RUNNING TITLE: Fair tests of model-based partitioning

BAMM at the court of false equivalency: a response to Meyer and Wiens

Daniel L. Rabosky

Museum of Zoology & Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109-1079, USA

Contact information:

Email: drabosky@umich.edu

Author contributions: DLR designed the study and wrote the paper.

Acknowledgements: I thank M. Alfaro, C. Ané, N. Lartillot, A. Rabosky, an anonymous reviewer, and members of the Rabosky lab for comments on the manuscript and associated discussion.

Data archival statement: Data and code will be archived in Dryad upon acceptance.

ABSTRACT

The software program BAMM has been widely used to study rates of speciation, extinction, and phenotypic evolution on phylogenetic trees. The program implements a model-based clustering algorithm to identify clades that share common macroevolutionary rate

dynamics and to estimate parameters. A recent simulation study by Meyer and Wiens (M&W) argued that (i) a simple inference framework ("MS") performs much better than BAMM, and (ii) evolutionary rates inferred with BAMM are poorly correlated with true rates. I address two statistical concerns with their assessment that affect the generality of their conclusions. These considerations are not specific to BAMM and apply to other methods for estimating parameters from empirical data where the true grouping structure of the data is unknown. M&W constrain roughly half of the parameters in their MS analyses to their true values, but BAMM is given no such information and must estimate all parameters from the data. This information disparity results in a substantial degrees-of-freedom advantage for the MS estimators. When both methods are given equivalent information, BAMM outperforms the MS estimators.

INTRODUCTION

Within ecology and evolution, there is great interest in model-based methods for data partitioning. Such methods allow researchers to infer hidden group structure from empirical data and to estimate associated parameters of interest. For example, model-based clustering is widely used to classify individuals into subpopulations that differ in phenotypes, allele frequencies, and other traits (e.g., STRUCTURE: Pritchard et al. 2000; BAPS: Corander et al. 2008; Gaussian mixture modeling: Cadena et al. 2017). In phylogenetics, model-based partitioning is widely used to identify and accommodate variation in the rate of molecular evolution among sites and across the branches of phylogenetic trees (Drummond and Suchard 2010; Heath et al. 2011; Lanfear et al. 2014). Model-based data partitioning can reveal heterogeneity in the processes of diversification and trait evolution and has thus been used extensively in macroevolutionary studies (Alfaro et al. 2009; Eastman et al. 2011; Venditti et al. 2011; Uyeda and Harmon 2014). Such analyses typically attempt to partition phylogenetic trees

into non-overlapping subclades that differ in parameters of interest related to either species diversification or to the tempo and mode of trait evolution.

The software program BAMM (Rabosky 2014) is a Bayesian framework for inferring heterogeneity in rates of species diversification and phenotypic evolution across phylogenetic trees. The underlying parametric model in BAMM assumes that phylogenetic trees have been shaped by a collection of distinct macroevolutionary rate regimes. The software implementation uses reversible jump Markov chain Monte Carlo to simulate posterior distributions of rate shift configurations that are consistent with the observed data (Rabosky, 2014). Although the mathematics and implementation algorithms underlying BAMM are complex, the method is essentially a cluster analysis that provides both parameter estimates and probabilistic measures of support for inferred group structures. The BAMM algorithm, model assumptions, and performance have been described in detail elsewhere (Rabosky 2014; Rabosky et al. 2014a; Mitchell and Rabosky 2016; Rabosky et al. 2017).

A recent article in Evolution (Meyer and Wiens 2017; hereafter, M&W) posed an important and timely problem in macroevolutionary inference: given a phylogenetic tree and a set of named higher taxa, how should we estimate diversification rates for those clades? M&W generated simulated phylogenetic datasets that contained both higher taxonomic categories and diversification rate heterogeneity, and they assessed how well several methods for rate estimation fared at recovering the true rates for each (simulated) higher taxon. Specifically, they compared parameter estimates from BAMM to those obtained from a much-simpler "method-of-moments" estimator (Magallon and Sanderson 2001) and found that these simple estimates (hereafter, "MS") performed substantially better than BAMM. The MS estimators can be used to compute the maximum likelihood estimate of a clade's net diversification rate under a constant-rate birth-death process, given the stem age and species richness for the clade. M&W concluded that BAMM provides unreliable parameter estimates and should not be used. This conclusion is

at odds with those drawn from a more extensive simulation study (Rabosky et al., 2017), which found that diversification inferences from BAMM were both reliable and consistent.

In this article, I demonstrate that the primary conclusions of M&W are not justified, due to several concerns in their experimental design that largely predetermine the outcome of their assessment. One issue in M&W involves a comparison between non-equivalent inference frameworks that differ substantially in the amount of information they are given by the investigators. Specifically, M&W provide the MS estimators with perfect information about the locations of rate shifts across the tree and simply compute the rate estimates for each true group in the dataset. BAMM is provided with no information about the true locations of rate shifts and must estimate group structure across the phylogeny prior to parameter estimation. The MS analyses are performed after M&W have set the values of many parameters to their true values; BAMM is forced to estimate the same parameters from the data, and the researchers do not account for this difference in parameters. M&W thus perform an uncorrected comparison between two modeling frameworks that differ substantially in their degrees of freedom, and the outcome is clear even before the experiment is performed.

A second concern regarding M&Ws experimental design is more nuanced, but involves the conflation of hypothesis testing and parameter estimation. Their analyses confound small effect sizes across treatment groups with error in parameter estimates and are thus unable to assess parameter reliability. In the extreme, this latter issue allows treatment groups with small effect sizes and/or highly-unbalanced experimental designs to generate low correlations between true and estimated parameter values, even as absolute error in the parameter estimates approaches zero. All methods for jointly inferring group structure and estimating parameters are susceptible to these assessment concerns, including nearly all model-based frameworks for data partitioning in evolution, ecology, and systematics.

SCOPE OF THE PRESENT ARTICLE

M&W includes a large number of analyses, most of which are affected by the two statistical issues I will describe. Hence, I will only revisit a subset of their results and will not repeat the same summaries across all combinations of parameters and simulation conditions. I will focus primarily on the comparison between net diversification rates (speciation minus extinction) as estimated with BAMM and those obtained with MS estimators. The results obtained below are available through supplementary tables that accompany M&W. I also repeated the BAMM analysis exactly as described by M&W for the first dataset (tree "A") in their article and use these results below. The BAMM results I obtained from my independent analysis yield nearly identical results to those reported by M&W. The relationship between subclade mean diversification rates obtained in my re-analysis versus those provided by M&W (rows 1-10 in Table S1) contain only trivial numerical discrepancies (linear regression: slope = 1.015, intercept = -0.002, $r^2 = 0.999$). None of the analyses and results performed below relate to technical aspects of the BAMM analyses, and M&W appear to have executed their BAMM analyses in a manner consistent with developer recommendations. All computer code and results from this article are available through Dryad, doi: ######).

EXPERIMENTAL DESIGN IN M&W

M&W compare the performance of BAMM and MS estimators across a set of 20 simulated phylogenetic datasets. Each simulated phylogeny was created by first generating a backbone tree of 10 tips. Each of these 10 tips was destined to represent a subclade with a unique speciation-extinction parameterization. For each of the 10 tips, M&W sampled speciation and relative extinction parameters from a uniform distribution. Complete species-level phylogenies were then simulated under the sampled parameters, such that the simulated subtrees had a stem clade age that was identical to the corresponding terminal branch length.

M&W then replaced each of the original 10 tips with a subtree generated under a unique speciation-extinction parameterization. Each phylogeny thus contains a backbone tree and exactly 10 "rate shifts", and each shift defines a subclade that contained between 10 and 1401 tips. For consistency of terminology, I refer to each clade with distinct rate parameters as a "rate class" or "true group"; there are exactly 10 true groups per tree that can be discovered by BAMM or any other method.

For each simulated phylogeny, M&W then simulated posterior distributions of macroevolutionary rate regimes using BAMM. They summarized their BAMM analyses by computing mean rates of speciation, extinction, and net diversification for each of the 10 true groups using summary functions from BAMMtools (Rabosky et al. 2014b). The mean rate for a given true group is simply the mean of the marginal posterior distribution of rates for all branches that belong to the group. For BAMM to accurately estimate distinct rates for each of the 10 groups, it would first be necessary for the program to correctly infer the grouping structure in the data (e.g., the locations of all 10 true groups). If BAMM fails to infer any groups (e.g., finds no shifts), then the rates estimated for all groups will be similar.

To determine whether the BAMM estimates are "good" or "bad", M&W perform a second analysis where they computed the analytical MS estimates of diversification rate for each of the 10 true groups. That is, they cut the tree into the 10 groups to which they have assigned distinct rate parameters, and estimate rates separately for each group. The MS estimates are far less complex than BAMM: for stem clades and with zero extinction, the MS estimate of diversification rate is simply the logarithm of species richness divided by time. To summarize results, M&W compute the proportional error of the rate estimates for each of the 10 true groups for each tree and express it as a percentage. This is computed as $(R_E - R_T) / R_T$, where R_T are R_E are the true and estimated rates for the focal group. They find that the MS estimators have lower error than the BAMM estimates (M&W: Figure 1). They also find that the slopes of

the relationships between true and estimated rates across all groups are more accurate for the simple MS estimators than for BAMM (M&W: Table 1), and that MS estimators are better able to detect true variation in diversification rates (M&W: Fig. 5).

NONEQUIVALENCE OF INFERENCE MODELS IN M&W

The statistical comparisons between BAMM and MS made in this fashion are not equivalent and strongly favor MS because the MS estimators are informed of the precise number and location of the true groups (e.g., rate shifts). Figure 1 summarizes the difference between these comparisons as performed by M&W. Consider a scenario in which a researcher is given a large set of body size measurements and asked to estimate the number of true populations from which the measurements were drawn, along with the means of those populations, in the absence of any other identifying information. To address this problem, we might perform a clustering and estimation analysis by modeling the distributions of sizes as a mixture of distributions (e.g., Gaussian clustering). Now, suppose that we are given additional information about each observation in the dataset: specifically, information about the precise subpopulation from which each of the measurements was drawn is now provided. We then perform a secondary analysis where the original data is simply partitioned by true subpopulation membership, and the sample means are computed for each of the true groups.

If we somehow knew the true means of each population, we would likely observe that the second approach – partitioning the data with true grouping information in hand – would provide greater accuracy than the mixture modeling approach, because the mixture model approach must estimate group structure from the data. This is precisely the comparison used by M&W: they provide the MS estimators, but not BAMM, with the true group structure of the data. For MS estimators, M&W compute averages after partitioning the data into subsets for each rate shift, and they only know where the shifts have happened because M&W created the simulation

scenarios. It is unsurprising that MS estimators outperform BAMM under such conditions, and the M&W comparison is equivalent to comparing statistical models that differ by a large number of parameters without controlling for the difference. Many statistical methods for partitioning data into groups and estimating population parameters would likely appear to perform poorly under such an assessment scenario.

Because true locations of rate shifts are generally unknown, assessing the performance of MS estimators under simple scenarios where rate shift locations are known without error – as in M&W – should provide a highly selective view of their performance. There is presently little evidence that named higher taxa (e.g., genera, families, phyla) are universally or even largely concordant with macroevolutionary rate shifts (Smith et al. 2011) so it is essential to understand how MS estimators perform relative to BAMM when applied to clades that may or may not be associated with rate shifts.

FAIR COMPARISONS BETWEEN MS AND BAMM

To determine the best method for estimating diversification rates for higher taxa, there are several approaches M&W could have used to perform more-or-less equivalent comparisons between MS estimators and BAMM. An obvious control experiment, which was not performed by M&W, is to repeat their analyses for clades other than the precise set that they have seeded with rate shifts. For empirical datasets, we typically have no knowledge of the potential rate shift locations. Hence, it is important to know how the M&W inference framework would fare if applied to clades that are sampled at random with respect to their "true group" assignment. To perform this comparison, I computed stem and crown MS estimates for all clades with at least 10 taxa from the first tree (tree "A") from M&W (Fig. 2, top row). The threshold of 10 taxa was chosen because M&W required their simulated shift clades to also contain at least 10 extant taxa. Then, using the results from a single BAMM analysis of the complete phylogeny (e.g.,

including all clades), I summarized the BAMM estimates of net diversification rate separately for all subclades exactly as in M&W using the BAMMtools getCladeRates function (Rabosky et al. 2014b). If BAMM found no evidence for rate variation at the scale of the full tree, then the mean rates computed for each subclade would be nearly identical. My results for this exercise are thus those that M&W would have obtained if they applied MS estimators to all clades, rather than selectively applying those estimators only to those clades to which they had assigned rate shifts. For the first tree (tree "A"), there are a total of 548 clades with at least 10 tips; by restricting their assessment to the 10 true groups, M&W tested estimation bias for a select set (2%) of potential higher taxa.

When MS estimators are applied to this more general set of clades, they perform far worse than BAMM (Fig. 2). The mean absolute proportional error in BAMM estimates for such clades is 12.2%, versus 38.7% for MS stem estimators and 47.3% for MS crown estimators. The reason for the poor performance of the MS estimators, relative to BAMM, is that the MS estimators are highly sensitive to stochastic variation in species richness due to the diversification process itself. If clades are simulated under a fixed speciation-extinction parameterization, one will observe stochastic variation in richness, and the MS estimators will track this variation closely. BAMM is more conservative because it uses information from the full tree when determining whether a given subclade is sufficiently distinct (e.g., significantly different) such that it should be assigned its own rate parameters.

The preceding exercise is not offered as a serious comparison of BAMM and MS estimators: it merely highlights the problems with comparing inference frameworks where one approach (MS) is given prior knowledge of true shift locations, but where the other (BAMM) is given only the data that would typically be available to researchers in practice. Several additional approaches could have used to determine which method performs better at estimating rates for higher taxa. First, M&W could have used formal data partitioning methods

to identify the clades to which MS estimators should be applied. For example, the MEDUSA method is, to some degree, a statistical approach for finding best-fit locations for applying MS estimators across a phylogeny. I performed such a comparison in the original BAMM description (Rabosky 2014), finding that BAMM performed at least as well as MEDUSA for the set of simulation scenarios considered. The eigengap method of Lewitus and Morlon (2016) is conceptually distinct from MEDUSA and BAMM, but nonetheless allows researchers to non-arbitrarily partition trees into clades that differ in their underlying diversification dynamics.

Second, M&W could have conditioned their BAMM analyses on the number and location of the true shifts, just as they have done for the MS estimators. In fact, this test is essentially what M&W did when they analyzed each true group (rate class) separately. Their results showed that BAMM performed very well for this test (M&W: figure 2). The BAMM and MS models are still non-equivalent, because they allowed BAMM to have multiple shifts within each true rate class. However, this decision should have made BAMM perform worse than the MS estimators, not better, because it imposes additional and unnecessary complexity on the BAMM model that is not present in the MS estimation framework.

Using results from M&W Table S5, I compared the proportional error in rate estimates from BAMM and from MS stem and crown estimators. Remarkably, BAMM outperforms the MS estimators under complete and incomplete taxon sampling (Table 1), directly contradicting the primary conclusions of M&W. The mean proportional error (bias) is lower for BAMM than for all MS stem or crown estimators. Furthermore, the mean absolute error is similar to or much better than all MS estimators used by M&W. It is worth noting that this comparison is irrelevant for empirical datasets, because researchers cannot condition on the location of the true shifts, which are unknown.

PRIOR SPECIFICATION: NOT THE SAME AS CONDITIONING

M&W imply that they have made a fair comparison, noting: "...we set the expected number of shifts to 10, given that each tree had 10 clades, each with random and independent diversification rates. Thus, we seeded the BAMM analyses with a number close to the actual number of rate regimes, even though this number would be unknown in empirical analyses." However, manipulation of a general tree-wide prior is not equivalent to conditioning the analysis on a specified number of shifts, for two reasons. First and most importantly, the posterior on the number of shifts is largely independent of the prior (Mitchell and Rabosky 2016; Rabosky et al. 2017) and specifying a prior is not seeding a tree with a specific number of shifts. In fact, M&W note that their estimates are largely independent of the prior, so they acknowledge that they are not seeding the analyses with 10 rate shifts. The mean number of shifts they found across each tree in their analyses with complete sampling was only 2.35, which rejects the idea that they are informing BAMM that there are 10 shifts in each dataset.

Second, even if M&W had conditioned their BAMM analyses on containing exactly 10 rate shifts, the comparison would be nonequivalent, because the MS estimators are given both the number of shifts and their precise locations. As an example, consider the first tree (tree "A") in the M&W dataset. This tree contains 5568 branches on which BAMM could place the 10 rate shifts. If we condition the analysis on exactly 10 rate shifts, the prior probability of a rate shift on any of the 10 true "shift branches" ranges from 0.0006 to 0.007 (Supporting information), and the prior odds that BAMM will place shifts on all 10 of these branches is the product of the 10 probabilities, or roughly 10^{-27}. For the MS estimators, these prior probabilities are 1, because the shifts are fixed to their true locations. Hence, even if BAMM was seeded with 10 shifts, the prior odds ratio favoring the MS estimators is on the order of $1 / 10^{-27} \approx 10^{27}$.

HYPOTHESIS TESTING VERSUS PARAMETER ESTIMATION

A seemingly obvious strategy for assessing the reliability of rates estimated using BAMM and other methods is to compare the correlation between the true evolutionary rates and the mean rates as inferred with BAMM. This approach is used by both M&W and by Moore et al. (2016) to assess the reliability of BAMM-estimated rates. However, such an assessment strategy suffers from a largely-unappreciated weakness that results when hypothesis testing is performed simultaneously with parameter estimation, as occurs implicitly with BAMM through Bayesian model averaging. For BAMM to obtain unconstrained parameter estimates for a particular subclade, the program must first sample a rate shift on the branch leading to the focal clade; the frequency with which such samples are obtained in the posterior is proportional to the evidence favoring such a shift. If shifts are not sampled on branches immediately ancestral to the focal clade, the corresponding rate estimates will not be independent of those rates inferred for the parent rate class. As the effect size among groups (e.g., clades or rate classes) decreases, the posterior estimates for specific shift groups will increasingly be influenced by information from other parts of the tree (e.g., the global mean rate). In Bayesian statistics, this phenomenon is referred to as shrinkage, whereby estimates for specific subgroups in a hierarchical model will "shrink back" towards some overall central tendency (Kruschke and Vanpaemel 2015). BAMM is effectively a Bayesian shrinkage method that uses a mixture model to determine the extent to which local estimates of rates (e.g., a specific subclade) should be informed by global (tree-wide) information. It is clear that, when clades with rate shifts are small, BAMM tends to overshrink: the method routinely fails to infer the presence of small rate classes and thus, the resulting estimates for the specific clade are driven by the "global" average As Rabosky et al. (2017) noted, the weak correlations that Moore et al (2016) observed between true and BAMM-estimated rates were largely driven by such overshrinking: most rate shifts in the Moore et al (2016) dataset led to clades with fewer than five tips. BAMM generally failed to

detect such rate shifts, such that rates for any local portion of the tree (e.g., a specific branch) largely reflected the tree-wide average rate.

For their main results (M&W Figure 1 and Table 1), M&W compare BAMM, which jointly infers the grouping structure of the data and associated parameter values, to the MS estimators, which simply estimate parameters for data partitions that have been defined *a priori*. For the MS estimators, M&W compute the values for each of the true groups, much as one might compute the arithmetic mean of a set of body size observations from a single true population. The non-comparability of these approaches is easier to understand if we consider that each of the true groups (rate classes) in the M&W phylogenies is essentially a treatment group, and each treatment group has an effect size that is a function of the corresponding phylogeny. For both BAMM and MS estimators, M&W then test whether the estimated group means are correlated with the true values for each of the treatment groups. There are multiple conditions under which this comparison will yield poor performance. If the effect sizes for individual treatments are small, such that rates estimated with BAMM shrink towards the tree-wide mean rate, then the program effectively estimates the overall rate and not a treatment (true group) mean. In the extreme, BAMM might recover no evidence for rate variation, and the correlation between true rates and estimated rates might equal zero even as rates are estimated with very high accuracy (Fig. 3). In contrast, there is no hypothesis testing associated with the MS estimators. The true groups are identified in advance and assumed to be different, and this information is only known because M&W created the simulation scenarios.

Another way of conceptualizing this issue is that M&W have imposed an effect size filter on their BAMM analyses but not their MS estimators. This means that issues of statistical power due to low effect sizes will compromise the performance of BAMM, but not the MS estimators. By implicitly performing hypothesis testing during the process of parameter estimation, through Bayesian model averaging, M&W induce strong non-independence across treatment

groups. If BAMM fails to infer the existence of a particular rate shift (e.g, the posterior probability of a shift is low), the resulting rate estimates for the shift clade will be correlated with or nearly identical to the rates inferred for the parent rate class. The number of rate shifts detected with BAMM is an approximation of the degrees of freedom (Rabosky and Huang 2015), and the mean number of shifts across all M&W full-tree analyses is only 1.87, indicating strong non-independence among group means as computed by M&W. In the extreme, BAMM will find no rate variation and all 10 true groups will have nearly identical rate estimates, meaning that M&W are essentially performing regression analyses with a single observation of the dependent variable. This is not a hypothetical scenario, because 12% of their BAMM analyses reported no detectable rate shifts. For these reasons, simple correlation analyses of true versus estimated rates for the full-tree BAMM analyses are not appropriate (Fig. 3). We have previously used correlation coefficients and regression slope analyses to assess BAMM's performance, but only with explicit consideration of the effect size (e.g., theoretical information content; sample size) of each true group (Rabosky et al. 2017).

There is a simple reason why researchers should be cautious about applying estimators to groups without using either model-based partitioning or an equivalent hypothesis-testing scheme. Sampling error (e.g., variance) is expected to result in numerical differences among groups even when the true (population) parameters are identical. As such, approaches that neglect this sampling error are potentially subject to a high frequency of false positives when characterizing rate variation among clades. In the analyses that underlie M&W Figure 5, the authors describe a statistical test for identifying differences in diversification rates among clades. They apply MS estimators to the small number of sister clade pairs to which they have assigned different rates of diversification, and they define success as any case where the numerical rate estimates are higher for the clade with the faster true rate. Because BAMM frequently returned similar or identical rate estimates for sister clade pairs, the authors

concluded that BAMM generally failed to correctly identify rate heterogeneity when it is present (M&W Fig. 5).

However, M&W do not perform an important control analysis, which is to test whether application of their framework will fail when applied to sister clades that do not vary in diversification rate. In fact, when sister clades have identical rates, the probability of a Type I error given the M&W assessment framework is very high: any stochastic difference in species richness between a pair of sister clades will lead to faster numerical MS estimates for one member of the pair, which they would interpret as a correct inference of differential diversification rate. Figure 4 demonstrates that the M&W assessment framework yields extreme Type I error rates when applied to sister clades with identical diversification rates. In general, numerical differences in means between treatment groups should not be used as a substitute for probabilistic hypothesis testing.

MS ESTIMATORS CAN BE USEFUL

This article is not a critique of MS estimators. Such estimators have proven extremely useful in the field and will continue to be useful, provided the assumptions of the estimators are met and/or the conditions under which they fail are adequately characterized (Rabosky 2009a, b). MS and related estimators allow researchers to extract valuable evolutionary insights from information on clade ages and species richness, even when taxon sampling in the underlying phylogenetic trees is limited (Raup 1985; Magallon and Sanderson 2001; Nee 2006; Ricklefs 2007). Simple methods frequently prove more robust than complex methods to violations of their underlying assumptions. Moreover, there are many groups of organisms for which species-level phylogenies are not presently available.

However, there is no evidence that simple MS estimators can outperform more complex models of diversification dynamics when lineage level phylogenies with at least 25% taxon

sampling are available. In practice, MS estimators might be expected to perform somewhat worse than suggested by M&W, because real phylogenies are likely to contain additional among-lineage or temporal rate heterogeneity within the focal higher taxa. By collapsing large phylogenies (average size: 2022 tips) to higher level phylogenies of just 10 tips, the approach of M&W discards data that can potentially provide greater insights into the nature of rate variation through time and among clades. It is unclear what useful information can be gained by ignoring within-taxon variation in diversification rates in any case where a suitable phylogeny is available for estimating such variation.

PROPER MODEL COMPARISON IS ESSENTIAL

The study by M&W raises a number of important statistical issues that are relevant to assessing any methods for clustering and parameter estimation. There is no question that simple estimators for population data, such as the MS estimators favored by M&W, have utility in ecology and evolution. However, M&W evaluate the performance of MS estimators by applying them only to groups with known (investigator-defined) differences between them, and they interpret any differences between groups as consistent with true variation in underlying parameters (e.g., M&W Figure 5). Because M&W neglect the sampling error (e.g., stochastic noise) associated with real data, their recommended approach performs poorly when applied to data when there are no differences between groups (Fig. 2, Fig. 4).

The approach used by M&W suffers from a second issue that is not widely appreciated. By computing numerical estimates of rates for individual clades, M&W implicitly assume that phylogenies can be carved up into an arbitrary number of higher taxa to serve as largely-independent units (data points) for downstream analyses. If a phylogeny or parts thereof are generated by a single underlying diversification process, the apparent numerical differences in diversification rate between constituent subclades are likely to reflect nothing more than

sampling error due to the inherent stochasticity of the diversification process (Fig. 4b). In light of this observation, it is perhaps unsurprising that some studies using MS estimators for higher taxa have obtained results that cannot be distinguished from a random splatter of data across the tips of the tree (see Rabosky and Adams 2012; Rabosky et al. 2012). BAMM, MEDUSA, and related methods (Morlon et al. 2011; Etienne and Haegeman 2012; Lewitus and Morlon 2016) may provide imperfect solutions for quantifying group structure across phylogenetic trees, but they do not suffer from the illusion of independence that comes from partitioning phylogenetic trees into subgroups that may have been generated under a common diversification process.

IS BAMM OVERLY CONSERVATIVE?

M&W observed that BAMM had low power to infer rate variation for some of their simulated datasets. As they discuss, one consequence of this conservatism is that rate estimates for small clades may essentially reflect a global average that need not be closely correlated with the true rates of the focal clade. BAMM clearly tends to underestimate the true number of rate shifts, and this conservatism has been discussed previously. Upon observing that correlations between true and estimated rates were zero for some fraction of datasets that were analyzed with BAMM, Rabosky (2014; see corresponding Figure 6) wrote: "... branch specific estimates of rates for a multiprocess model may be poor if model underfitting has occurred. In the extreme case, a tree that is estimated to have only a single process may have very similar rate estimates on each branch; the correlation between these rates and the true rates will necessarily be low if the true model includes multiple processes and considerable rate heterogeneity across the tree." There is considerable scope to clarify the causes of this conservatism and/or to determine whether BAMM is excessively conservative.

One way forward is to more explicitly assess BAMM's conservatism in light of the theoretical information content associated with each shift regime. Rabosky et al. (2017) used

such an approach to demonstrate that many of the purportedly rate-variable phylogenies simulated by Moore et al. (2016) were statistically indistinguishable from a constant-rate birth-death process. However, the "rate-shift" subclades generated by M&W are much larger than those in Moore et al. (2016); presumably, the subclades in M&W contain more information. Whether BAMM is overly conservative, relative to other methods that have been or might be devised, remains an open question. At this point, however, there is no evidence that BAMM's rates are unreliable (Rabosky et al. 2017): claims of unreliability in both M&W and in Moore et al (2016) are readily shown to result from BAMM's tendency to underestimate the true number of shifts (e.g., the program is conservative). In the analyses by Moore et al (2016), for example, the low correlation between true and estimated rates results almost entirely from the fact that BAMM failed to infer any rate shifts for most of their rate-variable datasets; in nearly all of these cases, the BAMM-estimated rates are approximately as good as the best tree-wide average rate that can be obtained with other methods (see Rabosky et al., 2017: Figures 11 - 13).

SUMMARY

In this article, I explain why the conclusions of Meyer and Wiens (2017) are not justified. Most significantly, M&W compare inference frameworks that differ substantially in the amount of information they are given by the investigators. By specifying the precise location of rate shifts for the MS calculations, M&W provide those estimators with an advantage that could never be present for real data, because the true location of rate shifts is unknown. The valid comparison in M&W involves a scenario where BAMM is constrained to the same (true) set of rate shifts as the MS estimators; as shown by M&W (M&W: supplementary table S5) and presented here (Table 1), BAMM performed equivalently to or better than both stem and crown MS estimators despite relying on a more complex inference model.

The results of this article should not be construed to imply that BAMM is inherently better at estimating diversification rates for higher taxa. The simulation design used by M&W allows us to compare BAMM and MS for a somewhat unusual scenario, whereby each named higher taxon is uniquely associated with a distinct speciation-extinction parameterization. However, there is no necessary reason why rate variation in real datasets need be associated with higher taxa. Despite the fact that BAMM outperforms MS when the locations of rate shifts are known (Table 1), it is essential to recognize the limits of these testing scenarios. In real empirical datasets, the nature and location of diversification rate variation is unknown. If most variation in diversification rates is partitioned among a set of higher taxa, then researchers should estimate rates separately for each taxon of interest, rather than perform a global BAMM analysis. However, if rate variation is largely decoupled from taxonomic categories, then it is possible that a global BAMM analysis will outperform taxon-specific rate estimates. To fairly determine the relative performance of BAMM and MS as applied to higher taxa, it is important to construct assessment scenarios where the association between rate variation and taxonomic groups is similar to the (largely unknown) relationship in real datasets.

Two additional caveats should be clearly stated. First, the results presented in M&W are limited to a comparison between MS estimators and BAMM. No other inference frameworks were considered, so no conclusions can be drawn about other models or software implementations that might have been used to analyze the same data (e.g., FitzJohn 2012; Morlon et al. 2016). Second, the results of this article pertain to the performance of BAMM when species-level phylogenies, potentially with incomplete sampling, are analyzed with the program and mean rates are then extracted for nested subclades. BAMM should generally not be used to analyze phylogenies of higher taxa, as might occur if a researcher applied BAMM to a phylogeny with single representatives of all family-level clades in a particular group of organisms. The BAMM likelihood function is not appropriate for such data because it describes the likelihood of

a particular branching pattern given the diversification parameters and taxon sampling. The more appropriate likelihood for terminally unresolved clades is the MEDUSA likelihood (Alfaro et al., 2009), which is based on the probability that a given diversification parameterization will produce a clade of the same size as the focal clade. However, incomplete sampling per se is not necessarily problematic for BAMM: as shown by M&W, BAMM performs well with low (25%) taxon sampling (Table 1). FitzJohn et al. (2009) discuss the distinction between skeletal trees with missing taxa (appropriate for BAMM) and trees with terminally-unresolved clades (not appropriate for BAMM).

This article is not intended to discourage independent performance assessments of BAMM and other methods: such testing should be strongly encouraged by the community. Major advances in methods development are often driven by studies that characterize the conditions under which existing methods perform poorly. However, studies that purport to test the relative performance of two methods must ensure the equivalency of the frameworks under consideration (Table 1) and also that adequate control experiments have been performed (Fig. 2, Fig. 4). In the case of Meyer and Wiens (2017), these concerns are sufficient to both overturn and reverse the conclusions presented in their article.

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.

Cadena, C. D., F. Zapata, and I. Jimenez. 2017. Issues and perspectives in species delimitation using phenotypic data: Atlantean evolution in Darwin's finches. Syst. Biol. doi.org/10.1093/sysbio/syx071.

Corander, J., P. Marttinen, J. Corander, and J. Tang. 2008. Enhanced Baeysian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9:539.

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

Eastman, J. M., M. E. Alfaro, P. Joyce, A. L. Hipp, and L. J. Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578-3589.

Etienne, R. S. and B. Haegeman. 2012. A conceptual and statistical Framework for adaptive radiations with a key role for diversity dependence. Am. Nat. 180:E75-E89.

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.

Heath, T. A., M. T. Holder, and J. P. Huelsenbeck. 2011. A Dirichlet process prior for estimating lineage-specific substitution rates. Mol. Biol. Evol. 29:939-955.

Kruschke, J. K., and W. Vanpaemel. 2016. Bayesian estimation in hierarchical models. Pp. 279-299 in J. T. Townsend and A. Eidels (eds). The Oxford Handbook of Computational and Mathematical Psychology. Oxford University Press, Oxford.

Lanfear, R., B. Calcott, D. Kainer, C. Mayer, and A. Stamatakis. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14:82.

Lewitus, E. and H. Morlon. 2016. Characterizing and comparing phylogenies from their Laplacian spectrum. Syst. Biol. 65:495-507.

Magallon, S. and M. J. Sanderson. 2001. Absolute diversification rates in angiosperm clades. Evolution 55:1762-1780.

Meyer, A. L. S. and J. J. Wiens. 2017. Estimating diversification rates for higher taxa: BAMM can give problematic estimate of rates and rate shifts. Evolution doi: 10.1111/evo.13378.

Mitchell, J. S. and D. L. Rabosky. 2016. Bayesian model selection with BAMM: effects of the model prior on the inferred number of diversification shifts. Methods Ecol. Evol. doi: 10.1111/2041-210X.12626.

Moore B.R., S. Hohna, M. R. May, B. Rannala, and J. P. Huelsenbeck. 2016. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. Proc. Natl. Acad. Sci. U.S.A. 113:9569–9574.

Morlon, H., E. Lewitus, F. L. Condamine, M. Manceau, J. Clavel, and J. Drury. 2016. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. Methods Ecol. Evol. 7:589-597.

Morlon, H., T. L. Parsons, and J. B. Plotkin. 2011. Reconciling molecular phylogenies with the fossil record. Proc. Nat. Acad. Sci. U.S.A. 108:16327-16332.

Nee, S. 2006. Birth-death models in macroevolution. Ann. Rev. Ecol. Evol. Syst. 37:1-17.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Rabosky, D. L. 2009a. Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. Ecology Letters 12:735-743.

Rabosky, D. L. 2009b. Ecological limits on clade diversification in higher taxa. American Naturalist 173:662-674.

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

Rabosky, D. L. and D. C. Adams. 2012. Rates of morphological evolution are correlated with species richness in salamanders. Evolution 66:1807-1818.

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

Rabosky, D. L., J. Mitchell, and J. Chang. 2017. Is BAMM flawed? Theoretical and practical concerns in the analysis of multi-rate diversification models. Syst. Biol. 66:477-498.

Rabosky, D. L., G. J. Slater, and M. E. Alfaro. 2012. Clade age and species richness are decoupled across the Eukaryotic tree of life. Plos Biology 10:e1001381.

Raup, D. M. 1985. Mathematical models of cladogenesis. Paleobiology 11:42-52.

Ricklefs, R. E. 2007. Estimating diversification rates from phylogenetic information. Trends in Ecology & Evolution 22:601-610.

Smith, S. A., J. M. Beaulieu, A. Stamatakis, and M. J. Donoghue. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. American Journal Of Botany 98:404-414.

Uyeda, J. C. and L. J. Harmon. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Syst. Biol. 63:902-918.

Venditti, C., A. Meade, and M. Pagel. 2011. Multiple routes to mammalian diversity. Nature 479:393-396.

Figure legends

Figure 1. Illustration of testing procedure used by M&W. Left: true phylogeny with three rate shifts (a, b, c), each with a distinct speciation-extinction parameterization. Middle: MS estimators are applied to the set of clades with rate shifts and no others. Right: BAMM is used to analyze the complete tree, but no information is provided about the number and location of rate shifts. Following completion of the analysis, an *a posteriori* summary is performed where the mean rate is extracted for the three true shift clades. If BAMM fails to identify significant differences in rates between true shift groups, as might occur in this example for clades (a) and (b), the mean rates for each clade will be similar and non-independent, because BAMM will assume that the clades were generated under a shared diversification process.

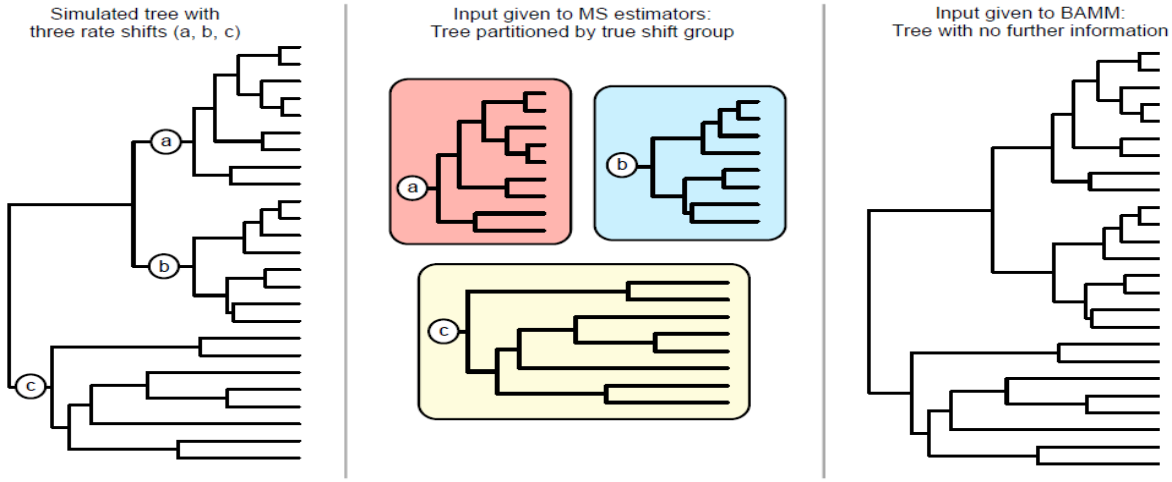

Figure 2. Diversification rates estimated with BAMM are far more accurate than those obtained with MS estimators when applied to clades that are selected without information about the presence or absence of rate shifts. M&W applied the MS estimators only to a small number of clades that were known in advance to be associated with rate shifts (Fig. 1), even though this information would be unknown for real datasets. (A) Illustration of revised testing procedure: MS estimates are computed for all clades, including those not associated with rate shifts. BAMM is applied to the complete phylogeny, and mean rates are extracted for each corresponding subclade. (B, C, D) Proportional error (top rows) and absolute proportional error (bottom rows)

for three estimators of net diversification rate (columns): (B) MS stem age estimator; (C) MS crown age estimator; (D) BAMM. Estimation error for BAMM is far lower than both the crown and stem age MS estimators; gray polygons indicate 10% and 90% limits on the distribution of proportional error estimates. Interquartile range in error for the BAMM estimates is (7.8%, 9.4%), versus (12.8%, 45.3%) for MS-stem and (11.7%, 58.0%) for MS-crown. Outliers with absolute error percentages exceeding 200% are omitted from the bottom panels, but the MS estimators contain many more such outliers than BAMM (MS-stem, 9 outliers; MS-crown, 21 outliers; BAMM, 1 outlier). This analysis uses the first tree (tree "A") from the M&W dataset; relative extinction for MS estimators was 0.5.

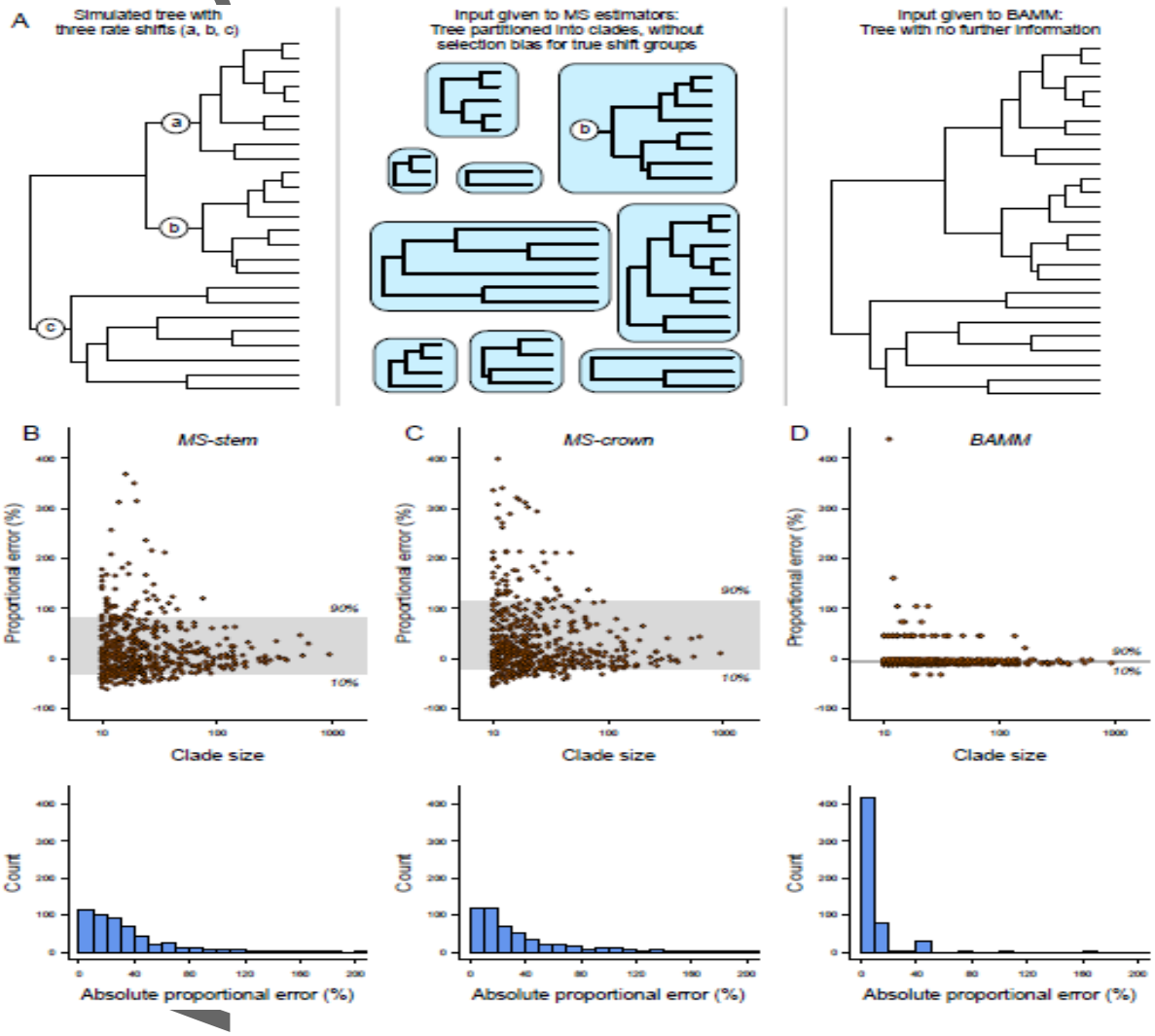

Figure 3. Two scenarios under which low effect sizes may compromise correlation-based assessments of BAMM and other clustering methods. Identity line is shown for reference (dashed). Left: true groups (black, white) show little variation in parameter values, such that the method assigns all groups to the same parameter class. The absolute error in this example may be low, but estimates can nonetheless be uncorrelated with the true values. Right: true parameter values differ substantially between groups, but the effect size of one or more groups is small due to highly unbalanced sampling. Even as parameter estimates are accurate across 99% of the combined data, the correlation coefficient is zero, because the estimated rates are identical for all groups. Diversification studies are particularly susceptible to unbalanced sampling across groups, because the amount of data within treatments (e.g., subclade size) will generally be correlated with the corresponding diversification parameters.

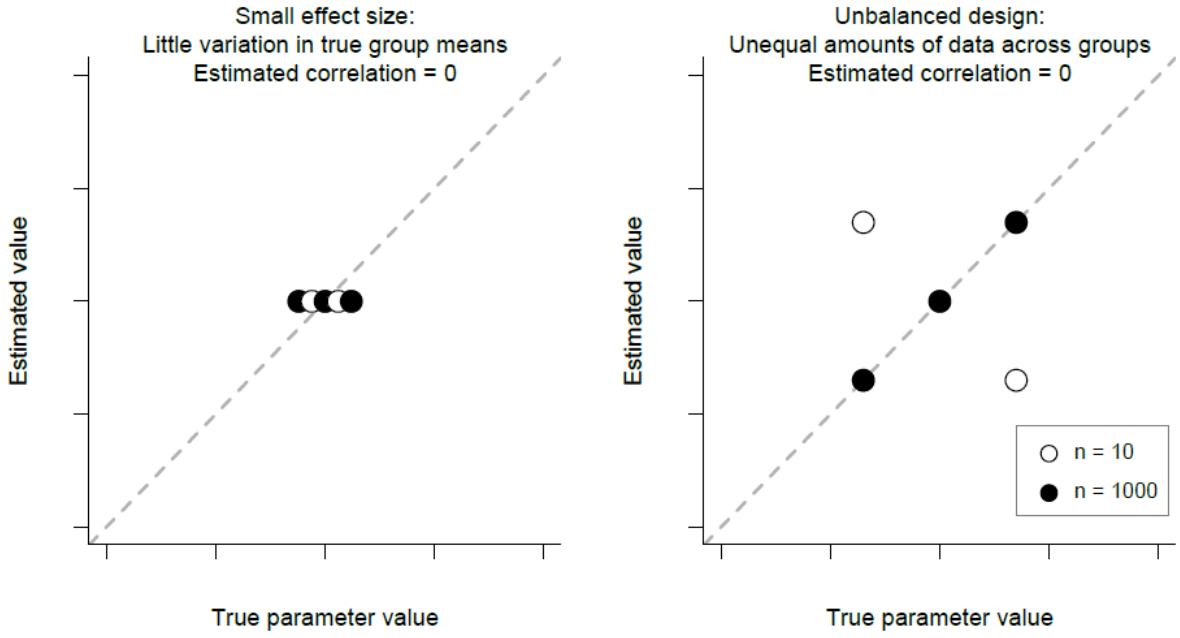

Figure 4. M&W testing scheme for determining whether diversification rates vary among sister clades is strongly affected by stochastic variation in species richness and leads to high Type I error rates. (A) Probability of rejecting a true null hypothesis (no variation in rates among sister clades) as a function of clade age, under the testing scheme described by M&W. Blue line shows

analytical probability that one member of a sister-clade pair of a given age (x-axis) will have a higher numerical MS value than the other, given that both clades have diversified under an identical net diversification rate. Red line shows Type I error rates under a more stringent threshold described by M&W, which requires MS estimates for clades to differ by 0.01 units or more in order to conclude that diversification rate variation is present. (B) MS-stem diversification estimates for 5000 replicates of an identical diversification process (rate = 0.12; stem age = 43.1), illustrating extensive variation in the value of MS estimators that can arise due to stochasticity in the diversification process itself. The variation illustrated in (B) is due to the inherent noisiness of the diversification process. For this parameterization, the 5% and 95% quantiles on the distribution of species richness are 11 and 598, respectively.

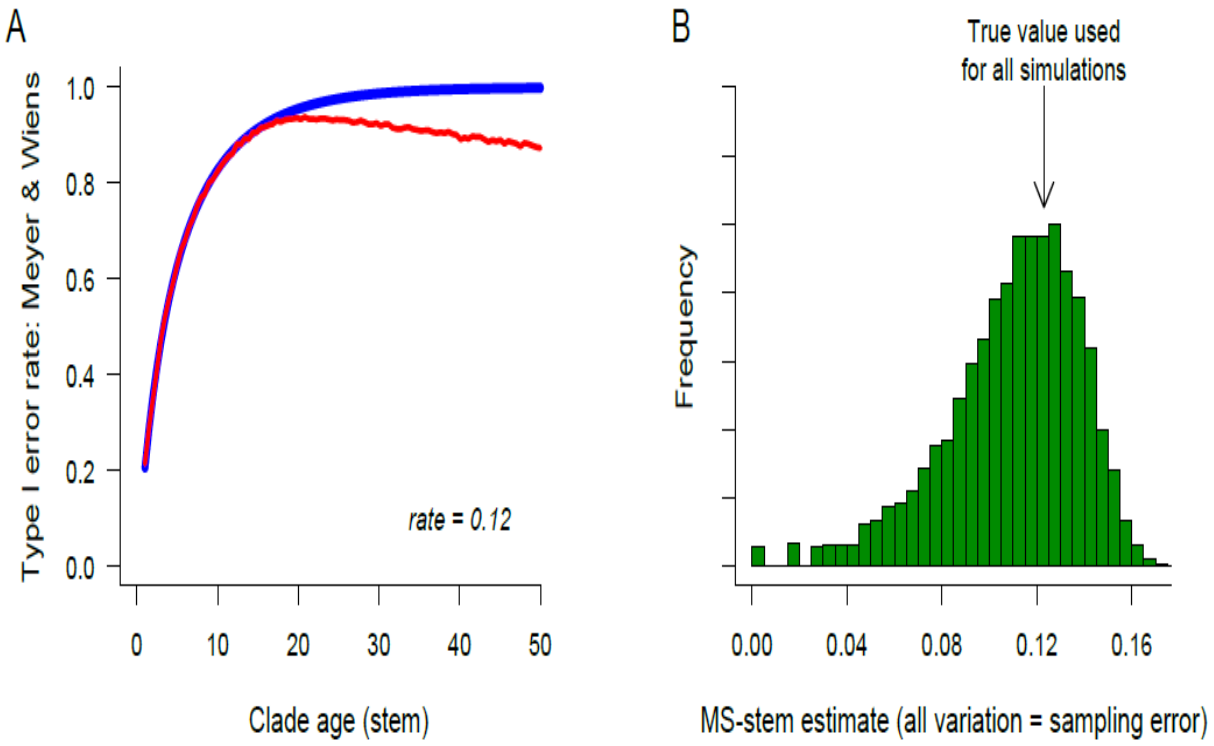

Table 1. Under equivalent comparison, BAMM estimates of diversification rate generally show lower bias and error than MS estimators; results are taken directly from M&W Tables S1 and S5. Models are ranked from best-performing to worst-performing (best = 1), by absolute error. Sampling, percent taxon sampling; ε, relative extinction rate; PE (bias, %), proportional error.

Method	*ε*	*% taxa sampled*	*rank*	*Mean absolute error (%)*	*PE (bias, %)*
BAMM	estimated	25%	1	22.4	-0.3
BAMM	estimated	50%	2	26.2	-7.4
MS-crown	0.9	-	3	26.3	-15.8
BAMM	estimated	100%	4	27.8	-10.1
MS-stem	0.5	-	5	30.5	15.9
MS-stem	0.9	-	6	32.8	-32.8
MS-crown	0.5	-	7	47.1	44.1
MS-stem	0	-	8	50.2	44
MS-crown	0	-	9	59.2	57.9