Received: 23 October 2017 | Accepted: 28 January 2018

DOI: 10.1111/2041-210X.12982

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

The influence of non-random species sampling on macroevolutionary and macroecological inference from phylogenies

Xia Hua | Robert Lanfear

Ecology and Evolution, Research School of Biology, Australian National University, Canberra, ACT, Australia

Correspondence
Xia Hua
Email: xia.hua@anu.edu.au

Funding information
Australian Research Council

Handling Editor: William Pearse

Abstract

1. Non-random species sampling is the rule rather than the exception in phylogenetics, but most phylogenetic methods to infer macroevolutionary and macroecological processes assume that the tips of the phylogenetic tree are either completely sampled or randomly sampled. In this study, we focus on extending the popular BiSSE framework to better account for non-random sampling of species. The existing BiSSE correction (which we describe hereafter as the unresolved clade correction) cannot be used on trees with clades of more than about 200 species, or when lineages that originate near the root are not sampled.
2. We propose new correction that does not have these two limitations. To assess the performance of our correction relative to the unresolved clade correction, we simulate trees using a common sampling strategy in which representative species of higher clades (e.g. genera) are sampled to include in a phylogeny.
3. Compared to the unresolved clade correction, we show that our new correction gives less biased parameter estimates; has higher power but a slightly elevated false positive rate to detect state dependence in speciation and extinction rates; and is less sensitive to a failure to sample all extant groups of taxa. Over all simulation scenarios, our correction perform equally well under conditions where the unresolved clade correction is applicable and conditions where the unresolved clade correction is inapplicable.
4. Given that both our correction and the unresolved clade correction have their own advantages and disadvantages, we suggest combining the two corrections. This can be done by applying our correction to groups that exceed the size limit of the unresolved clade correction or to account for the uncertainties in the placement of the lineages that originate near the root.

KEYWORDS
extinction, representative sampling, speciation, state-dependent diversification, trait evolution

1 | INTRODUCTION

Phylogenies of species or higher groups provide opportunities to test hypotheses about macroevolutionary and macroecological processes. For example they can be used to infer associations between geography (Ronquist & Sanmartín, 2011), life history (Weber & Agrawal, 2012) and diversification (Morlon, 2014), potentially throwing light on long-known but poorly understood patterns like

the latitudinal biodiversity gradient (Wiens & Donoghue, 2004). New methods (Morlon, 2014; O'Meara, 2012) provide increasingly rigorous statistical frameworks in which such hypotheses can be tested, but a number of important challenges remain. One of these involves accounting for the non-random sampling of the species from which the phylogenies are built. In this study, we focus on extending the popular BiSSE framework to better account for non-random sampling of species.

The BiSSE framework, formulated by Maddison, Midford, and Otto (2007), has been widely used to test hypotheses about state-dependent diversification. The framework has been extended to account for multiple states (MuSSE: FitzJohn, 2012), geographical states (GeoSSE: Goldberg, Lancaster, & Ree, 2011), quantitative states (QuaSSE: FitzJohn, 2010), unmeasured states that are correlates of the observed states (HiSSE: Beaulieu & O'Meara, 2016), as well as non-independence between speciation events and state changes (BiSSE-ness: Magnuson-Ford & Otto, 2012; and ClaSSE: Goldberg & Igić, 2012). However, all these BiSSE-type methods assume that the tips of the phylogenetic tree are either completely sampled or randomly sampled. This assumption can limit the power of the BiSSE framework because in reality we almost never sample tips completely or randomly.

Non-random species sampling is the rule rather than the exception in phylogenetics. For example one common approach to building phylogenies of large groups is to use representative sampling in which at least one representative is sampled for each higher taxonomic group (e.g. each genus). The sampling of discrete morphological data is also phylogenetically overdispersed (Guillerme & Cooper, 2016). The number of species sampled in each group can be determined by a number of factors, including the availability of existing sequence data in public databases, the availability of preserved tissues for sequencing, and the ease or difficulty of collecting new samples. The upshot is that the species sampling in any given phylogeny can be highly non-random. Failing to account for this can mislead downstream analyses.

FitzJohn, Maddison, and Otto (2009) introduced the unresolved clade correction to the BiSEE framework to account for non-random sampling. For an incompletely sampled monophyletic group, this correction calculates the probability that the group will have the same number of extant species in each state as observed, regardless of the phylogenetic relationships among these extant species. For example in Figure 1B, the unresolved clade correction calculates the probability that group *b* has two extant species in state 0 and one extant species in state 1, regardless of whether the two species in state 0 are each other's closet relative. The correction has two limitations: first, it is not applicable to groups that have more than about 200 species due to its computational burden; second, if no representative of a group is sampled, the correction collapses the group with its sister group into one unresolved group, thus losing the information on one group (e.g. groups *a*, *b*, *c*, *d* in Figure 2B are collapsed into one unresolved group). In certain situations, this can severely limit the power of the BiSSE framework. For example if the unsampled group originates near the root, the correction will collapse the whole tree into a single group.

In this study, we develop a new correction that does not have the two limitations of the unresolved clade correction. We first propose a correction for representative sampling that does not have a limit on group size. This correction is mathematically equivalent to an approach developed by FitzJohn (2012; "make.bisse.uneven" function in R package "DIVERSITREE") that splits groups with different sampling fractions into separate regions and fits a BiSSE model to each region, while constraining parameters in the model to be the same across regions. We then modify our correction to allow for non-random sampling at the group level, without collapsing these groups with their sister groups. We use simulations to assess the performance of our correction relative to the unresolved clade correction, as well as the performance of our correction for cases where the unresolved clade correction is inapplicable. We did not compare our correction to the make.bisse.uneven approach because the approach assumes that all groups have an equal chance of being sampled, so it cannot account for non-random sampling at the group level. The "make.bisse.uneven" function also does not accept groups with only one representative, which limits its application to most of the situations related to the

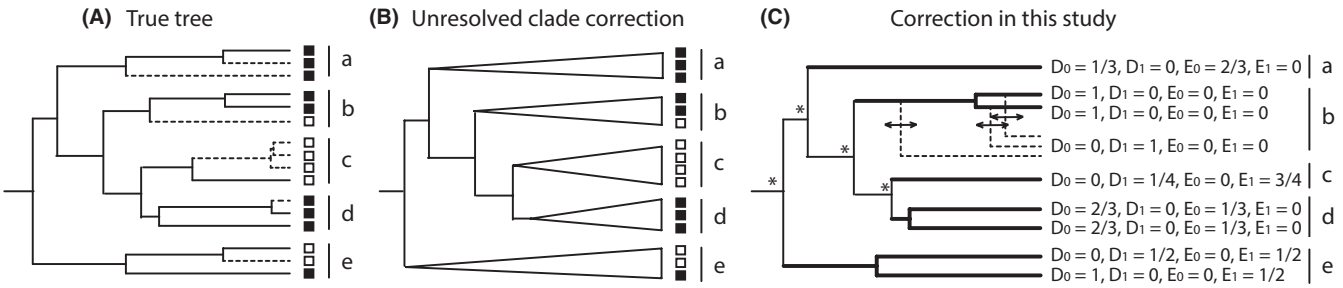

FIGURE 1 Corrections for representative sampling. In panel (A), the tree has five taxonomic groups (*a*–*e*) and two states (state 0 in black and state 1 in white). Solid tips are representatives sampled for each taxonomic group. In all the groups, sampling is not complete and dashed branches are not included in the tree. In panel (B), the unresolved clade correction collapses the groups with incomplete sampling into unresolved clade. Panel (C) illustrates our correction, where groups with incomplete sampling are not collapsed, but have adjusted initial values $D_i(0)$ and $E_i(0)$. In group *b*, state 1 is not sampled. This missing information is corrected by listing all possible locations where a representative of state 1, if sampled, would have attached to group *b*. $E_i(t)$ at the root of each group (marked with asterisk) is recalculated to reflect that all the groups have representatives in the tree

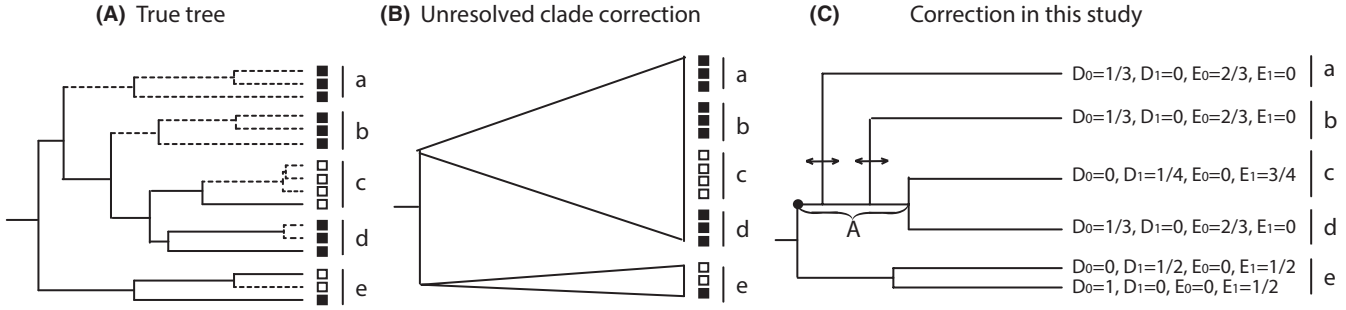

FIGURE 2 Corrections for unsampled groups. In panel (A), there should be 5 taxonomic groups, in which groups *a* and *b* are not sampled in the tree. In panel (B), the unresolved clade correction collapses groups *a*-*d* into one unresolved group. In panel (C), we know that groups *a* and *b* have split from branch *A* at some time points, so we calculate the weighted average of $D_i(t)$ at the older node of branch *A* (marked with black circle) under all possible combinations of split times of group *a* and *b* from branch *A*

cases we are interested in for this study. In Table 1, we summarize major differences among the unresolved clade correction, the make.bisse.uneven correction and our correction.

2 | MATERIALS AND METHODS

In the descriptions that follow, we assume that we have constructed a phylogeny of higher taxonomic groups (e.g. genera) by sampling zero or more representatives of each higher group, where the number of samples is unrelated to the number of species in the group except by an upper bound. In this situation, some groups may not be represented in the phylogeny (i.e. they have zero samples). We also assume that we know the number of species in each group, and can therefore calculate the sampling fraction for each group. The two challenges we have to address are: (1) accounting for variation in the fraction of each higher group that has been sampled (which we refer to hereafter as the group-specific sampling fraction); and (2) accounting for unsampled groups.

2.1 | Correction for group-specific sampling fraction

The BiSSE framework (Maddison et al., 2007) calculates likelihoods by tracking 2 probabilities, from the present to the root, for each state *i* (in this study, we assume two states, $i = 0$ or 1) along each branch in a tree: the probability that a lineage in state *i* at time *t* would evolve into a clade that has the same number of extant species and the same phylogenetic relationships among these species as observed in the present ($D_i(t)$) and the probability that a lineage in *i* at time *t* leaves no extant members ($E_i(t)$). When sampling is complete, the initial value $D_i(0)$ for each extant lineage is 1 if the lineage is in state *i* and 0 otherwise; the initial value $E_i(0)$ for each extant lineage is 0 as the lineage is known to be present. When sampling is random and the fraction f_i of extant lineages in state *i* is sampled, the initial value $D_i(0)$ for each extant lineage is f_i if the lineage is in state *i* and 0 otherwise; the initial value $E_i(0)$ for each extant lineage is $1-f_i$ as $1-f_i$ extant lineages are not sampled, leaving the same pattern in the tree as if these lineages went extinct.

TABLE 1 Comparison between the existing correction methods and our correction

	Unresolved clade	Make.bisse.uneven	This study
Allow non-random sampling within group	Yes	No	No
Allow different groups have different sampling fractions	Yes	Yes[a]	Yes
Allow non-random sampling of groups	Yes	No	Yes
Use phylogenetic relationships of sampled species within group	No	Yes	Yes
Use phylogenetic relationships of sampled groups	Yes[b]	Yes	Yes
Use phylogenetic relationships of unsampled groups	No	No	Yes

[a]The "make.bisse.uneven" function requires each group has at least two representatives in the tree.
[b]The unresolved clade correction uses phylogenetic relationships of sampled groups only when all the groups in a monophyletic clade have representatives in the tree. Otherwise, the correction will collapse the clade into an unresolved clade and lose all the information on the phylogenetic relationships of sampled groups in the clade.

We generalize this method to account for different sampling fractions in different groups within a single phylogeny. In our group-specific generalization, we first use the above correction to account for random sampling within each group, allowing each group to have a different sampling fraction for each state, f_i, and so different initial values $D_i(0)$ and $E_i(0)$ (e.g. the five groups in Figure 1C). We track $D_i(t)$ and $E_i(t)$ along each branch within each group (branches in bold in Figure 1C). The process results in $D_i(t)$ on the root of each group, that is the probability that the root of each group, if in state i, would evolve into the group we observe today.

In some cases, one state of a group may not have any representative in the tree (e.g. group b in Figure 1C). Using the above correction for the group will lose the information on the total number of lineages with that state in the group, so we need to modify the correction to account for the unsampled state. To do this, we start with a tip to represent the missing state and attach the tip to the clade of the group, as if the missing state had a representative in the tree (Figure 1C). Because we don't know the location in the tree to which this tip should attach, we consider all possible locations on each branch of the group clade where the tip can be attached (Figure 1C). Then, for each location, we attach the tip and apply the above correction to calculate $D_i(t)$ for the root of the group, which gives the probability that the root of each group, if in state i, would evolve into the group, where a random representative of the missing state, if sampled, could be attached to the group clade at the location. The weighted average of these $D_i(t)$ values gives the probability that the root of the group, if in state i, would evolve into the group in the situation where we have sampled a random representative of the missing state, but we don't know exactly where this representative is attached to the tree. The weight is $\frac{\sum_i D_i(t)^2}{\sum_i D_i(t)}$, which is the overall likelihood formula given in Appendix 1 of FitzJohn et al. (2009).

Now we have calculated $D_i(t)$ for the root of each group in the tree. When all the groups have at least one representative in the tree, all the branches connecting the root of these groups are included in the tree. In other words, at the group level, the sampling is complete. So, to further track $D_i(t)$ and $E_i(t)$ along each branch connecting the root of each group down to the root of the whole tree, we recalculate $E_i(t)$ on the root of each group (marked with stars in Figure 1C) as if the sampling across the whole tree is complete. This is done by numerically integrating $E_i(t)$ from the present time ($t=0$) to the root age of each group with initial values $E_i(0)$ equal to 0. Similarly, if we can assume random sampling at the group level, we simply integrate $E_i(t)$ with initial values $E_i(0)$ equal to 1 minus the sampling fraction at the group level. This is equivalent to the "make.bisse.uneven" approach developed by FitzJohn (2012).

2.2 | Correction for unsampled groups

When there are groups that have no representatives in the tree, we need to account for these unsampled groups in our analyses. This is possible as long as we know the topology that links all the groups, in other words, we know the internal branch of the tree to which each unsampled group attaches. For example unsampled groups a and b in Figure 2A attach to branch A (marked in Figure 2C). Of course, we don't know exactly where along branch A groups a and b attach, because groups a and b are not included in the tree. Here we propose a correction for this missing piece of information.

The correction starts with listing all possible combinations of time intervals that groups a and b can be added to the tree. Then, for each combination of time intervals, we attach a tip branch to branch A at each of the time intervals to represent a random species from each of the unsampled groups a and b (Figure 2C). Now that all the descendant groups a, b, c and d from the older node of branch A (marked with black circle in Figure 2C) have at least one representative in the tree, we can use our group-specific correction to track the $D_i(t)$ and $E_i(t)$ till the older node of branch A. The weighted average of these $D_i(t)$ values at the older node of branch A gives the probability that branch A in state i would evolve into groups a, b, c, and d in the situation where we don't know exactly when the groups a and b split from branch A. The weight for each split location equals $\frac{\sum_i D_i(t)^2}{\sum_i D_i(t)}$ (FitzJohn et al., 2009).

When there is only one unsampled group attach to branch A, we can numerically integrate $D_i(t)$ and $E_i(t)$ over the time when the unsampled group attaches to branch A. But when branch A has more than one unsampled descendent group, high-dimensional integrals need to be calculated, with each dimension corresponding to the time when each unsampled group is attached to branch A. We are not aware of any numerical integration algorithm that converges fast enough to be practical for parameter optimization. So, instead of numerical integration, we discretize branch A into many time intervals. In this study, we use 1 unit branch length as the length of each time interval. But the algorithm would be more efficient if we increased time intervals towards the root, because $D_i(t)$ and $E_i(t)$ values change less over time towards the root.

We illustrate our correction using the simplest example, where we know the full topology that links all the groups. In some cases, we can generalize our correction to account for incomplete knowledge of the topological relationships among groups, such as a situation in which we only know that groups a, b, c, and d in Figure 2C form a monophyletic clade, but we do not know the relationships within the clade. In this case, groups a and b, if sampled, could attach independently or as a monophyletic clade to the tree on any one of three branches: branch A; the branch that leads to group c; or the branch that leads to group d. Given that the order in which a and b attach to a branch matters if when attach to the same branch, this gives a total of 15 topologies where groups a and b could be attached in this example. The number of possible locations is then the number of possible combinations of branch lengths over the 15 topologies, which depend on the number of time intervals in each branch. As long as the number of possible locations is not too large to compute in reasonable time, the generalization of our correction is straightforward by listing all these possible locations, calculating $D_i(t)$ on the root of branch A, and averaging these $D_i(t)$ weighted by $\frac{\sum_i D_i(t)^2}{\sum_i D_i(t)}$. When the number of possible locations is too large, we probably have little knowledge on the topological relationships among groups. If a missing group can

be attached anywhere in the tree, our correction for unsampled groups is equivalent to assuming random sampling at the group level, so we can instead use our group-specific correction with random sampling at the group level, which we described in the previous section.

2.3 | Simulations

We simulated phylogenies to assess the performance of our correction and the unresolved clade correction. Our simulation starts by generating a BiSSE tree, with the root state fixed to 0. The simulation of each tree stops when we reach 500 species. We use this tree size because it is large enough that the BiSSE framework has relatively high power with complete sampling (Davis, Midford, & Maddison, 2013), and small enough that the unresolved clade correction is applicable for at least half of the simulated trees. We then identify each monophyletic clade as a group using two schemes:

1. "State" scheme: a monophyletic clade is a group if all of its extant members are in the same state. This scheme is often used to identify a group in biogeographical studies, where lineages inhabiting the same biogeographical area tend to be phylogenetically clustered and researchers tend to sample representatives from each biogeographical area within a taxonomic rank (e.g. Crottini et al., 2012).
2. "Time" scheme: a monophyletic clade is a group if its most recent common ancestor (MRCA) existed after half of the total evolutionary time of the tree. This scheme tries to simulate one aspect of how researchers tend to classify lineages into a taxonomic rank.

For each of the "State" and "Time" schemes, we include two sampling schemes:

1. "Complete group sampling" scheme: all the taxonomic groups have representatives in the tree, so sampling is complete at the group level.
2. "Incomplete group sampling" scheme: all the groups have a 50% chance of having representatives in the tree, so the tree has some unsampled groups.

If a group has no representatives, all its extant members are pruned from the tree. If a group has representatives, the identity of its representatives is randomly chosen from its extant members from the tree. The number of members kept in the tree is a random integer between 1 and 5 (or the size of the group if the total number of taxa in the group is smaller than 5).

We categorize the simulated trees into four schemes: (1) State scheme with Complete group sampling; (2) Time scheme with Complete group sampling; (3) State scheme with Incomplete group sampling; (4) Time scheme with Incomplete group sampling. In each scheme, we simulate trees under five sets of parameter values for speciation and extinction rates to reflect five possible ways that speciation and extinction rates can differ between different states (Table 2). For each set, we assume either equal state transition rates or unequal state transition rates (Table 2). We use the same parameter values as those used in Maddison et al. (2007) and FitzJohn et al. (2009) for comparison, but increase the ratio between parameter values of different state to 3. Under each set of parameter values, we simulate 200 trees. The five sets of parameter values are used to model:

1. Equal speciation and extinction rates between state 0 and state 1;
2. Faster diversification in state 0 than state 1 due to faster speciation in state 0;
3. Faster diversification in state 0 than state 1 due to slower extinction in state 0;
4. Slower diversification in state 0 than state 1 due to slower speciation in state 0;
5. Slower diversification in state 0 than state 1 due to faster extinction in state 0.

For each tree simulated with each set of parameter values under each simulation scheme, we use maximum likelihood (ML) to fit a BiSSE

TABLE 2 Summary of parameter values used in the simulation

Parameter sets	λ_0	λ_1	μ_0	μ_1	q_{01}	q_{10}
Equal speciation and extinction rates	0.1	0.1	0.03	0.03	0.01	0.01/0.001
Diversify faster in state 0 due to faster speciation	0.3	0.1	0.03	0.03	0.01	0.01/0.001
Diversify faster in state 0 due to slower extinction	0.1	0.1	0.01	0.03	0.01	0.01/0.001
Diversify slower in state 0 due to slower speciation	0.03	0.1	0.03	0.03	0.01	0.01/0.001
Diversify slower in state 0 due to faster extinction	0.1	0.1	0.09	0.03	0.01	0.01/0.001

Parameters include speciation rate of state 0 (λ_0), speciation rate of state 1 (λ_1), extinction rate of state 0 (μ_0), extinction rate of state 1 (μ_1), transition rate from state 0 to state 1 (q_{01}) and transition rate from state 1 to state 0 (q_{10}). For each parameter set, we assume either equal state transition rates ($q_{10} = 0.01$) or unequal state transition rates ($q_{10} = 0.001$)

model and a constrained BiSSE model (that constrains different states to have equal speciation rates and equal extinction rates) to the tree, using both the unresolved clade correction and our correction. For each correction, we compare the ML estimates to the true parameter values. We also compare the likelihoods of the best-fit model and the model used to simulate the tree. This helps us assess whether bias in the ML estimates reflect ridges on the likelihood surface, such that the true model does not necessarily fit worse than the best-fit model. Since the ML estimates are likely to be nearby the true parameter values, similar likelihoods also suggest that the confidence intervals of the ML estimates are likely to contain the true parameter values. We also estimate the power of both corrections to detect state-dependence in speciation and extinction rates, using likelihood ratio tests to compare the goodness-of-fit of the BiSSE model vs. the constrained BiSSE model. This helps us assess whether bias in the ML estimates causes the true process of state-dependent diversification to be rejected.

2.4 | Implementation

We simulate BiSSE trees and apply the unresolved clade correction using the R package "DIVERSITREE" (FitzJohn, 2012). We implement our group-specific correction and correction for unsampled groups in R, where $D_i(t)$ and $E_i(t)$ are numerically integrated along each branch of a tree using the R package "DESOLVE" (Soetaert, Petzoldt, & Setzer, 2010). To find the maximum likelihood (ML) estimates, we used the "subplex" method in the R package "NLOPTR" (Johnson, 2014). For both the unresolved clade correction and our correction, we started the search from the true parameter values and calculate the overall likelihood using the solution in Appendix 1 in FitzJohn et al. (2009).

3 | RESULTS

There is a general pattern of how non-random sampling affects parameter estimations of the BiSSE framework, regardless of the type of correction for the sampling or the simulation scheme we used in this study. With representative sampling, the BiSSE framework tends to underestimate the extinction rate and its estimation of speciation rate is often biased: speciation rate tends to be overestimated using the unresolved clade correction and underestimated using our correction (Figure 3; Figures S1–S9). BiSSE gives unbiased estimates of state transition rates when both states have the same speciation and extinction rates (Figure 3; Figures S1–S9). However, when lineages in state 0 diversify more rapidly than those in state 1, either because they have faster speciation rates or slower extinction rates, BiSSE tends to overestimate the transition rate from state 1 to state 0 (Figure 3; Figure S1–S9). In contrast, when lineages in state 0 diversify more slowly than state 1, either because they have slower speciation rates or faster extinction rates, BiSSE tends to overestimate the transition rate from state 0 to state 1 (Figure 3; Figures S1–S9).

There are three major differences in the performance between the unresolved clade correction and our correction for non-random sampling: bias, power, and performance. First, our correction gives consistently less biased parameter estimates than the unresolved clade correction for speciation rates, extinction rates, and state transition rates (Figure 3; Figures S1–S9). The bias in both corrections is a real problem, as suggested by the large difference in the likelihoods of the true model and the best-fit model (Figures S10, S11). This problem is more severe when applying our correction to trees simulated under "Time" scheme and with some groups not sampled (3rd and 4th columns in Figure S11). Under these conditions, confidence intervals of estimates from our correction are less likely to contain the true parameter values than the unresolved clade correction. Trees simulated under the "Time" scheme have fewer but larger groups (Figure S12) and so lower sampling fractions (Figure S13) for both states than trees simulated under "State" scheme.

Second, our correction has higher power and a slightly elevated false positive rate to detect state-dependence in speciation and extinction rates than the unresolved clade correction (Figure 4). When we define a group based on the age of its MRCA, our correction doubles the power of the unresolved clade correction (Figure 4). Our correction also has consistently higher power than the unresolved clade correction when some groups are not sampled in the tree (Figure 4).

Third, not sampling all the groups in the tree has less effect on the performance of our correction than the unresolved clade correction. Sampling half of the groups in the tree makes the unresolved clade correction inapplicable to more than 30% of the simulated tree, whereas our correction estimates parameters equally well under conditions where the unresolved clade correction is applicable (plotted in black in Figure 3 and Figures S1–S9) and conditions where the unresolved clade correction is inapplicable (plotted in white in Figure 3 and Figures S1–S9). Sampling half of the groups also reduces the power to detect state-dependence in speciation and extinction rates by a larger amount when using the unresolved clade correction than when using our correction (Figure 4).

4 | DISCUSSION

Non-random sampling, such as the representative sampling we discuss here, is common practice in phylogenetic studies. To minimize the influence of non-random sampling on macroevolutionary and macroecological inferences, we need to account for it and compare the performance of different corrections. The BiSSE framework represents a large family of phylogenetic methods. FitzJohn et al. (2009) introduced the unresolved clade correction to the BiSSE framework. The correction performs well, but is not applicable to phylogenies with large unresolved clades or with unsampled lineages that branch near the root of the tree. To overcome these limitations, we have proposed and tested alternative corrections in this study.

As with the existing unresolved clade correction, our correction cannot fully eliminate the influence of non-random sampling on macroevolutionary and macroecological inferences. Both corrections tend to give biased estimates of speciation rates, underestimates of

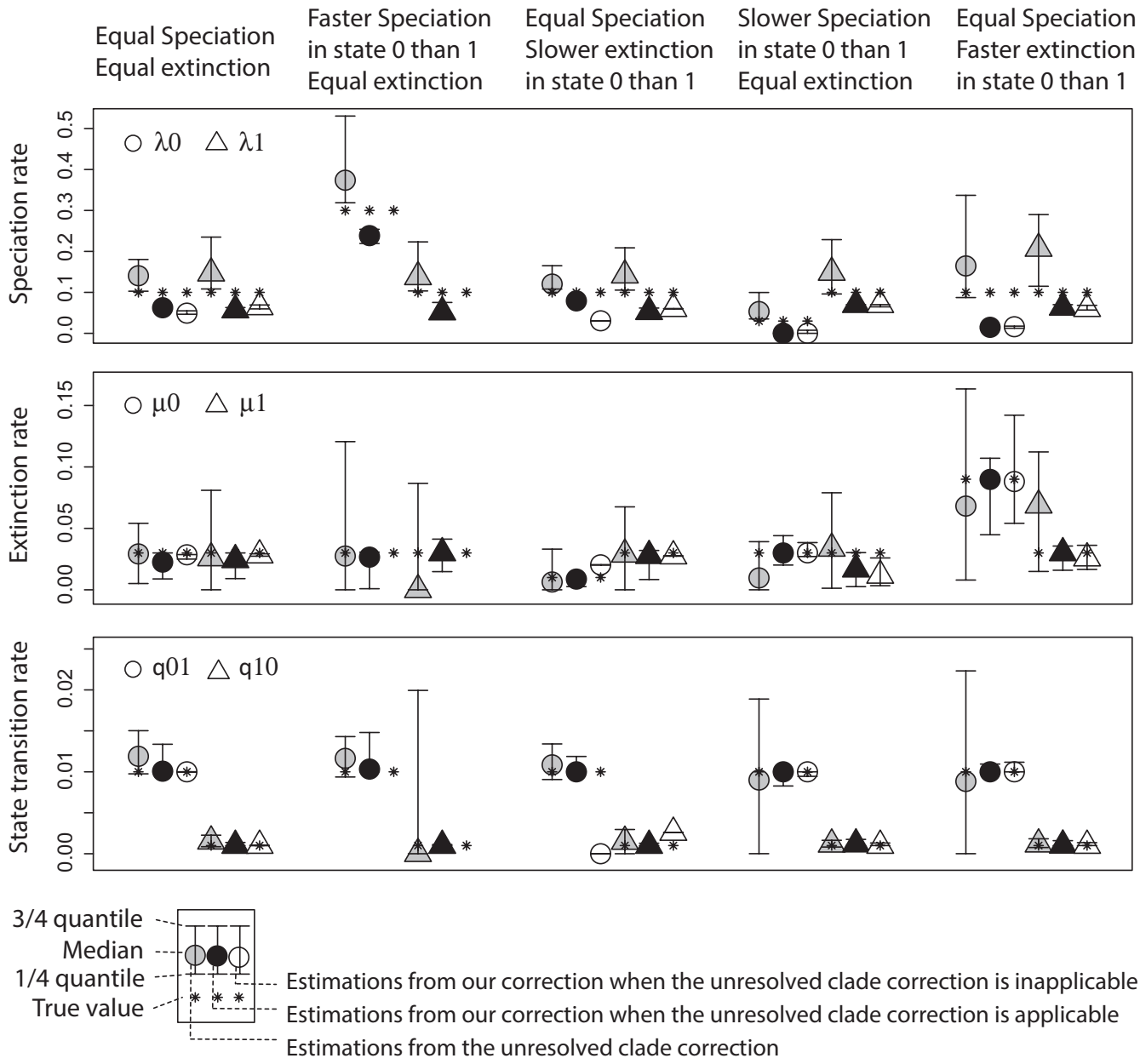

FIGURE 3 Parameter estimates from the unresolved clade correction and from our correction for trees simulated under the State scheme with Complete group sampling and unequal state transition rates. Panels from top to bottom plot the estimates for speciation rate, extinction rate, and state transition rate under the five parameter sets listed in Table 2. For each rate and under each parameter set, there are six point-and-whisker plots. Each plot shows the medium and the first and third quantiles of the estimate, with asterisk indicating the true parameter value. The three left plots are for the rate of state 0 (with the median plotted in circle) and the three right plots are for the rate of state 1 (with the median plotted in triangle). The three plots summarize the distribution of estimates from the unresolved clade correction (leftmost, coloured grey), from our correction for trees to which the unresolved clade correction is applicable (middle, coloured black) and from our correction for trees to which the unresolved clade correction is inapplicable (rightmost, coloured white). Figure S1 summarizes the parameter estimates under all the four schemes with unequal state transition rates. Figures S2–S5 show the scatterplots of estimations for the four schemes with unequal state transition rates. Figures S6–S9 show the scatterplots of estimations for the four schemes with equal state transition rates

extinction rates, and overestimates of the transition rate from the state that diversifies more slowly. Nevertheless, our correction leads to a large reduction in the overestimation of the transition rate. It has been suggested that even with complete sampling, the BiSSE framework has relatively low power for testing hypotheses about extinction and transition rates (Davis et al., 2013; Gamisch, 2016; Maddison et al., 2007). The bias in both corrections is real problem, suggesting that non-random sampling not only erases signals of the true evolutionary process, but also generates false signals that support other processes.

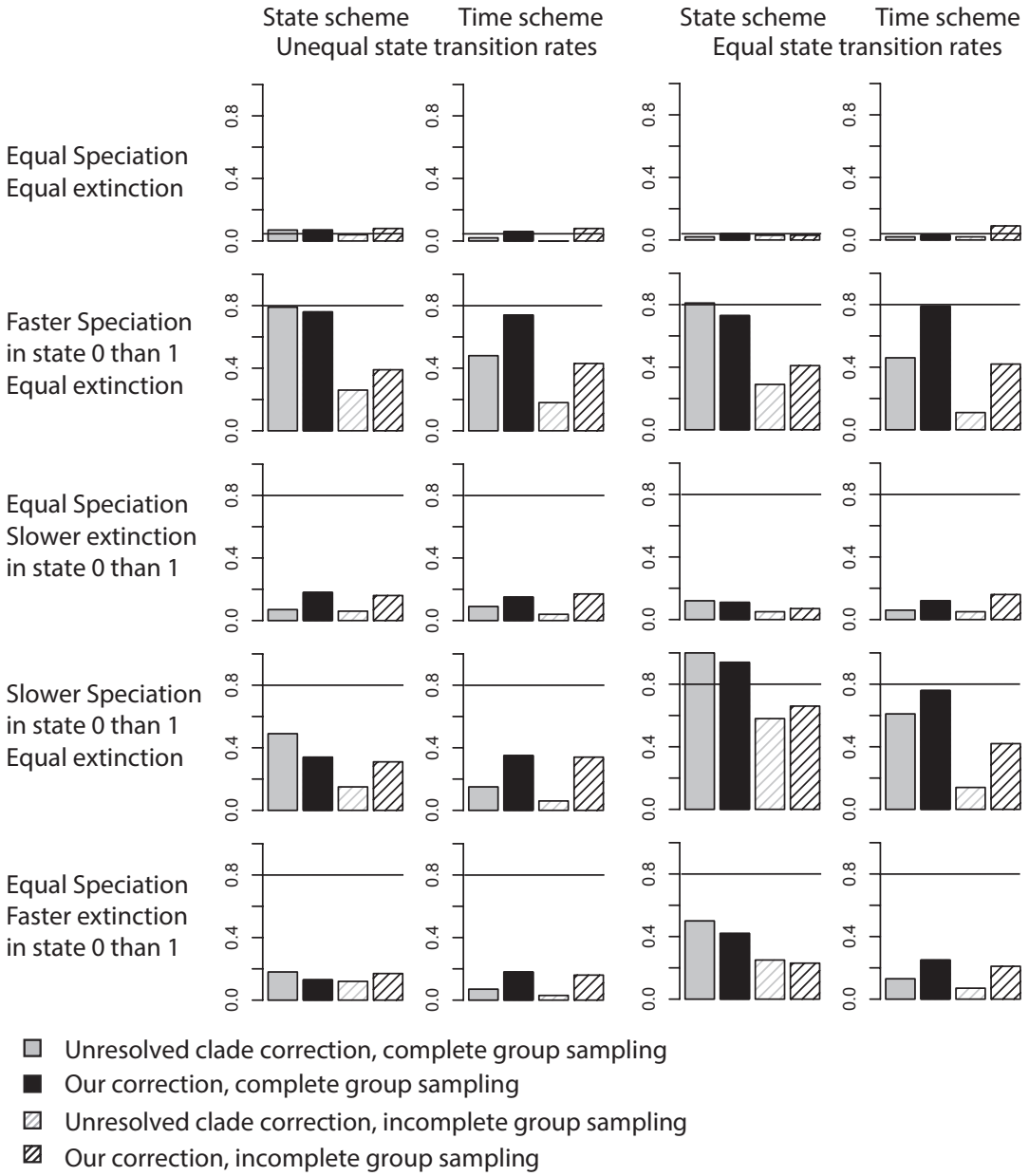

FIGURE 4 Statistical performance of the unresolved clade correction and the correction in this study. Each row of plots corresponds to each of the five parameter sets listed in Table 2, with the two columns of plots on the left using unequal state transition rates and the two columns of plots on the right using equal state transition rates. Each column of plots is under one simulation scheme: State scheme identifies a monophyletic clade as a group if all of its extant members are in the same state; Time scheme with Complete group sampling identifies a monophyletic clade as a group if its most recent common ancestor (MRCA) existed after half of the total evolutionary time of the tree. In each plot, bars from left to right show the percentage of simulated trees that BiSSE suggests state-dependent speciation and extinction rates when: all the groups are sampled in the tree and the unresolved clade correction is used (bar filled with solid grey); all the groups are sampled in the tree and our correction is used (bar filled with solid black); not all the groups are sampled in the tree and the unresolved clade correction is used (bar filled with grey lines); not all the groups are sampled in the tree and our correction is used (bar filled with black lines). Under equal speciation and equal extinction, the percentage of simulated trees that BiSSE suggests state-dependent speciation and extinction rates indicates the type I error of the method, which is expected to be 0.05 (the horizontal line). Under other parameter sets, the percentage of simulated trees that BiSSE suggests state-dependent speciation and extinction rates indicates the power of the method, which is expected to be above 0.8 (the horizontal line)

In cases where both corrections could be used (i.e. when the total number of species in an unresolved clade is <200), the best method will depend on the aim of the analysis. For example if we conduct BiSSE analyses to estimate speciation and extinction rates, then our correction will return estimates that have means closer to the true parameter values, but narrow confidence intervals that are less likely to contain the true values than using the unresolved clade correction. However, if we conduct BiSSE analyses to test

state-dependence in speciation and extinction rates, our correction has higher power to detect state dependence in speciation and extinction rates than the unresolved clade correction under most simulation schemes. This is because our correction is able to use more prior knowledge of the taxonomy and the topological relationships among groups than the unresolved clade correction.

An obvious advantage of our correction is that it performs equally well on trees to which the unresolved clade correction is applicable and to which the unresolved clade correction is inapplicable. A drawback of our correction is that it is less efficient than the unresolved clade correction when there are a large number of unsampled groups and/or unsampled states within each group, because our correction requires the listing of all possible combinations of time intervals when each unsampled group or state is added to the tree. When all the groups are sampled in a tree, our correction takes similar amount of time to compute the overall likelihood of the tree to the unresolved clade correction under the "State" scheme (1st and 2nd columns in Figure S14), because all extant members of each group have the same state under the scheme, so there is no need to correct for the unsampled state within each group. In contrast, under the "Time" scheme, our correction takes 10 times longer than the unresolved clade correction (1st and 2nd columns in Figure S15). When some groups are not sampled in the tree, our correction can take up to 100 times longer than the unresolved clade correction (3rd and 4th columns in Figures S14, S15).

It is important to note that the different corrections are not mutually exclusive. The unresolved clade correction can be combined with our correction to minimize the limitations of both approaches. For example we can use our correction for groups that exceed the size limit of the unresolved clade correction, whereas using the unresolved clade correction for the rest of the groups. When the unresolved clade correction is not applicable because some lineages that branch near the root are not sampled, we can use our correction to integrate out the uncertainties in the placement of these lineages. This combination of the two corrections would both increase the range of situations in which non-random sampling can be accounted for, and also reduce the amount of computing time required.

The BiSSE framework opens up an exciting way to study macroevolutionary and macroecological process. Its usage should not be limited by how we currently reconstruct phylogenies. Although no corrections so far can fully account for the impact of non-random sampling on the BiSSE framework, we show that it can still deliver meaningful results with the proper application of various corrections for non-random sampling. It is worthwhile to note that the BiSSE framework has its own limitations in inferring macroevolutionary and macroecological processes, including high sensitivity to model inadequacy and phylogenetic pseudoreplication (Maddison & FitzJohn, 2015; Rabosky & Goldberg, 2015). These limitations cannot be overcome even if we are able to fully account for the impact of non-random sampling. A recently developed nonparametric test "FiSSE" has been suggested as a promising complement to the BiSSE framework (Rabosky & Goldberg, 2017), however, the test has not yet been extended to account for non-random sampling.

ACKNOWLEDGEMENTS

We thank Lindell Bromham for her contribution to this project.

AUTHORS' CONTRIBUTIONS

X.H. developed and conducted the analyses. X.H. and R.L. wrote the manuscript. Both authors contributed critically to the drafts and gave final approval of publication.

DATA ACCESSIBILITY

R scripts are available from https://doi.org/10.5281/zenodo.1162969.

ORCID

Xia Hua http://orcid.org/0000-0003-3485-789X
Robert Lanfear http://orcid.org/0000-0002-1140-2596

REFERENCES

Beaulieu, J. M., & O'Meara, B. C. (2016). Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Systematic Biology*, *65*, 583–601. https://doi.org/10.1093/sysbio/syw022

Crottini, A., Madsen, O., Poux, C., Strauß, A., Vieites, D. R., & Vences, M. (2012). Vertebrate time-tree elucidates the biogeographic pattern of a major biotic change around the K-T boundary in Madagascar. *Proceedings of the National Academy of Sciences*, *109*, 5358–5363. https://doi.org/10.1073/pnas.1112487109

Davis, M. P., Midford, P. E., & Maddison, W. (2013). Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. *BMC Evolutionary Biology*, *13*, 38. https://doi.org/10.1186/1471-2148-13-38

FitzJohn, R. G. (2010). Quantitative traits and diversification. *Systematic Biology*, *59*, 619–633. https://doi.org/10.1093/sysbio/syq053

FitzJohn, R. G. (2012). Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, *3*, 1084–1092. https://doi.org/10.1111/j.2041-210X.2012.00234.x

FitzJohn, R. G., Maddison, W. P., & Otto, S. P. (2009). Estimating trait-dependent speciation and extinction rates from incomplete resolved phylogenies. *Systematic Biology*, *58*, 595–611. https://doi.org/10.1093/sysbio/syp067

Gamisch, A. (2016). Notes on the statistical power of the binary state speciation and extinction (BiSSE) model. *Evolutionary Bioinformatics*, *12*, 165–174.

Goldberg, E. E., & Igić, B. (2012). Tempo and mode in plant breeding system evolution. *Evolution*, *66*, 3701–3709. https://doi.org/10.1111/j.1558-5646.2012.01730.x

Goldberg, E. E., Lancaster, L. T., & Ree, R. H. (2011). Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Systematic Biology*, *60*, 451–465. https://doi.org/10.1093/sysbio/syr046

Guillerme, T., & Cooper, N. (2016). Assessment of available anatomical characters for linking living mammals to fossil taxa in phylogenetic

analyses. *Biology Letters*, *12*, 20151003. https://doi.org/10.1098/rsbl.2015.1003

Johnson, S. G. (2014). The NLopt nonlinear-optimization package (http://ab-initio.mit.edu/nlopt).

Maddison, W. P., & FitzJohn, R. (2015). The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic Biology*, *64*, 127–136. https://doi.org/10.1093/sysbio/syu070

Maddison, W. P., Midford, P. E., & Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, *56*, 701–710. https://doi.org/10.1080/10635150701607033

Magnuson-Ford, K., & Otto, S. P. (2012). Linking the investigations of character evolution and species diversification. *The American Naturalist*, *180*, 225–245. https://doi.org/10.1086/666649

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, *17*, 508–525. https://doi.org/10.1111/ele.12251

O'Meara, B. C. (2012). Evolutionary inferences from phylogenies: A review of methods. *Annual Review of Ecology, Evolution, and Systematics*, *43*, 267–285. https://doi.org/10.1146/annurev-ecolsys-110411-160331

Rabosky, D. L., & Goldberg, E. E. (2015). Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology*, *64*, 340–355. https://doi.org/10.1093/sysbio/syu131

Rabosky, D. L., & Goldberg, E. E. (2017). Fisse: A simple nonparametric test for the effects of a binary character on lineage diversification rates. *Evolution*, *71*, 1432–1442. https://doi.org/10.1111/evo.13227

Ronquist, F., & Sanmartín, I. (2011). Phylogenetic methods in biogeography. *Annual Review of Ecology, Evolution, and Systematics*, *42*, 441–464. https://doi.org/10.1146/annurev-ecolsys-102209-144710

Soetaert, K., Petzoldt, T., & Setzer, R. W. (2010). Solving differential equations in R: package deSolve. *Journal of Statistical Software*, *33*, 1–25.

Weber, M. G., & Agrawal, A. A. (2012). Phylogeny, ecology, and the coupling of comparative and experimental approaches. *Trends in Ecology and Evolution*, *27*, 394–403. https://doi.org/10.1016/j.tree.2012.04.010

Wiens, J. J., & Donoghue, M. J. (2004). Historical biogeography, ecology and species richness. *Trends in Ecology and Evolution*, *19*, 639–644. https://doi.org/10.1016/j.tree.2004.09.011

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Hua X, Lanfear R. The influence of non-random species sampling on macroevolutionary and macroecological inference from phylogenies. *Methods Ecol Evol*. 2018;9:1353–1362. https://doi.org/10.1111/2041-210X.12982