FiSSE: a simple non-parametric test for the effects of a binary character on lineage diversification rates

Daniel L. Rabosky[1, 2]

Emma E. Goldberg[3]

[1] Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, MI 48103

[2] Email: drabosky@umich.edu

[3] Department of Ecology, Evolution, and Behavior, University of Minnesota, Saint Paul, Minnesota, 55108

These authors contributed equally to this work.

Running title: Test for trait-dependent diversification

Key words: Speciation, extinction, BiSSE, species selection, key innovation, macroevolution

Data Archival Location: Dryad doi: #######

Abstract

It is widely assumed that phenotypic traits can influence rates of speciation and extinction, and several statistical approaches have been used to test for correlations between character states and lineage diversification. Recent work suggests that model-based tests of state-dependent speciation and extinction are sensitive to model inadequacy and phylogenetic pseudoreplication. We describe a simple non-parametric statistical test ("FiSSE") to assess the effects of a binary character on lineage diversification rates. The method involves computing a test statistic that compares the distributions of branch lengths for lineages with and without a character state of interest. The value of the test statistic is compared to a null distribution generated by simulating character histories on the observed phylogeny. Our tests show that FiSSE can reliably infer trait-dependent speciation on phylogenies of several hundred tips. The method has low power to detect trait-dependent extinction but can infer state-dependent differences in speciation even when net diversification rates are constant. We assemble a range of macroevolutionary scenarios that are problematic for likelihood-based methods, and we find that FiSSE does not show similarly elevated false positive rates. We suggest that non-parametric statistical approaches, such as FiSSE, provide an important complement to formal process-based models for trait-dependent diversification.

Introduction

Rates of lineage diversification are widely assumed to depend on biological properties of the lineages themselves. Mating system, trophic ecology, defense syndromes, population structure, and many other organismal and population-level attributes have been hypothesized to influence the rate at which lineages undergo speciation and extinction (Arnold and Fristrup 1982; Jablonski 2008; Ng and Smith 2014). Several statistical frameworks have been used to test hypotheses about the effects of traits on diversification rates, including non-parametric or semi-parametric sister-clade contrasts and probabilistic state-dependent diversification (SDD) models. Sister clade contrasts involve comparing the species richness of sister clades that show fixed differences in a character state of interest; the focal trait is inferred to influence diversification if a particular character state is consistently associated with higher (or lower) species richness (Mitter et al. 1988; Hodges 1997; Farrell 1998; Coyne and Orr 2004). Formal

SDD models describe a stochastic process that can jointly give rise to a phylogeny and character state data, where character states potentially differ in rates of speciation and/or extinction (Maddison et al. 2007; FitzJohn et al. 2009; FitzJohn 2010; Goldberg et al. 2011). These likelihood-based approaches enable researchers to conduct statistical comparisons of models where character states influence diversification to alternative models where the phenotypic evolutionary process is decoupled from speciation and extinction rates.

Sister-clade contrasts are intuitively appealing but suffer from several limitations (Maddison 2006; Kafer and Mousset 2014). Most importantly, asymmetric rates of character change can lead to ascertainment biases whereby one character state is consistently associated with increased species richness when sister clades are selected for analysis based on fixed trait differences (Maddison 2006). This effect can be observed even in the absence of any true relationship between traits and diversification. BiSSE (Maddison et al. 2007) and related SDD models were developed in part to enable researchers to disentangle asymmetries in character change from state-dependent diversification.

However, recent work has found that statistical comparison of SDD models is prone to incorrect inferences of state-dependent diversification, due to both phylogenetic pseudoreplication and model inadequacy (Maddison and FitzJohn 2015; Rabosky and Goldberg 2015). Rabosky and Huang (2015) proposed a structured permutation test for state-dependent diversification, but the method is only applicable to phylogenies that are large enough to infer lineage-specific variation in diversification rates independently of phenotypic information (using rate-shift models such as BAMM or MEDUSA; Alfaro et al. 2009; Rabosky 2014), and it has little power for rapidly-evolving traits. Beaulieu and O'Meara (2016) proposed an expanded SDD modeling framework that compares the fit of state-dependent models to those of a more complex set of models that includes the effects of latent variables on diversification rates. Their approach avoids issues associated with comparing state-dependent models to trivial (and likely incorrect) null hypotheses, but it may still be susceptible to phylogenetic pseudoreplication and model inadequacy if the "true" model is substantially different from those in the candidate set.

Here, we introduce a simple method for testing the effects of a binary character on diversification. The method, which we refer to as FiSSE (Fast, intuitive State-dependent Speciation-Extinction analysis), is effectively non-parametric and does not use an underlying model for character change or species diversification. We assess its performance on datasets simulated with or without state-dependent diversification and character change asymmetries, and also on simulated and empirical datasets known to reveal weaknesses of formal SDD models. We conclude that FiSSE is robust to both phylogenetic pseudoreplication and model inadequacy, and that it can be useful on even moderately-sized trees.

Methods

Description of FiSSE

FiSSE is a simple statistical test for the effects of a binary character on rates of lineage diversification. It provides estimates of "quasi-parameters" that are correlated with, but not identical to, underlying rates of speciation. The quasi-parameters can be interpreted intuitively because they are related to the distributions of branch lengths associated with each character state. Significance of the quasi-parameters is assessed by comparing the observed values to a null distribution that is generated through simulation. Construction of the null distribution is relatively fast and is limited primarily by the speed at which trait histories can be simulated on the tree.

The test involves several steps. To obtain the test statistic, we first compute an estimate of speciation rate for each tip in the tree using the inverse equal splits measure proposed by Jetz et al (2012; outlined below). We then compute the mean tip speciation rate associated with each character state, and the difference in these mean values is the test statistic. To obtain the null distribution of that test statistic, we first fit a one-parameter Markov model of symmetric character change (Mk1 model; Jukes and Cantor 1969; Lewis 2001) to the observed data. We then simulate histories of neutral characters on the empirical phylogeny using this parameter value. For each simulation, we count the number of inferred character changes under a parsimony criterion, and we accept only those simulations that have a parsimony score that is similar to the score computed for the empirical data. This procedure generates histories for characters that do not affect diversification, but that contain approximately the same number of state changes as the observed data. The test statistic is then computed for each simulated dataset, and the two-tailed significance is simply the proportion of simulations with values more extreme than the observed test statistic.

The equal splits (ES) measure was originally proposed as an index of evolutionary isolation that could be computed for each tip in a resolved phylogenetic tree (Redding and Mooers 2006). The ES metric for a given tip is computed as a weighted sum of branch lengths between the tip and the root of the tree. The weights are a simple downweighting of each successively more rootwards branch by a factor of 0.5. The ES metric for the *i*'th tip is computed as

$$ES_i = \sum_{j=1}^{N_i} l_j \frac{1}{2^{j-1}}$$

where N_i is the number of branches connecting the tip to the root, and l_j is the length of the *j*'th branch.

Jetz et al (2012) demonstrated that, under a pure-birth process, the reciprocal of ES is an estimate of the speciation rate, λ, and they proposed that this be used as an estimate of the tip-specific diversification rate. They referred to 1 / ES as the DR ("Diversification Rate") statistic. However, the metric does not explicitly account for extinction and is simply a measure of the splitting rate for surviving lineages. Jetz et al. (2012) and Belmaker and Jetz (2015) also noted that the metric is more closely related to the speciation rate than to net diversification. Hence, the reciprocal of ES is a quasi-parameter that imperfectly estimates the speciation rate at the tips of the tree. We refer to this quantity as the "inverse equal splits measure." For the i'th tip, we represent it symbolically as Λ_i^t, where the superscript indicates that it is the value for a single tip. The test statistic for FiSSE is computed as

$$\Delta\Lambda = \Lambda_0 - \Lambda_1 = \frac{1}{N_0} \sum_{i \in \{x_0\}} \frac{1}{\Lambda_i^t} - \frac{1}{N_1} \sum_{i \in \{x_1\}} \frac{1}{\Lambda_i^t}$$

where the N_k tips in state *k* comprise the set x_k, and Λ_k denotes the mean rate across all N_k tips in that state. The direction of the comparison is, of course, arbitrary (either character state can assume the label of 0 or 1).

To construct the null distribution, we first compute the number of parsimony changes on the observed data, denoted C_{obs}. We then estimate the transition rate *q* under a symmetric Mk1 model. A single simulation consists of the following steps. We first choose a root state (0, 1) with equal probability and simulate character histories with transition rate *q*. We then count the number of parsimony changes for each simulated character history, $C_{sim,}$ and we compute the absolute value of the difference between this quantity and the observed number of changes. If $|(C_{sim} - C_{obs})| / C_{obs}$ is less than some pre-defined threshold, we accept the simulation as valid. For the analyses below, we used a threshold of 0.1, thus requiring that simulated datasets have a parsimony score that is within 10% of the value of the empirical data. Using a threshold greater than 0 avoids imposing overly restrictive simulation conditions on the generation of the null distribution. For example, for datasets with large numbers of character transitions, requiring that C_{sim} exactly equal C_{obs} imposes a high computational burden on the analysis, as a high percentage of simulations will be rejected. Our additional tests (Supplementary Fig. S1) indicate that relaxing the threshold to 25% causes little difference in the significance assessment.

FiSSE is implemented in R. The analyses described below use diversitree (FitzJohn 2012) for simulation of discrete characters and phangorn (Schliep 2011) for reconstruction of character changes under parsimony. Code to reproduce these analyses is available through the Dryad submission that accompanies this article (DOI: ######) and through a dedicated GitHub repository (https://github.com/macroevolution/fisse).

Although FiSSE does not formally model the association between character states and diversification, the mean inverse ES metric computed for each character state, Λ_k, is an intuitive quantity related to the average branch length associated with a particular character state. Obviously, we do not know the true character states except at the tips of the tree. However, the weighting of the ES calculation ensures that branches closest to the tips contribute most to the overall value of ES, and it is this portion of the tree that is most likely to be identical in character state to the tip observations.

Performance of FiSSE

We assessed performance of FiSSE using three general strategies. First, we analyzed datasets simulated under parameters that loosely match those used in the original assessment of BiSSE's performance (Maddison et al. 2007; FitzJohn et al. 2009). Second, we repeated Rabosky and Goldberg's (2015) analysis of neutral characters simulated on empirical avian phylogenies under a range of transition rates. Finally, we performed a double-blind assessment of FiSSE and BiSSE on a wide range of datasets. The researcher who performed the FiSSE analysis was not provided with information about how the data were generated, and the researcher generating the datasets was not provided with information about the FiSSE algorithm.

For the first set of analyses, with parameters similar to those used by Maddison et al. (2007), we simulated datasets with (i) no state-dependent diversification ($\lambda_0 = \lambda_1, \mu_0 = \mu_1$), (ii) state-dependent speciation only ($\lambda_0 \neq \lambda_1$), (iii) state-dependent extinction only ($\mu_0 \neq \mu_1$), and (iv) state-dependent speciation and extinction such that net diversification rates were equal for both character states ($\lambda_0 - \mu_0 = \lambda_1 - \mu_1$). For each scenario, we simulated 1000 phylogenies using diversitree (FitzJohn 2012), 200 each for n = 100, 200, 300, 400, or 500 surviving tips per tree. For simulations without state-dependence, we used $\lambda_0 = \lambda_1 = 0.1$, $\mu_0 = \mu_1 = 0.03$, and $q_{01} = q_{10} = 0.01$. For simulations with state-dependent speciation only, we considered a two-fold increase in the rate of speciation, such that $\lambda_0 = 0.1$ and $\lambda_1 = 0.2$, and with all other parameters fixed to the values given above. For state-dependent extinction, we considered a scenario where the net diversification rate increased twofold, but where the increase was mediated solely by a change in the extinction rate: $\lambda_0 = \lambda_1 = 0.2$, $\mu_0 = 0.1$, and $\mu_1 = 0$. Finally, we considered two parameterizations where the net diversification rate was equal across character states but where the turnover rate varied. These parameterizations were $\lambda_0 = 0.1$, $\lambda_1 = 0.2$, $\mu_0 = 0.03$, $\mu_1 = 0.13$; and $\lambda_0 = 0.1$, $\lambda_1 = 0.3$, $\mu_0 = 0.03$, $\mu_1 = 0.23$. These scenarios involve 2x and 3x increases in the rate of speciation for state 1, respectively, but the net diversification rate is 0.07 for each state.

We then simulated the evolution of neutral characters on phylogenies sampled from a much larger time-calibrated phylogeny of birds (Jetz et al. 2012), following the analyses described in Rabosky and Goldberg (2015). We analyzed the maximum-clade credibility tree for the "Hackett" backbone of the Jetz et al (2012) phylogeny, after excluding all species that lacked genetic data; the resulting phylogeny contained 6670 taxa, about two-thirds of living bird species. We identified all rooted subtrees from this phylogeny that contained 200 to 500 descendant taxa, for a total of 60 subtrees. We simulated binary traits on each of these phylogenies under five phenotypic evolutionary scenarios, after rescaling the crown age of each subtree to 1.0 time units before the present. The first four scenarios specified a symmetric Markov model of character evolution, with transition rates of $q = 0.01$, $q = 0.1$, $q = 1$, and $q = 10$. The final scenario involved asymmetric rates of character change, with $q_{01} = 0.02$ and $q_{10} = 0.005$. This final scenario is especially important for FiSSE, whose null distribution is generated under a symmetric model of character change. For each simulation, we required that the rarer character state obtain a frequency of at least 10%. A total of 10 datasets were simulated for each of the 60 subtrees under the five evolutionary scenarios. Each dataset was analyzed with FiSSE as described above.

Double-blind assessment of FiSSE and BiSSE

Our third set of analyses is a double-blind assessment of FiSSE's performance. One author (DLR) implemented FiSSE but did not reveal any details of the method to the other author (EEG), other than to describe it as a statistical test for binary trait-dependent diversification. EEG then generated trial datasets under 42 distinct state-dependent and non-state-dependent diversification scenarios and parameterizations, but DLR had no knowledge of the generating scenarios. EEG provided DLR with a full set of phylogenies and character data, stripped of any attributes that might identify the simulation conditions or empirical sources. Each of the 42 simulation scenarios consisted of a set of phylogenies with binary trait data, with 50 such datasets per scenario. DLR analyzed all 2000 datasets with FiSSE and BiSSE and provided EEG with a summary of the results. EEG then prepared a report on the relative performance of the two approaches, focusing in particular on statistical power and the rate at which the methods inferred state-dependent diversification when no such relationship was present in the data. Following initial assessment and peer review of this article, we created and analyzed an additional eight datasets that were designed to test the limits of the FiSSE approach. These sets do not follow the double-blind procedure, and they are distinguished in the results.

Simulation scenarios and analysis summaries are provided in Tables S1 and S2. Set numbers are based on the results and thus were assigned after analysis. The test sets themselves are available from the Dryad submission that accompanies this article, and the generating scripts are available from the Phylogenetic Comparative Methods Benchmark database (PhyCoMB; https://github.com/eeg/PhyCoMB). As a partial list, the testing scenarios included (i) true BiSSE scenarios with fast, slow, and asymmetric rates of character change; (ii) cladogenic state-dependent change; (iii) continuous-valued state-dependent simulations with traits recoded as binary; and (iv) neutral character simulations (fast, moderate, slow, irreversible, heterogeneous, and continuous-valued recoded as binary) on both simulated and empirical phylogenies. For neutral character simulations (no state-dependence), the phylogenies on which characters were simulated included diversity-dependence, diversification rate shifts, mass extinctions, and other heterogeneity in speciation and extinction rates. In general, the conditions explored here (Table S1 -S2) greatly expand upon the general set of conditions from our previous assessment of BiSSE's performance (Rabosky & Goldberg 2015).

For comparison with FiSSE's performance, we also fit four BiSSE models to each simulated dataset: (i) the full 6 parameter BiSSE model; (ii) a five-parameter constrained model with $\mu_0 = \mu_1$, (iii) a five-parameter model with $\lambda_0 = \lambda_1$, and (iv) a four-parameter, character-independent model with $\lambda_0 = \lambda_1$ and $\mu_0 = \mu_1$. We performed a likelihood ratio test of the best state-dependent model against the four-parameter model with no state-dependence. Beaulieu and O'Meara (2016) described weaknesses in this commonly-used model comparison approach and suggested instead focusing on parameter estimates. We therefore also assessed the significance of state-dependent diversification for each dataset using MCMC to simulate posterior distributions of net diversification rates for each character state ($r_i = \lambda_i - \mu_i$) under the full BiSSE model. We summarized significance as the posterior probability (two-tailed) that $|r_1 - r_0| > 0$.

We also performed a second set of analyses where we expanded the candidate model set to include two "hidden-state" models (Beaulieu and O'Meara 2016). Beaulieu and O'Meara (2016) noted that support for an SDD model when the true generating process has no association between the character and speciation or extinction rate is not necessarily a "type I error" or "false positive" if the null non-SDD model is itself incorrect. This is a valid concern for BiSSE model comparisons when the data were not generated under the constant-diversification ($\lambda_0 = \lambda_1$, $\mu_0 = \mu_1$) process, and many of our testing sets included diversification rate heterogeneity that was unlinked to the focal character. Following Beaulieu and O'Meara (2016), we included a null model (CID-2) that allowed diversification rates to vary across the tree through association with an unobserved binary character state. We also included a full HiSSE model, which allows unobserved substates within each of the observed character states to influence the diversification process. Unlike CID-2, HiSSE is a state-dependent model because the observed states of the focal character are used to explain (in part) diversification rate differences. Both the CID-2 and HiSSE models had three transition rates, so that transitions between the states of the focal character were asymmetric and independent of the hidden state, and transitions between the hidden states were symmetric. We fit the CID-2 and HiSSE models to each test set using the R package "HiSSE" (Beaulieu and O'Meara 2016). We computed AIC scores for each model, including the four BiSSE models described previously. We concluded that state-dependent diversification was present if the best overall model with state dependence (HiSSE or any of the three SDD BiSSE models) was supported by ΔAIC > 2 relative to the best character-independent model (CID-2 or the four-parameter non-SDD BiSSE model). We also performed a second set of comparisons excluding the full HiSSE model, thus ensuring that the non-trivial null model (CID-2) has the same complexity as the most-complex SDD model (BiSSE). The CID-2 and HiSSE results were obtained when the testing regime was no longer blinded.

Results

For phylogenies simulated in the absence of diversification rate heterogeneity (non-SDD), we find that, like BiSSE, FiSSE rejects the non-SDD null at an appropriately low frequency (Fig. 1A). For trait-dependent speciation rates, under the parameter values tested by Maddison et al. (2007), we find that FiSSE can also reliably infer SDD, although power is modest for phylogenies with fewer than 300 tips (Fig. 1B). For this scenario, BiSSE has substantially greater power to infer SDD on small trees, but power to reject the non-SDD null hypothesis is similar for the two methods on phylogenies with at least 300 tips. FiSSE has low power to infer trait-dependent extinction, and even though this is also challenging for BiSSE, it performs much better overall (Fig. 1C). FiSSE and BiSSE both have high power to infer trait-dependent speciation, even when net diversification rates are constant across character states (Fig. 1D). Figure S2 shows the relationship between two-tailed p-values and the number of parsimony-inferred state changes for this set of analyses; in general, power to detect SDD increases as a function of the parsimony score. Power to detect true SDD was low when simulated datasets contained 5 or fewer parsimony changes, with SDD correctly inferred in only 22% of simulations. For datasets with more than 5 but fewer than 10 changes, power increased to 52%; datasets with more than 10 changes correctly inferred SDD in 85% of simulations.

The quasi-parameters Λ_k (the average of 1/ES for tips in state k) are not estimated using a formal diversification model, and we tested whether the state-specific estimates Λ_0 and Λ_1 were correlated with the true values of speciation in the generating model. Figures 2-3 illustrate the relationship between Λ and true speciation rates (λ) for each character state. For state-dependent speciation simulations with state-independent extinction, the Λ_i substantially overestimate the λ_i, but $\Delta\Lambda$ was only slightly more than the speciation rate difference. However, for simulations performed with constant net diversification but state-dependent speciation and extinction, estimates of $\Delta\Lambda$ were lower than the difference in speciation rates but higher than the difference in net diversification rates (Fig. 3). These results suggest than Λ is correlated with true speciation rates for character states, but also that the relationship between the quasi-parameters and the true rates may be complex. The overestimate of true speciation rates evident in Figures 2-3 may reflect an ascertainment bias similar to the "push of the past" discussed by Nee et al (1994), whereby phylogenies that survive to the present to be observed are characterized by an apparent excess of early speciation events (Phillimore and Price, 2008).

For neutral characters simulated on the empirical bird phylogenies (a non-SDD process), we previously showed that the BiSSE non-SDD model (constant λ and μ across the tree) is frequently rejected (Rabosky & Goldberg 2015; presumably because the null model of constant speciation and extinction rates is incorrect; Beaulieu & O'Meara 2015). For FiSSE, however, we do not find elevated false positive rates with this set of trees (Fig. 4), even for high transition rates that exacerbated the problem with BiSSE (see Figure 7 from Rabosky and Goldberg 2015). Furthermore, even when the FiSSE null model is violated by asymmetric transition rates, the FiSSE test does not return a statistically significant result. It thus appears that FiSSE is robust to violation of its assumptions about the underlying process of character change.

To investigate this robustness further, we used a double-blind performance assessment. We found that FiSSE and BiSSE had broadly comparable power to infer true state-dependent diversification (Fig. 5A), although BiSSE performed better in most simulation scenarios. FiSSE had greater power than BiSSE in one scenario (0.28 versus 0.02; scenario 1 in Fig 5A; Table S2), entailing trait-dependent diversification under a cladogenetic model of character change (Magnuson-Ford and Otto 2012; Goldberg & Igic 2012). FiSSE's power relative to BiSSE was lowest when character state changes were very rare (scenario 12). For this scenario, the rate of character state change was approximately two orders of magnitude lower than the speciation rate, and most (80%) of the simulated trees contained only a single parsimony-inferred state change. Because of the lack of replication in diversification rate shifts in this scenario, we question whether recovering the generating model, by inferring SDD, is the desired outcome for evolutionary inference.

False positive rates with FiSSE were generally acceptable across the range of non-SDD simulation scenarios considered (Fig. 5B). The mean proportion of datasets that were incorrectly inferred to show SDD across all 34 non-SDD scenarios was 0.055. No scenario showed a rejection rate in excess of 0.18. Six scenarios had rejection rates of 0.1 or more; these included both simple birth-death trees and trees with diversification rate shifts, but they tended to be scenarios with slow, erratic, or asymmetric trait change (although other scenarios with these trait change properties fared better). The elevated false positive rates in at least several of these scenarios are not simply due to the relatively small number of simulations (50) per scenario. We verified this by creating an additional 500 datasets under the two testing scenarios where FiSSE showed the highest false positive rates; repeating FiSSE on these expanded sets yielded false positive rates of 0.21 and 0.19 (for scenarios 37 and 47, respectively). For the BiSSE-only comparisons (no HiSSE / CID-2), the mean proportion of significant SDD inferences across the 34 non-SDD simulation scenarios was 0.35. The highest values with BiSSE occurred when neutral characters were simulated on empirical phylogenies or phylogenies that had been generated under compound (multi-regime) diversity-dependent processes. We obtained generally congruent results when making inferences from the rate parameter estimates rather than model comparisons (probability that the r_1 - r_0 difference excludes zero, inferred from MCMC). The results were qualitatively similar to those presented in Figure 5, although both statistical power and rates of incorrectly inferring SDD were somewhat lower (Fig. S3).

In the results thus far reported, BiSSE was compared against a very simple null model that allows no diversification rate heterogeneity. We relaxed this restriction by comparing BiSSE against the CID-2 model, which allows for diversification rate shifts tied to a hidden character rather than the focal character. This often substantially reduced BiSSE's false positive rate (it decreased by 0.3 or more in 11 of the 34 scenarios) while maintaining statistical power (Fig. 6, BiSSE; open triangles). BiSSE's highest false positive rates (> 0.4) when CID-2 was included as a null model involved neutral characters simulated on empirical supertrees (scenarios 41-42, Fig. 6). However, several other simulation conditions---even on the same empirical phylogenies---were markedly less problematic for BiSSE when it was compared against CID-2 rather than against the four-parameter null model alone. The scenarios with next-most-elevated BiSSE false positive rates (0.3 and 0.24 for scenarios 36 and 34) involved slowly-evolving neutral traits, suggesting that phylogenetic pseudoreplication remains a challenge for this class of model.

Finally, we added the HiSSE model to the comparison, allowing hidden substates contained within the focal characters to affect diversification. Power to detect true SDD scenarios was similar for BiSSE + HiSSE as for BiSSE, when the null models included both the constant-rate scenario (as in Fig. 5) and the CID-2 model. Including HiSSE, however, frequently increased the false positive rate relative to the scenario where BiSSE was evaluated against CID-2 and constant-rate models (20 out of 34 scenarios) and never decreased it. Note that "HiSSE + BiSSE" in Figure 6 reflects all simulations where one of the two true SDD models (BiSSE or HiSSE) provided a better fit to the data than all other models in the candidate set (Tables S1-S2). One scenario stands out as causing all three methods (FiSSE, BiSSE with or without CID-2, BiSSE + HiSSE) to incorrectly infer SDD more than 10% of the time. This is a symmetric neutral trait simulated on an empirical tree, with a rapidly-diversifying clade then fixed (manually) to a single value of the trait (scenario 37).

Discussion

We have described a simple non-parameteric method, called FiSSE, that can reliably test hypotheses about the effects of a binary character on lineage diversification rates. We found the method can detect state-dependent differences in diversification rates on phylogenies with a modest number of tips, although the method is demonstrably less powerful than a formal state-dependent model (BiSSE) across many simulation scenarios we considered. However, FiSSE also appears to be largely robust to spurious inferences of state-dependent diversification. This is true even on datasets generated under a broad range of empirically-relevant diversification scenarios that are problematic for the BiSSE framework as traditionally applied (Fig. 4. & 5B).

Importantly, we also found that including a non-trivial null model (CID-2 from the HiSSE framework) in the candidate set for BiSSE analyses dramatically reduces the overall false positive rate for BiSSE, while maintaining statistical power (Fig. 6). Nonetheless, FiSSE's false positive rates were generally lower than those observed for the expanded BiSSE+HiSSE modeling framework for a range of empirically-relevant diversification scenarios. Given the substantial reduction in false positive rates obtained by including CID-2, we agree with Beaulieu and O'Meara (2016) that CID-2 (or similar) should be included as a null model when performing BiSSE analyses. On the other hand, we found that use of HiSSE, BiSSE, and CID-2 in concert frequently yielded incorrect conclusions, with 9 of 34 non-SDD scenarios having false positive rates in excess of 0.25. These results are consistent with those presented by Beaulieu and O'Meara (2016; e.g., their Figure 6).

Advantages of BiSSE-style models (including CID-2 and HiSSE) relative to FiSSE include explicit parameter estimation and increased statistical power, but FiSSE offers further reductions in false positive rates. There is thus a reason to view FiSSE as providing an important check on the reliability of results obtained with formal state-dependent models. Although we do not believe that methods should generally be chosen based on computational speed, we also note that model-based SDD analyses are computationally intensive relative to FiSSE analyses. The most complex model-based analyses we performed required approximately 100x - 1000x more CPU time to complete than the corresponding FiSSE analyses.

Why is FiSSE generally robust to phylogenetic pseudoreplication? Consider, for example, the extreme case of a single increase in speciation rate and a single change in character state along the same branch. When character histories are simulated under a process with a very low transition rate, there will be only one or few trait changes, and they could occur anywhere on the tree. The null distribution of $\Delta\Lambda$ will thus have high variance, making it difficult to detect true SDD when it is present. Conversely, this high variance also makes it difficult to find significant evidence for SDD when it is not present, thus reducing the influence of phylogenetic pseudoreplication on the false positive rate. We see this phenomenon in scenario 12, in which FiSSE fares poorly in identifying SDD on trees simulated under the BiSSE model with low *q*. In contrast, scenario 11 had the same state-dependent speciation and extinction as scenario 12 but a much higher transition rate, and FiSSE performed nearly as well as BiSSE (the null distribution of $\Delta\Lambda$ had standard deviations 1.9 and 5.3 for scenarios 11 and 12, respectively). FiSSE fails to recognize pseudoreplication in scenario 37, however, because there are many character changes on the tree outside of the rapidly-diversifying clade that has fixed state. Why is FiSSE generally robust to complex diversification rate heterogeneity? With many shifts in diversification across the tree, there will be much variation in any subset of the Λ^t tip values. Regardless of how this variation is partitioned into the two states -- whether a neutral trait is evolving slowly or quickly -- $\Delta\Lambda$ will have high variance, again correctly reducing inferences of SDD.

In the interpretation of results obtained with FiSSE, we caution that the Λ_k quasi-parameter is an imperfect measure of the speciation rate and does not directly reflect extinction or net diversification. The method has much less power than BiSSE to infer trait-dependent extinction (Fig. 1C). However, whether BiSSE or any other method can usefully infer extinction rates when they are as heterogeneous as in nature remains controversial (Rabosky 2010; Davis et al. 2013; Beaulieu and O'Meara 2015; Rabosky 2016). In general, we recommend that researchers compare the values of the Λ_k quasi-parameters to speciation and extinction rates obtained from a formal state-dependent model. We suggest that the strongest inference of state-dependent diversification is one where FiSSE and BiSSE results are in agreement, where BiSSE has been evaulated against the non-trivial CID-2 null model, and where BiSSE's speciation or net diversification estimates and the FiSSE quasi-parameters are generally congruent.

Conceptually, FiSSE is related to the framework developed by Bromham et al. (2016), who proposed a set of summary statistics to assess the adequacy of the BiSSE model and various constrained submodels (e.g., $\lambda_1 = \lambda_0$, $\mu_1 > \mu_0$, $q_{01}=q_{10}$). Their procedure involves fitting a set of full and constrained BiSSE models to the observed data and then simulating null distributions of phylogenies under each of the candidate models (Day et al. 2016; Hua and Bromham 2016). This approach is substantially more complex than FiSSE, which uses a fixed topology and conditions only on an estimate of the number of state changes. One advantage to the Bromham et al. (2016) framework is that it provides an absolute test of model adequacy and can lead to rejection of all models under consideration. The FiSSE approach is also related to the test proposed by Freckleton et al. (2008) for continuous characters. The Freckleton et al. (2008) test involves computing a tip-specific measure of speciation rate from the density of nodes along the path leading from the root to the tips of the tree. The relationship between those tip-specific rates and a trait is assessed using PGLS.

Just as fitting models by approximate Bayesian computation (Beaumont 2010) requires seemingly-arbitrary decisions about summary statistics, so does the FiSSE procedure involve arbitrary (but intuitively-motivated) decisions, such as the test statistic definition, the use of parsimony, and a symmetric-rates model for trait evolution. Consequently, there are many other methods that could be constructed along these same lines. For example, alternative test statistics could describe the difference in diversification rates between two character states, as in Bromham et al (2016). The encouraging results from FiSSE suggest that exploration of such methods could be a worthwhile line of investigation to continue. But because *ad hoc* methods like this cannot be rigorously justified on theoretical grounds, they can only be assessed based on their performance. A comprehensive suite of testing scenarios is therefore especially important. We created such a suite here, which we hope will be useful for and extended during the testing of future methods.

Conclusion

We have developed a simple test for the effects of a binary character on lineage diversification rates. Using a double-blind testing procedure, we demonstrated the method has reasonable performance across a range of simulation scenarios (Fig. 5-6). Our results suggest two substantive recommendations for testing hypotheses about trait-dependent diversification involving discrete character states. First, it seems clear that hypothesis tests with BiSSE should incorporate one or more non-trivial null models, following Beaulieu and O'Meara (2016). As we have shown, the incorporation of one such model (CID-2) into the candidate set of BiSSE-type models led to a dramatic reduction in false positive rates across the range of testing scenarios. Second, we recommend that hypothesis tests with FiSSE be included as a complement to formal state dependent models. For BiSSE+HiSSE analyses, we found that false positive rates were appreciably elevated in several testing scenarios even when CID-2 was included as a null model. We have shown that FiSSE can provide an additional check on results obtained with the BiSSE family of models. We believe that there is considerable value in further development of non-parametric and semi-parametric approaches for testing hypotheses about trait-dependent diversification (Freckleton et al 2008; Rabosky and Huang 2015; Bromham et al. 2016). Such approaches provide a valuable complement to formal process-based models in the quest to identify methods that are both powerful and robust to phylogenetic pseudoreplication and model inadequacy.

Acknowledgements

Data Archiving

Analysis scripts, benchmark datasets, and supplementary tables and figures have been archived in the Dryad repository, doi: ########. FiSSE can also be found on GitHub at https://github.com/macroevolution/fisse.

LITERATURE CITED

Alfaro, M. E., F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, and L. J. Harmon. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Nat. Acad. Sci. U.S.A. 106:13410-13414.

Arnold, A. J. and K. Fristrup. 1982. The theory of evolution by natural selection: a hierarchical expansion. Paleobiology 8:113-129.

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

Beaulieu, J. M. and B. C. O'Meara. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65:583-601.

Beaumont, M. A. 2010. Approximate Bayesian Computation in Evolution and Ecology. Ann. Rev. Ecol. Evol. Syst. 41:379-406.

Belmaker, J. and W. Jetz. 2015. Relative roles of ecological and energetic constraints, diversification rates and region history on global species richness gradients. Ecology Letters 18:563-571.

Bromham, L., X. Hua, and M. Cardillo. 2016. Detecting Macroevolutionary Self-Destruction from Phylogenies. Syst. Biol. 65:109-127.

Coyne, J. A. and H. A. Orr. 2004. Speciation. Sinauer, Cambridge.

Davis, M. P., 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.

Day, E. H., X. Hua, and L. Bromham. 2016. Is specialization an evolutionary dead end? Testing for differences in speciation, extinction, and trait transition rates across diverse phylogenies of specialists and generalists. J. Evol. Biol. 29:1257-1267.

Farrell, B. D. 1998. "Inordinate fondness" explained: Why are there so many beetles? Science 281:555-559.

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

FitzJohn, R., W. P. Maddison, and S. P. Otto. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595-611.

FitzJohn, R. G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution 3:1084-1092.

Freckleton, R. P., A. B. Phillimore, and M. Pagel. 2008. Relating traits to diversification: A simple test. American Naturalist 172:102-115.

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

Hodges, S. A. 1997. Floral nectar spurs and diversification. Int. J. Plant. Sci. 158:S81-S88.

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.

Jablonski, D. 2008. Species Selection: Theory and Data. Annual Review of Ecology Evolution and Systematics 39:501-524.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. Mooers. 2012. The global diversity of birds in space and time. Nature 491:444-448.

Kafer, J. and S. Mousset. 2014. Standard sister-clade comparison fails when testing derived character states. Syst. Biol. 63:601-609.

Maddison, W. P. 2006. Confounding asymmetries in evolutionary diversification and character change. Evolution 60:1743-1746.

Maddison, W. P. and R. 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. Systematic Biology 56:701-710.

Magnuson-Ford, K. and S. P. Otto. 2012. Linking the investigations of character evolution and species diversification. Am. Nat. 180:222-245.

Mitter, C., B. Farrell, and B. Wiegmann. 1988. The phylogenetic study of adaptive zones - has phytophagy promoted insect diversification? American Naturalist 132:107-128.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994. Extinction rates can be estimated from molecular phylogenies 344:77-82.

Ng, J. and S. D. Smith. 2014. How traits shape trees: new approaches for detecting character state-dependent lineage diversification. J. Evol. Biol. 27:2035-2045.

Phillimore, A. B., and T. D. Price. 2008. Density-dependent cladogenesis in birds. PLoS Biology. 6:e71.

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

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

Rabosky, D. L. 2016. Challenges in the estimation of extinction from moelcular phylogenies: a response to Beaulieu and O'Meara. Evolution 70:218-228.

Rabosky, D. L. and H. Huang. 2015. A robust semi-parametric test for detecting trait-dependent diversification. Systematic BIology 65:181-193.

Redding, D. W. and A. O. Mooers. 2006. Incorporating evolutionary measures into conservation prioritization. Conservation Biology 20:1670-1678.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592-593.

Figure legends

Figure 1. Proportion of simulated datasets where significant state-dependent diversification was detected using FiSSE (circles) and BiSSE (diamonds). (A) Control: no state-dependence in simulation model. (B) State-dependent speciation only. (C) State-dependent extinction only. (D) State-dependent speciation and extinction, but net diversification rate constrained to be constant ($r_0 = r_1 = 0.1$, $\lambda_0 = 0.1$, $\lambda_1 = 0.2$).

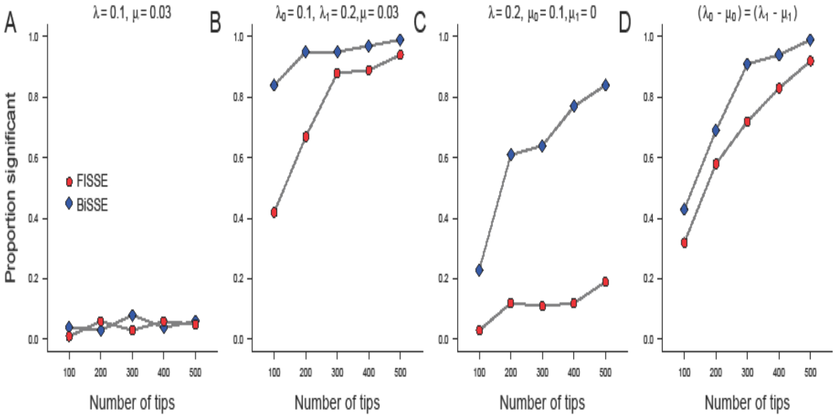


Figure 2. Relationship between mean tip-specific Λ estimates for two character states for phylogenies simulated with (filled circles) and without (open circles) state-dependent diversification (SDD). True speciation rates are illustrated with solid (non-SDD) and dashed (SDD) gray lines. Panel (A) shows all simulated trees, and panel (B) shows only those datasets where FiSSE reported a significant association between the character state and diversification. Parameters for non-SDD phylogenies: $\lambda = 0.1$, $\mu = 0.03$, $q=0.01$; SDD parameters: $\lambda_0 = 0.1$, $\lambda_1 = 0.2$, $\mu = 0.03$, $q=0.01$). For SDD phylogenies, mean estimates for Λ_0 and Λ_1 were 0.140 and 0.253, respectively.

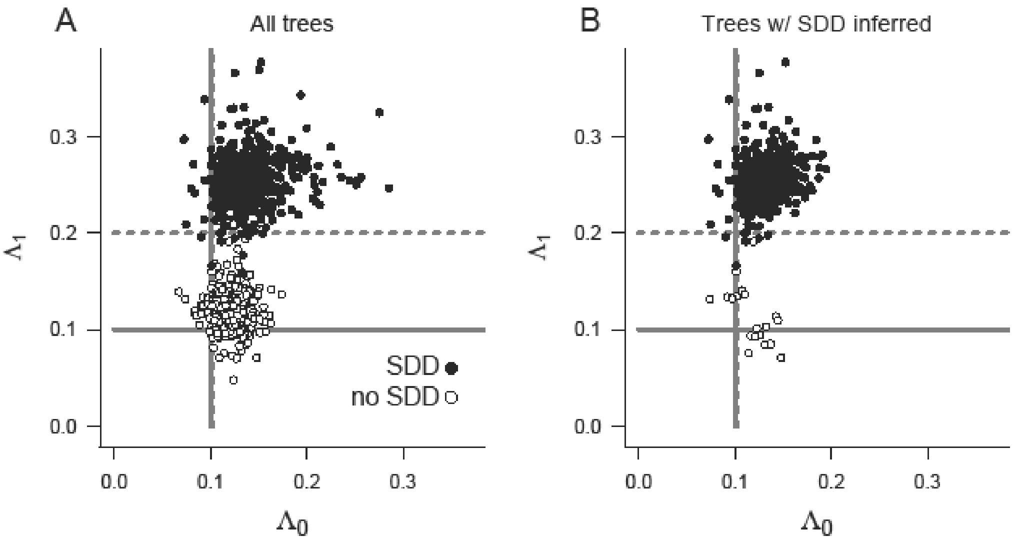


Figure 3. Relationship between mean tip-specific Λ estimates for two character states for phylogenies simulated with two-fold (A) and three-fold (B) increases in the speciation rate for the derived character state while holding net diversification rates constant ($r_0 = r_1$). True speciation rates for each state are illustrated by dashed lines. Results in (A) are based on the same set of phylogenies that underlie results shown in Figure 1D. For simulations with a two-fold increase in speciation, the mean estimate for ΔΛ was 0.063 (compare with true Δλ = 0.1); with a threefold increase in speciation rate, the mean estimate for ΔΛ was 0.130 (true Δλ = 0.2).

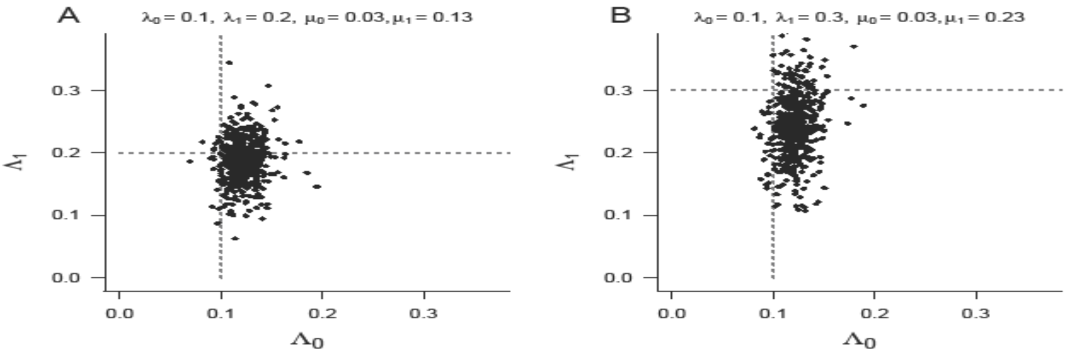


Figure 4. False positive rates for FiSSE when neutral characters are simulated on the avian empirical phylogenies (a non-SDD process). Five transition rates are illustrated; a total of 600 simulated datasets (60 phylogenies, 10 replicates per tree) were analyzed for each transition

rate. The "Asymm" scenario specified a four-fold difference in the relative transition rate between the two character states (q_{01} = 0.02, q_{10} = 0.005).

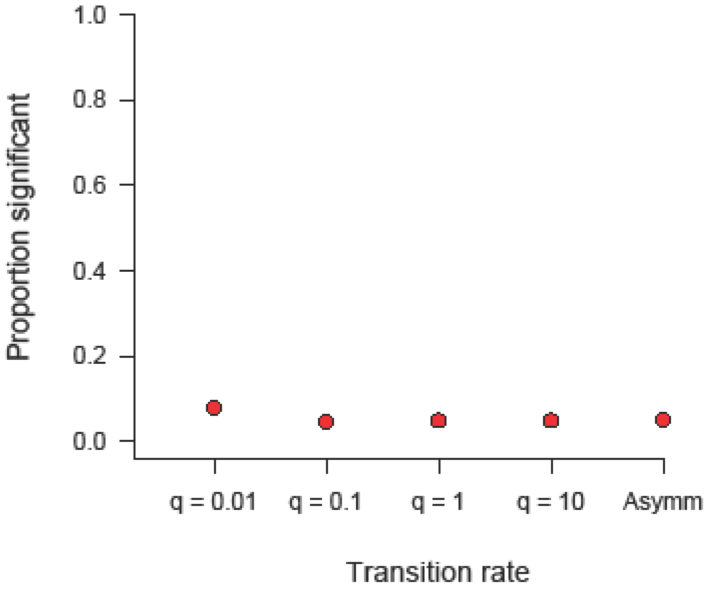


Figure 5. Performance assessment of FiSSE (circles) and BiSSE (diamonds) across scenarios with (A) and without (B) state-dependent diversification (SDD). All scenarios are described in Tables S1-S2. The eight scenarios tested when the assessment was no longer double-blind are marked with asterisks. Proportion significant for (A) is power to detect a true relationship between traits and diversification. Proportion significant for (B) is the fraction of simulated datasets where FiSSE or BiSSE reported a significant association for neutral characters simulated independent of the diversification process. FiSSE generally has lower power than BiSSE to detect state-dependent diversification when it is present, but it is characterized by a substantial reduction in the false positive rate. These results compare BiSSE against the simple four-parameter constant-rate null model.

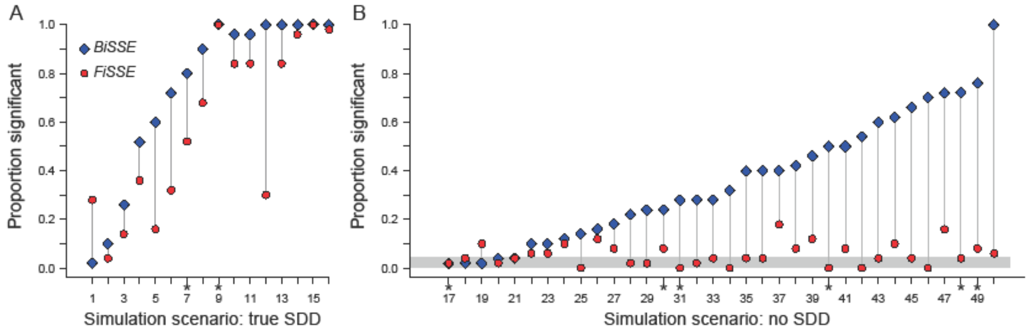
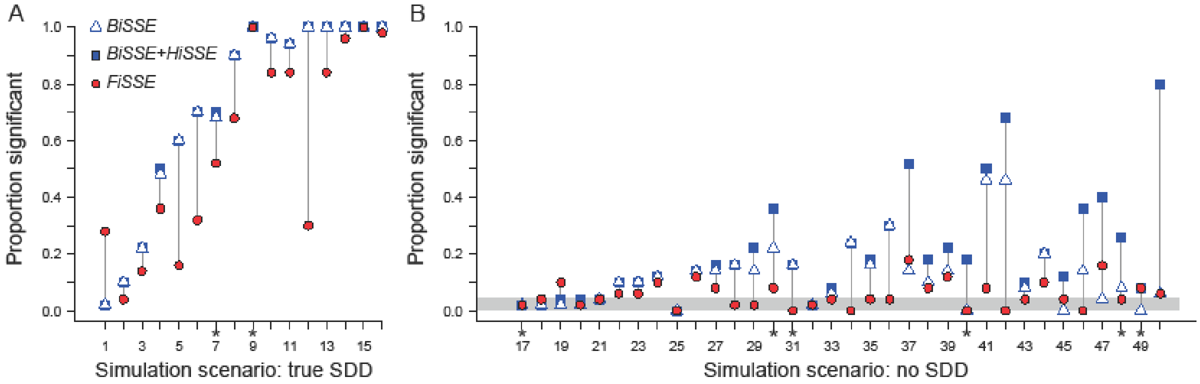


Figure 6. Performance of expanded set of BiSSE-class models across testing scenarios with (A) and without (B) state-dependent diversification (SDD) when the null model set is expanded to include a character-independent model with two unobserved diversification states (CID-2). Scenarios are the same as in Figure 5; see Tables S1 and S2 for details. BiSSE (triangles) is the proportion of simulations where the BiSSE model was substantially favored (ΔAIC > 2) over both character-independent models (constant-rate and CID-2). BiSSE + HiSSE (squares) is the proportion of simulations where either BiSSE or HiSSE identified significant state-dependent diversification associated with the focal character, relative to the constant-rate and CID-2 null models. State-dependent diversification is concluded when either the BiSSE or HiSSE model fits the data better than the two null models, and we thus present the combined proportion of simulations where SDD was inferred. Table S2 further breaks down the BiSSE +HiSSE category into "BiSSE best" and "HiSSE best" subcategories. Results for FiSSE are identical to those shown in Fig. 5 and are included here for comparison. In panel (A), BiSSE and BiSSE+HiSSE typically had identical statistical power, such that symbols are overplotted.


A
B
△ BiSSE
■ BiSSE+HiSSE
● FiSSE
Proportion significant
1.0
0.8
0.6
0.4
0.2
0.0
1
3
5
7
9
11
13
15
Simulation scenario: true SDD
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Simulation scenario: no SDD