Systematic Biology Advance Access published September 22, 2015

A Robust Semi-Parametric Test for Detecting Trait-Dependent Diversification

Daniel L. Rabosky* and Huateng Huang

Department of Ecology and Evolutionary Biology and Museum of Zoology

University of Michigan, Ann Arbor, MI 48109

*Email: drabosky@umich.edu

Corresponding author:

Daniel L. Rabosky

Phone: 510-610-9082

ABSTRACT

Rates of species diversification vary widely across the tree of life and there is considerable interest in identifying organismal traits that correlate with rates of speciation and extinction. However, it has been challenging to develop methodological frameworks for testing hypotheses about trait-dependent diversification that are robust to phylogenetic pseudoreplication and to directionally biased rates of character change. We describe a semi-parametric test for trait-dependent diversification that explicitly requires replicated associations between character states and diversification rates in order to detect effects. To use the method, diversification rates are reconstructed across a phylogenetic tree with no consideration of character states. A test statistic is then computed to measure the association between species-level traits and the corresponding diversification rate estimates at the tips of the tree. The empirical value of the test statistic is compared to a null distribution that is generated by structured permutations of evolutionary rates across the phylogeny. The test is applicable to binary discrete characters as well as continuous-valued traits and can accommodate extremely sparse sampling of character states at the tips of the tree. We apply the test to several empirical datasets and demonstrate that the method has acceptable Type I error rates.

[Speciation, Extinction, BAMM, State-Dependent Diversification, Key Innovation, Statistical Methodology, Comparative Analysis]

INTRODUCTION

There is widespread interest in identifying the organismal traits that are associated with differential rates of species diversification. Some groups of organisms are far more species rich than other groups, and at least some of this variation can be explained by associations between species traits and diversification rates (Mitter et al. 1988; Farrell et al. 1991; Coyne and Orr 2004; Jablonski 2008). Formal statistical analysis of the relationship between character states and diversification rates began with the pioneering work of Mitter et al. (1988), who introduced a method of contrasts for testing whether a particular character state is associated with consistent differences in species richness for sister clades that differ in character state. The sister-clade contrasts method has been used to identify a number of correlates of differential diversification, including phytophagy, plant defense syndromes, floral traits associated with pollinator specificity, and other traits (Mitter et al. 1988; Hodges and Arnold 1995; Coyne and Orr 2004; Kay et al. 2006). More recently, a set of state-dependent speciation-extinction models have been developed that provide a probabilistic framework for reconstructing the history of character evolution and diversification on phylogenetic trees (Maddison et al. 2007; FitzJohn et al. 2009; FitzJohn 2010; Goldberg et al. 2011). Character state-dependent speciation-extinction models have been widely used in the recent literature to detect trait-dependent diversification (Goldberg et al. 2010; Pyron and Burbrink 2014; Rolland et al. 2014).

State-dependent models of speciation and extinction have recently been found to have high Type I error rates when applied to real datasets (FitzJohn 2012; Machac 2014;

Rabosky and Goldberg 2015), suggesting an acute need for robust alternative methods of analysis. The Type I error rate arises, at least in part, because formal state-dependent models do not account for the number of phylogenetically independent shifts in diversification that are associated with shifts in character state (Maddison and FitzJohn 2014). As such, a single shift in diversification rates that occurs anywhere within a phylogenetic tree can lead to significant relationships between diversification rates and many character states on a given phylogeny. Figure 1 illustrates several types of character state distributions that result either in overinflated statistical support for trait-dependent diversification (Fig. 1a) or — more worryingly — potentially spurious associations between traits and diversification (Fig. 1b, c). Rabosky and Goldberg (2015) suggested that this propensity for false positives has led to a substantial excess of trait-diversification associations in the literature. They provided simulation and analysis recommendations to enable researchers to assess whether significant effects detected with BiSSE, QuaSSE, and related models can be distinguished from trait-independent diversification.

Here, we introduce an alternative method of testing hypotheses of trait-dependent diversification that explicitly accounts for the number of independent shifts in diversification that occur across a phylogeny. The method described here uses an evolutionary model to estimate rates of species diversification on a time-calibrated phylogenetic tree, but it does not attempt to jointly reconstruct the history of character change and diversification. Rather, the method uses structured permutations to generate null distributions of evolutionary rates across the tips of a phylogenetic tree. The null

distributions generated from this procedure can be used to test whether a particular character state is associated with differential rates of species diversification. The method is applicable to continuous traits as well as to binary discrete characters.

MATERIALS AND METHODS

Permutation Test for Continuous Traits

The method described here for testing hypotheses about trait-dependent diversification involves three steps. The first step involves the estimation of diversification rates across a phylogenetic tree. Second, an empirical test statistic is computed between character states and diversification rates. Finally, the observed test statistic is compared to a null distribution obtained by structured permutations of the evolutionary rate distribution across the tips of the phylogeny. We refer to our implementation of this test as STRAPP ("STructured RAte Permutations on Phylogenies").

We will describe application of the test to evolutionary rates estimated using BAMM, a software tool for inferring evolutionary rates on phylogenetic trees (Rabosky et al. 2013; Rabosky 2014). The methods described here have been implemented in the BAMMtools R package (Rabosky et al. 2014b). In principle, this method can be applied to evolutionary rates estimated using a variety of software platforms and modeling approaches (e.g., MEDUSA) (Alfaro et al. 2009). The method requires that diversification rates be estimated under a multi-process model such that any lineage on

the tree can be assigned to one of *k* distinct rate classes. This is typically the case for evolutionary models that posit the occurrence of discrete "rate shifts" in species diversification or phenotypic evolution. In our examples below, we focus solely on the relationship between speciation and phenotypic traits, because speciation rates are typically estimated with much greater accuracy than extinction rates on phylogenetic trees (Nee et al. 1994; Maddison et al. 2007; Davis et al. 2013). However, the method is applicable to net diversification rates and even extinction rates alone, provided these rates are estimated under a multiprocess diversification model such as BAMM. Our implementation of STRAPP in BAMMtools explicitly allows analysis of speciation, extinction, and net diversification rates.

BAMM provides estimates of evolutionary rates across phylogenetic trees under a Poisson process model of rate variation. Each sample from the posterior distribution simulated using BAMM consists of a mapping of rate shifts to the focal phylogeny, plus the parameters associated with those shifts. The model allows the explicit estimation of speciation rates at each tip of a phylogenetic tree. However, these rates cannot be treated as independent data points for the purposes of statistical hypothesis testing. In the BAMM framework, any two species that share a common macroevolutionary rate regime will share similar diversification rates at the tips of the tree (Rabosky 2014; Rabosky et al. 2014a). As such, species that are consistently assigned to the same macroevolutionary rate regime will have highly correlated rates. In effect, the number of distinct evolutionary rate regimes across a phylogeny as estimated with BAMM is one estimate

of the number of independent data points that can be used to detect the effect of a character state on diversification.

We use the Spearman correlation coefficient as a test statistic for the relationship between diversification rates and continuous-valued traits. The null distribution of the test statistic is constructed using block permutations of evolutionary rates on a phylogenetic tree. The rates are reshuffled across the tips of a phylogeny in such a way that the pairwise covariances in evolutionary rate regimes between species are maintained (Figure 2). Formally, let $S_1, S_2 \ldots S_K$ denote subsets of species that share common macroevolutionary rate parameters: e.g., they have been assigned to the same diversification rate regime and all species within a rate class (e.g., S_1) are characterized by identical rates. Let $\theta_1, \theta_2 \ldots \theta_K$ denote the corresponding rate parameters associated with each set of taxa. The permutation test involves permuting the indices of the rate parameters assigned to each taxon block. For example, each species in taxon set S_1 will be assigned evolutionary rate parameters sampled from the observed collection of parameters $\theta_1, \theta_2 \ldots \theta_K$. Figure 2 illustrates a distribution of tip rates on a phylogenetic tree, with a set of fast-diversifying lineages (dark colors) and slow-diversifying lineages (light colors), largely resulting from a single diversification rate shift in the common ancestor of a single labeled clade. The reshuffling procedure randomizes rates across the tips, but ensures that all taxa with a "fast" rate in the original dataset share an identical rate in each permuted dataset.

The strength of this approach is that it explicitly accounts for the number of independent shifts in diversification that have occurred and requires repeated associations between a

character state and diversification in order to detect an effect. Suppose, in the illustrative example (Figure 1), that the "fast rate" clade (black clade) is associated with some character state x_0, and that the "slow rate" clade (gray clade) is associated with some state x_1. Given that only a single shift in diversification rates occurred in the dataset, it would be difficult to find a significant relationship between traits and diversification using the approach here. Because the dataset contains evidence for only a single shift in diversification rates, the permutation test will generate a null distribution that contains a high frequency of associations between x_0 and fast rates. Because multiple independent shifts in diversification are required to detect an effect, the permutation test here is expected to be robust to the detection of false positives in the analysis of character states and diversification rates.

The null model that underlies our permutations is not designed to produce realistic patterns of diversification rates in the permuted rate distribution. The procedure is only intended to maintain the covariance structure in rate regimes among taxa while sampling new rates for taxa from an empirically-parameterized distribution. As such, taxa with (true) fast diversification rates can be assigned slow rates in the permuted rate distribution, and taxa with (true) slow speciation rates can be assigned fast rates in the permuted distribution. The phylogeny is only used during the actual estimation of rates with BAMM. Following analysis with BAMM, the tree itself is irrelevant: our analysis relies only on the set of diversification rates, the phenotypic trait values, and the rate class index to which each taxon belongs.

The permutation test described here uses rates of lineage diversification that are estimated in the absence of any information about character states. As such, the method is not susceptible to several biases that confound traditional sister clade contrasts. As first pointed out by Maddison (2006), asymmetric rates of character change can compromise sister clade contrasts when sister clades are selected on the basis of fixed character state differences. A related bias was identified by Kafer and Mousset (2014), who showed that a homogeneous speciation-extinction process could result in spurious associations between traits and diversification when the ancestors of sister clade pairs are characterized by a common root state. However, in both of these cases, the bias emerges because sister clades are selected based on fixed character state differences (e.g., clades that lack fixed character state differences are not selected as focal clades for analysis). Diversification rates can be estimated in the absence of trait data, because the tree structure itself contains information about the history of speciation and extinction that can be inferred without reference to trait data.

In the examples below, we used the posterior distribution of macroevolutionary rate shift configurations sampled using BAMM as the basis for the permutation test. The permutation test thus entails two stages (see Figure 2). First, a mapping of evolutionary rates is randomly sampled from the posterior distribution simulated with BAMM; each sample in the posterior from a BAMM analysis represents such a mapping (Figure 2a). We then block-permute the rates to obtain a set of permuted rates (Figure 2a). For this focal sample, we could compute the empirical (i.e., "true") correlation coefficient between evolutionary rates observed at the tips of the phylogeny and a character state of

interest (Figure 2b), and also the permuted correlation—an instance of a "null" correlation between the trait and diversification (Figure 2c). This procedure is repeated many times (Figure 2d-f) to generate a set of observed correlation coefficients, each of which is paired with a null coefficient where character states and diversification rates are uncorrelated (Figure. 2e-f). The proportion of times that the null correlation coefficient is higher than the corresponding observed coefficient is the one-tailed *p*-value for assessing the significance level of a positive correlation. Likewise, the proportion of randomly sampled mappings in which the null is more extreme than the observed correlation (i.e., comparing the absolute values) is equivalent to the *p*-value under a two-tailed test. It is not necessary that the test be performed across a full posterior distribution of rate shift configurations; one could also apply the method to any single rate shift configuration (e.g., Fig. 2a), such as the configuration with the highest maximum *a posteriori* probability. In this case, the analysis should entail repeated permutations of rates from a single shift configuration.

It is worth noting that this permutation approach can yield significant associations between traits and diversification even if no individual rate shifts achieve a particular threshold level of significance. BAMM simulates a posterior distribution of macroevolutionary rate shift configurations, and any given node may be associated with a relatively low marginal shift probability even if the overall probability of rate heterogeneity is high (Rabosky et al. 2014). For example, in the cetacean example (Fig. 1, Fig. 2), the marginal probability of a shift on any particular branch does not exceed 0.57, but the posterior probability of a rate shift on one of several ancestral branches leading to

the dolphin clade exceeds 0.98 (Rabosky 2014). However, STRAPP will be unable to detect significant associations between traits and diversification if the mean number of inferred shifts is low. Statistical power is a function of the number of diversification rate regimes inferred across a phylogeny but not of our confidence in the precise location of any particular rate regime.

Analysis of Binary Traits

The framework described above can be applied to binary discrete traits by modifying the test statistic used to measure the association between diversification and trait value. We used the Mann-Whitney U-test statistic as a basis for comparing rates among lineages that vary in the value of a binary discrete character. To account for correlations among evolutionary rates for individual species that are attributable to the BAMM model itself, we obtained null distributions of the U statistic through structured permutations of rates as described above. The proportion of BAMM posterior samples with a U statistic more extreme than that of its permuted sample indicates whether the two different states of the binary trait are significantly differ in regards to tip-level speciation rate. STRAPP for both discrete (binary) and continuous characters has been included in the BAMMtools package for R (Rabosky et al. 2014b).

An obvious question concerns why we permute evolutionary rates rather than character states. For discrete characters, it would be possible to devise a permutation scheme that maintained the covariance in state identity across taxa, such that all taxa with identical

character states in the real data are characterized by identical character states in each permutated trait distribution. However, it is not clear how to apply this approach to continuous traits. One could fit a model of continuous character evolution to the trait data and simulate character states under the fitted model, but this would entail additional steps and relies on the accuracy of both the fitted diversification model as well as the fitted model of character evolution. The test proposed here is only as good as the fitted BAMM model, but there is no additional need to model the evolution of character states to generate a null distribution.

Example: Speciation and Phenotypic Evolution in Fishes

We illustrate this approach by re-analyzing the relationship between rates of speciation and rates of morphological evolution across ray-finned fishes. Rabosky et al. (2013) found that rates of body size evolution were correlated with rates of speciation across a time-calibrated phylogenetic tree of 6760 fish species. In the original BAMM model described by Rabosky et al. (2013), a shift in evolutionary rates of species diversification was assumed to occur in concert with a shift in phenotypic evolutionary rates. Rabosky et al. (2013) found that shifts were positively correlated: an increase in speciation rate was consistently associated with an increase in phenotypic evolutionary rates. One drawback to the original BAMM model is that it forced a coupling (positive or negative) between rates of diversification and trait evolution. If a dataset contains evidence for multiple rates of species diversification but only a single rate of phenotypic evolution, the analysis would necessarily force a separate model of phenotypic evolution onto every subset of

the data where a distinct diversification regime occurred. Hence, the model could potentially lead to overparameterization of either diversification or trait evolution. For this reason, BAMM now forces the user to model the species diversification and phenotypic evolution as separate processes.

We performed 2.5 x 10^8 generations of MCMC sampling using constant-rate models of both speciation and phenotypic evolution on the fish phylogenetic and body size datasets analyzed by Rabosky et al. (2013). Although BAMM accommodates analyses of time-varying rates of species diversification and phenotypic evolution (Rabosky 2014; Rabosky et al. 2014a), we assumed that rates were constant in time for comparability with Rabosky et al. (2013), as well as to facilitate convergence of BAMM analyses on large phylogenetic trees. We used an exponential distribution with a rate parameter of 0.02 as a hyperprior on the Poisson rate parameter governing the number of rate shift events across the phylogeny. We treated phenotypic evolutionary rates at the tips of the phylogenetic tree as character states and assessed the correspondence between tip-specific rates of phenotypic evolution and rates of speciation using the permutation test described above. We analytically accounted for missing taxa, using family-specific sampling probabilities as described in Rabosky et al. (2013).

We tested whether the analysis described above was characterized by high Type I error rates by assessing the relationship between BAMM-estimated speciation rates in fishes and phenotypic data simulated under a Brownian motion model of character evolution. We generated 1000 phenotypic datasets by simulating an unconstrained Brownian motion

process on the fish phylogeny, assuming a root state of 0 and a trait variance (σ^2) of 1.0 per million years. We performed the permutation test on each simulated dataset with STRAPP to assess the relationship between rates of speciation and a randomly generated but phylogenetically structured trait that lacked any association with rates of speciation. Using BiSSE, we previously found that this phylogenetic dataset is susceptible to false positives: neutral characters simulated on the fish phylogeny tend to result in significant but spurious associations between trait and diversification (Rabosky and Goldberg 2015).

Permutation Test with Incomplete Character State Data

In the example above, character state data (rates of phenotypic evolution) were available for majority of the tips on the phylogenetic tree used to estimate speciation rates. This is unlikely to be the case for many datasets, where researchers may have highly incomplete character state data. A strength of the present method is that character state data can be extremely incomplete: the permutation test can easily be applied to datasets where character states have been assessed for only a small fraction of the total taxa. We performed subsampling analyses to mimic a situation where a high-resolution phylogenetic tree is paired with an incomplete character state dataset. We randomly selected subsets of species from the fish phylogeny and computed correlations between phenotypic evolutionary rates and speciation rates; we then performed structured permutations of rates across the fish phylogeny as described above to generate a null distribution of correlations. We repeated the subsampling process 100 times for a range of taxon counts (25, 50, 100, 200, 400, 800, 1600, and 3200 species). This subsampling

analysis was also performed on 100 phenotypic datasets simulated with an unconstrained Brownian model to examine the effect of incomplete character data on Type I error rates.

Example: Plumage Dichromatism and Speciation in Birds

Huang and Rabosky (2014) analyzed the relationship between bird plumage dichromatism and their macroevolutionary rate of speciation. In that study, we performed a PGLS (Grafen 1989; Martins and Hansen 1997) analysis using average tip-level speciation rates estimated from BAMM and reflectance-based measurements of sexual dichromatism (Armenta et al. 2008) to assess whether dichromatism predicted phylogenetic speciation rates. PGLS regression is a widely used method in comparative phylogenetic studies to account for phylogenetic correlation. However, the error structure in PGLS is potentially inappropriate for modeling correlations between BAMM-estimated rates and character state data. With BAMM-estimated rates, species that tend to share common macroevolutionary rate regimes will also have highly correlated evolutionary rates. If all members of a clade are assigned to the same rate regime, they will always have identical rates, and PGLS will underestimate the extent to which speciation rates are autocorrelated. Conversely, for any species consistently assigned to different rate regimes, their rates will be entirely decoupled. In this situation, PGLS will overestimate the extent to which rates are autocorrelated, because the true correlation for species in different rate regimes is zero.

We re-analyzed the relationship between dichromatism and speciation rate using the permutation test described above. Due to minor updates to the BAMM program since performing the analyses in Huang and Rabosky (2014), we repeated the BAMM analysis of speciation rates on the Hackett backbone MCC phylogeny from Jetz et al. (Jetz et al. 2012) after pruning all species lacking genetic data (leaving 6,670 species). We analytically accounted for incomplete taxon sampling in the BAMM analyses using a global sampling fraction of 0.67. We conducted 2.5 x 10^8 generations of MCMC sampling, discarding the first 10% of samples as burn-in and selecting 1000 evenly spaced samples from the posterior distribution of rate shift configurations for the permutation test.

The measurements of plumage dichromatism used in the previous study are continuous-valued traits based on spectrophotometer data, originally published by Armenta et al. (2008). As the analyses in Huang and Rabosky (2014) found little difference among the three different indices of reflectance-based measure of dichromatism (Armenta et al. 2008), we focused only on the first index—PCA—in this re-analysis. We tested the relationship between continuous-valued dichromatism measurements and speciation rates using STRAPP. We also coded dichromatism as a binary character to illustrate the analysis of discrete-valued traits using the structured permutation framework described here. Because the data published by Armenta et al. (2008) were standardized across species, we chose two different quantitative thresholds for determining whether a given species was dichromatic. We chose two quantiles of the distribution of PCA (50% and 80%) and coded all species with values lower than these as monochromatic.

We performed simulations to assess type-I error rates for the analysis of discrete traits on the bird phylogeny. Traits were simulated under the binary logistic model from Ives and Garland (2010) using the *R* package *phylolm* (Ho and Ane 2014). We excluded all species without dichromatism measurements from the bird phylogeny (i.e., leaving 904 species) to generate a pruned phylogeny for simulation. This ensured that simulated datasets comprised the same amount of missing data as the empirical dataset. We simulated binary traits under a symmetric evolutionary process with different levels of phylogenetic signal by varying the *a* parameter from Ives and Garland (2010). The parameter *a* is inversely correlated with the transition rate. Traits simulated with lower *a* evolve more rapidly and are expected to show reduced phylogenetic signal. We simulated binary traits with high (a =1), intermediate (a = 0; similar to Brownian motion), and weak (a = -1) levels of phylogenetic autocorrelation.

We estimated the level of phylogenetic signal from the empirical dichromatism dataset after treating dichromatism as a binary character; we found relatively weak phylogenetic signal under both PCA threshold values (a = -2.26 and -2.10 for the 50% and 80% quantile cutoffs, respectively). Hence, we also simulated binary traits with a level of phylogenetic signal chosen arbitrarily to be intermediate between these values (a = -2.19). In addition to the four levels of phylogenetic signal (a = -2.19, -1.0, 0.0, and 1.0), we also examined the effects of biased transition rates by simulating binary traits with an asymmetric transition matrix. We assumed that one rate was four times faster than the reverse transition, which predicts a 0.8:0.2 ratio in state frequency at equilibrium. This

also corresponds to the 80% quantile cutoff used to code species as dichromatic or monochromatic in the empirical dataset. In total, we simulated 100 datasets for each combination of phylogenetic autocorrelation parameter (*a*) and equilibrium state frequency.

RESULTS

Speciation and Phenotypic Evolution in Fishes

Separate analyses of speciation and phenotypic evolution in fishes both found strong evidence for multiple macroevolutionary rate regimes. For both analyses, we analyzed 1000 samples from the posterior distribution, after discarding 10% as burn-in. The median of the posterior distribution of the number of rate shifts in speciation and extinction was 71 (0.025 and 0.975 quantiles of 63 and 79, respectively), while that of the rate shift in body size evolution was 96 (0.025 and 0.975 quantiles of 85 and 108, respectively. We calculated the average phenotypic evolutionary rates at the tips of the phylogeny, which were used as character states for downstream analyses. In total, we obtained estimates of speciation rates and rates of body size evolution for 6,670 species. The relationship between the estimated rates for speciation and phenotypic evolution across all fishes in the dataset is shown in Figure 3. A permutation test recovered a significant correlation between these two rates (p=0.01 for two-tailed test; see Figure 4a). We examined the Type I error rate of STRAPP by simulating traits under Brownian evolution model on the observed fish phylogeny, and applying the permutation test to

each simulated set of neutral characters. These simulations thus preserve the heterogeneity in speciation rates observed in the empirical phylogeny, but trait data are simulated under a trait-independent model of diversification. The estimated type-I error rate was 0.03 (Figure 4b), suggesting that the method is not prone to false positives.

Permutation Test with Incomplete Character State Data

Subset analyses were performed with the fish dataset to examine how power to detect the correlation is affected by the number of taxa with the measured character state. Note that, in each case, we are using speciation rate information drawn from the full BAMM analysis of 6,760 taxa. As expected, when the number of species with trait data is small (e.g., 25), power to detect an association is low (Fig. 5a). However, reducing the number of taxa with character state data to 800 species (i.e., 11.8% of the total taxa in the fish phylogeny) resulted in a minimal reduction in power, and 80% of analyses with 400 randomly chosen species (~ 6%) obtained a significant correlation at the 0.05 significance level (Figure 5a). In contrast, subset analyses on traits simulated under Brownian evolution model found no increase in Type I error rates for subsampled datasets (Figure 5b). Hence, relationships between character states and diversification can be detected using the method described here, even if character state data are available for a small fraction of total species. Note, however, that speciation rates were estimated across the full phylogeny.

Speciation and Sexual Dichromatism in Birds

Across the time-calibrated phylogeny of 6,670 species of birds, we found a mean of 69.4 diversification shifts (0.025 and 0.975 quantiles of 60 and 80, respectively). While the correlation between a PCA-based dichromatism index and tip-level speciation rates is positive, it does not differ significantly from zero under a permutation test ($p = 0.14$). Likewise, treating dichromatism as a binary trait found no significant difference in rates with respect to character state, regardless of the cutoff used for the dichromatic state (Figure 6a). As dichromatism data were only available for 919 species (13.7% of avian species in the phylogeny), it is possible that additional dichromatism data might increase power to detect a significant positive correlation.

Rabosky and Goldberg (2015) documented a high false positive rate for the BiSSE model with simulated binary characters, in particular, for neutral characters that evolve rapidly on empirical phylogenies with diversification rate heterogeneity. Using binary traits simulated on the bird phylogeny, we found that the structured permutation test has low type-I error rates with binary data (Figure 6b). The Type I error rate was less than 0.05 for every combination of state frequency and evolutionary rate, suggesting that character change asymmetries are not problematic for the structured permutation test.

DISCUSSION

In this article, we describe a new method for testing hypotheses about trait-dependent lineage diversification rates. To use the method, researchers must be able to estimate

evolutionary rates across the tips of a phylogenetic tree under a multi-process model such as BAMM or MEDUSA. The key feature of the statistical approach is the random reshuffling of rates across the tips of the tree in a fashion that preserves the covariances in evolutionary rate regimes among taxa. To detect a significant correlation between traits and diversification rates, the method requires multiple independent associations between character states and diversification and is thus robust to phylogenetic pseudoreplication (Maddison and FitzJohn 2014). Inferences drawn using this method are necessarily limited by the accuracy of the diversification rate reconstructions, which in turn are contingent on the statistical framework used for inference, the shape of the underlying phylogenetic tree, and our confidence in the phylogeny. The method is suitable for both continuous and discrete character states.

The method does not rely on a formal model for coupling between traits and diversification, as in BiSSE and related models (Maddison et al. 2007; Goldberg et al. 2011; Magnuson-Ford and Otto 2012). As such, the method does not enable researchers to reconstruct ancestral character states and does not provide a means to test, via likelihood, the relative fit of alternative models linking character states to diversification. Likewise, the approach does not provide quantitative estimates of evolutionary rate parameters associated with specific character state values. Researchers can visualize the relationship between character states and evolutionary rates at the tips of the tree (Fig. 3), but the statistical test for trait-dependent diversification is presently limited to an assessment of the extent to which the observed correlation deviates from a null distribution generated by structured permutations of rates across the tips of the phylogeny.

Recent studies have found that formal state-dependent models can be characterized by high Type I error rates (FitzJohn 2012; Machac 2014; Rabosky and Goldberg 2015). For neutral characters simulated on real phylogenies, the frequency of false positive associations between traits and diversification can approach 100% (Rabosky and Goldberg 2015). This effect appears to be attributable to a combination of phylogenetic pseudoreplication (Maddison and FitzJohn 2014) and violations of the assumptions of the underlying statistical model (Rabosky and Goldberg 2015). For example, the BiSSE model (Maddison et al. 2007) assumes that all of the variation in diversification rate across a phylogenetic tree can be explained by the two character states under consideration. Rabosky and Goldberg (2015) found that diversification rate variation unlinked to a character state of interest can lead to substantially elevated Type I error rates. Beaulieu and O'Meara (2015) developed a "hidden state" speciation and extinction model (HiSSE), which allows the possibility that diversification dynamics within a phylogeny reflect the effects of an unobserved discrete trait. This model provides researchers with the ability to relax BiSSE's assumption that all diversification rate variation is associated with the trait of interest, but - as for other discrete character models - HiSSE does not address the issue of phylogenetic pseudoreplication.

The permutation test described in this article does not require that character states explain all variation in diversification rate across a phylogenetic tree. Variation in diversification rates, as identified using BAMM or similar modeling framework, is a prerequisite for detecting relationships between traits and diversification, but the method requires no

special provisions to accommodate this variation if the variation is unlinked to the focal character state. Our analyses of neutral characters simulated on the empirical phylogenies of birds and fishes suggests that the permutation test is not characterized by elevated Type I error rates despite considerable rate heterogeneity in the empirical phylogenies. Rabosky and Goldberg (2015) simulated neutral (binary) characters on subtrees drawn from these same phylogenies and found that a majority of such simulations resulted in significant (spurious) associations between traits and speciation.

Tree-Size Constraints

The permutation test described here requires rather large phylogenies to detect significant associations between traits and diversification. The nature of the randomization is such that the effective number of data points with which to detect an effect is a function of the number of distinct evolutionary rate regimes across a phylogenetic tree. For example, the radiation of modern whales (~90 species) can be reasonably well-characterized by a single increase in speciation rates that occurred along one of the stem lineages leading to modern dolphins (Steeman et al. 2009; Rabosky 2014). There are thus, on average, just two diversification rate partitions: the dolphins, and the non-dolphins. Using the permutation test described here, it would not be possible to detect a significant association between traits and diversification rates for this dataset. Even a perfect association between trait and diversification (e.g., Fig. 1a) would have a p-value of approximately 0.5, as roughly 50% of permuted datasets would assign the "fast" speciation rate to the dolphin clade, and 50% would assign the "fast" rate to the non-

dolphin lineages. In light of this consideration, it seems likely that the method will primarily be applicable to phylogenies that include at least several thousand tips. For the phylogenies of fishes and birds (N > 6500), we identified approximately 60 - 80 distinct evolutionary rate regimes. At least for fishes, this was sufficient to detect a significant relationship between speciation and the rate of body size evolution.

However, it is not clear whether any statistical framework will be able to robustly detect trait-dependent diversification on phylogenetic trees of sizes similar to those that have been used previously in many empirical studies (e.g., < 1000 tips). Interestingly, a previous study (Davis et al. 2013) demonstrated that the BiSSE method has low power to detect trait-dependent diversification under many evolutionary scenarios, even if the number of tips is moderate (< 500). Many published studies have reported significant trait-dependent diversification using phylogenetic trees containing fewer than 500 tips. In light of recent concerns (Maddison and FitzJohn 2014; King and Lee 2015; Rabosky and Goldberg 2015), it seems likely that many trait-dependent diversification relationships reported in the literature are not real. As such, we presently lack an understanding of the phylogenetic scope over which we could, in principle, detect associations between character states and diversification rates.

The method described here and its STRAPP implementation is limited by power to infer distinct evolutionary rate regimes. Power to detect distinct rate regimes (shifts) using BAMM is a function of the number of lineages in the regime as well as the extent to which the rates differ from those of the parental rate class. For example, a lineage that

undergoes a massive increase in diversification rates in the very recent past and which consequently has only a handful of descendant lineages may be difficult to detect with BAMM. Likewise, a derived character state that is consistently associated with a substantial drop in speciation will have low power. In the extreme, a character state might lead to speciation rates of zero and at most a single surviving lineage at the tips of the tree possessing the character state. BAMM has very low power to identify evolutionary rate regimes from single lineages, and scenarios such as this would be largely undetectable. Power to detect rate shifts in BAMM is also influenced by the completeness of phylogenetic data. Although the model includes a formal correction for incomplete taxon sampling, power of the approach is maximized when phylogenetic data are complete. BAMM was intended for use on well-sampled phylogenies and we do not recommend application of the method to extremely incomplete phylogenies of higher taxa (e.g., family-level phylogenies of beetles or angiosperms).

Other Concerns

STRAPP is necessarily dependent on all assumptions of the Poisson process model of diversification rate variation currently implemented in BAMM. As pointed out by a reviewer of this manuscript, the "discretization" of diversification rates with BAMM discards useful information about diversification dynamics when rates have varied continuously across the tree. Lineage-specific measures of diversification, such as the weighted path diversification rate estimate of Jetz et al. (2012), potentially capture variation that is ignored by BAMM. However, we presently lack a formal model that can

describe the covariance structure of these rates across the tree. The advantage to the present approach is that it maintains covariances in rate regimes as inferred using a parametric model of diversification rate variation. It is important to recognize that BAMM may be overly conservative in its ability to infer evolutionary rate heterogeneity. Until we understand limits to inference using BAMM, we caution researchers that failure to detect a significant trait-dependent diversification may simply reflect low statistical power.

Experimental Design for Macroevolutionary Analysis

A desirable feature of the present method is that it does not require complete character state data. In fact, the method can be applied to problems where phenotypic data are available for a small fraction of species in the phylogeny. For the analyses of rates of body size evolution in fishes, we were able to detect significant associations with relatively small subsets of the total data (Fig. 5). Note, of course, that diversification rates were estimated from the full (N = 7824) phylogeny. Hence, power to detect diversification shifts will be critically dependent on the size of the phylogeny, but power to detect trait-rate correlations may be high even when trait data are available for relatively few of those species.

This statistical approach enables researchers to assess the relationship between new types of traits and diversification, particularly traits that cannot be obtained without considerable investment of time and money. For example, suppose we are interested in

testing whether phenotypic plasticity in plants promotes speciation. Researchers could design an experiment to measure phenotypic plasticity in several hundred species in response to a common set of environmental perturbations. These species could even be selected, *a priori*, to maximize the power of the planned diversification analyses. Many types of traits fall into this category: researchers could conceivably use a common experimental design to quantify adaptive genetic variation, responses to a common selective agent, thermal acclimation capacity, and so on.

This raises the possibility of performing future diversification analyses around an optimal experimental design. To perform such analyses, researchers would quantify traits from a subset of taxa that maximize statistical power to detect an effect, provided such an effect exists. A rough guideline for maximizing statistical power can be formulated as follows. Consider a phylogenetic tree of N taxa, from which a researcher can select a subset of size k from which to quantify some variable of interest. Power is maximized by choosing the set of taxa k such that they span the maximum number of reconstructed diversification rate regimes. For example, if a BAMM analysis identified 100 evolutionary rate regimes across a phylogeny of all land plants, a good experimental design would involve quantifying phenotypes from k focal species that were selected to represent as many distinct rate regimes as possible. Clearly, if all k species were sampled from one or two rate regimes, we would not have power to detect an effect of traits on diversification, because all species would essentially belong to the same evolutionary rate class.

Conclusions

We introduced a permutation-based method for testing whether organismal traits are associated with differential rates of species diversification. The method can be applied to discrete and continuous data. It is limited by the amount of diversification rate heterogeneity that occurs across a phylogenetic tree and by the accuracy of diversification rate estimates. At present, the method cannot be applied to discrete data with more than two character states. In our example analyses, we accounted for parametric uncertainty in evolutionary rates by conducting analyses across a posterior distribution of macroevolutionary rate configurations simulated using BAMM. However, it would be straightforward to extend this to account for phylogenetic uncertainty, by conducting analyses across the joint posterior distribution of phylogenetic tree topologies and macroevolutionary rate configurations. The method is robust to asymmetric rates of character change and has low Type I error rates for neutral characters simulated on empirical phylogenies. Finally, the method provides researchers with the potential to design comparative studies such that they include sets of taxa that maximize statistical power to detect trait-diversification relationships.

SUPPLEMENTARY MATERIAL

Data are available from the Dryad Digital Data Repository: doi:10.5061/dryad.kp93h

FUNDING

This research was supported in part by the University of Michigan, the David and Lucile Packard Foundation, and the U.S. National Science Foundation [DEB-1256330].

ACKNOWLEDGEMENTS

We thank R. FitzJohn, E. Goldberg, W. Maddison, and S. Otto for discussions of the general problem of testing hypotheses about trait-dependent diversification. We thank F. Anderson, R. Ree, V. Savolainen, and two anonymous reviewers for comments on the manuscript.

REFERENCES

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.

Armenta, J. K., P. O. Dunn, and L. A. Whittingham. 2008. Quantifying avian sexual dichromatism. J. Exp. Biol. 211:2423-2430.

Beaulieu, J.M., and B.C. O'Meara. 2015. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. bioRxiv doi: http://dx.doi.org/10.1101/016386

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.

Farrell, B. D., D. E. Dussourd, and C. Mitter. 1991. Escalation of plant defense - do latex and resin canals spur plant diversification? American Naturalist 138:881-900.

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.

Goldberg, E. E., J. R. Kohn, R. Lande, K. A. Robertson, S. A. Smith, and B. Igic. 2010. Species selection maintains self-incompatibility. Science 330:493-495.

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.

Grafen, A. 1989. The phylogenetic regression. Phil. Trans. R. Soc. Lond. B 326:119-157.

Ho, L. S. T. and C. Ane. 2014 A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Systematic Biology 63: 397-408,

Hodges, S. A., and M. L. Arnold. 1995. Spurring plant diversification: Are floral nectar spurs a key innovation? Proceedings of the Royal Society of London Series B-Biological Sciences 262:343-348.

Huang, H., and D. L. Rabosky. 2014. Sexual selection and diversification: Reexamining the correlation between dichromatism and speciation rate in birds. Am. Nat. 184:E101-14

Ives, A. R., and T. Garland, Jr. 2010. Phylogenetic Logistic Regression for Binary Dependent Variables. Systematic Biology 59:9-26.

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. In press.

Kay, K., C. Voelckel, J. Yan, K. Hufford, D. Kaska, and S. A. Hodges. 2006. Floral characters and species diversification *in* L. Harder, and S. Barrett, eds. Ecology and Evolution of Flowers. Oxford Univ. Press, Oxford.

King, B., and M. S. Y. Lee. 2015. Ancestral state reconstruction, rate heterogeneity, and the evolution of reptile viviparity. Syst. Biol. 64: 532 - 544.

Machac, A. 2014. Detecting trait-dependent diversification under diversification slowdowns. Evol. Biol. 41:201-211.

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

Maddison, W. P., and R. FitzJohn. 2014. 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.

Martins, E. P., and T. A. Hansen. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646-667.

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. Phil. Trans. R. Soc. B 344:77 - 82.

Pyron, R. A., and F. T. Burbrink. 2014. Early origin of viviparity and multiple reversions to oviparity in squamate reptiles. Ecol. Lett. 17:13-21.

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

Rabosky, D. L., S. C. Donnellan, M. Grundler, and I. J. Lovette. 2014a. Analysis and Visualization of Complex Macroevolutionary Dynamics: an Example from Australian Scincid Lizards. Systematic Biology 63:610-627.

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

Rabosky, D. L., M. Grundler, C. Anderson, P. Title, J. Shi, H. Huang, J. W. Brown, and J. Larson. 2014b. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. Methods Ecol Evol 5:701-707.

Rabosky, D. L., F. Santini, J. M. Eastman, S. A. Smith, B. Sidlauskas, J. Chang, and M. E. Alfaro. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nature Communications 4:10.1038/ncomms2958.

Rolland, J., F. L. Condamine, F. Jiguet, and H. Morlon. 2014. Faster speciation and reduced extinction in the tropics contribute to the mammalian latitudinal diversity gradient. PLoS Biology 12:e1001775.

Steeman, M. E., M. B. Hebsgaard, R. E. Fordyce, S. Y. W. Ho, D. L. Rabosky, R. Nielsen, C. Rahbek, H. Glenner, M. V. Sorensen, and E. Willerslev. 2009. Radiation of Extant Cetaceans Driven by Restructuring of the Oceans. Systematic Biology 58:573-585.

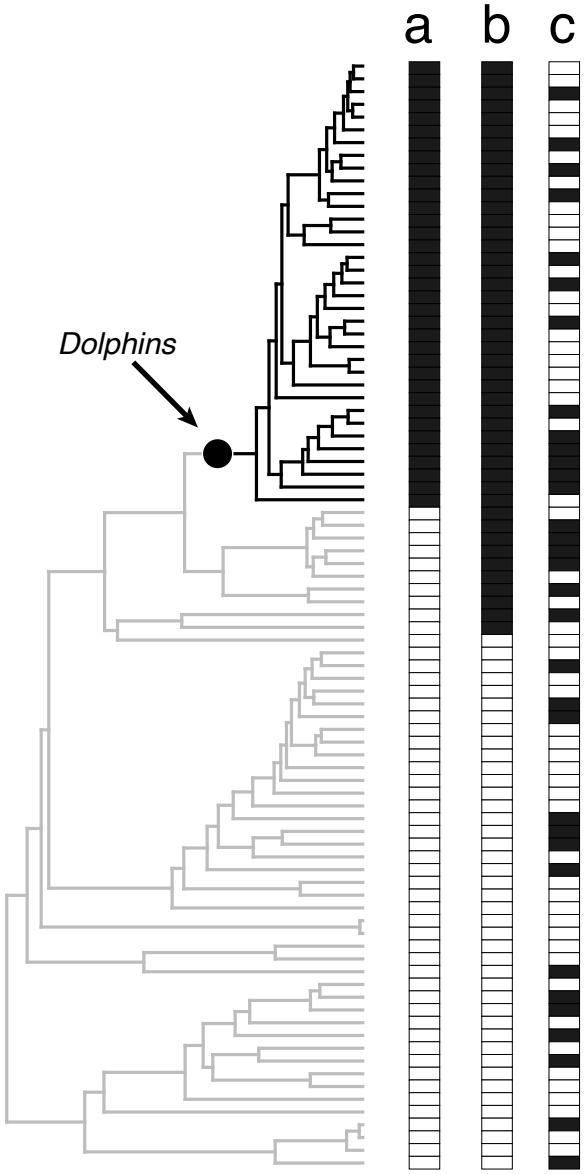

Figure 1. Character state distributions that result in overinflated statistical support for trait-dependent diversification (a) or spurious associations between traits and diversification (b, c), when using formal state-dependent speciation-extinction models. Example phylogeny is for 87 species of living whales; BAMM analysis of this tree finds strong support for a diversification rate increase in the ancestor of the dolphin clade (indicated by the black dot on the phylogeny; Rabosky 2014). Open and filled rectangles

represent the two states of a binary trait. Any synapomorphy of the dolphin clade (a) would be significantly associated with diversification, as discussed by Rabosky and Goldberg (2015). However, this is a form of phylogenetic pseudoreplication (Maddison and FitzJohn 2014), as there is only a single independent association between character state and diversification. False positive results may also occur with the traits that are weakly associated with the rate shift (b) and even with traits that lack phylogenetic signal (c).

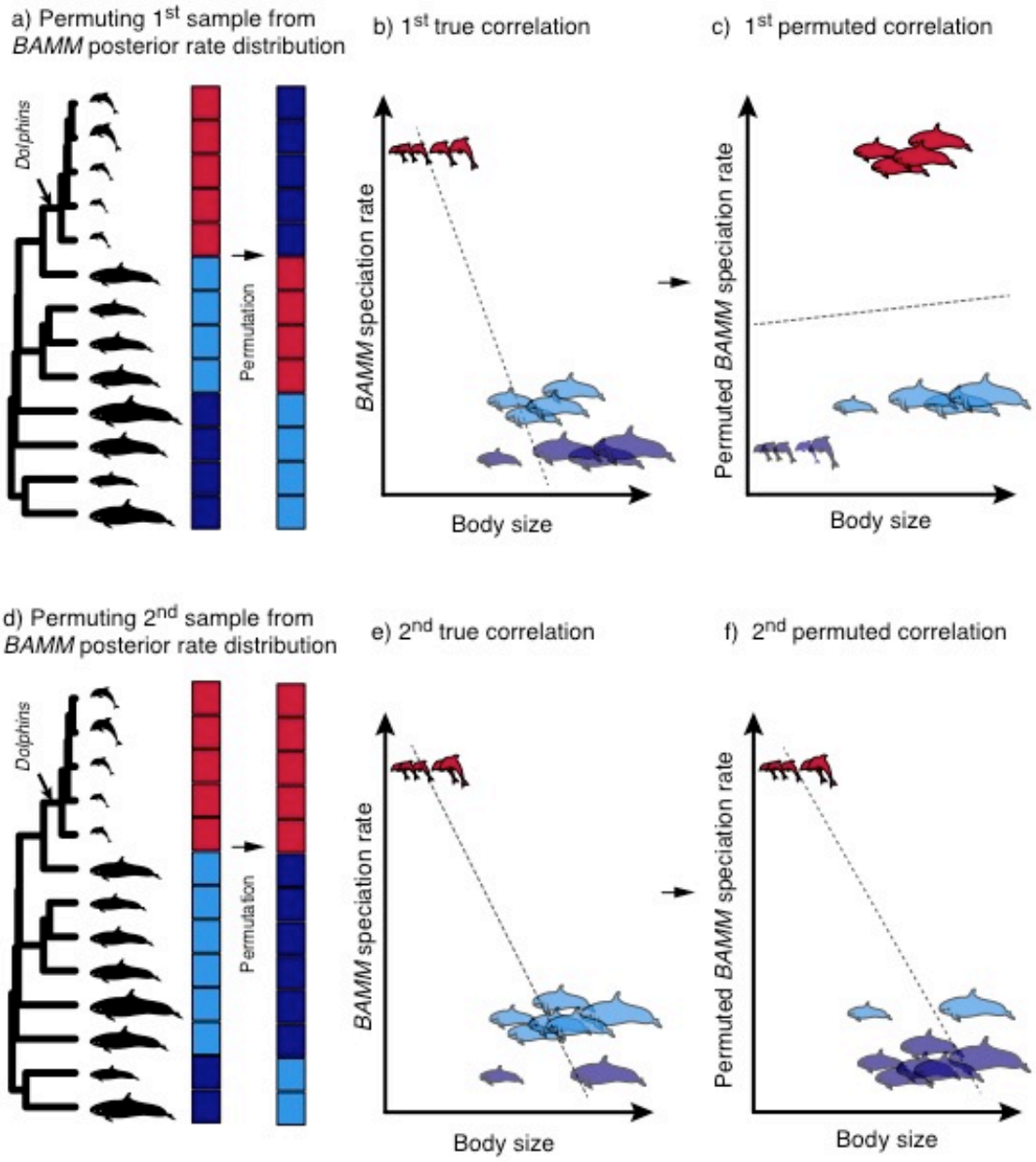

Figure 2. Procedure for the structured permutation test (STRAPP) illustrated using simplified whale phylogeny with species' body sizes represented by the corresponding icon size. For each replicate of the permutation test, we randomly draw one sample from the BAMM posterior rate distribution (a; first column with colored cells). Estimated present-day speciation rates for each taxon are represented as colored cells in a column,

with warmer and cooler colors denoting faster and slower rates of speciation, respectively. For any given sample from the posterior, all species assigned to the same diversification regime have identical rates. On average, species in the dolphin clade are characterized by faster rates (red) relative to other whales (Rabosky 2014). To perform a STRAPP analysis, the speciation rates are permuted among tips while maintaining the covariance in evolutionary rate regimes among species (a, second column): any pair of cells with the same color in the true distribution will also have identical colors in the permuted distribution, but the colors (rates) are randomly assigned to blocks of species. Test statistics (e.g., Spearman's correlation coefficient for continuous traits) are calculated for the relationship between the focal trait and the evolutionary rates for both the observed (b) and permuted (c) rate distributions. For the first posterior sample top row, the empirical correlation (b) is stronger than permuted null correlation (c), but this is not the case for the second posterior sample (bottom row). We estimate the significance of the relationship by tabulating the percentage of permutations where the null correlation is more extreme than the observed correlation. As can be seen from this example, one is unlikely to obtain a significant result if the phylogeny can be described by a small number of diversification rate regimes because such associations can arise frequently in the permuted distribution by chance alone (bottom row). Detecting a significant association between a trait and speciation rate would require that the data contain many more shifts than are illustrated here.

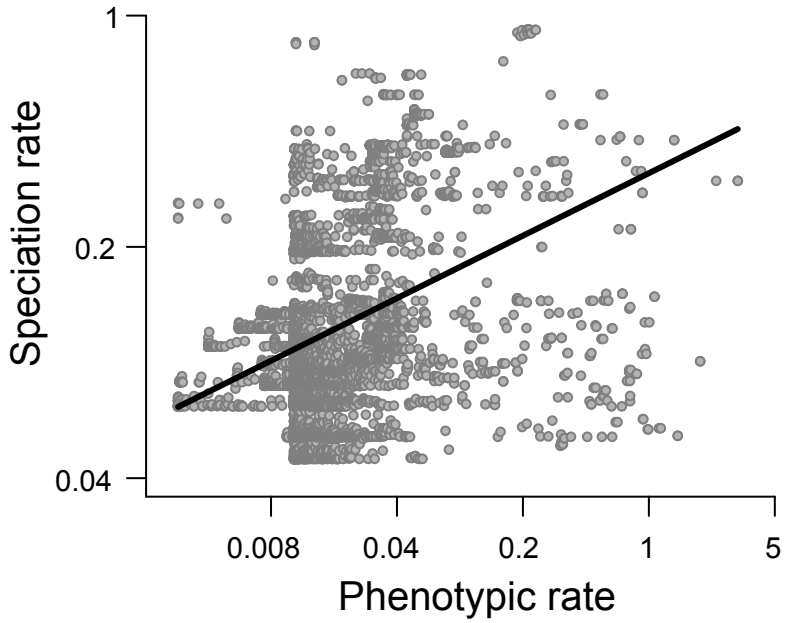

Figure 3. Relationship between tip-specific estimates of the rate of phenotypic evolution (body size) and the rate of speciation for 6,670 species of ray-finned fishes. Black line is OLS regression with no correction for non-independence of data. Tip rates were estimated using BAMM; each value is the mean of the marginal posterior distribution of rates for individual species. Rates are shown on logarithmic scales.

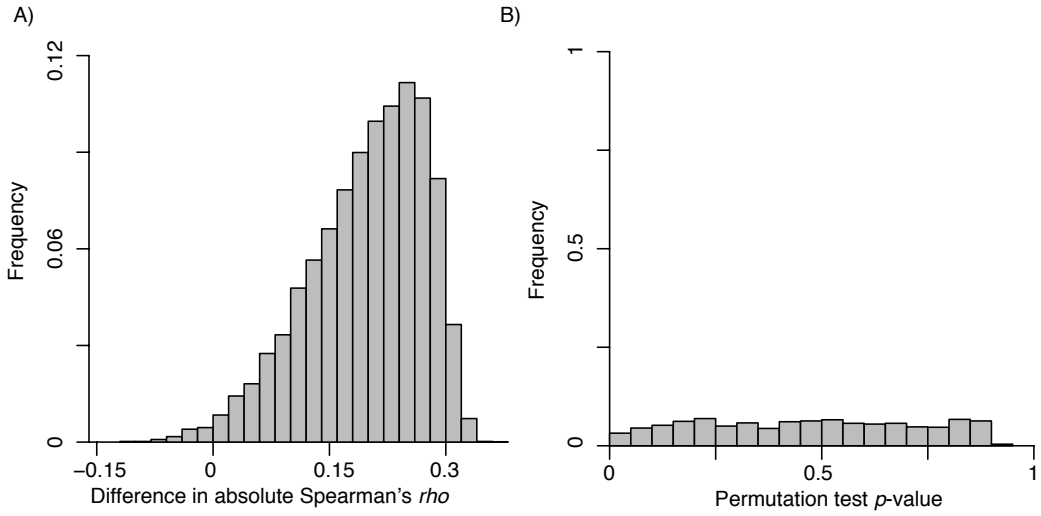

Figure 4. Structured permutation tests for the effect of rate of body size evolution on speciation rate in ray-finned fishes. a) Spearman's *r* between the focal trait and speciation rate was calculated across 1,000 pairs of BAMM posterior and permuted rate distributions, as shown in Figure 2. Plot shows the frequency distribution the absolute difference in *r* for each pair of observed and permuted samples from the posterior. The distribution is strongly right-biased and not centered on zero, suggesting that — for most samples from the BAMM posterior — the observed correlation is greater than expected under random rate permutations ($p = 0.01$). b) Distribution of *p*-values from application of STRAPP to neutral characters simulated on the fish phylogeny under Brownian motion (n = 1000 datasets), indicating that the method is not characterized by elevated Type I error rates.

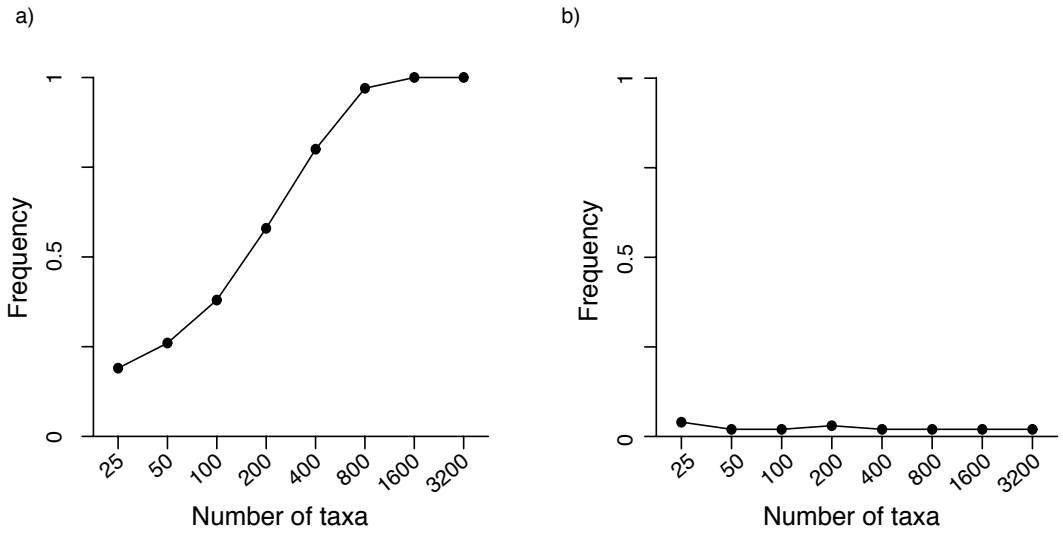

Figure 5. Subsampling analyses. a) Power to detect a significant correlation between speciation rate and the rate of body size evolution decreases as a function of the number of taxa included included in the analysis. Power is the proportion of subsampled datasets (N = 100 per subsampling level) that show a significant association ($p < 0.05$) between rate of speciation and body size evolution. Analyses performed with 800 randomly-sampled tips have approximately the same power as the full dataset (N = 6670 species). b) Type-I error rates (assessed at $\alpha = 0.05$) are invariant with respect to levels of subsampling.

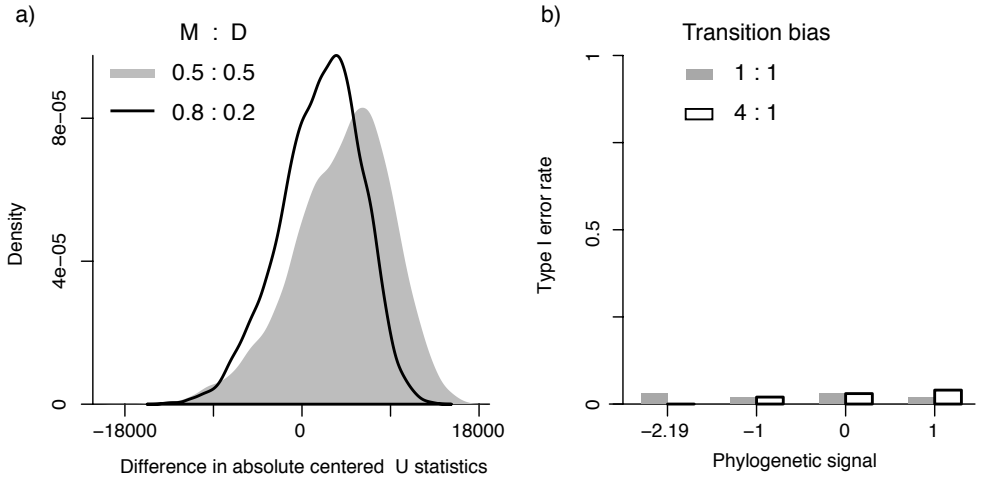

Figure 6. STRAPP analysis of binary trait data using speciation rates inferred with BAMM across a time-calibrated phylogeny of 6,670 species of extant birds. a) Effects of avian plumage dichromatism on speciation rate. Density curves show the distributions of the differences between Mann-Whitney U statistics computed for observed and permuted speciation rate distributions; values greater than zero indicate a positive association between speciation and dichromatism, and values less than zero indicate a negative association. For two-tailed test, the *U* statistic was centered by its theoretical mean—half of the product of the number of species in different trait state. Two cutoffs were used to code the dichromatism data, resulting in two ratios of monochromatic vs. dichromatic species – 50: 50 (gray shaded area) or 80:20 (black line). b) Type-I error rates for simulated (neutral) binary traits on the bird phylogeny. There is no appreciable effect of phylogenetic signal on Type I error rates. Bias in transition rates—1:1 (i.e., no bias, gray bars) and 4:1 (open bars)—does not influence the propensity for false positives either.