doi:10.1111/evo.14406

Improving inference and avoiding overinterpretation of hidden-state diversification models: Specialized plant breeding has no effect on diversification in frogs

Daniel S. Moen[1,2]

[1]*Department of Integrative Biology, Oklahoma State University, Stillwater, Oklahoma 74078*

[2]*E-mail: daniel.moen@okstate.edu*

Received January 22, 2021

Accepted May 4, 2021

The hidden-state speciation and extinction (HiSSE) model helps avoid spurious results when testing whether a character affects diversification rates. However, care must be taken to optimally analyze models and interpret results. Recently, Tonini et al. (TEA hereafter) studied anuran (frog and toad) diversification with HiSSE methods. They concluded that their focal state, breeding in phytotelmata, increases net diversification rates. Yet this conclusion is counterintuitive, because the state that purportedly increases net diversification rates is 14 times rarer among species than the alternative. Herein, I revisit TEA's analyses and demonstrate problems with inferring model likelihoods, conducting post hoc tests, and interpreting results. I also reevaluate their top models and find that diverse strategies are necessary to reach the parameter values that maximize each model's likelihood. In contrast to TEA, I find no support for an effect of phytotelm breeding on net diversification rates in Neotropical anurans. In particular, even though the most highly supported models include the focal character, averaging parameter estimates over hidden states shows that the focal character does not influence diversification rates. Finally, I suggest ways to better analyze and interpret complex diversification models—both state-dependent and beyond—for future studies in other organisms.

KEY WORDS: **Anura, breeding strategies, HiSSE, likelihood optimization, phytotelmata, state-dependent diversification.**

> "We hope HiSSE is viewed as a step away from [thinking that a focal trait acts in isolation], as we no longer have to necessarily focus analyses, or even interpret the results, by reference to the focal trait by itself, but can instead estimate how important it is as a component of diversification overall."
>
> Beaulieu and O'Meara (2016)

In the last 15 years, increasingly detailed methods have been developed to model the effect of character states on diversification rates (i.e., speciation and extinction rates). These models include the effects of binary traits (Maddison et al. 2007), multistate discrete traits (FitzJohn 2012), quantitative traits (FitzJohn 2010), and geographic traits (Goldberg et al. 2011) or other traits that can be split by speciation (Goldberg and Igic 2012; Magnuson-Ford and Otto 2012). The methods have seen wide use. For example, the previously cited six articles have been cited a combined 2472 times as of the writing of this article (Google Scholar, 6 May 2021).

Despite their utility, these methods also come with challenges (FitzJohn 2012; Davis et al. 2013; Maddison and FitzJohn 2015; Rabosky and Goldberg 2015). A key problem is that analyses of empirical phylogenies tend to show a statistically significant fit between diversification and the tested character even when that character has nothing to do with diversification (Rabosky and Goldberg 2015). This happens when a model with diversification rates held constant across the tree is compared to

a model that allows rates to vary with character states. When a tree has any kind of variation in diversification rates, the latter model will often be favored even if rate variation is unrelated to the tested character (FitzJohn 2012; Beaulieu and O'Meara 2016). The state-dependent model is favored simply because the alternative—constant rates across the whole tree—is untenable. Beaulieu and O'Meara (2016) introduced the hidden-state speciation and extinction model (HiSSE) as a solution to this problem. HiSSE allows heterogeneity in diversification rates beyond that explained by the focal character alone. Beaulieu and O'Meara (2016) also developed character-independent diversification (CID) models, which model diversification-rate heterogeneity completely independent of the focal character. When testing the effect of a focal character, these CID models provide a more realistic null model than the constant-rates model. Overall, HiSSE models represent reasonable alternatives to the hypothesis that the focal character alone affects rates. As indicated in the quote at the beginning of this article, the framework emphasizes considering focal characters as one of many factors that likely affect diversification (Caetano et al. 2018). Moreover, recent extensions to these methods allow the analysis of multiple focal traits (Herrera-Alsina et al. 2019; Nakov et al. 2019).

This solution is helpful and hopeful. However, care must be taken when implementing and testing these models, as their increased number of parameters make them potentially complex. For example, in the simplest possible case of a single binary character, models can have as few as three parameters (one speciation, one extinction, and one transition rate) or up to 16: two states of a focal character and two states of a hidden character allow four speciation rates, four extinction rates, and eight transition rates. With so many options, choosing one's models to test, finding their maximum-likelihood parameter estimates, and interpreting results can challenge researchers.

In a recent article published in this journal, Tonini et al. (2020; hereafter TEA) used HiSSE models to test the impact of breeding in water-filled holes in plants, called phytotelmata, on diversification rates in Neotropical frogs. They compared 47 models and found that most statistical support fell on two models in which diversification rates varied by both observed and hidden states. TEA interpreted these results as indicating higher diversification rates in taxa that breed in phytotelmata than those that do not, but they greatly underemphasized the role of the hidden states and their effect on diversification. Herein, I reanalyze TEA's data and find that even though the top models for their data include the focal character of phytotelma breeding, parameter distributions show that this character has no discernible effect on diversification rates. Moreover, I show that TEA's analyses demonstrate some of the challenges of testing HiSSE models. These challenges include finding the optimal models for the data, conducting post hoc analyses of results from HiSSE analyses, and interpreting results. I conclude by suggesting ways to improve future analyses of HiSSE models specifically and complex diversification models more generally.

Background

Nearly half of all living species of anurans (frogs and toads) occur in the Neotropics (AmphibiaWeb 2020). Most species breed terrestrially or in bodies of water at ground level (Duellman and Trueb 1986; Gomez-Mestre et al. 2012). However, specialized breeding in phytotelmata occurs across many families of anurans. Therefore, TEA studied the evolution of breeding in phytotelmata and its potential effects on diversification in Neotropical anurans. They asked two key questions: (1) What are the frequencies of changes among breeding strategies (e.g., how many times has breeding in phytotelmata evolved; is its origination more common than its reversals)? (2) Does the evolution of breeding in phytotelmata affect net diversification rates?

To address these questions, TEA considered many models for the joint evolution of diversification and phytotelma breeding. These models varied in three types of parameters: speciation rate, extinction rate, and rate of character-state transitions. They included both the observed character (phytotelma breeding) and hidden states. TEA then used the results of these analyses to estimate both the frequency of transitions and their potential effects on diversification rates. TEA also conducted complementary analyses based on the methods of Bromham et al. (2016) and Hua and Bromham (2016). Given the distinctness of these latter models and their inconclusive results for TEA's dataset, I do not discuss them in this note. I focus exclusively on the use and interpretation of the state-dependent diversification models, as these latter models drove their article's title and the major conclusions in the abstract, results, and discussion.

A comparison of TEA's data with their conclusions suggests cause for concern. They inferred that net diversification rates were higher for phytotelm-breeding taxa, yet phytotelm breeding is a relatively rare state, often associated with single terminal branches. By contrast, one would generally expect a state that increases net diversification rates to be somewhat common and associated with many branching events in the tree. Moreover, TEA concluded from additional analyses that phytotelm breeding has independently driven higher diversification rates within many families. However, as in the whole-tree analysis, many of those same families showed very few branching events associated with phytotelm breeding.

This nonintuitive relation between TEA's data and their conclusions warrants a closer look. Herein, I examine their analyses and results by focusing on three issues: obtaining maximum-likelihood estimates of model parameters, conducting down-

stream analyses with the rate estimates from these models, and interpreting results. I then perform several additional analyses to illustrate how likelihood optimization may be sensitive to search parameters. Finally, based on the results of these analyses, I suggest how to improve inference in complex diversification analyses.

Data Analyses

I first attempted to replicate TEA's HiSSE analyses to examine parameter values and better understand their results. I only conducted analyses of TEA's 1579-taxon reduced dataset, given that all their presented results are based on this dataset. I assumed a sampling fraction for the two phytotelma-breeding states based on the differences in state totals between the full, 3105-taxon dataset and the reduced, 1579-taxon dataset. For all other search parameters, I assumed default values unless otherwise indicated. I later extended these analyses to explore potential problems with TEA's results, as well as to show more general challenges in likelihood optimization of HiSSE models (see *Optimal Searching of Model Likelihoods* below). I conducted all analyses in R version 4.0.2 (R Core Team 2020) with the package *hisse* version 1.9.8 (Beaulieu and O'Meara 2016), using the function "hisse.new" and its associated functions. Note that although recent work suggested the likelihood in *hisse* improperly conditions on clade survival (Herrera-Alsina et al. 2019), subsequent analyses showed that different conditioning schemes have little effects on results (Nakov et al. 2019). Thus, I used *hisse* with the default conditioning to maximize consistency with the analyses of TEA. Finally, I provide all analysis code and results in the Supplementary Information.

Problems with Post Hoc Tests and Interpretation of Results

A key advance of the HiSSE framework is that candidate models can accommodate diversification-rate variation that is partly related or even unrelated to the focal trait (Beaulieu and O'Meara 2016; Caetano et al. 2018). This property is particularly relevant for large phylogenies, which will almost certainly have some rate variation across taxa (O'Meara 2012; Beaulieu and Donoghue 2013; Rabosky 2014; Beaulieu and O'Meara 2016). A key consequence of this framework, therefore, is that support for a state-dependent model may not necessarily indicate that a focal character is important for explaining rate variation.

TEA tested 47 total models belonging to seven classes, primarily derived from Bromham et al. (2016): full, baseline, dead-end, suicide, lonely, irreversible, and CID. They reported that two versions of the suicide class, Suicide 4 and 6, were the models most highly supported by their data. What did these models say about the effect of phytotelm breeding on diversification rates? Parameter values are often instructive, but they can also defy simple interpretation when hidden states give mixed signals. For example, with all possible combinations of the two focal states (0, 1) and two hidden states (A, B), there are four sets of rates (i.e., one rate each for 0A, 1A, 0B, and 1B). Imagine that net diversification rates are highest under 0B, lowest under 0A, and intermediate in 1A and 1B. In this situation, one cannot determine whether rates are generally higher for focal states 0 or 1, because those rates depend on the underlying (and unobserved) hidden states. In this situation, Caetano et al. (2018) suggested estimating diversification rates for the tip taxa. Because we know the focal states for tip taxa, but we do not know which hidden states they have, we can weight their focal states' diversification rates—two estimates each, one for each hidden state—by the marginal likelihood of the hidden states at the tips. For example, imagine that a hypothetical species with focal state 1 has a likelihood of 0.25 of hidden state A and 0.75 of B. If the net diversification rate of state 1A is 0.2 and state 1B is 0.4, then the estimated rate for state 1, integrating over the two hidden states, is $0.25 \times 0.2 + 0.75 \times 0.4 = 0.35$ (Caetano et al. 2018). Doing this for all taxa produces a distribution of rates for the two focal states, which can be plotted to visualize their potential differences. Such distributions also show how differing likelihoods of hidden states along branches allows HiSSE to model continuous variation in realized diversification rates, even though the focal states are discrete (Nakov et al. 2019).

To assess whether phytotelm breeding affected diversification rates, TEA presented means and variation in net diversification rates for their focal states, which came from tip rates calculated as I described above. On average, phytotelm-breeding species had higher net diversification rates than nonphytotelm breeding taxa. TEA then tested this mean difference with a *t*-test, presenting $P < 2.2 \times 10^{-16}$ to indicate a significantly higher rate in phytotelm-breeding taxa.

Yet this test and interpretation of tip rates are highly problematic for two reasons. First, TEA's *t*-test is statistically unsound because it assumed phylogenetic independence of taxa. Even if the test incorporated phylogeny, however, it would still be problematic because of a more important source of nonindependence: the rates were estimated and assigned to taxa based on a diversification analysis of the whole tree, yet TEA analyzed the tip rates as if they were data collected independently for each species. Most common phylogenetic comparative tests of continuous characters account for the nonindependence of taxa by specifying that the more closely related taxa are, the more similar their phenotypes will be (Felsenstein 1985; Hansen and Martins 1996; O'Meara 2012). This is not necessarily true in the case of state-dependent diversification, because similarity of taxa in their

estimated diversification rates depends only on whether they have the same character states. And although closely related taxa may often have the same character states, the diversification rates estimated for these states depend on their distribution across the whole tree. Species A may have an entirely different net diversification rate than its sister species B because they have different character states, but A may have the same rate as a distant species Z with the same character state. That said, hidden-state models can show phylogenetic clustering of rates, as closely related species with the same observed states also tend to have similar probabilities of the underlying hidden states, meaning they will also have similar estimated diversification rates (Caetano et al. 2018; Nakov et al. 2019). A phylogenetic test (e.g., a phylogenetic ANOVA or PGLS; Garland et al. 1993; Rohlf 2001) will thus still likely perform better than the nonphylogenetic *t*-test of TEA. However, such tests are more generally unnecessary, because the formal comparison of diversification models with AICc (or any other model-comparison framework) indicates clearly the support for one model versus another. When one model allows rates to be higher for one character state than another, and that model is highly supported statistically, then downstream tests are superfluous at best and possibly misleading.

Second, the statistical significance of the *t*-test seemingly misled TEA about the biological significance of their results. Although a very large sample size (here, 1579 species) often leads to statistical significance, that significance may not be biologically meaningful (Johnson 1999; Anderson et al. 2000; Stephens et al. 2007). TEA's figure 3—a histogram of tip net diversification rates for the focal states—shows this well (see also Fig. 3 herein, based on reanalysis). Even though the mean rate of phytotelm breeders is higher than that of nonphytotelm breeders, the distribution of rates in the latter group completely overlaps the distribution of the former. Moreover, about one-third of nonphytotelm-breeding taxa were inferred to have the highest rates across all taxa. So although some nonphytotelm-breeding taxa have lower rates, many have much higher rates. Simply plotting the data thus suggests no effect of phytotelm breeding on net diversification rates.

TEA next used a family-by-family breakdown of these same tip rates to conclude that eight of nine families showed higher rates in phytotelm-breeding taxa. Graphically, one again notices a contrast between TEA's conclusions and their data: the overlap in distributions across different families (TEA's fig. 4) is similar to the overlap in analysis of the full phylogeny (TEA's fig. 3), with the rate distributions of phytotelm-breeding taxa broadly overlapped by those with the alternative state. Phytotelm-breeding taxa have net diversification rates outside the central 50% of the nonphytotelm-breeding distribution in only three of these families (Craugastoridae, Leptodactylidae, and Microhylidae). Even in these three cases, TEA's figure 2 (a map of phytotelm breeding on their phylogeny) shows an additional problem. Although they treat these family-level analyses as independent support of their whole-tree results, the family results are clearly not independent. For example, no phytotelm-breeding lineage in Craugastoridae shows branching (i.e., speciation), and those three lineages have relatively long branches. Such data cannot support a higher net diversification rate than in nonphytotelm-breeding taxa, which in the same family have hundreds of such branching points and many short branches. This problem resembles TEA's *t*-test of the whole tree: the family-level tip rates come from an analysis of the entire tree, so the rates for phytotelm breeding in one family are linked to the rates in another. To test whether families independently supported an overall pattern, such families would need to be analyzed independently (e.g., doing HiSSE analyses at the family level). However, most families were likely too small to robustly conduct HiSSE analyses (Beaulieu and O'Meara 2016), making the utility of such analyses doubtful.

In summary, TEA used an analysis of the whole, 1579-species tree to estimate net diversification rates for each species. They then compared these rates for states of their focal character using nonphylogenetic *t*-tests, for both the whole tree and individually for nine families, which were statistically invalid for multiple reasons. TEA also plotted rate distributions that showed no difference in rates for the two focal states, neither at the whole-tree nor family levels, yet concluded in both cases that phytotelm breeding increased rates. Notably, TEA's complementary, simulation-based analyses of summary metrics (Bromham et al. 2016; Hua and Bromham 2016) were more consistent with my conclusion than theirs: both model adequacy and power simulations of three metrics showed that the dataset contained insufficient information to support any particular diversification model.

Optimal Searching of Model Likelihoods

Perhaps a more basic problem with interpreting the results of TEA is that their analyses did not reach maximum likelihood peaks for many models. Likelihood statistics function by searching for the parameter values that, for a given model, maximize the probability of the observed data given the model (i.e., the likelihood). Complex diversification models can be difficult to optimize, meaning finding the parameter values that maximize the likelihood. The search may fail to converge, or it may become trapped on a likelihood peak lower than the global optimum. Thus, one must recognize that model-fitting results may represent suboptimal, local likelihood peaks. A simple way to check this possibility is to compare nested models; a general model must have an equal or higher likelihood than a more specific nested model (Edwards 1972; Huelsenbeck and Crandall 1997).

If the nested model has a higher likelihood, then one cannot have reached the global optimum for the more general model.

This problem often occurs in TEA's model results, including their models with the highest statistical support. The most supported model Suicide 4 is nested within Full 2, and the second most supported model Suicide 6 is nested within Full 3. In both cases, the more parameter-rich models (Full 2 and 3) had lower likelihoods than the more constrained, nested models (Suicide 4 and 6), meaning that TEA's results may be an artifact of underestimating the likelihoods of competing, more complex models. In AICc-based model comparison, these complex models were doomed to fail: their likelihood was underestimated and thus could not compensate for the AICc penalty of having more parameters.

Because of these problems in estimating likelihoods, I aimed to more thoroughly search the likelihood surface of TEA's top models. I initially considered two simple changes to the search algorithm. First, I adjusted the parameter bounds to values much closer to those estimated in preliminary analyses, as doing so may improve optimization (J. Beaulieu, pers. comm.). Here, I bounded turnover rate to a maximum of 10, which is 1000 times lower than the default bound but nearly 100 times higher than any estimate in my preliminary analyses. By contrast, I increased the extinction fraction bound to 10, given that some preliminary estimates approached the default upper bound of 3. I left the transition-rate bound to the default of 100, given one preliminary estimate of 85. Second, I used a two-step optimization procedure, which starts with simulated annealing (Bertsimas and Tsitsiklis 1993) to more broadly explore the likelihood surface, then refines initial results with the standard subplex search. Simulated annealing can greatly slow likelihood optimization, but it may be more effective at finding the highest likelihood peaks. Note that both these strategies are addressed in *hisse*'s vignettes. Moreover, the two-step optimization with simulated annealing became the default likelihood search option in version 1.9.9. This version is more recent than that used by me and TEA, although the option was available at the time of our analyses.

To show the sometimes-drastic effects of these two procedures alone, I started by only considering TEA's seven Binary-State Speciation and Extinction models (BiSSE; Maddison et al. 2007). These were the simplest models (i.e., no hidden states), which one would hope are most likely to be estimated well with default search options. (Note that "default" hereafter indicates default search options in *hisse* at the time of TEA's analyses.) I started with a search under the default search options to confirm correspondence between TEA's results and the conditions under which I conducted analyses. For all but one model, the likelihoods I obtained under default searches were nearly identical to those of TEA (Table 1; see table legend for why Mk2 did not match). Next, I found that bounding parameter space markedly

Table 1. Comparison of log-likelihoods of BiSSE models under different likelihood-optimization conditions.

Model	Nested within	k	TEA	Default search	Bounded parameters	Simulated annealing	Difference	AICc	w_i
Full	None	6	–6719.07	–6719.10	–6703.55	**–6694.36**	24.74	13,400.78	0.140
Suicide	Full	5	–6728.75	–6728.60	–6702.69	**–6694.80**	33.80	13,399.65	0.247
Lonely	Full	4	–6733.57	–6733.99	–6737.07	**–6718.07**	15.93	13,444.16	0.000
Dead-end	Full, Suicide	4	–6733.11	–6733.06	–6708.07	**–6707.52**	25.54	13,423.07	0.000
Mk2*	Full, Suicide	4	–6681.71	**–6694.90**	–6699.39	**–6694.90**	0.00	13,397.83	0.613
Irreversible	Full, Mk2, Suicide	3	–6772.16	–6772.15	–6757.44	**–6753.13**	19.03	13,512.27	0.000
Baseline	All but Irreversible	3	–6738.64	–6738.58	–6723.89	**–6719.40**	19.18	13,444.82	0.000

"Nested within" = models in this table in which each row's model is nested. Note that nested models must have log-likelihoods lower than their more general counterparts (Edwards 1972; Huelsenbeck and Crandall 1997); absence of this property means that parameter estimates of the more general models are not at their maximum-likelihood values. "*k*" = number of parameters estimated for each model in my searches. "TEA" = log-likelihood published in Tonini et al. (2020). "Default search" = log-likelihood I obtained under *hisse*'s default likelihood search (in version 1.9.8), indicating similar results as TEA for all models but Mk2. (*TEA mistakenly implemented a different transition-rate matrix for Mk2 than they described in their Table S1; the matrix they used included a hidden state [i.e., a HiSSE, rather than BiSSE, model]. This difference explains the model's particularly high log-likelihood and six parameters, rather than four, in TEA's results.) "Bounded parameters" = log-likelihood estimated when adjusting parameter bounds in the search, as described in the main text. "Simulated annealing" = log-likelihood estimated when adding a simulated annealing step to the likelihood search. "Difference" = improvement in log-likelihood values between simulated annealing and the default search. Likelihoods obtained for searches with both simulated annealing and bounded parameters were identical to those from simulated annealing alone and so are excluded here. Maximum likelihood across my searches for each model is indicated in bold and is used in the AICc and associated weights (w_i) columns. In contrast, the AICc weight of Mk2 is 1.0 under default search conditions, showing the somewhat different conclusions reached when altering search conditions.

improved inference for five of seven models (Table 1). Adding simulated annealing to the likelihood search improved inference even more, with six of seven models showing a log-likelihood improvement of at least 15 units over default searches (Table 1). In terms of model comparison, this means that any one of the first five models estimated under simulated annealing (worst: Baseline BiSSE AICc = 13,444.82) would have bested all AICc values presented for these same models by TEA (their best: Full BiSSE AICc = 13,450.19). When both bounding parameter search space and using simulated annealing, I found identical likelihoods as simulated annealing alone (Table 1), which suggests that at least for simple models, simulated annealing may consistently reach the same likelihood peaks under different conditions.

Overall, neither speciation nor extinction rate is affected by phytotelm breeding in the optimal BiSSE model (Mk2). Moreover, the two other models (Full and Suicide) that have appreciable AICc support are more generalized versions of Mk2, with nearly identical likelihoods (Table 1), suggesting scarce information in the dataset to support them. Their AICc weights result almost exclusively from parameter penalization: an AICc comparison of any three models with this sample size and log-likelihood, but differing by one or two parameters, would give nearly the same model weights as here ($w_n = 0.667$, $w_{n+1} = 0.244$, $w_{n+2} = 0.089$, where n indicates a baseline number of free parameters; compare to Table 1). Finally, BiSSE models have a high propensity to spuriously assign rate variation to focal characters when no other option is given (Rabosky and Goldberg 2015; Beaulieu and O'Meara 2016), yet they did not do so here. These results all suggest little to no signal in TEA's dataset for a relationship between phytotelm breeding and diversification in frogs.

The BiSSE analyses illustrate how adjustment of search options can improve likelihood estimates of even the simplest state-dependent diversification models. However, they do not address whether or how underestimating likelihoods may have affected TEA's results for their most supported models and thus their study's conclusions. So, I next reconsidered the statistical support for TEA's top models. To conduct many searches to thoroughly explore each model's likelihood surface, I restricted my comparison to six models: the two TEA report as their top models (Suicide 4 and Suicide 6), the two more general versions of these models (Full 2 and Full 3), and two versions of CID. The latter two models corresponded to CID2 and CID4 in Beaulieu and O'Meara (2016). CID2 only allowed diversification-rate differences due to hidden states ($\tau 0A = \tau 1A \neq \tau 0B = \tau 1B$; $\epsilon 0A = \epsilon 1A \neq \epsilon 0B = \epsilon 1B$) and allowed all nondual transitions between states to differ. CID4 had four hidden states affecting turnover rates ($\tau 0A = \tau 1A \neq \tau 0B = \tau 1B \neq \tau 0C = \tau 1C \neq \tau 0D = \tau 1D$), four analogous extinction fractions that differed by hidden state, and eight transition rates (all transitions from 0 to 1 equal, regardless of hidden state; same for 1 to 0; and six rates of hidden-state change, symmetric between each pair of hidden states, regardless of focal state). Although complex, this latter model performed best of eight candidate CID4 models I considered in preliminary analyses. It also matched the complexity of the most parameter-rich model with the focal character (Full 2), an important property of null models tested in the HiSSE framework (Beaulieu and O'Meara 2016).

As above, I conducted default searches (following TEA), bounded the parameter search space, and used simulated annealing. Additionally, I conducted many searches with variable starting parameter values (Nakov et al. 2019), which can improve inference when multiple local optima occur on the likelihood surface (Rabosky and Goldberg 2015). I conducted searches for each model under 48 sets of starting parameter values. For the first 24 searches, I randomly sampled starting values from exponential distributions whose rate parameters I based on the mean parameter estimates of the highest-supported models in preliminary analyses (Full 2 and Full 3). I chose the exponential distribution for sampling because most parameter estimates in my preliminary analyses were small but occasionally large. For the remaining 24 searches, I randomly sampled starting values from uniform distributions on the interval (0,1] (for turnover and transition rates) and (0,4] (for extinction fractions). I determined these bounds from extremes found in my preliminary analyses of all 47 models considered by TEA. I estimated the likelihood of each of the six models under all 48 sets of starting parameter values. Moreover, I bounded parameter space for one set of searches, then added simulated annealing (still bounding parameter space) for another, leading to 96 total searches for each of the six models. Starting values are included in the Supporting Information.

In the first set of analyses with default starting values, likelihood searches using all default options in *hisse* produced uniformly suboptimal results. Bounding parameter space improved inference for some models, as did using simulated annealing (Table 2). Generally, relative model support varied dramatically under different search conditions (Fig. 1), and all three modifications (bounding parameters, using simulated annealing, and doing both) were necessary to find the maximum likelihood peaks across all six models (Table 2).

Different starting parameter values also affected the likelihood searches, sometimes dramatically. Across the 48 different sets of starting values, all six models had at least one set of starting values where that model was the most strongly favored (Fig. 2). Most searches favored Suicide 4, Suicide 6, or CID4, the latter two of which showed very low weights in the final results (Table 3). The optimal model across all searches, Full 2, was only the optimal model under 12.5% of the starting values (Fig. 2). The dramatic variation in likelihoods and model support across these

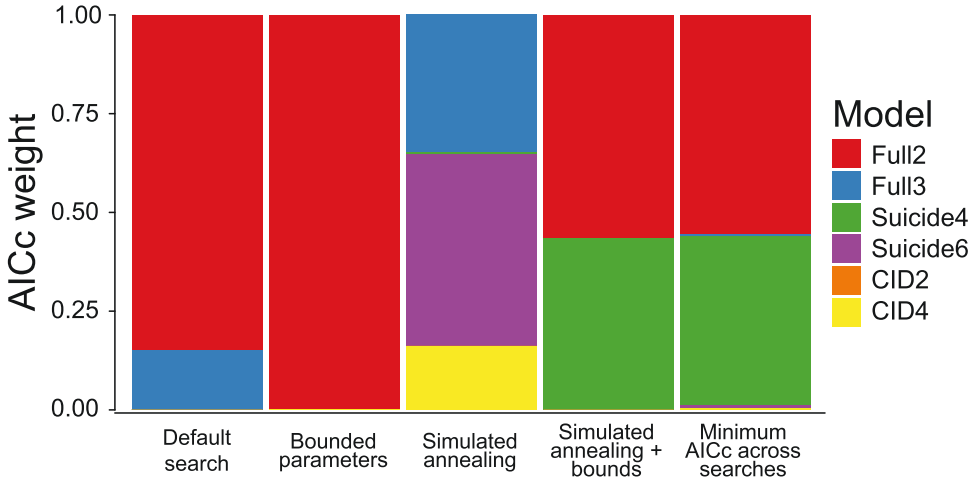

Figure 1. Variation in statistical support for six models as a function of likelihood search conditions, all estimated with default starting parameter values. "Minimum AICc across searches" results from taking the minimum (i.e., optimal) AICc for each model across the first four searches and recalculating AICc weights.

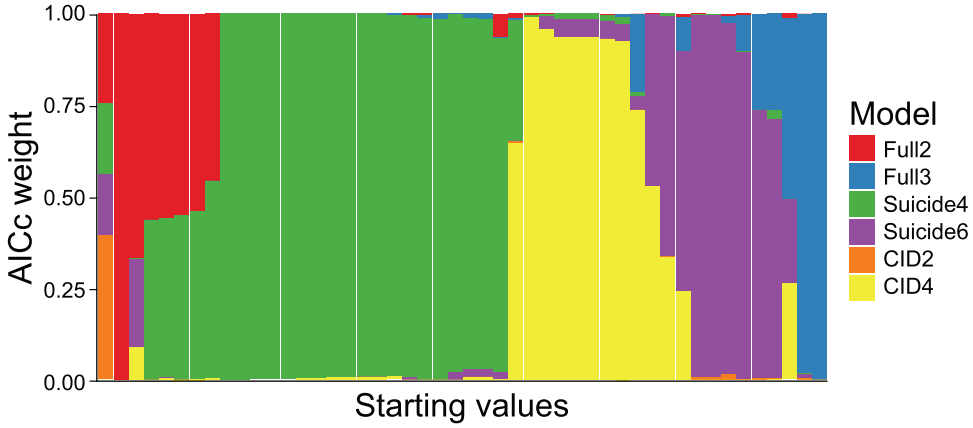

Figure 2. Variation in statistical support for six models as a function of likelihood search starting conditions, showing the drastic variation in support for different models based only on variation in starting parameter values of the likelihood search. Each bar represents the weights calculated across the six models for one of 48 sets of starting parameter values. For each model under each set of starting parameter values, I present AICc weight calculated from the highest likelihood across two searches: one in which bounds of parameter values were adjusted, and another that both adjusted bounds and also used a two-step optimization procedure with simulated annealing. Order of bars on the horizontal axis is arbitrary, as I manually grouped searches with similar model weights.

searches shows the sensitivity of searches to starting parameter values, at least for this dataset. This sensitivity is not inherently problematic when conducting HiSSE analyses; in practice, one considers different sets of starting values in an attempt to reach a model's maximum likelihood peak across all sets (e.g., Nakov et al. 2019). However, I demonstrate the dramatic variation in results here because the default starting values in *hisse* should be considered simply as representing one of these 48 sets. In other words, if all models are only estimated under the single set of default starting parameter values, one could obtain any of these outcomes, even though very few of them match the final results (Table 3).

Across all likelihood searches I conducted for this article, the models Full 2 and Suicide 4 had the lowest AICc values and thus highest AICc weights (Table 3), with comparable support for the two models. These models are similar, with Full 2 allowing additional variation in turnover rates. Model-averaged parameter estimates were hard to interpret, as nonphytotelm breeding had the highest net diversification rate under one hidden state but the lowest under the other (Table 4). Thus, I estimated diversification rates for the phylogeny's tips, as described in the previous section. These tip estimates allowed me to average over models and hidden states to address potential differences in focal states. Net diversification rates of phytotelm-

Table 2. Likelihoods obtained under different search optimization conditions with default starting parameter values.

Model	Nested within	k	TEA	Default search	Bounded parameters	Simulated annealing	Bounds + simulated annealing
Full 2	None	16	–6601.66	–6582.36	**–6579.35**	–6591.86	**–6579.35**
Full 3	Full 2	14	–6602.50	–6586.13	–6593.19	**–6585.85**	–6588.61
Suicide 4	Full 2	14	–6597.81	–6592.21	–6592.07	–6590.02	**–6581.65**
Suicide 6	Full 2, 3; Suicide 4	13	–6599.45	–6594.40	–6591.22	**–6586.53**	–6591.22
CID2	Full 2, 3; Suicide 4, 6	12	NA	–6608.63	–6604.06	–6603.96	**–6593.83**
CID4	None	16	NA	–6606.34	–6585.74	**–6584.58**	–6605.79

"Nested within" = models in this table in which each row's model is nested. "*k*" = number of parameters estimated. "TEA" = log-likelihood published in Tonini et al. (2020), who considered CID2 and CID4 models that differed from those I tested here. "Default search" = log-likelihood obtained under *hisse*'s default likelihood search (in version 1.9.8). "Bounded parameters" = log-likelihood estimated when adjusting parameter bounds in the search, as described in the main text. "Simulated annealing" = log-likelihood estimated when adding a simulated annealing step to the likelihood search. Maximum likelihood for each model across searches is indicated in bold, showing that no single strategy always produces optimal results.

breeding taxa fell squarely within the center of the distribution of nonphytotelm-breeding taxa (Fig. 3), as found by TEA (their fig. 3). Moreover, speciation and extinction rates showed the same patterns (Fig. 3). In contrast to TEA, however, I interpret these distributions as showing that net diversification, speciation, and extinction rates are all largely insensitive to variation in the focal character of phytotelm versus nonphytotelm breeding.

This conclusion seems nonintuitive. If state-dependent models that include the focal character are favored, how can the focal character *not* affect diversification rates? The answer to this question is unclear, but I suspect that the focal character functions simply to increase the number of different diversification-rate regimes that are fit to the tree (four in the top two models), which is favored because of the complex heterogeneity in rates across a large phylogeny. In other words, there is nothing special about the focal character per se, but in combination with the hidden character it allows for four sets of speciation and extinction rates to be fit to the tree. To confirm this interpretation of no direct effect of the focal character, I also conducted analyses with FiSSE, an alternative, simulation-based method for testing the hypothesis of character state-dependent diversification (Rabosky and Goldberg 2017). With FiSSE, one first calculates a test statistic analogous to the mean difference in speciation rate under the two states of a focal character. Binary characters unrelated to diversification are next simulated on the observed phylogeny. The same test statistic is then calculated for these null characters, and finally the observed test statistic is compared to the null test statistics to calculate a *P*-value. I conducted FiSSE analyses in R with code provided by Rabosky and Goldberg (2017), using most default options and 999 simulation replicates. The observed test statistic suggested slightly higher speciation rates in nonphytotelm-breeding taxa, but it fell squarely in the middle of simulated (null) test statistics, with a two-tailed $P = 0.974$. This test thus supports similar rates of speciation between the two breeding states, which confirms that phytotelm breeding has no effect on diversification rates in Neotropical anurans.

Given these results, why did HiSSE models that included the focal character (Full 2, Suicide 4) outperform CID4, which was explicitly designed to detect a complex pattern of CID (Beaulieu and O'Meara 2016; Caetano et al. 2018) and contained four sets of turnover rates and extinction fractions, as in Full 2? One possibility is that the transition-rate matrix heavily influences the likelihood of models for this dataset, and the matrices I specified for CID4 more poorly modeled the data than those of Full 2 and Suicide 4. Davis et al. (2013) showed with simulations that when one character state is much rarer than another, BiSSE (and by extension, possibly other SSE methods) tends to assign the source of state rarity to asymmetric transition rates, even when the source

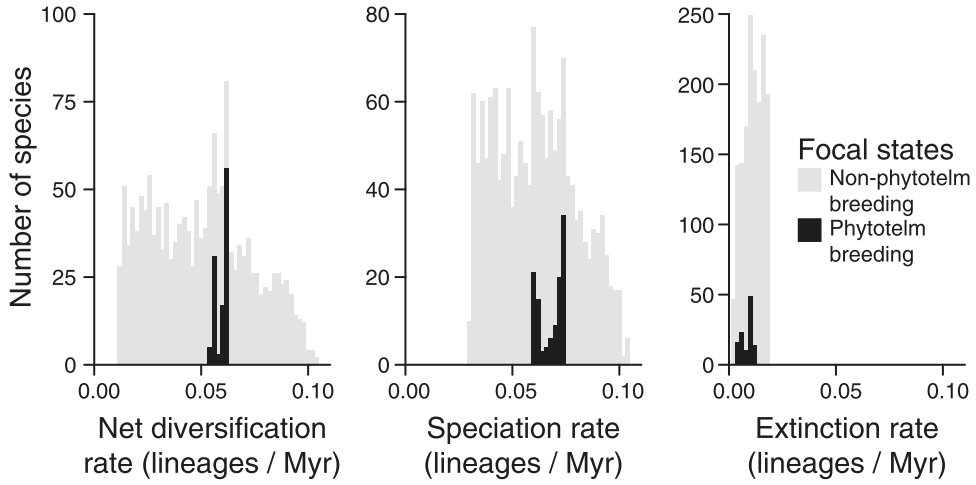

Figure 3. Histograms of estimated tip diversification rates. Rates were averaged across the top two models I estimated, Full 2 and Suicide 4 (Table 3). Note that tip branches have observed focal states (i.e., phytotelm breeding and nonphytotelm breeding) and unobserved hidden states, each with their own diversification rates (Table 4). Thus, the presented tip-rate estimates (given observed focal states) account for hidden states by weighting diversification rates by the marginal likelihood support for each hidden state (Caetano et al. 2018).

of rarity is due to differential speciation or extinction rates. They found this occurred when the rarer state had a frequency of less than 10%; here, phytotelm breeding occurs in 7.1% of taxa. Thus, this dataset may be particularly challenging for SSE methods, and its likelihood may be strongly affected by the exact form of the transition matrix. Given the size of this transition matrix—with up to 32 distinct rates under CID4—deciding how to optimally specify it will remain a future challenge.

To summarize, I found that optimization of state-dependent diversification models is particularly challenging for this dataset. This meant that even the simplest state-dependent models (BiSSE) analyzed by TEA were not estimated well under default

Table 3. Maximum likelihood estimates of HiSSE models across all searches in this article, with AICc support.

Model	Nested within	k	$ln\ L$	AICc	w_i
Full 2	None	16	–6579.35	13,191.05	0.554
Full 3	Full 2	14	–6585.85	13,199.97	0.006
Suicide 4	Full 2	14	–6581.65	13,191.57	0.427
Suicide 6	Full 2, 3; Suicide 4	13	–6586.50	13,199.23	0.009
CID2	Full 2, 3; Suicide 4, 6	12	–6592.64	13,209.48	0.000
CID4	None	16	–6584.45	13,201.24	0.003

"Nested within" = models in this table in which each row's model is nested. "k" = number of free parameters. "$ln\ L$" = maximum log-likelihood found across all analyses and conditions described in this article. "AICc" = Akaike Information Criterion adjusted for small sample size. "w_i" = AICc weights.

Table 4. Parameter estimates averaged across the top two models, Full 2 and Suicide 4.

Rate	Nonphytotelm A	Phytotelm A	Nonphytotelm B	Phytotelm B
Speciation	0.029	0.055	0.108	0.074
Extinction	0.019	0.002	0.000	0.011
Net diversification (speciation – extinction)	0.010	0.053	0.108	0.063
Net turnover (speciation + extinction)	0.048	0.057	0.108	0.085
Extinction fraction (extinction/speciation)	0.652	0.041	0.000	0.132

Parameter estimates indicate events per million years. The letters (A, B) next to each observed state (nonphytotelm breeding and phytotelm breeding) refer to hidden states.

search conditions in *hisse*. This optimization difficulty extended to TEA's top models, which included hidden states. I found that implementing three changes to the search—adjusting parameter bounds, searching with simulated annealing, and conducting searches under many different starting parameter values—greatly improved the search. The latter two strategies produced the biggest improvement. These changes led me to a different model, Full 2, as optimal for TEA's data. Nonetheless, the resulting parameter estimates and distributions of tip rates were very similar to those of TEA, showing no effect of phytotelm breeding on net diversification rates. Finally, a nonparametric alternative test gave similar results.

Ways Forward

What can be done to avoid some of the pitfalls that I highlight in this article? My critique and analyses suggest a few best practices when estimating hidden-state diversification models. However, I should first emphasize that some of the drastic variation in results under different analysis conditions may be specific to the dataset presented by TEA. For example, speciation and extinction rates were positively related in phytotelm-breeding taxa (Fig. 3; Table 4); the tight clustering of net diversification rates in this state resulted from taxa that achieved such rates by either having low speciation and low extinction rates or high values for both. This observation suggests that the distribution of phytotelm breeding on the phylogeny may be explained equally well by many combinations of speciation and extinction rates (i.e., there is a ridge in parameter space that produces similar likelihoods and net diversification rates). Moreover, the rarity of phytotelm breeding may also challenge SSE methods. As I described above, SSE methods may tend to attribute state rarity to heterogeneous transition rates rather than differential diversification rates (Davis et al. 2013). This property could explain the surprising result that phytotelm breeding was estimated with strong support as first originating relatively deep in the tree, as the ancestral state for most major clades (TEA's fig. 2). This result is biologically counterintuitive, as >90% of species do not breed in phytotelma. Moreover, other studies suggest arboreality—which may be important for breeding in the often vertically distributed phytotelma (Lannoo et al. 1987; Lehtinen et al. 2004; Ferreira et al. 2019)—originated much more recently in anurans (Moen et al. 2016; Feng et al. 2017). Thus, the properties of the phytotelm-breeding data, as well as the counterintuitive results relative to frog biology, indicate that this dataset presents significant hurdles for likelihood optimization. Such hurdles may be less substantial in other datasets.

Regardless of the details of this particular dataset and my focus on HiSSE, my results indicate some general ways forward for any state-dependent diversification analysis. First, the potential complexity of these models and searches for their global likelihood peaks necessitate transparent methods, including published R code (which TEA did provide). Reproducibility is particularly important here because so many search conditions can affect one's results. Explanation for why certain models are tested and what statistical support for them would mean biologically is important for setting forth general ideas on how diversification unfolds, to be tested for many other characters and taxa. Beyond choosing specific models, it is also important to recognize that a given dataset may be equally well explained by diversification scenarios that will never be considered or accurately estimated, including those that imply very different relationships between traits, speciation, and extinction (Louca and Pennell 2020). This is particularly true when using phylogenies of only extant species to estimate extinction rates, whose accuracy is debated (Rabosky 2010, 2016; Beaulieu and O'Meara 2015). Regardless, given time constraints, the number of models tested will be inversely related to how thoroughly one can test them, and my results show that thorough testing should not be underestimated. At a minimum, one should not end analysis if likelihoods of simpler nested models are higher than their more general counterparts, which surely indicates failing to reach the latter's global likelihood peaks (Edwards 1972; Huelsenbeck and Crandall 1997).

Second, the sometimes-complicated parameter spaces of state-dependent diversification models require diverse search strategies. I have shown here that default searches are best considered preliminary. One should consider bounding parameter space, as it can improve inference. The two-step likelihood search with simulated annealing is now *hisse*'s default; although much slower than the alternative subplex-only search, my analyses show that the resulting improved inference offsets its increased computational time. Moreover, many independent searches should be started from different parameter values. That said, I do not have a sense for the optimal way to vary these starting points, so my approach should be considered one of many options (e.g., Herrera-Alsina et al. 2019; the MuHiSSE vignette in *hisse*). Minimally, authors should describe how they decided upon the various starting parameter values and provide them in supplementary data appendices. All three of these search strategies are likely to be important for any parameter-rich diversification model, including recent developments that further expand the hidden-state approach (Herrera-Alsina et al. 2019; Nakov et al. 2019).

Third, one should not expand on diversification results with downstream methods (e.g., nonphylogenetic *t*-tests) that ignore the phylogenetic correlation in rates induced by the inference model itself. Such tests are at best not independent of the original analyses and at worst highly misleading. Furthermore, breaking down results for subtrees (e.g., families, genera) using rate estimates from the whole tree is likewise unsound. More generally, post hoc tools—such as those suggested by Caetano et al. (2018)

and Nakov et al. (2019)—should only be used for visual exploration of results, rather than as a source of data for additional tests. Ultimately, one must consider the biological relevance of results beyond their statistical significance (Yoccoz 1991; Johnson 1999; Anderson et al. 2000; Stephens et al. 2007). This includes presenting parameter estimates and their confidence intervals (Beaulieu and O'Meara 2016), an essential part of interpreting the often-complicated outputs of these models (Caetano et al. 2018). It also may include testing the adequacy of the most-supported model through simulations (Pennell et al. 2015; Rabosky and Goldberg 2015; Hua and Bromham 2016). Simulation methods of model adequacy for hidden-state models are yet to be developed and tested; they remain an important priority for future work.

Conclusions

Studies of diversification have exploded in recent years, and state-dependent models are key tools in such efforts. Herein, I stress the importance of carefully analyzing these models and interpreting results. Tonini et al. (2020) presented an impressive set of analyses, compiling data on phytotelm breeding for thousands of species of frogs and analyzing its evolution using diverse methods. They also discussed many biological hypotheses for how this character might affect diversification. Nonetheless, my reanalysis shows that phytotelm breeding has no effect on frog diversification, echoing TEA's observation that "few lineages of phytotelma-breeding frogs appear to have diversified extensively" (Tonini et al. 2020; abstract). Although some of my overall results are similar to those of TEA, I emphasize the counterintuitive result that even if the top HiSSE model(s) includes a focal character, it does not necessarily mean that the character influences diversification rates. Careful consideration of parameter estimates remains vital for avoiding such errors in interpretation, as does eschewing additional, nonphylogenetic tests on those estimates. I hope that my additional suggestions for analyses will be useful to researchers evaluating complex diversification models.

Finally, previous studies have shown that many ecological factors impact anuran diversification, including sexual-size dimorphism, climate, and microhabitat use (Pyron and Wiens 2013; De Lisle and Rowe 2015; Gómez-Rodríguez et al. 2015; Moen and Wiens 2017). My analyses show that phytotelm breeding—like other life-history traits analyzed to date (Gomez-Mestre et al. 2012)—is not among these factors that have an effect. Future work will productively consider and weight the many factors that might affect diversification (e.g., Moen and Wiens 2017; Hernández-Hernández and Wiens 2020). Recent developments of state-dependent models that allow analysis of multiple characters will no doubt play an important role in this endeavor (Caetano et al. 2018; Herrera-Alsina et al. 2019; Nakov et al. 2019).

AUTHOR CONTRIBUTIONS

DSM conducted analyses and wrote this article.

ACKNOWLEDGMENTS

I thank R. Bonett, A. Pyron, and J. Tonini for comments on an earlier version of this manuscript. I thank J. Beaulieu for help troubleshooting HiSSE analyses. I also thank Associate Editor D. Rabosky and two anonymous reviewers for constructive comments that improved this article. I conducted this work on the High Performance Computing Center facilities of Oklahoma State University at Stillwater, supported by the National Science Foundation (MRI-1531128). Preparation of this manuscript was also supported by National Science Foundation award DEB-1655812.

DATA ARCHIVING

Data, R code, and results files are provided as Supporting Information S1–S13. Data and results are deposited on the Dryad Digital Repository (Moen 2021; http://doi.org/10.5061/dryad.w9ghx3fpp), while R code is available by link on Dryad but hosted on Zenodo (https://doi.org/10.5281/zenodo.4708174).

LITERATURE CITED

AmphibiaWeb. 2020. AmphibiaWeb: information on amphibian biology and conservation. Available at http://amphibiaweb.org. Accessed October 5, 2020.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. J. Wildlife Manage. 64:912–923.

Beaulieu, J. M. and M. J. Donoghue. 2013. Fruit evolution and diversification in campanulid angiosperms. Evolution 67:3132–3144.

Beaulieu, J. M. and B. C. O'Meara. 2015. Extinction can be estimated from moderately sized molecular phylogenies. Evolution 69:1036–1043.

———. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65:583–601.

Bertsimas, D. and J. Tsitsiklis. 1993. Simulated annealing. Stat. Sci. 8:10–15.

Bromham, L., X. Hua, and M. Cardillo. 2016. Detecting macroevolutionary self-destruction from phylogenies. Syst. Biol. 65:109–127.

Caetano, D. S., B. C. O'Meara, and J. M. Beaulieu. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution 72:2308–2324.

Davis, M., P. E. Midford, and W. P. Maddison. 2013. Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. BMC Evol. Biol. 13:38.

De Lisle, S. P. and L. Rowe. 2015. Independent evolution of the sexes promotes amphibian diversification. Proc. R. Soc. B Biol. Sci. 282:20142213.

Duellman, W. E. and L. Trueb. 1986. Biology of amphibians. Johns Hopkins Univ. Press, Baltimore, MD.

Edwards, A. W. F. 1972. Likelihood. Cambridge Univ. Press, Cambridge, U.K.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Feng, Y. J., D. C. Blackburn, D. Liang, D. M. Hillis, D. B. Wake, D. C. Cannatella, and P. Zhang. 2017. Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs

at the Cretaceous-Paleogene boundary. Proc. Natl. Acad. Sci. USA 114:E5864–E5870.

Ferreira, R. B., A. T. Mônico, C. Z. Zocca, M. T. T. Santos, F. C. F. Lírio, J. F. R. Tonini, L. T. Sabagh, R. S. Cipriano, C. Waichert, M. L. Crump, et al. 2019. Uncovering the natural history of the bromeligenous frog *Crossodactylodes izecksohni* (Leptodactylidae, Paratelmatobiinae). South Am. J. Herpetol. 14:136–145.

FitzJohn, R. G. 2010. Quantitative traits and diversification. Syst. Biol. 59:619–633.

———. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods Ecol. Evol. 3:1084–1092.

Garland, T. J., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. Syst. Biol. 42:265–292.

Goldberg, E. E. and B. Igic. 2012. Tempo and mode in plant breeding system evolution. Evolution 66:3701–3709.

Goldberg, E. E., L. T. Lancaster, and R. H. Ree. 2011. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. Syst. Biol. 60:451–465.

Gomez-Mestre, I., R. A. Pyron, and J. J. Wiens. 2012. Phylogenetic analyses reveal unexpected patterns in the evolution of reproductive modes in frogs. Evolution 66:3687–3700.

Gómez-Rodríguez, C., A. Baselga, and J. J. Wiens. 2015. Is diversification rate related to climatic niche width? Global Ecol. Biogeogr. 24:383–395.

Hansen, T. F. and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417.

Hernández-Hernández, T. and J. J. Wiens. 2020. Why are there so many flowering plants? A multiscale analysis of plant diversification. Am. Nat. 195:948–963.

Herrera-Alsina, L., P. van Els, and R. S. Etienne. 2019. Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. Syst. Biol. 68:317–328.

Hua, X. and L. Bromham. 2016. PHYLOMETRICS: an R package for detecting macroevolutionary patterns, using phylogenetic metrics and backward tree simulation. Methods Ecol. Evol. 7:806–810.

Huelsenbeck, J. P. and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Ann. Rev. Ecol. Syst. 28:437–466.

Johnson, D. H. 1999. The insignificance of statistical significance testing. J. Wildlife Manage. 63:763–772.

Lannoo, M. J., D. S. Townsend, and R. J. Wassersug. 1987. Larval life in the leaves: arboreal tadpole types, with special attention to the morphology, ecology, and behavior of the oophagous *Osteopilus brunneus* (Hylidae) larva. Fieldiana Zool. 38:1–31.

Lehtinen, R. M., M. J. Lannoo, and R. J. Wassersug. 2004. Phytotelm-breeding anurans: past, present and future research. Misc. Pub. Mus. Zool. Univ. Mich. 193:1–9.

Louca, S. and M. W. Pennell. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature 580:502–505.

Maddison, W. P. and R. G. FitzJohn. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst. Biol. 64:127–136.

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.

Magnuson-Ford, K. and S. P. Otto. 2012. Linking the investigations of character evolution and species diversification. Am. Nat. 180:225–245.

Moen, D. S. 2021. Supplementary files from: improving inference and avoiding over-interpretation of hidden-state diversification models: specialized plant breeding has no effect on diversification in frogs. Evolution, Dryad Digital Repository. https://doi.org/10.5061/dryad.w9ghx3fpp.

Moen, D. S. and J. J. Wiens. 2017. Microhabitat and climatic niche change explain patterns of diversification among frog families. Am. Nat. 190:29–44.

Moen, D. S., H. Morlon, and J. J. Wiens. 2016. Testing convergence versus history: convergence dominates phenotypic evolution for over 150 million years in frogs. Syst. Biol. 65:146–160.

Nakov, T., J. M. Beaulieu, and A. J. Alverson. 2019. Diatoms diversify and turn over faster in freshwater than marine environments. Evolution 73:2497–2511.

O'Meara, B. C. 2012. Evolutionary inferences from phylogenies: a review of methods. Ann. Rev. Ecol. Evol. Syst. 43:267–285.

Pennell, M. W., R. G. FitzJohn, W. K. Cornwell, and L. J. Harmon. 2015. Model adequacy and the macroevolution of angiosperm functional traits. Am. Nat. 186:E33–50.

R Core Team. 2020. R: a language and environment for statistical computing, version 4.0.2. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Pyron, R. A. and J. J. Wiens. 2013. Large-scale phylogenetic analyses reveal the causes of high tropical amphibian diversity. Proc. R. Soc. B Biol. Sci. 280:20131622.

Rabosky, D. L. 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution 64:1816–1824.

———. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS ONE 9:e89543.

———. 2016. Challenges in the estimation of extinction from molecular phylogenies: a response to Beaulieu and O'Meara. Evolution 70: 218–228.

Rabosky, D. L. and E. E. Goldberg. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Syst. Biol. 64:340–355.

———. 2017. FiSSE: a simple nonparametric test for the effects of a binary character on lineage diversification rates. Evolution 71: 1432–1442.

Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160.

Stephens, P. A., S. W. Buskirk, and C. Martinez del Rio. 2007. Inference in ecology and evolution. Trends Ecol. Evol. 22:192–197.

Tonini, J. F. R., R. B. Ferreira, and R. A. Pyron. 2020. Specialized breeding in plants affects diversification trajectories in Neotropical frogs. Evolution 74:1815–1825.

Yoccoz, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bull. Ecol. Soc. Am. 72:106–111.

Associate Editor: D. Rabosky
Handling Editor: T. Chapman