**	**1**	-	**1**	-	-	**1**	**1**	5.0	5.0
Cettiidae	4	4	-	-	-	-	-	-	-	-	0	0.0	0.0
Cheirogaleidae	12	12	-	-	-	-	-	-	-	-	0	0.0	0.0
Ctenosaura	4	4	-	**1**	-	-	**1**	-	-	-	**1**	25.0	25.0
Etheostoma	4	4	-	-	-	-	-	-	**1**	-	**1**	25.0	25.0
Herpystichum	10	10	**1**	-	-	**1**	-	**1**	**1**	-	**1**	10.0	10.0
Lepus	14	14	-	-	-	-	-	-	-	-	0	0.0	0.0
Liolaemus	20	20	**6**	**1**	**1**	**7**	**1**	**7**	**8**	-	**10**	50.0	50.0
Locustellidae	5	5	-	-	-	**1**	-	-	**1**	-	1	20.0	20.0
Malurus	17	17	-	-	-	-	**1**	-	-	-	**1**	5.9	5.9
Manacus	5	5	-	-	-	-	-	-	-	-	0	0.0	0.0
Myotis	7	7	**1**	-	-	-	-	**1**	-	**1**	**1**	14.3	14.3
Phocidae	16	40	-	**2**	**2**	-	**2**	-	**2**	**1**	**4**	10.0	25.0
Phyllomedusa	3	3	-	-	-	-	-	-	-	-	0	0.0	0.0
Psittacidae	8	8	-	**1**	**1**	-	**2**	-	-	**1**	**2**	25.0	25.0
Sceloporus	8	8	-	-	**1**	-	**1**	-	**1**	-	**2**	25.0	25.0
Sistrurus	19	19	-	**1**	**1**	-	**1**	-	-	**2**	**3**	15.8	15.8
Sitta	17	17	-	**1**	**1**	-	**1**	-	-	**2**	**2**	11.8	11.8
Tamias	5	5	-	**1**	**1**	-	**1**	-	-	-	**1**	20.0	20.0
Thomomys	7	7	-	-	-	-	-	**1**	**2**	-	**3**	42.9	42.9
Ursus	14	14	-	**1**	**1**	-	-	**1**	-	**4**	**4**	28.6	28.6
Total	**240**	**267**	**10**	**12**	**10**	**12**	**13**	**13**	**20**	**13**	**45**	**16.9**	18.3

FIGURE LEGENDS – MOVE TO AFTER REFERENCES WHEN DONE WITH ENDNOTE

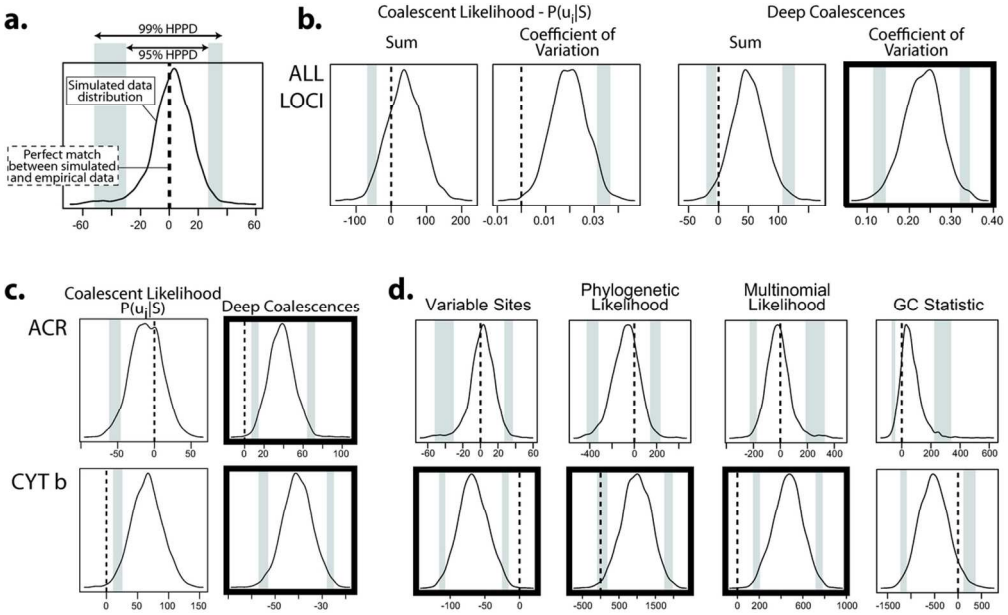
Figure 1: A schematic of the *BEAST analysis model and how our PPS model checks relate to it. We check the fit of two parts of the model independently: (a) the fit of the coalescent genealogies to the species tree and (b) the fit of DNA sequence data to the gene trees. Information from the data filters up through the model (black lines), while prior information filters down (dashed lines). Simulated data (gray) are influenced by the empirical data, the model structure, and the prior.

Figure 2: PPS model check results for *Tamias*. (a) Key for the figure. Distributions of test statistics are shown, where the dashed line is the expectation (0), and gray bars indicate the boundaries of the 95% and 99% highest posterior predictive density intervals. (b) and (c) give results for the test of the fit of coalescent genealogies; boxes with bold black lines indicate poor fit. A single gray bar indicates a one-tailed test. (b) Coalescent genealogy tests for all loci; (c) Coalescent genealogy tests for the two loci that fit poorly, ACR and Cyt *b*; (d) Sequence data tests for the same two loci.

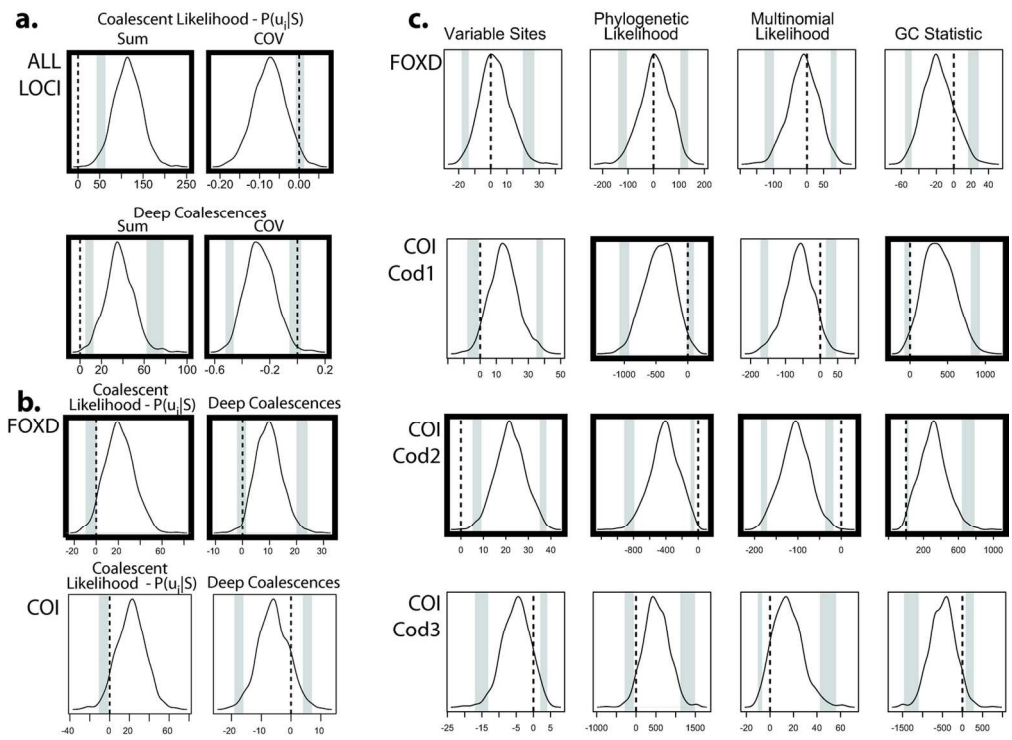
Figure 3: PPS model check results for *Aliatypus*. See Figure 2 for interpretation. (a) Coalescent genealogy tests for all loci. (b) and (c) show two of three loci for which the data were a poor fit; (b) Coalescent genealogy tests; (c) Sequence data tests, with the COI locus partitioned by codon.



125x27mm (300 x 300 DPI)



112x68mm (300 x 300 DPI)



140x102mm (300 x 300 DPI)

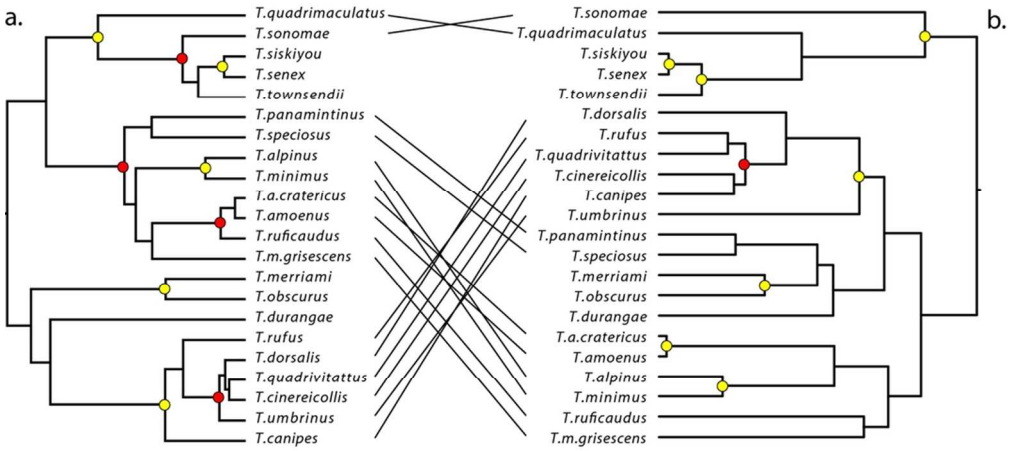


Figure S1. *Tamias* species trees estimated in *BEAST using 5 markers. (a) includes the mitochondrial locus Cyt *b* and (b) excludes it. Circles indicate 0.95 posterior probability. Red red indicates nodes are incompatible with the opposite tree.

90x50mm (300 x 300 DPI)