Journal of Phylogenetics & Evolutionary Biology

McVay and Carstens, J Phylogen Evolution Biol 2013, 1:3
DOI: 10.4172/2329-9002.1000114

Research Article Open Access

Phylogenetic Model Choice: Justifying a Species Tree or Concatenation Analysis

John David McVay[1] and Bryan C. Carstens[1,2]*

[1]*Department of Biological Sciences, Louisiana State University, 202 Life Sciences, Baton Rouge, LA 70803, USA*
[2]*Evolution, Ecology and Organismal Biology, The Ohio State Univeristy, 318 W. 12th Ave., Columbus, OH 43210, USA*

Abstract

There are two paradigms for the phylogenetic analysis of multi-locus sequence data: one which forces all genes to share the same underlying history, and another that allows genes to follow idiosyncratic patterns of descent from ancestral alleles. The first of these approaches (concatenation) is clearly a simplified model of the actual process of genome evolution while the second (species-tree methods) may be overly complex for histories characterized by long divergence times between cladogenesis. Rather than making an *a priori* determination concerning which of these phylogenetic models to apply to our data, we seek to provide a framework for choosing between concatenation and species-tree methods that treat genes as independently evolving lineages. We demonstrate that parametric bootstrapping can be used to assess the extent to which genealogical incongruence across loci can be attributed to phylogenetic estimation error, and demonstrate the application of our approach using an empirical dataset from 10 species of the Natricine snake sub-family. Since our data exhibit incongruence across loci that are clearly caused by a mixture of coalescent stochasticity and phyogenetic estimation error, we also develop an approach for choosing among species tree estimation methods that take gene trees as input and those that simultaneously estimate gene trees and species trees.

Introduction

There are two primary paradigms for estimating phylogeny from multi-locus sequence data [1]. The conventional method, which developed from arguments in favor of total evidence [2], estimates phylogeny by concatenating data across multiple genes collected from exemplar samples. In this approach, the data are treated as a single locus, and essentially the estimate of genealogy from each locus is averaged across genes. Underlying this method is the intuition that phylogenetic accuracy improves with an increase in the number of variable sites [3]. While this assumption certainly holds within a particular locus, applying this method across multiple loci requires the assumption that the gene trees across loci share a similar topology. When this is demonstrably not the case, incongruence across loci is attributable to phylogenetic estimation error rather than to coalescent processes (e.g., the independent sorting of alleles across loci). Recently, the primacy of concatenation has been challenged on several fronts [4-8], and methods that estimate phylogeny while allowing for incongruence across loci due to coalescent processes have been proposed. These coalescent-based approaches to phylogeny inference estimate species tree either given gene trees [9,10], or estimate gene trees and species tree topologies simultaneously [11,12]. Either approach accounts for population-level processes, such as the incomplete sorting of ancestral polymorphism that can cause gene tree discordance.

Given the growing criticism of concatenation, empiricists are faced with a vexing decision regarding the choice of phylogenetic method to apply to their system. Coalescent-based approaches are often favored *a priori* in phylogeographic investigations, where the incomplete sorting of ancestral polymorphism can be dramatically evident across loci [4,13-17], while concatenation continues to be favored among those working at deeper taxonomic levels [18-20]. However, it is clear that population level processes such as the sorting of ancestral polymorphism have occurred throughout the history of life; further, one of the central theses of the modern synthesis is the expectation that evolutionary processes within populations ultimately produce phylogenetic patterns [21]. This led Edwards [1] to argue that species tree approaches are preferable on first principles. Philosophical implications aside, the question of phylogenetic method choice is also of dramatic practical importance because the ideal sampling schemes for concatenation and coalescent-based approaches are quite different. Since the former assumes that population-level processes do not have an effect on phylogeny estimation, systematists who concatenate their data benefit from sampling as many genes as possible and fewer individuals per species. Alternatively, coalescent-based approaches appear to be most accurate with intermediate numbers of loci and multiple individuals sampled within species [22-24]. This places an empiricist in a difficult position; optimally they need to recognize which of these approaches appears to be appropriate given their data before all of it is collected in order to employ the optimal sampling scheme. It is also the position we found ourselves in some months ago, and in this study we propose an approach to answering this question using a preliminary data set of 7 genes from 1-2 individuals for each of 10 species of thamnophiine snakes. Given our data, how should we determine which of the competing phylogenetic paradigms to employ?

Perhaps the most important evidence available to empiricists who seek to objectively determine whether to concatenate their data or use species-tree methods is the degree of incongruence among loci. If the gene trees are mostly congruent, this is evidence that the branch lengths of the species tree are sufficiently long to have allowed lineage sorting to reach completion, and thus concatenation may be justified. Alternatively, incongruence among gene trees may be caused by coalescent processes and would suggest that coalescent-based methods are required. One approach would simply be to measure the incongruence across gene trees using a metric for tree comparison such as the Robinson-Foulds distance [25]. Distributions of the pairwise R-F distances can be substantial at shallow phylogenetic depths; this incongruence can

***Corresponding author:** Bryan C. Carstens, Evolution, Ecology and Organismal Biology, The Ohio State Univeristy, 318 W. 12th Ave., Columbus, OH 43210, USA, Tel: 614-292-6587; E-mail: carstens.12@osu.edu

Received May 02, 2013; **Accepted** July 24, 2013; **Published** July 29, 2013

Citation: McVay JD, Carstens BC (2013) Phylogenetic Model Choice: Justifying a Species Tree or Concatenation Analysis. J Phylogen Evolution Biol 1: 114. doi:10.4172/2329-9002.1000114

also persist at deeper levels (Figure 1). However, observed discordance among gene tree estimates can arise from other neutral sources such as mutational stochasticity, as well as phylogenetic estimation error, and thus a major challenge for empiricists is determining if the observed incongruence across gene trees can be attributed to phylogenetic estimation error alone. It is reasonable to conclude that concatenation is appropriate when the level of discord is of a magnitude that can be attributed to phylogenetic estimation error, here the substitutions across loci will provide valuable information regarding ancestral nodes. Conversely, gene tree estimates that are incongruent to a greater extent than would be expected due to phylogenetic estimation error alone is an indication that coalescent uncertainty has caused the discord, and therefore must be accounted for through the use of species tree estimation approaches. Here we use parametric bootstrapping to conduct a series of pair wise tests to ascertain whether the incongruence across genealogies estimated from our empirical data can be attributed to phylogenetic estimation error alone.

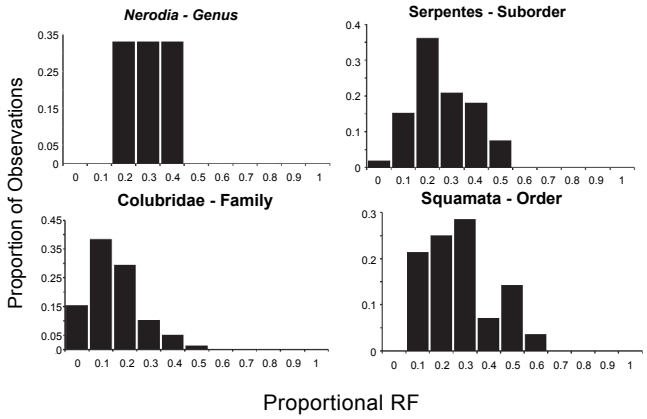

Figure 1: Histograms of pairwise symmetric difference distance among gene trees from four multi-locus empirical datasets. To assess the degree to which individual gene trees share the same topology for multigene datasets representing four depths of phylogeny (Table 1), we used the symmetric difference distance (RF distances; Robinson and Foulds, 1981) to compare all trees in a pairwise fashion using PAUP*. Observations of zero indicate no topological incongruence between two trees.

Study	Taxon	T(mya)	Loci*
Jennings and Edwards (2005)	Poephila (sister species)	<1	25
Wiens et al. (2008)	Colubridae (family)	40	18
Wiens et al. (2008)	Serpentes (suborder)	90	15
Vidal and Hedges (2005)	Squamata (order)	160	9

*Number of loci used in this study; some loci data sets were incompleteand thus not used.

Table 1: Datasets from literature used in this study. Shown for each dataset are the citation, the focal taxon, the level of phylogenetic divergence (T, in millions of years) and the number of loci.

We explore these questions using data collected from the North American colubrid snake tribe Thamnophiini, which consists of ~65 species representing nine genera. Previous studies [26,27] have estimated partial phylogenies of this group with differing results, yet to date no complete phylogeny has been estimated. Most recently, Alfaro [28] published a Bayesian estimate of phylogeny in which the *Nerodia*, the water snakes, were not monophyletic. Specifically, two species of the genus *Regina* and the monotypic *Tropidoclonion* were nested within *Nerodia*. While this result was not strongly supported by the data, the relationships within the Thamnophiini remain unsatisfactorily resolved. As empiricists, our ultimate research goal is to resolve the phylogeny of this group; however, our proximate goal, and the primary goals of this study is to determine which phylogenetic method is appropriate so that we can identify the optimal sampling scheme.

Methods

Empirical

Data collection: Molecular sequence data from five nuclear and two mitochondrial gene fragments were collected from seven species of the snake genus *Nerodia*, two North American relatives (*Regina grahamii* and *Tropodoclonion lineatum*), and an old world relative (*Natrix natrix*). The additional North American species were included to include an individual that has been previously suggested to be within *Nerodia* (*R. grahamii*) and to include one from a putatively deeper split (*Tropidoclonion*; Alfaro, 2003). DNA was extracted from tissue or blood using DNeasy kit (Qiagen, Hilden, Germany) following manufacturer's protocols. Each fragment was amplified via polymerase chain reaction using standard protocols: 25-50 ng template, 5 pmoles each primer (Table 2), 1.25 nmoles each dNTP, 1X PCR Buffer (New England Biolabs, Ipswich, MA), 0.5 units Taq polymerase, and nuclease-free H_2O to 25 µl. Amplicons were purified with Exonuclease I and

Primer	Gene	Oligo (5'-3')	Reference
BDNF-F	BDNF	GACCATCCTTTTCCTKACTATGGTTATTTCATACTT	Leache and McGuire (2006)
BDNF-R		CTATCTTCCCCTTTTAATGGTCAGTGTACAAAC	
FSHR_f1	FSHR	CCDGATGCCTTCAACCCVTGTGA	Wiens et al. (2008)
FSHR_r2		RCCRAAYTTRCTYAGYARRATGA	
Lglu	CYTB	TGATCTGAAAAACCACCGTTGTA	Alfaro and Arnold (2001)
H15544		AATGGGATTTTGTCAATGTCTGA	
G482	MC1R	TCAGCAACGTGGTGGA	Austin et al. (2009)
G480		ATGAGGTAGAGGCTGAAGTA	
ND4	ND4	TGACTACCAAAAGCTCATGTAGAAGC	Forstner et al. (1995)
M246		TTTTACTTGGATTTGCACCA	Skinner et al. (2006)
NTF3_F1	NT3	ATGTCCATCTTGTTTTATGTGATATTT	Wiens et al. (2008)
NTF3_R1		ACRAGTTTRTTGTTYTCTGAAGTC	
L75 F	R35	TCTAAGTGTGGATGATYTGAT	Fry et al. (2006)
H792 R		CATCATTGGRAGCCAAAGAA	

Table 2: Primer pairs used in this study. Shown for each primer is the targeted gene, the primer sequence, and the source of the primer.

Antarctic phosphatase following Glenn and Schable [29]. Fragments were sequenced following manufacturer's protocol, and sequences were analyzed on an ABI 3130 Sequence Analyzer (Applied Biosystems, Foster City, CA). When heterozygotes were detected, we first attempted to determine phase based on sample parameters using Phase [30,31]. For those whose estimated phase had a posterior probability less than 0.95, amplicons were cloned using a Qiagen cloning kit, and sequenced multiple clones per heterozygous individual to determine the exact phase.

Gene tree estimation: For each dataset, we generated a maximum likelihood estimate of genealogy for each nuclear gene and the concatenated mitochondrial data. After checking alignment by eye, DT-ModSel [32] was used to select the model of evolution that best fit each fragment, and a heuristic search was performed in PAUP* [33] to estimate the ML tree. Support for each gene tree was assessed by performing 1000 heuristic search bootstrap replicates. Using Bayes Factors [34] based on stepping-stone (ss) estimates of marginal likelihood [35], we tested three models of evolutionary rate: 1) under a strict clock, 2) under an uncorrelated relaxed clock (independent gamma rates), and 3) under a non-clock model, in MrBayes. Two stepping stone MCMC chains ($2x10^6$ generations) were run for each model for each gene fragment to ensure convergence; significance of marginal likelihood disparities between competing models were assessed following Kass and Raftery [34].

Concatenated phylogenetic analyses: Phylogeny was estimated for Nerodia were using both a likelihood and Bayesian approach. A maximum likelihood phylogeny was estimated using PAUP*. The best model for the concatenated dataset was chosen using DT-ModSel [32], subsequently a heuristic search was performed using estimated model parameters. Statistical support was assessed with 1000 heuristic search bootstrap replicates. The Bayesian phylogeny estimate was produced using BEAST 1.7 [36]. The dataset was partitioned into genes and the each gene was allowed to evolve under its own estimated substitution model (Table 3) and under an uncorrelated lognormal relaxed molecular clock model. Two indentical runs of 10^9 generations, sampling every 10^4 generations were performed, and proper mixing of the MCMC was assessed using Tracer 1.5 [37]

Species tree estimation: The methods that currently exist for estimating species trees can be placed into two categories: those which estimate a species given estimated gene trees as input (eg., MDC [10,38] and STEM [9]) and those which simultaneously estimate the gene trees and species tree (BEST [8], MrBayes [39],*BEAST [12]). The former class relies on simple algorithms to estimate the species tree, whereas the latter uses Markov chains to approximate the posterior probabilities of trees and parameters. These Bayesian methods are often computationally intensive, thus the "gene tree input" approaches (i.e., MDC, STEM) may be preferred when no a priori reason for method choice exists. However, these methods rely on the assumption that the gene trees are well-estimated, which may not be the case in many empirical datasets, and inclusion of poor estimates of gene trees into studies may decrease the accuracy of species tree estimates. Empiricists are in a difficult position, as there is no simple measure of accuracy for gene trees estimated from empirical data because the actual genealogy is unknowable. We proceed here by estimating species trees using both approaches (i.e., species tree from gene trees and simultaneous estimation of species and gene trees) and conducting several simulation studies to enhance our understanding of how accurate we can expect various methods to be given our data.

Gene	Length (bp)	var. sites	Model
BDNF	572	18	K2P
FSHR	512	25	K81uf + G
mtDNA	1133	319	GTR + G
MC1R	435	32	HKY+I
NT3	561	38	K2P + I
R35	659	28	HKY + G
Total	3857	460	N/A

Table 3: Descriptive statistics of sequenced loci. Positive value of Bayes Factor indicates magnitude of preference for relaxed clock model; at minimum, each gene shows substantial support (>3) for the relaxed clock model.

Species tree from gene trees estimation: Mesquite [40] was used to estimate the species tree by minimizing the number of deep coalescences [23]. Since Mesquite produces an estimate of the topology but not the branch lengths, STEM [9] was used to identify the ML estimate of the species tree (with branch lengths) given the gene trees. For both analyses, maximum likelihood estimates of the gene trees generated in PAUP were used.

Bayesian species tree estimation: Two methods for estimating species trees in a Bayesian framework are currently available, both of which simultaneously approximate the posterior distribution of the gene trees and the species tree, given multi-locus datasets and distributions of parameter priors. For BEST [11,41], we conducted two runs of seven chains (one for each gene tree, species tree), for 10^8 generations, sampling every 10^4 generations. We used an inverse gamma distribution with shape parameters $\alpha=3$, $\beta=0.003$ ($\Theta = 0.0015$) for the theta prior and a uniform genemu prior with bounds 0 and 5, with the upper bound corresponding to k-1 independent loci (D. Rabosky, pers. comm.). Convergence of chains was assessed using the program Tracer [37]. We also used *BEAST [12], which is implemented in the BEAST 1.7 software package. *BEAST allows use of a single or multiple MCMC chains to estimate both the species tree and the gene trees; here a single chain was allowed to run for 10^9 generations, sampling every 10^5 generations. The first 2000 samples were discarded as burn-in, and each parameter was checked for autocorrelation using the program Tracer provided in the BEAST package. A maximum clade credibility tree was created using Tree Annotater, also provided with the BEAST package.

Simulations

Quantifying the lingering effects of coalescent variance: To better understand how the coalescent processes that acted on the ancestral nodes of phylogenetic trees can influence phylogeny estimation, we conducted a series of analyses using data simulated in Mesquite 2.7.2 [40]. For each of ten ten-species topologies simulated under a birth/death process, we simulated 20 coalescent gene trees (20 alleles) contained within each species for four depths: 1N, 10N, 100N, 1000N. For each of these topologies, we made pairwise comparisons of topology (RF distances) using PAUP*. Then, for each gene tree at all depths, DNA sequence data was simulated using the average fragment lengths and models of evolution from the empirical datasets that most closely resemble the each species tree depth (Table 1). For these simulations, effective population size was set to N_e=10,000 and a generation time of 2.5 years was used [43]; estimated node ages based on fossil data [44-46] were converted to N generations. We estimated a ML tree in PAUP for each simulated dataset under the model of evolution used to simulate the data, and once again compared the topologies using RF distances. To measure how much phylogenetic estimation error affects the topology, comparisons of distributions of RF distances of the simulated gene trees and their respective estimated gene trees were performed. Finally, in order to discern the effect of gene tree

discordance on phylogenetic inference, a concatenated estimate was produced for each species tree and compared to the simulated topology using RF distances and the metric employed by Kuhner and Felsenstein [47], implemented in Ktreedist [48], which calculates relative Kuhner-Felsenstein (KF) distances for trees of differing total length. Because all fragments were simulated under the same model parameters, we did not partition the data for analysis.

Identifying the cause of gene tree incongruence: For any two gene trees estimated from independent loci, some combination of phylogenetic estimation error and coalescent uncertainty can account for observed topological discordance. Determining the relative contributions of these processes is a vital step towards determining which of the competing paradigms to use to estimate the species tree. To test whether incongruence among gene trees could be attributed solely to phylogenetic estimation error, we use the parametric bootstraps [49], an approach that utilizes simulation to construct a null distribution of the amount of phylogenetic error expected under a null model of no difference in topology across genes. We conducted pairwise test for all loci; in each we constrained the ML tree search of gene "A" to trees that matched the topology of gene "B", then measured the deterioration of the likelihood score between the topologically unconstrained and constrained trees ($-lnL_{uncon} - -lnL_{con} = \delta lnL$) using PAUP. We then simulated 1000 datasets under the model and parameters of gene "A" on the topology of gene "B" using Seq-Gen [50] and built a null distribution of δlnL to examine our test statistic. Since parametric bootstrapping depends on an adequate fit of the model of sequence evolution to the data [51], we conducted an absolute goodness-of-fit tests on each gene and corresponding model with a modified method of Sullivan et al. [52], using 1000 simulated datasets. For both tests, significance was assessed using a Bonferroni corrected α (0.05/*n* comparisons).

Gene tree support and species tree accuracy: The quality of a maximum likelihood phylogenetic estimate is typically assessed by calculating non-parametric bootstrap for each node of the phylogeny [53]. To test if our set of gene trees estimates (with assessed nodal support) were sufficiently accurate for STEM to recover the correct species tree, a series of simulations were conducted. Starting with the topology of the species tree estimated from the empirical data using *BEAST, we simulated 1000 coalescent gene trees, consisting of two alleles per species with an N_e of 10,000 (based on estimates of maximum likelihood estimates of N_e [54] from Nerodia erythrogaster; data not shown) and a total tree depth of 50N (a depth loosely based on observed mitochondrial mutation rate and overall tree length), using Mesquite. For each gene tree, sequence data was simulated using the program Seq-Gen [50], under the HKY model of sequence evolution, with nucleotide frequencies and length (551 bp) of each fragment taken from a mean of the empirical data. Each dataset was simulated on a tree with a length drawn from an exponential distribution with mean 0.05 substitutions/site (i.e., the mean tree length of the nuclear gene tree estimates). Maximum likelihood gene trees were then estimated under the same model under which they were simulated, and 100 "fastsearch" bootstrap replicates were performed for each tree. Gene tree quality was assessed in two ways. First, a measure of average nodal support (ANS) was calculated for each tree as the sum of all nodal support values above 50 divided by the total number of potentially supported nodes [18].

We used STEM to estimate a species tree from 1000 subsets of 6 randomly chosen ML gene trees; then the Kuhner-Felsenstein and Robinson-Foulds metric was calculated between the estimated and actual species tree to assess accuracy. This number was the compared with linear regression to both the mean and variance of ANS. Perl scripts were written to automate these simulations and are available by request from the author (JDM).

Results

Data collection and gene tree estimation

A total of 3857 bp of phased DNA sequence data was collected for 13 individuals representing 10 species. Gene tree estimates for each gene are shown in Figure S1. Average nodal support across all gene trees as 47.6. This number is proportional to the number of segregating sites in each gene (data not shown). Descriptive statistics for each gene, model of evolution and molecular clock model selected can be seen in Table 3; both clock-like models were greatly preferred to a non-clock model (data not shown)

Quantifying the lingering effects of coalescent variance

For simulated species trees, coalescent gene trees showed some level of discordance at all depths (Figure S2 and Figure 2), in 9/10 and 1/10 topologies at 100N and 1000N respectively, while there was some discordance among estimated gene trees in all cases. Comparisons of RF distributions of actual and estimated gene trees indicate that in most cases, the primary source of topological incongruence is phylogenetic error. In one of ten 100N and three of ten 1000N trees, concatenated ML estimates of the species tree differed from their respective simulated topologies.

Identifying the cause of gene tree incongruence

Discordance between topologies was significant ($p < 0.002$) in 18 of 25 pairwise tests (Table 4; 23/25 were "significant" prior to correction

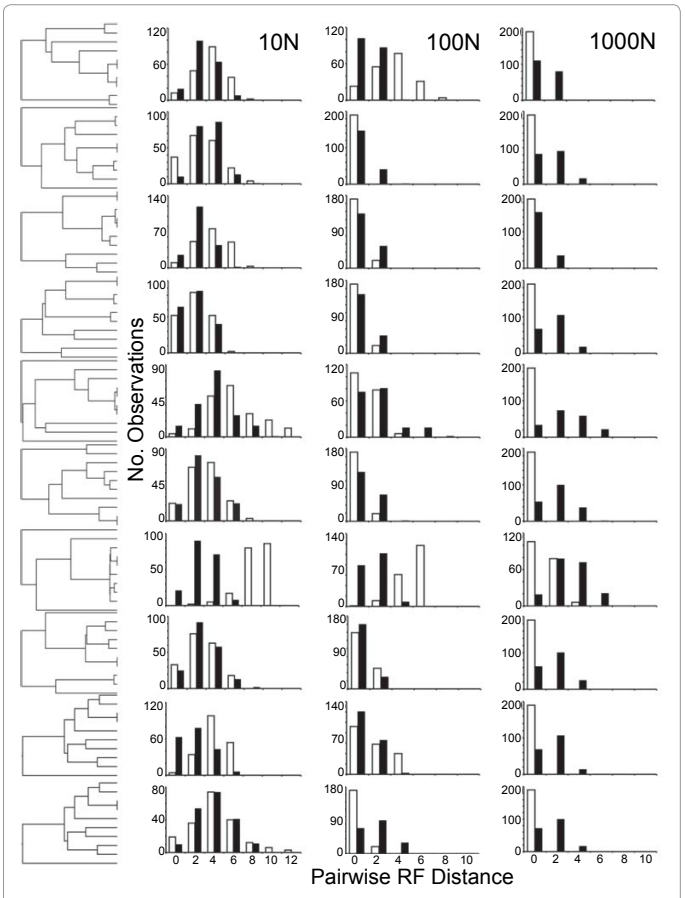

Figure 2: Distributions of Robinson-Foulds distances of actual (white) and estimated (black) gene trees. Trees on left indicate actual species tree under which coalescent genealogies were simulated.

for multiple comparisons). Comparisons testing fit of the FSHR gene to other topologies were not conducted, as the model was a poor ($p < 0.001$) fit to the data.

Gene tree support and species tree accuracy

Results of this simulation exercise (Figure S3 and Figure 3) are consistent with the prediction that accuracy of species tree estimation is directly correlated with quality of gene tree estimation. Average nodal support for the empirical dataset was 47.6, which was below the lowest simulated ANS for which the gene tree subset yielded the correct topology (Figure 3). Based on these results, results of gene tree-based species tree estimators are not presented.

Species tree estimation

The maximum clade credibility tree obtained from *BEAST can be seen in figure 4. After 10^9 generations, effective sample sizes (ESS) of all parameters were greater than 200 (the minimum suggested by the authors for publication). BEST results are not shown. After 10^8 generations, standard deviation of split frequencies for all gene tree chains were above 0.07; stationarity is assumed when these values are below 0.01. Convergence was not assessed as stationarity had not been reached.

Genes	BDNF	FSHR	MT	R35	NT3	MC1R
BDNF	-	NA	0.008	0.006	<0.001*	<0.001*
FSHR[a]	NA	-	NA	NA	NA	NA
MT	<0.001*	NA	-	0.001*	<0.001*	<0.001*
R35	0.001*	NA	0.41	-	<0.001*	0.043
NT3	<0.001*	NA	<0.001*	<0.001*	-	<0.001*
MC1R	0.002	NA	<0.001*	<0.001*	<0.001*	-

Table 4: P-values of pairwise parametric bootstrap test of topological congruence. The P-value of the pairwise comparison between each locus is shown for all comparisons.

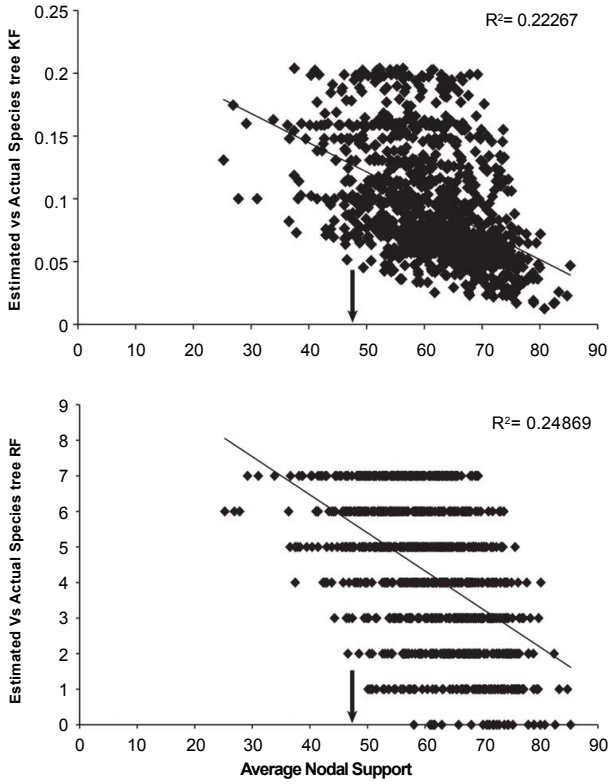

Figure 3: Comparison of average nodal support across a subset of gene trees and their utility in species tree estimation, using a) Kuhner-Felsenstein distances and b) Robinson- Foulds distances. Arrow indicates average nodal support of empirical dataset.

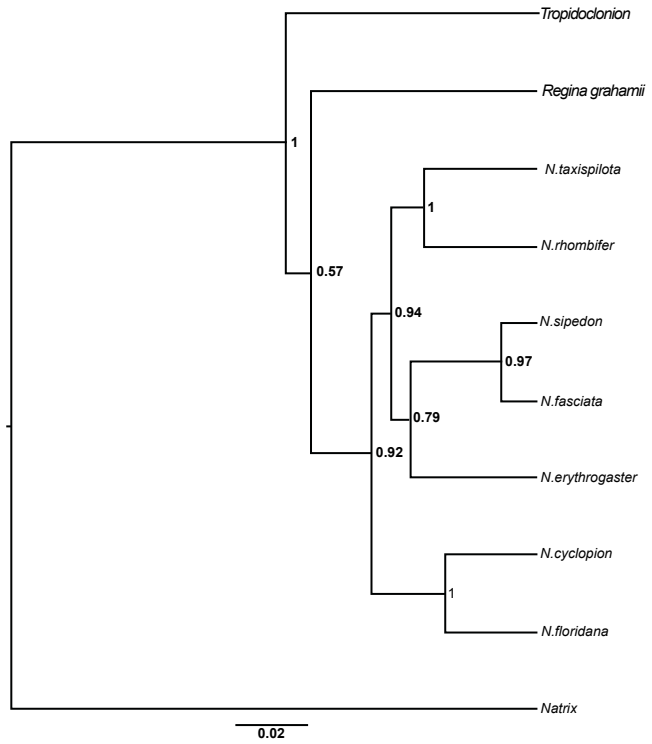

Figure 4: Maximum clade credibility tree based on 5 nuclear and one mitochondrial locus from *BEAST.

Discussion

We provide a simple framework which will allow researchers to make an *a priori* decisions about which model of phylogeny is best to use given their data: a simpler model in which all genes share a topologically identical history, or more complex models which allow genealogies to vary due to coalescent processes. In the first step, we compare the topologies of gene trees using a parametric approach. If topologies are not significantly different we could safely estimate our species tree using a concatenation approach, and additional loci can be gathered at the expense of within-species sampling. However, if gene trees exhibit an amount of incongruence that can not be attributed to phylogenetic estimation error alone, then a coalescent-based approach is preferred. For our data this is clearly the case as some 18/25 of our comparisons were able to reject the null hypothesis (i.e., that there is no difference in the topology of gene A and gene B) even using the conservative Bonferroni correction. Faced with these results, we attempted to determine if our gene trees were estimated sufficiently well to produce accurate results using STEM, since this program produces accurate estimates of species phylogenies when the gene trees are estimated without error [9,24]. We proposed a procedure based on the calculation of the average nodal support; trees that are estimated with little error will tend to have highly supported nodes as measured by non-parametric bootstrapping. Our results indicate that the ANS is low for our system, suggesting to us that we can not be sure of the accuracy of the species tree estimate from STEM. Therefore, we simultaneously

estimated the posterior distributions of our gene trees and species tree using *BEAST.

Relationships among Nerodia

Results of the *BEAST analysis recovers *Nerodia* as a monophyletic clade, with reasonable support at most nodes (Figure 4). Disagreement between our estimation and that of Alfaro [28] may be due to several factors (e.g., taxon sampling in terms of the species representing the ingroup, the total number of individuals per species, and the total number of taxa included in the analysis). Results of our analysis will likely change as taxa and individuals are added, as our ultimate goal is to estimate phylogeny of Thamnophiini. Given our findings from these preliminary data; we are presently collecting data from multiple individuals in all (~60) of the Thamnophiini species and will present a more densely-sampled and rigorous phylogeny at a later date. However, our results illustrate several striking patterns that have clear implications for empirical systematists.

Quantifying the lingering effects of coalescent variance and phylogenetic estimation error

Our first set of simulations suggest that anomalous lineage sorting due to coalescent stochasticity can result in gene tree discordance even in phylogenies with large total tree depths. This is perhaps not surprising, since discordance should be expected within any species trees that possess internode less than ~6N generations in length [55] somewhere within the tree. However, we determined that, at deeper phylogenetic levels, phylogenetic estimation error was a more common source of estimation error than coalescent uncertainty. While these would seem to imply that concatenation would perform well, it is generally the case that *both* of these processes contribute to decreased accuracy in phylogeny estimation. We advocate parametric bootstrapping as a method for determining whether the observed incongruence across multiple loci can be attributed to phylogenetic estimation error alone.

Gene tree support and species tree accuracy

One of the approaches used by empiricists to measure the quality of their phylogeny estimates is the bootstrap support of particular nodes in the phylogeny. We extend this convention and measure the overall quality of our gene tree estimates by averaging the nodal support, and then used regression to demonstrate that the accuracy of species trees estimated using STEM are correlated to the ANS. Based on results of the third set of simulations, we consider the information contained within our gene trees inadequate to estimate gene genealogies sufficiently for use in gene tree-based species tree estimators. Species tree estimates using STEM and Mesquite differed in topology from both the concatenated and *BEAST results, with STEM not recovering *Nerodia* as a monophyletic group (Figure 5a and Figure 5b). It is possible that these estimates will improve when numbers of alleles per species are increased [56] or when more substitution-rich loci are incorporated in the analysis.

Causes of discordance

While we have evoked the explanation for discordance among our gene trees as coalescent stochasticity, there are other phenomena, such as hybridization and gene duplication/extinction, which may cause similar patterns in the observed data. While a theoretical and practical framework are currently being developed that incorporate hybridization into coalescent-based analyses of phylogenetic estimation [9,28,57], this long-standing problem remains a difficult one. A key to useful incorporation of hybrid mechanisms may be a better understanding of the system-specific patterns of introgression. Within

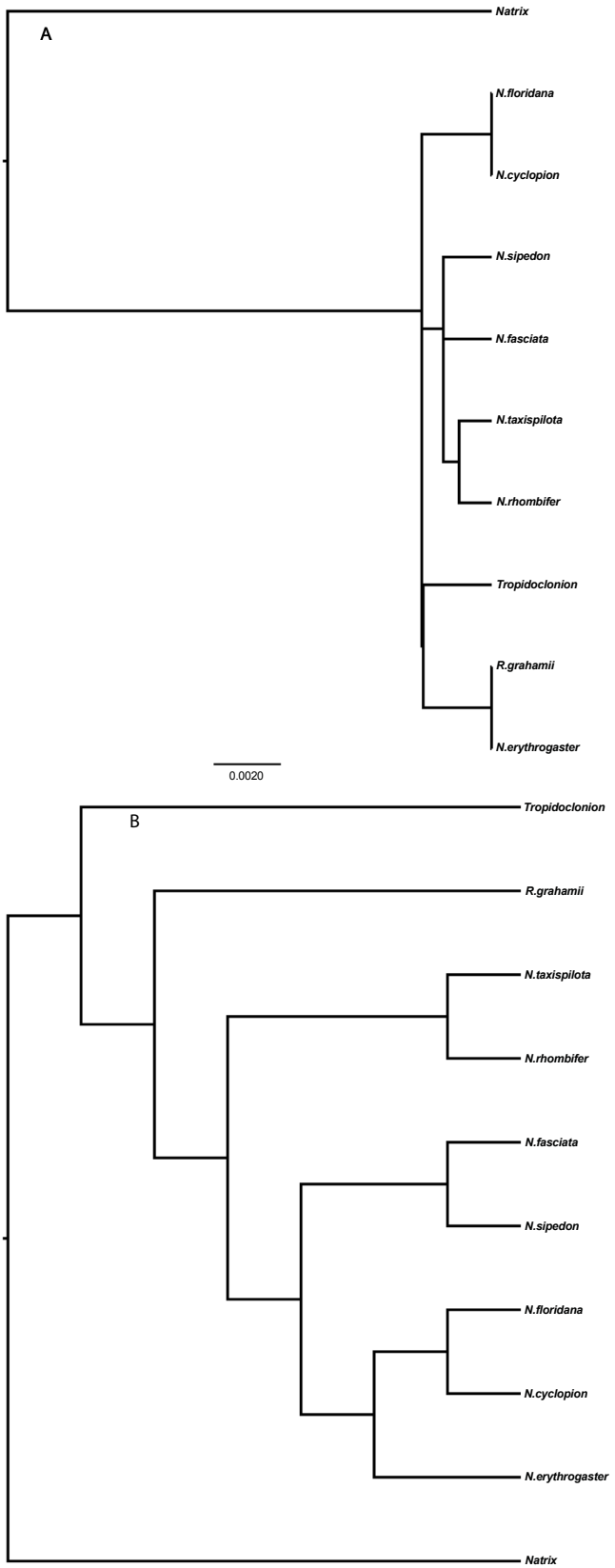

Figure 5: Species tree estimates obtained from a) STEM and b) Mesquite (MDC).

Thamnophiini, hybridization has been shown to occur between both sister and non-sister pairs of species [58,59]; consequently we cannot ignore hybridization as a possible mechanism influencing the patterns we observe.

Rate heterogeneity and species tree estimation

An advantage to using species tree estimators that require gene trees as input is that each gene tree contributes equally to the likelihood of the species tree, thus no single gene tree topology can disproportionally influence the species tree estimation. This is not the case with concatenation. Disconcertingly, concatenated multilocus phylogenetic estimations often include one or more mitochondrial loci, and the sheer bias in the number of variable sites is likely to result in an estimation in which the signal from the nuclear data is treated a noise, overwhelmed by the information in the plasmid loci [60]. Even if the mitochondrial genealogy is concordant with other gene trees or the species phylogeny, the concatenated estimate will suffer from a bias in branch length estimates [61], which can result in incorrectly estimated node ages, or bias in ancestral character state reconstruction.

Conclusions

We demonstrate a useful and direct approach to choosing among the two dominant phylogenetic models; concatenation and species tree estimation. Central to our description of the issues related to choosing among these models is the assumption that it is important to have an *a priori* expectation of model performance in order to avoid a *post hoc* evaluation of the phylogenies. We also contend that the optimal sampling design differs for these competing models; for our data we show clearly that coalescent processes are likely to produce incongruence across loci and therefore future efforts will be focused on increasing the number of individuals included in the analysis.

Acknowledgements

We thank S. Hird, M. Koopman, T. Pelletier, and N. Reid for helpful discussion. We also thank C. Austin, R. Brumfield, D. Dittmann and J. Maley for facilitating access to tissues.

References

1. Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63: 1-19.
2. Kluge AG (1989) A Concern for Evidence and a Phylogenetic Hypothesis of Relationships among Epicrates (Boidae, Serpentes). Systematic Zoology 38: 7-25.
3. Hillis DM, Huelsenbeck JP, Cunningham CW (1994) Application and accuracy of molecular phylogenies. Science 264: 671-677.
4. Carstens BC, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst Biol 56: 400-411.
5. Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2: e68.
6. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24: 332-340.
7. Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56: 17-24.
8. Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV (2009) Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53: 320-328.
9. Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25: 971-973.
10. Maddison W P, Maddison D R, Mesquite: a modular system for evolutionary analysis. 2011.
11. Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24: 2542-2543.
12. Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol Biol Evol 27: 570-580.
13. Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. Syst Biol 56: 887-895.
14. Leaché AD (2009) Species tree discordance traces to phylogeographic clade boundaries in North American fence lizards (*Sceloporus*). Syst Biol 58: 547-559.
15. 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.
16. King MG, Roalson EH (2009) Discordance between phylogenetics and coalescent-based divergence modelling: exploring phylogeographic patterns of speciation in the *Carex macrocephala* species complex. Molecular Ecology 18: 468-482.
17. Godinho R, Crespo EG, Ferrand N (2008) The limits of mtDNA phylogeography: complex patterns of population history in a highly structured Iberian lizard are only revealed by the use of nuclear markers. Mol Ecol 17: 4670-4683.
18. Wiens JJ, Kuczynski CA, Smith SA, Mulcahy DG, Sites JW Jr, et al. (2008) Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. Syst Biol 57: 420-431.
19. Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, et al. (2009) Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. BMC Evol Biol 9: 71.
20. Li C, Ortí G (2007) Molecular phylogeny of Clupeiformes (Actinopterygii) inferred from nuclear and mitochondrial DNA sequences. Mol Phylogenet Evol 44: 386-398.
21. Schopf JW (1994) Disparate rates, differing fates: tempo and mode of evolution changed from the Precambrian to the Phanerozoic. Proc Natl Acad Sci U S A 91: 6735-6742.
22. Corl A, Ellegren H (2013) Sampling strategies for species trees: the effects on phylogenetic inference of the number of genes, number of individuals, and whether loci are mitochondrial, sex-linked, or autosomal. Mol Phylogenet Evol 67: 358-366.
23. Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55: 21-30.
24. McCormack JE, Huang H, Knowles LL (2009) Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst Biol 58: 501-508.
25. Robinson DF, Foulds LR (1981) Comparison of Phylogenetic Trees. Mathematical Biosciences 53: 131-147.
26. Alfaro ME, Arnold SJ (2001) Molecular systematics and evolution of *Regina* and the thamnophiine snakes. Mol Phylogenet Evol 21: 408-423.
27. de Queiroz A, Lawson R, Lemos-Espinal JA (2002) Phylogenetic relationships of North American garter snakes (*Thamnophis*) based on four mitochondrial genes: how much DNA sequence is enough? Mol Phylogenet Evol 22: 315-329.
28. Alfaro ME (2003) Sweeping and striking: a kinematic study of the trunk during prey capture in three thamnophiine snakes. J Exp Biol 206: 2381-2392.
29. Glenn TC, Schable NA (2005) Isolating microsatellite DNA loci. Methods Enzymol 395: 202-222.
30. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978-989.
31. Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73: 1162-1169.
32. Minin V, Abdo Z, Joyce P, Sullivan J (2003) Performance-based selection of likelihood models for phylogeny estimation. Syst Biol 52: 674-683.
33. Swofford D L, PAUP* (2003) Phylogenetic Analysis Using Parsimony (*and Other Methods) Sinauer Associates: Sunderland, Massachusetts.
34. Kass R E, Raftery A E (1995) Bayes factors. J American Statistical Association 90: 773-795.
35. Xie W, Lewis PO, Fan Y, Kuo L, Chen MH (2011) Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst Biol 60: 150-160.
36. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969-1973.

37. Rambaut A, Drummond A J (2009) Tracer v 1.5.

38. Maddison W (1997) Gene trees in species trees. Systematic Biology 46: 523-536.

39. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61: 539-542.

40. Maddison WP, Maddison DR (2009) Mesquite: a modular system for evolutionary analysis.

41. Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proc Natl Acad Sci U S A 104: 5936-5941.

42. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.

43. Gibbons J W, Dorcas M E (2004) North American watersnakes: a natural history. Animal natural history series v. 8. Norman: University of Oklahoma: 438.

44. Apesteguía S, Zaher H (2006) A Cretaceous terrestrial snake with robust hindlimbs and a sacrum. Nature 440: 1037-1040.

45. Holman J A (2000) Fossil snakes of North America: origin, evolution, distribution, paleoecology. Life of the past. Bloomington: Indiana University Press: 357 p.

46. Evans SE (2003) At the feet of the dinosaurs: the early history and radiation of lizards. Biol Rev Camb Philos Soc 78: 513-551.

47. Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11: 459-468.

48. Soria-Carrasco V, Talavera G, Igea J, Castresana J (2007) The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23: 2954-2956.

49. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference, in Molecular systematics D.M. Hillis, C. Moritz, and B.K. Mable, Editors Sinauer: Sunderland, Massachusetts 407–514.

50. Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13: 235-238.

51. Goldman N, Thorne JL, Jones DT (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J Mol Biol 263: 196-208.

52. Sullivan J, Arellano E, Rogers DS (2000) Comparative Phylogeography of Mesoamerican Highland Rodents: Concerted versus Independent Response to Past Climatic Fluctuations. Am Nat 155: 755-768.

53. Felsenstein J (1985) Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. Evolution 39: 783-791.

54. Beerli P, Migrate N, (2008) Version 2.4.

55. Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. Evolution 56: 1557-1565.

56. 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.

57. Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol 75: 35-45.

58. Fitzpatrick BM, Placyk JS Jr, Niemiller ML, Casper GS, Burghardt GM (2008) Distinctiveness in the face of gene flow: hybridization between specialist and generalist gartersnakes. Mol Ecol 17: 4107-4117.

59. Mebert K (2008) Good species despite massive hybridization: genetic research on the contact zone between the watersnakes *Nerodia sipedon and N. fasciata* in the Carolinas, USA. Mol Ecol 17: 1918-1929.

60. Carstens BC, Knowles LL (2007) Shifting distributions and speciation: species divergence during rapid climate change. Mol Ecol 16: 619-627.

61. 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.