doi:10.1111/evo.14

Modeling colonization rates over time: Generating null models and testing model adequacy in phylogenetic analyses of species assemblages

Xia Hua[1,2,3] **and Lindell Bromham**[2]

[1]*Mathematical Sciences Institute, Australian National University, Canberra, ACT 2601, Australia*

[2]*Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, ACT 2601, Australia*

[3]*E-mail: xia.hua@anu.edu.au*

Received November 5, 2019

Accepted August 19, 2020

A central theme connecting macroevolutionary processes to macroecological patterns is the shaping of regional biodiversity over time through speciation, extinction, migration, and range shifts. The use of phylogenies to explore the dynamics of diversification due to variation in speciation and extinction rates has been well-developed and there are established methods for inferring speciation times from phylogenies and generating its null distributions (as represented by node heights on molecular phylogenies). But inferring colonization events from phylogenies is more challenging. Unlike speciation events, represented by nodes, colonization events could occur at any point along a branch connecting species in the assemblage to the regional pool. We account for uncertainty in identification of colonization lineages and timing of colonization events by using an efficient analytical solution to inferring the distribution of colonization times from an assemblage phylogeny. Using the same solution, we efficiently derive the null distribution of colonization times, which provides us with a general approach to testing the adequacy of a model to describe colonization events into the assemblage. We illustrate this approach by demonstrating how the movement of squamate lineages into Madagascar has been uneven over time, peaking in the early Cenozoic when ocean conditions favored colonization.

KEY WORDS: **Community phylogenetics, diversification, macroevolution, Madagascar, model adequacy, molecular phylogeny.**

The rate of addition of species to a biological community is a concept of central importance in many fields of ecology and evolutionary biology. Can new species be continually added to a given area, or do ecological opportunities saturate, reducing the opportunity for new species to establish? (e.g., Harmon and Harrison 2015; Rabosky and Hurlbert 2015). Studies of the accumulation of biodiversity within a region by the processes of diversification of lineages in situ and migration of lineages into the area have a long history (e.g., Von Humboldt 1847; Wallace 1855), and there have been a range of approaches to addressing this question, ranging from theoretical models of species addition and subtraction (e.g., MacArthur and Wilson 1967) to population-based modeling (e.g., Cabral et al. 2019) to palaeotontological studies of changes in community composition over time (e.g., Jablonski et al. 2006). In recent decades, the popularity of phylogenetic analyses of community formation has soared (Webb et al. 2002; Vellend 2010), as the expansion of molecular databases has put assemblage-level phylogenies within reach for a broad range of biological communities (Emerson and Gillespie 2008; Cavender-Bares et al. 2009; Valente et al. 2015; Onstein et al. 2016; Valente et al. 2017a,b; Matos-Maraví et al. 2018; Valente et al. 2020). In this article, we consider a current limitation of phylogenetic community assembly methods, concerning the inference of colonization times from phylogenies and formal tests of model adequacy in modeling colonization events over time.

Studies in phylogenetic community ecology usually track the formation of an assemblage over time using DNA sequences from living species to reconstruct a phylogeny representing their evolutionary history, where an "assemblage" is typically taken to mean all of the members of a particular high-level clade that are found in a given area of interest. The assemblage area is typically a spatial unit of study, such as a geographical feature or a continent, or a kind of habitat, such as a rainforest. For example, we might be interested in the assemblage of cichlids in Lake Tanganyika (Salzburger et al. 2002), carnivorous mammals in Africa (Cardillo 2011), trees in an Australian rainforest (e.g., Kooyman et al. 2012), or the human gut microbiome (Shafquat et al. 2014). Taxa will usually be species, but may be higher (such as genera) or lower (such as subspecies). Critically, the investigation of assemblage formation must encompass not only speciation and extinction events, but also biogeographic events representing range shift (e.g., Emerson 2002; Ricklefs and Bermingham 2004, 2008; Valente et al. 2020). New lineages can be added to an assemblage through range expansion or colonization from another area, and lineages can be lost from an assemblage through range contraction or local extinction in the assemblage area. Therefore, tracking patterns of biodiversity over time in a given area using phylogenies requires the inclusion of taxa from the "regional pool" that represent potential colonists into the assemblage area, to identify edges in the phylogeny that represent the movement of a lineage into the assemblage area (colonizing edges in Fig. 1: an edge is any branch in the phylogeny, connecting two nodes, whether it joins two internal nodes or connects a node to a taxon at the tip). The regional pool will typically be defined not simply as a geographic area but also on the basis of lineages that would have the ability to expand their range over time into the assemblage area. We must also model the way the composition of the regional pool changes over the history of the assemblage, rather than considering the pool as a static assemblage, based on current biodiversity.

It is now tractable to construct a phylogeny not only of all of the species in an assemblage, but also of most or all of the members of the regional pool that represent the source of colonizing lineages (Webb et al. 2002). A comprehensive assemblage-level phylogeny, including all species currently found in the assemblage, will encapsulate information about speciation events that have occurred within the assemblage (in situ speciation), represented by nodes in the phylogeny. The accuracy and precision of the estimated dates of speciation events is influenced by many factors, including the methods used, calibrating information, and patterns of variation in rate of molecular evolution (Welch and Bromham 2005; Bromham et al. 2018). Many studies on patterns of diversification assume that speciation times are error free, for example, by fitting a diversification model to a single phylogeny with point estimates of node ages (Rangel et al. 2015). Uncertainty in speciation times can be incorporated by fitting the model to a range of alternative phylogenetic solutions (De Villemereuil et al. 2012; Rangel et al. 2015). To test if this model gives an adequate description of the diversification processes that shape the assemblage phylogeny, we need to compare the inferred speciation times to a null distribution, derived from an array of possible phylogenies that could have produced this assemblage, given the model of diversification. Null models for speciation rates can be produced using a "birth-death" model that simulates the growth of a phylogeny by speciation and extinction. Such models are standard in the macroevolutionary literature (e.g., McPeek 2008; Phillimore and Price 2008; Rabosky 2009; Morlon et al. 2010) and are used to detect shifts in diversification rate above and beyond the expected noise in tree shape generated by stochastic birth-death processes (Raup et al. 1973; Hartmann et al. 2010; Stadler 2011; Höhna 2013).

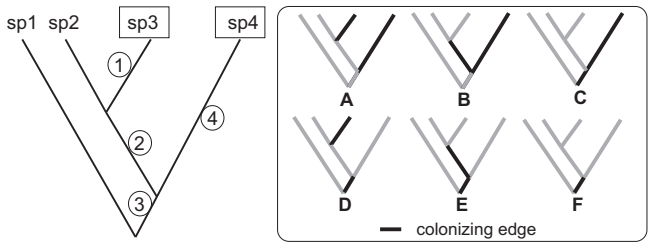

Figure 1. **Identifying colonizing edges by including close relatives of native taxa in the phylogeny of an assemblage. For each native taxon (species 3 and 4) in the assemblage, we can identify edges on which a colonization event could have happened that led to the native taxon. For example, species 3 could have been derived from a colonization event on edge 1 (A and D), edge 2 (B and E), or edge 3 (C), independently of a colonization event on edge 4 (A, B, and C) or edge 3 (D and E) that led to species 4. Alternatively, the two species could have been derived from the same colonization event on edge 3 (F). In total, there could be six colonization histories led to the assemblage phylogeny. The relative likelihoods of these colonization histories depend on the occurrence rate of various events (described in Fig. 3) that could add or remove a taxon from the assemblage. Adding close relatives from the regional pool to the native taxa helps us identify potential colonizing edges. For example, adding species 1 provides information on the window of time for colonization, represented by edge 3, which informs the likelihood of potential colonization event on edge 3.**

But in situ speciation events are not the only way that diversity can be added to an assemblage. Lineages can be added to the assemblage by colonization, when a species migrates from another area and establishes a stable breeding population within the assemblage. Many models have been developed to describe these various ways of adding species into an assemblage, including the DIVA (Dispersal-Vicariance Analysis) model (Ronquist 1997), the DEC (Dispersal-Extinction-Cladogenesis) model (Ree and Smith 2008), the GeoSSE (Geographic State Speciation and

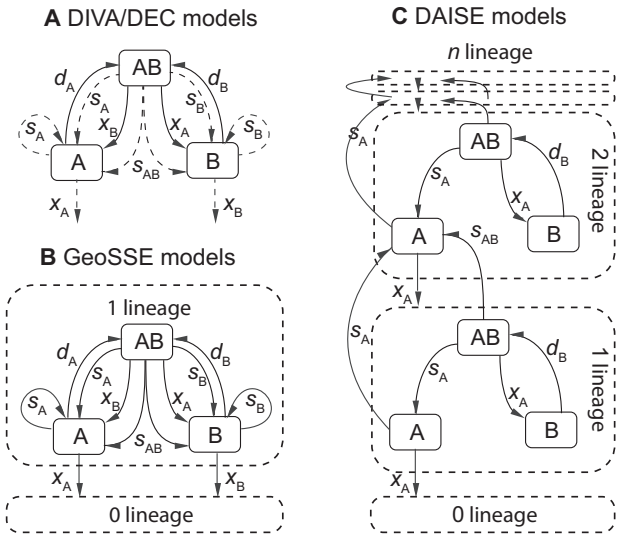

Figure 2. Comparisons among existing models of assemblage formation. All the models have three basic states: regional pool (state B), widespread (state AB), endemic (state A). The GeoSSE model allows for all possible transitions among the three states: including colonization (d_B), range expansion (d_A), extinction in the regional pool (x_B), extinction in the assemblage area (x_A), vicariance speciation (s_{AB}), in situ speciation in the regional pool (s_B), and in situ speciation in the assemblage area (s_A). The DIVA/DEC model assumes a fixed assemblage phylogeny, so it only allows speciation to occur at each node of the assemblage phylogeny and does not model global extinction properly (processes in dashed lines). The GeoSSE model relaxes this assumption by explicitly calculating: (1) the probability that a lineage stays as a single lineage (1 lineage) from the ancestral node to the descendent node of an edge on the assemblage phylogeny and (2) a lineage leaves no extant descendent (0 lineage), so the lineage is not included in the assemblage phylogeny. The GeoSSE model assumes that state transitions on a lineage are independent to other lineages. The DAISE model relaxes this assumption by calculating the probability of having a specific number of lineages (*n* number of lineages) in the assemblage. However, the DAISE model assumes static regional pool, so it does not model events that affect the number of regional pool taxa.

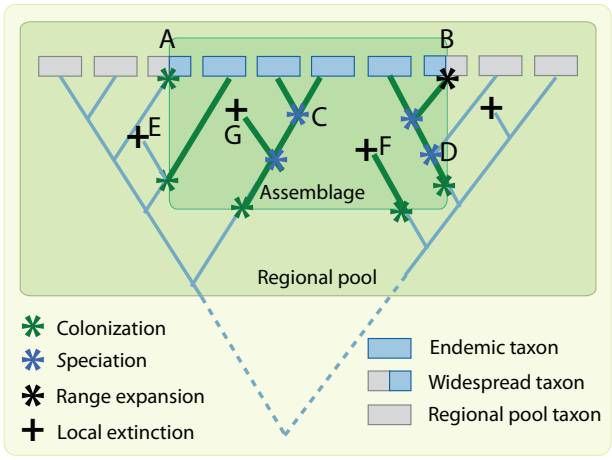

Figure 3. Events that can add a taxon to or remove a taxon from an assemblage. A widespread taxon can be added to the assemblage by colonization of the assemblage area by a regional pool taxon (A). An endemic taxon can become a widespread taxon by range expansion to the regional area (B). An endemic taxon can be added to the assemblage by speciation within the assemblage area (C). A widespread taxon can become an endemic taxon if the population within the assemblage area becomes isolated from the populations in the regional area (D), or by the extinction of its populations outside the assemblage area (E). A widespread taxon can be removed from the assemblage by the extinction of its populations within the assemblage area (F). An endemic taxon can be removed from the assemblage by extinction (G). Combinations of these events lead to various histories of assemblage formation. For example, an endemic taxon could be removed from the assemblage by first becoming a widespread taxon through range expansion (B), followed by the extinction of its populations within the assemblage area (F). When the origin of the higher taxon is much earlier than the formation of the assemblage, the assemblage phylogeny includes noninformative edges (dashed) that provide little information to the evolutionary dynamics in the regional pool. We need to prune out these edges, so the phylogeny ends up in two subtrees.

Extinction) model (Goldberg et al. 2011), and the DAISE (Dynamic Assembly of Islands through Speciation, Immigration, and Extinction) model (Valente et al. 2015). These models share a common mathematical foundation by describing processes of assemblage formation as a continuous-time discrete-state Markov chain with similar structures of transitions among three basic states of a taxon. The three states are as follows: "Regional pool taxa," taxa that make up the regional pool consisting of potential colonists into the assemblage (state B in Fig. 2); "Widespread taxa," taxa found in both the assemblage area and in the regional pool (state AB in Fig. 2); and "Endemic taxa," taxa found only within the assemblage area and nowhere else (state A in Fig. 2).

A full picture of all possible transitions among the three states is given in Figure 2A, and the equivalent events mapped on a phylogeny are given in Figure 3. A widespread taxon may have resulted from the colonization of the assemblage area by a taxon from the regional pool (colonization rate d_B in Fig. 2; event A in Fig. 3). Alternatively, a widespread taxon is created by an endemic taxon expanding its range outside the assemblage area (range expansion rate d_A in Fig. 2; event B in Fig. 3). Three events can generate an endemic taxon: speciation within the assemblage area by an endemic taxon (in situ speciation rate s_A in Fig. 2; event C in Fig. 3), or a widespread taxon splits into two lineages, one falling entirely within the assemblage area (vicariance speciation rate s_{AB} in Fig. 2; event D in Fig. 3), or the widespread taxon goes locally extinct outside the assemblage but

persists within the assemblage area (extinction rate x_B in Fig. 2; event E in Fig. 3). Local extinction within the assemblage area can remove taxa from the assemblage, including widespread taxa that become extinct within the assemblage but persist elsewhere (extinction rate x_A in Fig. 2; event F in Fig. 3), and endemic taxa that become globally extinct (extinction rate x_A in Fig. 2; event G in Fig. 3). The diversification of the regional pool is also governed by speciation (s_B in Fig. 2) and extinction (x_B in Fig. 2). These rates can have different properties in different species assemblage models. For example, all these rates can vary over time in the three models in Figure 2, vary across lineages in the GeoSSE model (Fig. 2B), and depend on the number of taxa in the assemblage in the DAISE model (Fig. 2C). Therefore, it is possible to select an appropriate species assemblage model to account for a wide range of scenarios of assemblage formation. The commonality among these models of assemblage formation suggests that it should be possible to develop a general approach to use these models to infer times for colonization events from assemblage phylogenies and to generate the null distribution of times for colonization events to test model adequacy, as has been previously done for times of speciation events. In this article, we present this general approach.

Although procedures for inferring speciation times and their null distributions from phylogenies are well established, there are four complications that make it much harder to infer times for colonization events from phylogenies and to generate null distribution of times for colonization events. First, we must add species from the regional pool to our phylogeny to identify the potential colonizing edges (Fig. 1). The uncertainty in identifying the edges is increased if we are unsure of the relationships between potential source lineages and assemblage colonists, or if we are missing data on those colonizing lineages due to incomplete sampling or extinction. Second, even a well-sampled phylogeny of the assemblage plus the regional pool does not allow us to unambiguously identify the source of colonization, because there could be multiple potential colonizing edges (Fig. 1). Which edges are more likely to be the source of colonization depends on not only the processes occurring within the assemblage, but also the changing nature of the regional pool over time. Third, identifying the lineages that represent the addition of a colonizing lineage to the assemblage does not provide us with a date of colonization, but a window of time in which a colonization event has occurred (Fig. 1), because the colonization event could have occurred any time between the split with the nearest known relative from the regional pool and the first node occurring within the assemblage (either an in situ speciation event or a terminal node representing an extant taxon in the assemblage). Fourth, although there are well-established methods for generating a phylogenetic null distribution for speciation times, a phylogenetic null distribution for colonization times is more complex because it must include not only the processes occurring within the assemblage, but also the processes in the regional pool. For meaningful comparisons to observed assemblage phylogeny, phylogenies in the null distribution must also be constrained to give the same number of species and the same proportions of endemic and widespread taxa as the observed assemblage. In this article, we present a general approach that solves all four challenges, and we demonstrate how to apply our approach to a real dataset using a case study on squamate assemblage of Madagascar.

Methods

Our approach consists of a number of steps, which we summarize here, and following sections provide details. First, we define the members of the assemblage, and we construct a phylogeny that consists of all of the members of the assemblage and taxa representing the regional pool of potential colonists over the history of the assemblage. Such a phylogeny contains important information on the accumulation of diversity within the assemblage area by speciation and colonization, but it may also contain edges that represent lineages that existed in a different time or place, such as the deep edges of the tree or diversification events that do not give rise to any regional pool taxon or members of the assemblage. When this happens, we prune out the parts of the phylogeny that do not relate to the assemblage formation. This leaves us with a set of subtrees that are informative for the formation of this assemblage. Then, we choose any appropriate model of assemblage formation (e.g., a GeoSSE model) and we fit the model to the phylogeny (or simultaneously to all the subtrees if the phylogeny is pruned). Given the fitted model, we analytically derive the null distribution of times of colonization events as expected by the model and infer the distribution of times of colonization events on the phylogeny. Last, we assess the adequacy of the fitted model by testing significant departure in the inferred distribution against the null distribution of colonization times, using test statistics that summarize the null distribution. All methods described here are implemented in R code, which is available on GitHub and as an Appendix.

CONSTRUCTING AN ASSEMBLAGE PHYLOGENY

Our aim is to construct a phylogeny that encapsulates the key events that contribute to the accumulation of biodiversity within an assemblage. We consider an assemblage to refer to all of the taxa of a given higher-level clade occurring within a specified region or habitat (as outlined above). This phylogeny must contain not only the taxa that form consistent resident populations within the assemblage (which we refer to as "native taxa," including both endemic taxa and widespread taxa), but also taxa that make up the regional pool of potential colonists. Ideally, we

want to include all taxa in the extant regional pool that would have the ability to colonize the assemblage area based on their dispersal range and habitat similarity to the environmental conditions in the assemblage area. An alternative way is to include all closely related taxa to the native taxa as the potential pool of colonists. Doing so will reduce the uncertainty in identifying the colonizing edges (Fig. 1), but is prone to nonrandom incomplete sampling of regional pool taxa, which will be accounted for by the model of assemblage formation.

Next, we create a phylogeny that contains all the native taxa and identified regional pool taxa. For example, we might generate this phylogeny by sampling all of the relevant taxa from a large phylogeny (e.g., Valente et al. 2015), or by using DNA sequences from the relevant taxa to estimate a phylogeny (e.g., Valente et al. 2020). The branch lengths of this phylogeny are considered to represent evolutionary time, whether in absolute terms (e.g., a molecular phylogeny dated using fossil or biogeographic calibrations) or relative terms (e.g., where node heights can be taken as relative to the initial formation of the assemblage).

Once we have our phylogeny of the assemblage and the regional pool, the next step is to remove the parts of the phylogeny that do not relate to the formation of the assemblage, which we refer to as noninformative edges. If an assemblage contains disparate taxa with deep relationships, then the phylogeny may include many events that occurred before the formation of the assemblage, because it describes the shared history of all taxa in the assemblage (Fig. 3). For example, the bird assemblage of Hawaii includes many different orders such as ducks and eagles, so a phylogeny of birds of Hawaii will include edges connecting the duck and eagle lineages to their last shared common ancestor. But this last common ancestor lived tens of millions of years before the Hawaiian islands came into existence, so these basal edges do not describe events that occurred within the bird assemblage of Hawaii. The same will also be true of continental assemblages, if they contain taxa whose relationships are deeper than the history of the assemblage. For example, the age of the split between the gymnosperms and angiosperms is older than the Australian rainforests they co-occur in, and the human gut microbiome contains taxa that are far more ancient than the host species they inhabit. We have two additional reasons to remove these noninformative edges from our phylogeny: (1) the likelihood of a colonization event occurring on these edges is so low that removing these edges will not influence our estimates of colonization times, but will make the following analyses less time-consuming; (2) because these edges dated before the formation of the assemblage and descendants of these edges are often poorly sampled, including these edges may bias our inferences on the evolutionary dynamics of the regional pool during the assemblage formation. As a result of pruning noninformative edges, our assemblage phylogeny may now consist of several separate subtrees, each of which includes both native taxa and their regional pool taxa (Fig. 3).

ANALYTICAL METHOD FOR DERIVING A NULL DISTRIBUTION OF COLONIZATION TIMES

We can now use the assemblage phylogeny (or the set of subtrees) to find a model that best describes the processes that shape the assemblage phylogeny. As outlined above, many models are available and these models share the same mathematical foundation. But to test whether the model gives an adequate description of the processes, we need to be able to derive a null distribution that summarizes an aspect (x), for example, colonization times, of all possible phylogenies generated by the model of assemblage formation that have the same number of extant taxa in A, B, and AB states as the observed assemblage:

$$f\left(x|n_{\mathrm{A}}, n_{\mathrm{B}}, n_{\mathrm{AB}}, \mathrm{model}\right) = \sum_{\mathrm{phylogeny}} f\left(x|\mathrm{phylogeny}, \mathrm{model}\right) \times p\left(\mathrm{phylogeny}|\mathrm{model}\right).$$

Because of this, the model needs to be able to calculate p (phylogeny | model). Both GeoSSE model and DAISE model give p (phylogeny | model), but not DIVA/DEC models, as they are conditional on the assemblage phylogeny. But we note that, assuming no global extinction, DIVA/DEC models can be considered special cases of the GeoSSE model (see Supporting Information). For the purposes of illustrating our method, we will focus on the GeoSSE model in this article; however, we also explain how our method is applicable to the DAISE model in the Supporting Information.

The GeoSSE model calculates, for a certain time along a certain edge of a phylogeny (t), the probability of observing the part of the phylogeny descended from the edge after t, given that the edge is in geographic state s at t, that is, p(phylogeny after $t \mid s$, model), where $s \in \{\mathrm{A}, \mathrm{B}, \mathrm{AB}\}$. By convention, this term is denoted as $D_s(t)$ and is dependent on $E_s(t)$ that is the probability of a lineage at time t in state s leaving no descendent at the present (Goldberg et al. 2011). Integrating the probability from each tip taxon (boundary conditions defined by tip state) along each edge in the phylogeny till the root gives $D_s(T)$, where T is the root age. Then p(phylogeny | model) $= D(T) = \Sigma_s D_s(T) p$(root s | model). When there is no prior knowledge on the root state, p(root s | model) is best set as $D_s(T)/\Sigma_s D_s(T)$ (see FitzJohn et al. 2009).

Using Bayes' theorem, $f(x \mid n_{\mathrm{A}}, n_{\mathrm{B}}, n_{\mathrm{AB}}, \mathrm{model}) \propto p(n_{\mathrm{A}}, n_{\mathrm{B}}, n_{\mathrm{AB}} \mid x, \mathrm{model})\, f(x \mid \mathrm{model})$. We first show how to calculate $p(n_{\mathrm{A}}, n_{\mathrm{B}}, n_{\mathrm{AB}} \mid \mathrm{model})$, which is the sum of the probabilities of all possible phylogenies that could connect the extant taxa. It may seem impractical to list all possible phylogenies, but we can flip the logic by asking what is the probability of having observed the extant taxa if we have no knowledge of the true phylogeny. In

other words, the null distribution ignores the topology and the branch lengths of the assemblage phylogeny (or subtrees). In the GeoSSE model, ignoring the actual phylogeny is equivalent to randomly sampling one extant native taxon from the assemblage without any knowledge of the state of any other taxa or the relationships between taxa (FitzJohn et al. 2009). The sampled native taxon effectively becomes a tree with a single edge connecting the taxon to the root of the assemblage, so $p(n_A, n_B, n_{AB} \mid \text{model})$ is $D(T)$ of the single edge, which can be calculated by integrating $D_s(t)$ from the initial state of the sampled native taxon, that is, $D_s(0)$ and $E_s(0)$ (see R code), along the single edge till the root. For $D_s(0)$, because the taxon starts as a single edge at the present, so the sum of $D_s(0) = 1$, and because the taxon is native, the probability of the taxon being endemic at the present is $D_A(0) = n_A/(n_A + n_{AB})$, being widespread at the present is $D_{AB}(0) = n_{AB}/(n_A + n_{AB})$, and being from the regional pool at the present is $D_B(0) = 0$. For $E_s(0)$, because we only sample one native taxa, the probability that an endemic taxon is not sampled in the phylogeny is $E_A(0) = 1 - 1/n_A$, and the probability that a widespread taxon is not included in the phylogeny is $E_{AB}(0) = 1 - 1/n_{AB}$. Also, because the tree implicitly includes one regional pool taxon to define its root, so $E_B(0) = 1 - 1/n_B$.

Next, let's choose an x of interest and calculate $f(x \mid \text{model})$ and $p(n_A, n_B, n_{AB} \mid x, \text{model})$. Here, we choose x as the colonization time that lead to a native taxon in the assemblage. In the Supporting Information, we also give the solution for x being tip branch length of a native taxon. $f(x \mid n_A, n_B, n_{AB,} \text{model})$ is now the probability density of a colonization event occurring at time x along the single edge connecting a randomly chosen native taxon to the root of the assemblage. $f(x \mid \text{model})$ is then the probability density of a colonization event occurring on an edge at time x given the model, which is how the model defines colonization rate d_B as a function of time. For example, if a GeoSSE model assumes d_B decreasing exponentially over time, then $f(x \mid \text{model})$ is the exponentially decreasing function of x. Similarly, for a constant rate GeoSSE model, $f(x \mid \text{model})$ is a uniform distribution. We have already shown that $p(n_A, n_B, n_{AB} \mid \text{model})$ is $D(\text{root})$ along the single edge connecting a randomly chosen native taxon and the root of the assemblage, so $p(n_A, n_B, n_{AB} \mid x, \text{model})$ is $D(\text{root} \mid x)$, which is the probability of the single edge given that the colonization event that lead to the randomly chosen native taxon occurs at time x.

If the native taxon is added to the assemblage by a colonization event at time x, then two conditions must be satisfied:

Condition (1): Any colonization event that occurs along the edge between the present and x does not lead to the taxon. The probability of the single edge under this condition, which we denote as $D(T \mid \tau \geq x)$, can be calculated by setting $D_{AB}(\tau) = 0$ in the term $d_B D_{AB}(\tau)$ in the differential equation for $D_B(\tau)$ when $\tau < x$, because $d_B D_{AB}(\tau)$ describes the probability of the single edge from τ to the present, in other word, the probability of observing the sampled native taxon, after a colonization event at τ (see Goldberg et al. 2011).

Condition (2): Any colonization event that occurs along the edge before time x does not lead to the taxon. The probability of the single edge under this condition equals $D(T) - D(T \mid \tau > x)$.

Then, as the joint probability of conditions (1) and (2):

$$\begin{aligned} p(n_A, n_B, n_{AB} | x, \text{model}) &= D(T | \tau \geq x)\,[D(T) - D(T | \tau > x)] \\ &= D(T | \tau \geq x) - D(T | \tau > x). \end{aligned}$$

Because this has no closed solution, we discretize the null distribution into bins of size Δt, so that the null distribution becomes:

$$\begin{aligned} &p(t \leq x < t + \Delta t | n_A, n_B, n_{AB}, \text{model}) \\ &\propto p(n_A, n_B, n_{AB} | t \leq x < t + \Delta t, \text{model}) \int_t^{(t+\Delta t)} f(x | \text{model})\, dx, \end{aligned}$$

where $p(n_A, n_B, n_{AB} \mid t \leq x < t + \Delta t, \text{model}) = D(T \mid \tau \geq t) - D(T \mid \tau > t + \Delta t)$.

To compare our analytical method with the conventional simulation approach to deriving null distributions, we carry out simulations under a range of GeoSSE models, collect the time of the colonization events that lead to each extant native taxon in the simulations, and compare the distribution of these simulated colonization times to the null distribution derived from our analytical method. Our analytical method provides an effective way of deriving the null distribution of colonization times under different GeoSSE models without requiring time-intensive simulations. The null distribution of colonization times derived by our simulation-free method closely matches the distribution of colonization times in the simulations (Figs. S1–S3). But our analytical method takes less than a second, about 0.5–50% of the time taken to run the simulations depending on the model. More importantly, the computation time of our method does not depend on how often a model can generate the same number of taxa in different states as in the observed assemblage. Simulation details and results are available in the Supporting Information.

INFERRING COLONIZATION EVENTS TO THE ASSEMBLAGE

To test model adequacy, we also need to infer from the observed assemblage phylogeny the probability distribution of the timing of colonization events that led to each native taxon in the observed assemblage, that is, $p(t \leq x < t + \Delta t \mid \text{phylogeny, model})$. We can use the same arguments that derive $p(t \leq x < t + \Delta t \mid n_A, n_B, n_{AB}, \text{model})$ to derive $p(t \leq x < t + \Delta t \mid \text{phylogeny, model})$, because $f(x \mid \text{phylogeny, model}) \propto p(\text{phylogeny} \mid x, \text{model})\, f(x \mid \text{model})$. The only difference is that we calculate $D(T \mid \tau \geq t)$ and $D(T \mid \tau > t + \Delta t)$ by integrating along each edge of the observed

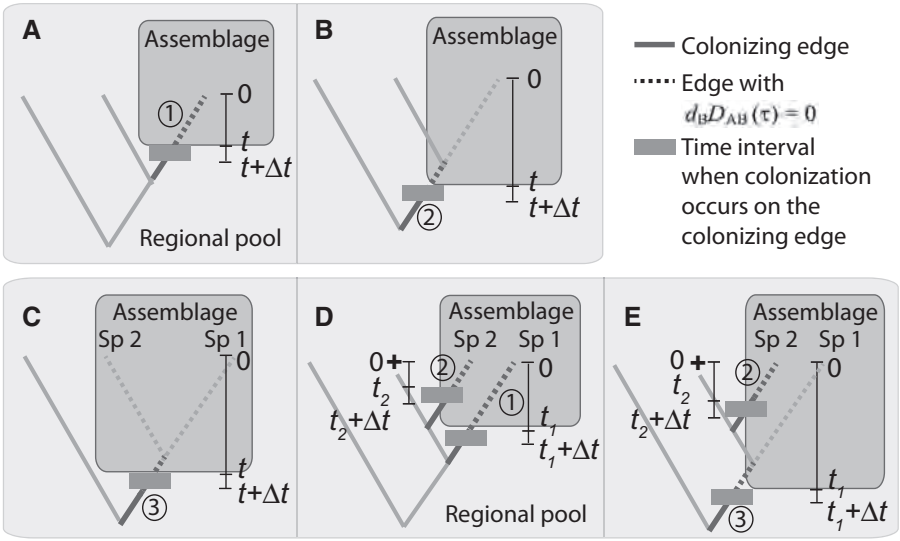

Figure 4. Inferring the distribution of colonization times from the phylogenetic information on assemblage formation. (A and B) The simplest case is where a subtree has only one native taxon. The colonization event that led to the native taxon could have occurred at any time interval on edge 1 (A), or edge 2 (B; e.g., if the native taxon was widespread and split into an endemic species and a regional pool species). (C to E) A subtree with two native taxa, which could have been derived from the same colonization event on edge 3 (C), or from separate colonization events on edge 1 and edge 2 (D), or from separate colonization events on edge 2 and edge 3 (E; same for separate colonization events on edge 1 and edge 3).

assemblage phylogeny from tips to root, rather than along a single edge connecting a randomly chosen native taxa to the root of the assemblage.

Let's start with the simplest case, where a subtree has only one native (endemic or widespread) taxon (Figs. 4A and 4B). We know that this taxon is the descendant of a colonizing lineage that, at some point in the past, colonized the assemblage from the regional pool. We can identify two colonizing edges on the subtree in Figures 4A and 4B, but we do not know at which point on which colonizing edge the colonization event actually occurred. Considering a time t along edge 1 (Fig. 4A), the probability of observing the phylogeny given that the colonization event that led to the native taxon had occurred in Δt right before time t is $p(\text{phylogeny} \mid t \leq x < t + \Delta t \text{ on edge 1, model})$, which equals $D(T \mid \tau \geq t \text{ on edge 1}) - D(T \mid \tau > t + \Delta t \text{ on edge 1})$, and $D(T \mid \tau \geq t$ on edge 1) is integrated from tips to root with the term $d_B D_{AB}(\tau)$ in the differential equation for $D_B(\tau)$ set to 0 along edge 1 when $\tau < t$. Similarly, considering a time t along edge 2 (Fig. 4B), the probability that the colonization event occurred in Δt right before time t along edge 2 equals $D(T \mid \tau \geq t \text{ on edge 1 and 2}) - D(T \mid \tau > t + \Delta t \text{ on edge 1 and 2})$, and $D(T \mid \tau \geq t \text{ on edge 1 and 2})$ is integrated from tips to root with the term $d_B D_{AB}(\tau)$ set to 0 along edge 1 and 2 when $\tau < t$.

When there are multiple native (endemic or widespread) taxa in a subtree, we sum the probabilities of the assemblage phylogeny under all the possible combinations of colonization times for each taxon. For example, if the subtree has two native species that are sister species (Figs. 4C–E), then there are three potential colonizing edges on the subtree. If the two native species are descended from a single colonizing lineage that speciated within the assemblage, then the lineage that connects both native species to the regional pool taxon (edge 3) is the colonizing edge (Fig. 4C). If the two native species are descended from two distinct lineages in the regional pool, then the tip edge of each native species are both colonizing edges (edge 1 and 2 in Fig. 4D). If one native species (say species 1) was widespread and split into an endemic species and a regional pool species (event D in Fig. 3), and the same regional pool species colonized the assemblage again and become species 2, then edge 3 and edge 2 are colonizing edges (Fig. 4E). Using the same method that we used for subtrees with one native taxon, we can calculate the probability of the assemblage phylogeny under these three possible colonization histories of two native taxa. For example, in the first history (Fig. 4C), the probability of the assemblage phylogeny under the condition that the colonization event occurred in Δt right before time t along edge 3 equals $D(T \mid \tau \geq t \text{ on edge 1,2,3}) - D(T \mid \tau > t + \Delta t$ on edge 1,2,3).

After calculating the probability of the assemblage phylogeny under the condition that a colonization event occurring at a time interval on a colonizing edge in a subtree, we sum this probability, for every time interval, over all edges in all subtrees and get $p(t \leq x < t + \Delta t \mid \text{phylogeny, model})$. The relative probability of this, weighted by $\int_t^{t+\Delta t} f(x \mid \text{model})\, dx$, gives the inferred distribution of colonization times of our assemblage,

given the phylogenetic information we have. Thus, although our phylogeny was pruned into subtrees to remove the noninformative edges, the solution we obtain is for the set of all subtrees describing the addition of species to the assemblage.

TESTING MODEL ADEQUACY

So far, we have inferred the distribution of possible timing of colonization events into the assemblage under a model of assemblage formation and have derived the null distribution of colonization times for an assemblage containing the same number of endemic, widespread, and regional pool taxa as the observed assemblage under the same model of assemblage formation. Now, we can ask if this model provides a reasonable description of our observed assemblage. To do this, we need to test whether our observed assemblage phylogeny is significantly different from the phylogenies that could have been generated by the model. If it is, then we reject the model as failing to provide an adequate representation of the process that produced our assemblage. As mentioned before, we need to summarize some relevant aspect (x) of the assemblage phylogeny that is expected to be influenced by the processes of interest and its null distribution can be easily derived (Hua and Bromham 2016), and we have shown how to derive the null distribution for colonization times.

To compare our inferred distribution from real data to the null distribution to ask whether the observed pattern departs significantly from the expectation, we need a statistic that can summarize patterns in any distribution and the expected distribution of the statistic (sampling distribution) is easily derived, so that we can test the statistic value of the inferred distribution against the sampling distribution. Quantiles are a group of generally applicable statistics whose sampling distribution can always be derived analytically. Using quartile as an example, its sampling distribution from a sample of size N

$$F_{Q_a} = \sum_{i=1}^{Q_a} \sum_{j=0}^{\mathrm{ceil}\left(\frac{aN+a}{4}\right)-1} \sum_{k=0}^{N-\mathrm{ceil}\left(\frac{aN+a}{4}\right)} \frac{N!}{j!k!\,(N-j-k)!} \left[\sum_{m=0}^{i-1} f\,(m)\right]^{j} \left[1-\sum_{n=0}^{i} f\,(n)\right]^{k} f(i)^{N-j-k},$$

where $f(\cdot)$ is the probability mass function of the null distribution, N is the number of native taxa in the assemblage phylogeny, and Q_a is the ath quartile of the inferred colonization time, or the minimum colonization time that is larger than $a/4$ proportion of the inferred colonization times. Using the first quartile as an example, $a = 1$. The term to be summed over i gives the probability that, if we randomly draw N colonization times from the null distribution, how likely we can get the first quartile of the N draws equal to i. The first quartile of the N draws equals i when $j\,(j < \lceil\frac{N+1}{4}\rceil)$. number of drawn colonization times are younger than i, $k(k \leq N - \lceil\frac{N+1}{4}\rceil)$ number of drawn colonization times are older than i, and the remaining $N - j - k$ number of drawn colonization times are at i. Summing the term from $i = 1$ to Q_1 gives F_{Q_1}, the probability that the first quartile of N random draws from the null distribution is lower or equal to the first quartile of the inferred colonization times. The F values are related to p-values for the three quartile statistics, which help us evaluate whether the inferred colonization times in the observed assemblage are consistent with random draws from the null distribution. The ath quartile of the inferred colonization times is significantly younger than that of the null distribution of colonization times if $F \leq 0.05$, and significantly older than that of the null distribution if $F \geq 0.95$.

Case Study: Squamate Assemblage of Madagascar

Now we want to know if our approach can detect departures from an assumed model of assemblage formation in a real dataset, focusing on colonization rates. We chose the Malagasy squamate (lizard and snake) assemblage as an ideal data-rich case study in assemblage formation. Madagascar is known for its strikingly high biodiversity and endemism, with approximately 85% of species on the island found nowhere else (Goodman and Benstead 2005). Although Madagascar has a rich fossil record, it is discontinuous: gaps in the record limit the degree to which colonization rates over time can be inferred from paleontological data alone (Samonds et al. 2013). Molecular phylogenies have been used to explore the assembly of Madagascar's unique fauna (Nagy et al. 2003; Townsend et al. 2009; Crottini et al. 2012; Samonds et al. 2012; Scantlebury 2013; Erens et al. 2017; Burbrink et al. 2019). Previous phylogenetic studies derived colonization times from the stem age of the separation between Malagasy clades and their closest non-Malagasy sister group, which were compared to a null model based on a priori assumptions about the time periods over which colonization rates might have changed (Yoder and Nowak 2006; Crottini et al. 2012). Madagascar is therefore a useful case study to illustrate how to apply our methods to test for departures from the expectation of a GeoSSE model that assumes constant colonization rates without requiring colonization events to be assigned to phylogenetic nodes or making a priori assumptions about the patterns of colonization over time.

We uniformly sampled 300 assemblage phylogenies from the posterior distribution generated by Crottini et al. (2012). We included 31 species to represent all the genera in the main island, plus 30 regional pool lineages named as "sister groups" in the original study on the basis of existing taxonomic and phylogenetic studies (Table S2). We first pruned the noninformative edges out of each of the 300 sampled phylogenies. This turns

each sampled phylogeny into 14 subtrees, each representing at least one colonization event into the assemblage (Fig. S1). We then accounted for incomplete sampling of the regional pool lineages by assigning sampling fractions to each edge in a subtree and using a modified GeoSSE model by Hua and Lanfear (2018) to allow sampling fractions vary across edges. Then, we fit the GeoSSE model to all the subtrees by calculating the likelihood of all the subtrees as the product of the likelihoods of individual subtrees. This results in the posterior distribution of all the parameter values in the GeoSSE model (Fig. 2B). Last, we uniformly sampled 1000 sets of parameter values from their posterior distribution and, for each set of parameter values, we applied our approach to infer the distribution of colonization times on the phylogeny, derive the null distribution of colonization times, and compare the inferred distribution to the null distribution to detect departures from constant colonization rates. Details for the analyses are available in the Supporting Information and the R code.

RESULTS

For the case study of Malagasy squamates, our approach detects significant departures in the distribution of colonization times from that expected by constant colonization rates over time. According to the inferred distribution of colonization times, more Malagasy squamates originated from colonization events in the middle period of the assemblage history than expected under the best-fit GeoSSE models (Fig. 5). We plotted the F values of each set of parameter values for each phylogeny in Figure 5 and used the median of these F values to test significant departure from the null distribution of colonization times. We found that the first quartile ($F = 0.946$) and the median ($F = 0.984$) of inferred times of colonization events are significantly older than expected, and the third quartile ($F = 0.006$) of inferred times of colonization events is significantly younger than expected.

If the dates inferred from this molecular phylogeny are accurate, our results indicate that squamate colonization of Madagascar peaked in the Early Cenozoic, a pattern that is not expected under a model of constant rate of species addition over the history of the assemblage. Instead, the pattern is consistent with a paleo-oceanographic model (Ali and Huber 2010) that predicts the occurrence of two surface currents, one in the Early Cenozoic and the other in the Late Cenozoic. The Early Cenozoic surface currents favored overseas dispersals from Africa to Madagascar between 66 and 16 million years ago, whereas the Late Cenozoic surface currents favored overseas dispersals from Madagascar to Africa from 16 million years ago to present (Ali and Huber 2010). The two surface currents, when considered together, could cause an Early Cenozoic peak time in colonization, by favoring movement of lineages from Africa to Madagascar in the early Cenozoic but reducing the chance of colonization when the currents moved in the opposite direction.

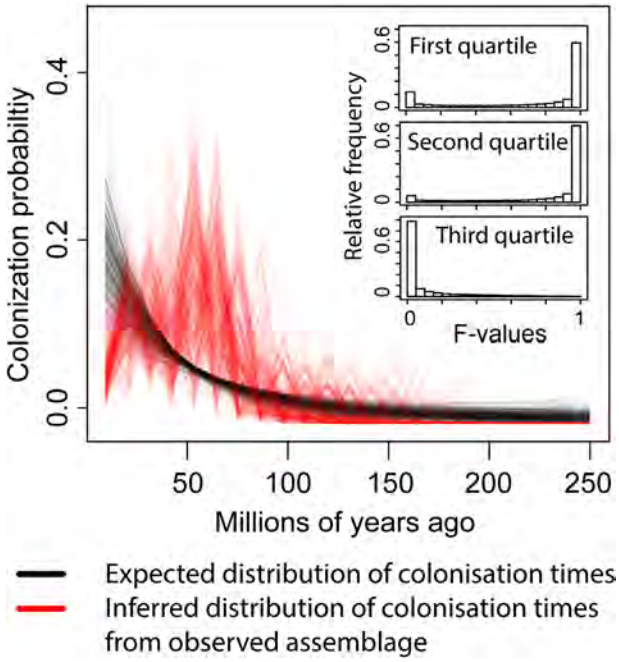

Figure 5. Comparisons between the null and the inferred distribution of colonization times for Malagasy squamates. Each black line shows the null distribution generated by our analytical method for the best-fit GeoSSE models to each of the 300 sampled phylogenies for the assemblage. Each red line shows the distribution of colonization times inferred on each phylogeny by the best-fit GeoSSE models. Peaks in the red lines are related to time intervals with the most colonizing edges where colonization events are likely to occur. The three histograms in the plot show the distribution of *F*-values for the three quartiles of the inferred distribution of colonization times. Each *F*-value is related to the probability that the inferred colonization times could have been drawn from the null distribution of colonization times. $F \leq 0.05$ suggests significantly younger colonization times than expected by the null distribution, and $F \geq 0.95$ suggests significantly older colonization times than expected.

Discussion

Although the development of phylogenetic methods for inferring community assembly has blossomed and they have been applied to an ever-increasing range of case studies, there has been a notable limitation of available methods. Although there are well-established methods for generating expected distributions of speciation times, given an assemblage phylogeny and a model for assemblage formation, methods for inferring changes in colonization rates have remained subject to a number of key limitations. In particular, there has been a lack of general method that simultaneously accounts for several key sources of uncertainty in inferring colonization times from phylogenies and generating null distribution of colonization times under various models of assemblage formation. Here, we provide such a general approach that explicitly accounts for these sources of uncertainty yet is tractable for

real datasets. Our approach analytically derives the null distribution of colonization times for an assemblage, given the observed number of endemic, widespread, and regional pool taxa, and it provides a way to infer colonization events on the assemblage phylogeny and compare the inferred distribution of colonization times on phylogeny to the null distribution, allowing us a means of testing the adequacy of models for assemblage formation.

Applying the method to study the formation of Malagasy squamates, we get results that agree with recent studies of the phylogenetic and biogeographic patterns of extant Malagasy vertebrate groups (Crottini et al. 2012; Samonds et al. 2012). But our approach provides a more robust test for the paleo-oceanographic model than previous studies, because it does not rely on a priori hypotheses, explicitly accounts for uncertainties in inferring colonization times from phylogenies, and compares the inferred colonization times with the null expectation under constant species addition over time. Our approach puts the test firmly in a hypothesis testing framework in which the detection of peak time in colonization events does not rely on a priori assumptions about particular historical events that might have influenced colonization rate. Using this approach, we can demonstrate that more squamate lineages were added to the Malagasy assemblage during the middle period of the assemblage history than we would expect if colonization rates had been constant over time.

In our case study of the squamate assemblage of Madagascar, native taxa are clearly nested within related lineages that are diversifying in other places, so we are able to identify subtrees that represent independent colonization events into the assemblage. But for some assemblages, such as the rodent assemblage of South America (Fabre et al. 2012), native taxa are not all nested within lineages from other places. This makes it impossible to break the assemblage phylogeny into small subtrees. Our method to infer colonization times requires listing all possible locations of colonization events that could have led to each extant native taxon in a subtree, so the larger the subtree is, the more possible locations of colonization events, and the more computational burden. This is not a problem when most native taxa in the subtree form a monophyletic clade, as it is unlikely that each native taxon in the clade was added to the assemblage by independent colonization events, so we can consider colonization events only along the edge that connects the native clade to its closest neighboring relatives. But when some nested lineages are not native, there could be multiple independent colonization events in the clade, so we need a more efficient algorithm to infer colonization times. One solution to this problem is to use Markov chains (Hobert et al. 2006) to approximate the inferred distribution of colonization times and use more efficient algorithm to calculate the likelihood of the tree (e.g., Louca and Pennell 2020). Alternatively, we can use stochastic character mapping to simulate the distribution of transitions from regional pool to widespread or endemic state on the assemblage phylogeny (e.g., Freyman and Höhna 2019).

In this study, we focus on testing model adequacy about colonization rates, so we use summary statistics of the null distribution of colonization times as our test statistics. But these statistics may not be optimal for testing the model adequacy about other processes of assemblage formation, such as the extinction rates. To test model adequacy about all processes of assemblage formation, we need to use as many test statistics as possible. Technically speaking, our approach can develop any test statistics related to times of changes in the assemblage. For example, we demonstrate how it can derive the null distribution of tip lengths of endemic or widespread taxa, which is more sensitive to extinction rates, in the Supporting Information.

In conclusion, we offer a flexible and efficient approach to testing changes in colonization rates, which can be used as a general approach to testing the adequacy of a model to describe the formation of an observed assemblage, using its phylogeny. Together with standards in testing speciation rates, we hope that this method will offer a general pipeline for investigating the dynamics of biodiversity accumulation over time for a wide range of biological communities.

AUTHOR CONTRIBUTIONS

XH and LB conceived the study and wrote this article. XH devised, developed, and tested the methods. XH conducted analyses. Both authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank T. Bennett, R. Lanfear, and M. Cardillo for their contributions to this project, and members of the MacroEvoEco group for their comments on the manuscript, particularly R. Dinnage, A. Ritchie, and A. Skeels. We are grateful to C. Poux and M. Vences for providing the data for the Madagascar case study.

DATA ARCHIVING

The R code used in this article is available at https://github.com/huaxia1985/TestingConstantSpeciesAddition.

LITERATURE CITED

Ali, J. R., and M. Huber. 2010. Mammalian biodiversity on Madagascar controlled by ocean currents. Nature 463:653–656.

Bromham, L., S. Duchêne, X. Hua, R. M. Ritchie, D. A. Duchêne, and S. Y. W. Ho. 2018. Bayesian molecular dating: opening up the black box. Biol. Rev. 93:1165–1191.

Burbrink, F. T., S. Ruane, A. Kuhn, N. Rabibisoa, B. Randriamahatantsoa, A. P. Raselimanana, M. S. M. Andrianarimalala, J. E. Cadle, A. R. Lemmon, E. M. Lemmon, et al. 2019. The origins and diversification of the exceptionally rich gemsnakes (Colubroidea: Lamprophiidae: Pseudoxyrhophiinae) in Madagascar. Syst. Biol. 68:918–936.

Cabral, J. S., K. Wiegand, and H. Kreft. 2019. Interactions between ecological, evolutionary and environmental processes unveil complex dynamics of insular plant diversity. J. Biogeogr. 46:1582–1597.

Cardillo, M. 2011. Phylogenetic structure of mammal assemblages at large geographical scales: linking phylogenetic community ecology with macroecology. Phil. Trans. R. Soc. B 366:2545–2553.

Cavender-Bares, J., K. H. Kozak, P. V. Fine, and S. W. Kembel. 2009. The merging of community ecology and phylogenetic biology. Ecol. Lett. 12:693–715.

Crottini, A., O. Madsen, C. Poux, A. Strauß, D. R. Vieites, and M. Vences. 2012. Vertebrate time-tree elucidates the biogeographic pattern of a major biotic change around the K–T boundary in Madagascar. Proc. Natl. Acad. Sci. USA 109:5358–5363.

De Villemereuil, R., J. A. Wells, R. D. Edwards, and S. P. Blomberg. 2012. Bayesian models for comparative analysis integrating phylogenetic uncertainty. EMC Evol. Biol. 12:102.

Emerson, B. C. 2002. Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. Mol. Ecol. 11:951–966.

Emerson, B. C., and R. G. Gillespie. 2008. Phylogenetic analysis of community assembly and structure over space and time. Trends Ecol. Evol. 23:619–630.

Erens, J., A. Miralles, F. Glaw, L. W. Chatrou, and M. Vences. 2017. Extended molecular phylogenetics and revised systematics of Malagasy scincine lizards. Mol. Phylogenet. Evol. 107:466–472.

Fabre, P., L. Hautier, D. Dimitrov, and E. J. P. Douzery. 2012. A glimpse on the pattern of rodent diversification: a phylogenetic approach. BMC Evol. Biol. 12:88.

FitzJohn, R. G. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595–611.

Freyman, W. A., and S. Höhna. 2019. Stochastic character mapping of state-dependent diversification reveals the tempo of evolutionary decline in self-compatible Onagraceae lineages. Syst. Biol. 68:505–519.

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

Goodman, S. M., and J. P. Benstead. 2005. Updated estimates of biotic diversity and endemism for Madagascar. Oryx 39:73–77.

Harmon, L. J., and S. Harrison. 2015. Species diversity is dynamic and unbounded at local and continental scales. Am. Nat. 185:584–593.

Hartmann, K., D. Wong, and T. Stadler. 2010. Sampling trees from evolutionary models. Syst. Biol. 9:465–476.

Hobert, J. P., G. L. Jones, and C. P. Robert. 2006. Using a Markov chain to construct a tractable approximation of an intractable probability distribution. Scand. J. Stat. 33:37–51.

Höhna, S. 2013. Fast simulation of reconstructed phylogenies under global time-dependent birth-death processes. Syst. Biol. 29:1367–1374.

Hua, X., and L. Bromham. 2016. Phylometrics: an R package for detecting macroevolutionary patterns, using phylogenetic metrics and backward tree simulation. Methods Ecol. Evol. 32:2633–2645.

Hua, X., and R. Lanfear. 2018. The influence of non-random species sampling on macroevolutionary and macroecological inference from phylogenies. Methods Ecol. Evol. 9:1353–1362.

Jablonski, D., K. Roy, and J. W. Valentine. 2006. Out of the tropics: evolutionary dynamics of the latitudinal diversity gradient. Science 314:102–106.

Kooyman, R., M. Rossetto, C. Allen, and W. Cornwell. 2012. Australian tropical and subtropical rain forest community assembly: phylogeny, functional biogeography, and environmental gradients. Biotropica 44:668–679.

Louca, S., and M. W. Pennell. 2020. A general and efficient algorithm for the likelihood of diversification and discrete-trait evolutionary models. Syst. Biol. 69:545–556.

MacArthur, R. H., and E. O. Wilson. 1967. The theory of island biogeography. Princeton Univ. Press, Princeton, NJ.

Matos-Maraví, P., N. J. Matzke, F. J. Larabee, R. M. Clouse, W. C. Wheeler, D. M. Sorger, A. V. Suarez, and M. Janda. 2018. Taxon cycle predictions supported by model-based inference in Indo-Pacific trap-jaw ants (Hymenoptera: Formicidae: *Odontomachus*). Mol. Ecol. 27:4090–4107.

McPeek, M. A. 2008. The ecological dynamics of clade diversification and community assemblage. Am. Nat. 172:E270–E284.

Morlon, H., M. D. Potts, and J. B. Plotkin. 2010. Inferring the dynamics of diversification: a coalescent approach. PLoS Biol. 8:e1000493.

Nagy, Z. T., U. Joger, M. Wink, F. Glaw, and M. Vences. 2003. Multiple colonization of Madagascar and Socotra by colubrid snakes: evidence from nuclear and mitochondrial gene phylogenies. Proc. R. Soc. B 270:2613–2621.

Onstein, R. E., G. J. Jordan, H. Sauquet, P. H. Weston, Y. Bouchenak-Khelladi, R. J. Carpenter, and H. P. Linder. 2016. Evolutionary radiations of Proteaceae are triggered by the interaction between traits and climates in open habitats. Glob. Ecol. Biogeogr. 25:1239–1251.

Phillimore, A. B., and T. D. Price. 2008. Density-dependent cladogenesis in birds. PLoS Biol. 6:483–489.

Rabosky, D. L. 2009. Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. Ecol. Lett. 12:735–743.

Rabosky, D. L., and A. H. Hurlbert. 2015. Species richness at continental scales is dominated by ecological limits. Am. Nat. 185:572–583.

Rangel, T. F., R. K. Colwell, G. R. Graves, K. Fučíková, C. Rahbek, and J. A. F. Diniz-Filho. 2015. Phylogenetic uncertainty revisited: implications for ecological analyses. Evolution 69:1301–1312.

Raup, D. M., S. J. Gould, T. J. Schopf, and D. S. Simberloff. 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol. 81:525–542.

Ree, R. H., and S. A. Smith. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Syst. Biol. 57:4–14.

Ricklefs, R. E., and E. Bermingham. 2004. Application of Johnson et al.'s speciation threshold model to apparent colonisation times of island biotas. Evolution 58:1664–1673.

———. 2008. The West Indies as a laboratory of biogeography and evolution. Philos. Trans. R. Soc. B 363:2393–2413.

Ronquist, F. 1997. Dispersal-Vicariance Analysis: a new approach to the quantification of historical biogeography. Syst. Biol. 46:195–203.

Salzburger, W., A. Meyer, S. Baric, E. Verheyen, and C. Sturmbauer. 2002. Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. Syst. Biol. 51:113–135.

Samonds, K. E., L. R. Godfrey, J. R. Ali, S. M. Goodman, M. Vences, M. R. Sutherland, M. T. Irwin, and D. W. Krause. 2012. Spatial and temporal arrival patterns of Madagascar's vertebrate fauna explained by distance, ocean currents, and ancestor type. Proc. Natl. Acad. Sci. USA 109:5352–5357.

———. 2013. Imperfect isolation: factors and filters shaping Madagascar's extant vertebrate fauna. PLoS ONE 8:e62086.

Scantlebury, D. P. 2013. Diversification rates have declined in the Malagasy herpetofauna. Proc. R. Soc. B 280:20131109.

Shafquat, A., R. Joice, S. L. Simmons, and C. Huttenhower. 2014. Functional and phylogenetic assembly of microbial communities in the human microbiome. Trends Microbiol. 22:261–266.

Stadler, T. 2011. Simulating trees on a fixed number of extant species. Syst. Biol. 60:676–684.

Townsend, T. M., D. R. Vieites, F. Glaw, and M. Vences. 2009. Testing species-level diversification hypotheses in Madagascar: the case of microendemic Brookesia leaf chameleons. Syst. Biol. 58:641–656.

Valente, L. M., A. B. Phillimore, and R. S. Etienne. 2015. Equilibrium and non-equilibrium dynamics simultaneously operate in the Galápagos islands. Ecol. Lett. 18:844–852.

Valente, L., R. S. Etienne, and L. M. Dávalos. 2017a. Recent extinctions disturb path to equilibrium diversity in Caribbean bats. Nat. Ecol. Evol. 1:0026.

Valente, L., J. C. Illera, K. Havenstein, T. Pallien, R. S. Etienne, and R. Tiedemann. 2017b. Equilibrium bird species diversity in Atlantic islands. Curr. Biol. 27:1660–1666.

Valente, L., A. B. Phillimore, M. Melo, B. H. Warren, S. M. Clegg, K. Havenstein, R. Tiedemann, J. C. Illera, C. Thébaud, T. Aschenbach, et al. 2020. A simple dynamic model explains the diversity of island birds worldwide. Nature 579:92–96.

Vellend, M. 2010. Conceptual synthesis in community ecology. Q. Rev. Biol. 85:183–206.

Von Humboldt, A. 1847. Cosmos: a sketch of a physical description of the universe. John Murray, Lond.

Wallace, A. R. 1855. On the law which has regulated the introduction of new species. Ann. Mag. Nat. Hist. 16:184–196.

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and community ecology. Annu. Rev. Ecol. Syst. 33:475–505.

Welch, J. J., and L. Bromham. 2005. Molecular dating when rates vary. Trends Ecol. Evol. 20:320–327.

Yoder, A. D., and M. D. Nowak. 2006. Has vicariance or dispersal been the predominant biogeographic force in Madagascar? Only time will tell. Annu. Rev. Ecol. Evol. Syst. 37:405–431.

Associate Editor: M. Bonsall
Handling Editor: T. Chapman

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Sets of parameter values used in simulations to test our analytical solution over macroevolutionary models.
Table S2. Native species sampled in the Malagasy squamates case study to represent their higher taxa.
Figure S1. Distribution of colonization times for endemic taxa in simulated phylogenies and that derived from analytical method.
Figure S2. Distribution of colonization times for widespread taxa in simulated phylogenies and that derived from analytical method.
Figure S3. Distribution of colonization times for both endemic and widespread taxa in simulated phylogenies and that derived from analytical method.
Figure S4. Distribution of tip branch length for endemic taxa in simulated phylogenies and that derived from analytical method.
Figure S5. Distribution of tip branch length for widespread taxa in simulated phylogenies and that derived from analytical method.
Figure S6. Distribution of tip branch length for both endemic and widespread taxa in simulated phylogenies and that derived from analytical method.
Figure S7. The maximum clade credibility tree of Malagasy squamates from Crottini et al. (2012) showing the identified subtrees.