Received: 15 October 2021 | Revised: 22 August 2022 | Accepted: 26 August 2022

DOI: 10.1111/jbi.14489

Journal of Biogeography WILEY

METHOD

The missing link in biogeographic reconstruction: Accounting for lineage extinction rewrites history

Leonel Herrera-Alsina[1] | Adam C. Algar[2] | Lesley T. Lancaster[1] | Juan Francisco Ornelas[3] | Greta Bocedi[1] | Alexander S. T. Papadopulos[4] | Cecile Gubry-Rangin[1] | Owen G. Osborne[4] | Poppy Mynard[1] | I. Made Sudiana[5] | Pungki Lupiyaningdyah[6] | Berry Juliandi[7] | Justin M. J. Travis[1]

[1]School of Biological Sciences, University of Aberdeen, Aberdeen, UK

[2]Lakehead University, Thunder Bay, Ontario, Canada

[3]Departamento de Biología Evolutiva, Instituto de Ecología, A.C. (INECOL), Xalapa, Mexico

[4]School of Natural Sciences, Bangor University, Bangor, UK

[5]Research Center for Biology, Indonesian Institute of Sciences, Jakarta, Indonesia

[6]Zoology Division, Museum Zoologicum Bogoriense, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Indonesia

[7]Department of Biology, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia

Correspondence
Leonel Herrera-Alsina, School of Biological Sciences, University of Aberdeen, Aberdeen, UK, AB24 2TZ.
Email: leonelhalsina@gmail.com

Funding information
nerc/newton

Handling Editor: Erin Saupe

Abstract

Aim: In the most widely used family of methods for ancestral range estimation (ARE), dispersal, speciation and extirpation events are estimated from information on extant lineages. However, this approach fails to consider the geographic distribution of extinct species and their position on the phylogenetic tree, an omission that could compromise reconstruction. Here, we present a method that models the geographic distribution of extinct species and we quantify the potential inaccuracy in ancestral range estimation when extinction rates are above zero.

Location: Global applications, with an example from the Americas.

Taxon: All taxa, with an example from hummingbirds (*Amazilia*).

Methods: Methods capable of explicitly modelling extinct branches along with their reconstructed geographic information (GeoSSE) have been overlooked in ARE analysis, perhaps due to the inherent complexity of implementation. We develop a user-friendly platform, which we term LEMAD (Lineage Extinction Model of Ancestral Distribution) that generalizes the likelihood described in GeoSSE for any number of areas and under several sets of geographic assumptions. We compare LEMAD and extinction-free approaches using extensive simulations under different macroevolutionary scenarios. We apply our method to revisit the historical biogeography of *Amazilia* hummingbirds.

Results: We find that accounting for the lineages removed from a tree by extinction improves reconstructions of ancestral distributions, especially when rates of vicariant speciation are higher than rates of in situ speciation, and when rates of extinction and range evolution are high. Rates of in situ and vicariant speciation are accurately estimated by LEMAD in all scenarios. North America as the most likely region for the common ancestor of hummingbirds.

Main conclusions: Methods that neglect lineage extinction are less likely to accurately reconstruct true biogeographic histories of extant clades. Our findings on an

empirical dataset reconcile the Eurasian origin of *Amazilia* with biogeographic reconstructions when lineage extinction is considered.

KEYWORDS
ancestral distribution, BioGeoBEARS, centre of origin, diversification events, extinction rates, hummingbird evolution, in situ speciation, vicariance

1 | INTRODUCTION

Identifying the geographic centre-of-origin for diverse clades has long been of interest in biogeography. This endeavour is made difficult because the presence or absence of a species at a given location varies over time and, over longer time-scales, species continuously appear and disappear from the Earth (Barraclough & Vogler, 2000; Jablonski & Sepkoski, 1996; Losos & Glor, 2003). The distribution of clades is the result of shifts in the distribution of constituent species via range shifts and speciation and extinction, but, in many cases, these processes may leave little fossil or other tangible evidence of their history, meaning that inferences of centres-of-origin must be inferred from data on extant species and extant ranges. A foundational field in modern biogeographic research has been the reconstruction of the geographic distributions of ancestral lineages, in order to relate biogeographic processes to extrinsic events (e.g. geological shifts, onset of ice ages) while increasingly taking intrinsic, evolutionary processes into account.

For Ancestral Range Estimation (ARE), the two popular methods (DIVA, Dispersal Vicariance Analysis; Ronquist, 1997 and DEC Dispersal-Extinction-Colonization model; Ree & Smith, 2008) use the term 'extinction' to refer to extirpation (i.e. local extinction), while true lineage extinction is ignored. These approaches (hereafter Extinction Free approaches; EF) consider the following events: dispersal, extirpation and speciation, and are therefore appropriate when all lineages and speciation events are represented in the phylogeny, that is, no branches are missing due to extinction. However, the vast majority of available phylogenetic trees are reconstructions where extinction has removed many branches, such that a pair of extant species that appear as sister species (or clades) in a reconstructed tree might not be true sisters due to missing nodes. Using the geographic distributions of the apparent pair of sister clades to infer whether in situ speciation or vicariance occurred at the node where they diverged (the putative common ancestor) may be unreliable because any extinct, intermediary lineages are not only absent from the tree but any information on their geographic distribution is also missing. Thus, attempting to infer in situ speciation and vicariance events across a phylogenetic reconstruction without accounting for extinct lineages could compromise the ancestral range estimation (Figure 1). Although the problem of extinct lineages in macroevolution and biogeography has been pointed out by Sanmartín and Meseguer (2016) and more specifically for ancestral range estimation by Crisp et al. (2011), the consequences of ignoring extinct lineages for ARE are still unknown and unquantified.

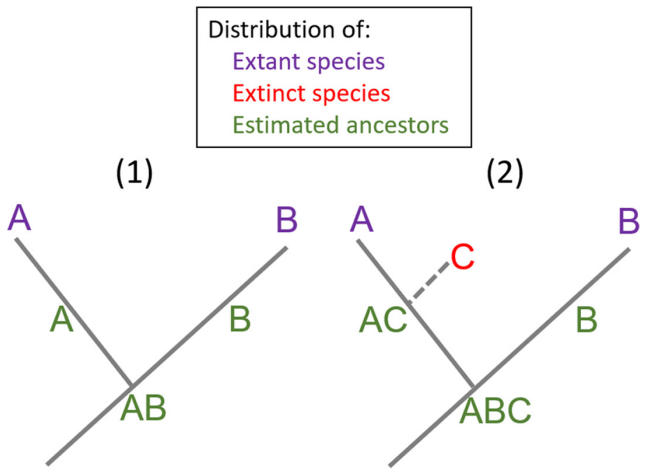

FIGURE 1 The reconstruction of the ancestral distribution for a two-species clade inhabiting a region with subregions A, B and C. We show the main difference between (1) extinction free (EF) method and (2) lineage extinction model of ancestral distribution (LEMAD). In contrast to EF, LEMAD considers the missing lineages due to extinction and their geographic distribution in the analysis.

One way forward is to model the potential past existence of lineages at any point of a tree branch, which extinction subsequently removed, and to account for all the possible geographic distributions of those extinct lineages. This is achieved in ClaSSE (Cladogenetic State change Speciation and Extinction) and GeoSSE (Geographic State Speciation and Extinction) models (Goldberg et al., 2011; Goldberg & Igić, 2012) where speciation, lineage extinction, dispersal and extirpation events are part of the biogeographic dynamics. These models have been used for describing how biodiversity accumulates over time in a dynamic context, and in particular, to explore spatial differences in diversification rates (e.g. Ding et al., 2020; Meseguer et al., 2020). However, their potential for estimating ancestral distributions has been almost overlooked (but see Lancaster & Kay, 2013, Caetano et al., 2018). This is surprising, because the spatial distribution of ancestors is estimated during SSE likelihood calculation. Matzke (2014) and Ree and Sanmartín (2018) acknowledged the utility of -SSE models but found that existing implementations were not easy to use, especially with more than two regions. Here, we use extensive simulations to quantify the impact of including lineage extinction in ancestral range estimation by comparing the outcome of

EF approaches to that of our new -SSE implementation, which we have also made available as a user-friendly R package. We simulated biogeographic scenarios that differed in the relative rates of in situ and vicariant speciation along with different rates of lineage extinction to document variation in performance of both approaches. Finally, we apply our approach to estimate the biogeographic history of *Amazilia* hummingbirds. Evolutionary studies using DIVA and DEC have supported South America as the most likely location of the first speciation event in hummingbirds (McGuire et al., 2014); however, the fossil record points to Eurasia as the source region of the hummingbird lineage (Louchart et al., 2008; Mayr, 2004). This leaves a time gap of several million years and a geographic gap of thousands of kilometres. Our approach shows that this spatiotemporal gap is explained when lineage extinction is no longer neglected.

2 | MATERIALS AND METHODS

2.1 | Extinction free models: Differences between DIVA and DEC

Extinction free methods (EF) require that the distribution of a clade be divided into regions (letters are used for convention) so that the geographic distribution of a given species is coded by its presence in those regions, and occupancy of multiple regions is indicated by the combination of corresponding letters. A time-calibrated tree that includes all extant species is also needed (our approach requires the same data, see below). Matzke (2013) upgraded DIVA (Ronquist, 1997) from parsimony to likelihood in his BioGeoBEARS R package while maintaining its properties, whereas DEC is a model originally developed in a likelihood framework (these methods differ in some assumptions, see below). In such models, speciation is typically modelled as in situ speciation (occurring within a region, increasing local diversity) or vicariance (geographically mediated divergence resulting in allopatry, i.e. complementary ranges). Here, we use the notation $DIVA_{events}$ and DEC_{events} to refer to the two sets of biogeographic assumptions and leave DIVA and DEC to refer to the models of ancestral range estimation implemented in BioGeoBEARS. $DIVA_{events}$ assumes that widespread species can split their ranges (vicariance) in any combination regardless of the number of areas where daughter lineages inhabit (e.g. a species presents in region A, B, C and D can split in AB-CD or A-BCD; widespread vicariance sensu Matzke, 2013) while DEC_{events} assumes that one of the daughter lineages will be present at a single region (e.g. ABCD species splits in A-BCD or B-ACD; narrow vicariance). For in situ speciation and in contrast with $DIVA_{events}$, DEC_{events} allows widespread lineages to speciation by having one population (i.e. one of the regions where it is present) diverging from the rest and coexisting with the parental lineage: for instance, ABCD species produces one daughter lineage which is present at ABCD and the other daughter which is restricted to region A (in situ subset hereafter; sympatry subset sensu Matzke, 2013).

2.2 | Lineage extinction model of ancestral distribution (LEMAD)

We use the area/trait-dependent diversification framework (State-dependent Speciation and Extinction, -SSE models; Maddison et al., 2007; Goldberg et al., 2011; Herrera-Alsina et al., 2019) to model past changes in species' geographic distributions. We generalize the computation of the likelihood described in GeoSSE (Goldberg et al., 2011) for any number of areas and under several sets of geographic assumptions that facilitate its use in ancestral range estimation (ARE). Notice that GeoSSE and ClaSSE (Goldberg & Igić, 2012) models have the same system of equations. During the R package building process, we calculated the likelihood under GeoSSE (from diversitree package) and LEMAD for a dataset (model parameters, tree and geographic distribution of species in two areas) to confirm that the likelihoods are identical (Fitzjohn, 2012). Unlike EF methods, the -SSE framework considers that, at any point along a tree branch, a lineage could have been present but went extinct, with or without first producing (also extinct) descendants. To this end, the algorithm uses two coupled differential equations (Appendix S1), where one accounts for the probability of a lineage being at a given region (or set of regions), and the other reflects the probability of a lineage going extinct for the same region (or set of regions). These equations are numerically integrated to obtain a likelihood value for the data given the model with its parameters (dispersal/contraction, in situ and vicariant speciation). Different parameter combinations are tested to find the best combination (likelihood optimization). With the parameters that maximize the likelihood, we compute the change in probability for a lineage to be at each distribution from the present (tree tips) to the past (root) and extract those probabilities at the nodes. Ancestral range probabilities were estimated by taking the partial likelihoods from the downpass and rescaling them so that they summed to 1 at each node (Nguyen, 2011). In summary, the model simultaneously considers the probabilities of dispersal, extirpation and speciation (via in situ or vicariance) for extant and extinct lineages. The likelihood of the model is optimized, and the rates of geographic change, in situ speciation and vicariance are estimated. Lineage extinction can be estimated or fixed to a specific rate by the user. In short, Lineage Extinction Model of Ancestral Distribution (LEMAD) computes the likelihood of the current distribution of species (given the parameters of the model) where lineage extinction is a fundamental part of the calculation. The R package lemad is available at https://github.com/leonelhalsina/lemad.

2.3 | General assumptions in LEMAD

Although the LEMAD model can account for differences in diversification rates across regions (like in GeoSSE/ClaSSE original application), in LEMAD, the rates of speciation and extinction are constant across regions. This is achieved by assigning the same rate of speciation and extinction to each area or combination of areas during the parameter

setup. This simplification is necessary to reduce the otherwise immense complexity of parameter space when the analysis is performed for many regions; note that this assumption is the same in DIVA and DEC models. In LEMAD, we assume that shifts in the geographic distribution of species are the product of expansion and contraction. For example, a species present in region A cannot instantaneously change to region B. It has first to expand to region B (to be present in AB) followed by an extirpation event in A. These assumptions are the same as in EF methods. Lineage extinction can be modelled in two ways: extinction by extirpation and instantaneous extinction. In the former case, a lineage can undergo extirpation events in different regions of its distribution (range contraction) and eventually go extinct when it is extirpated from its last remaining region. This is similar to the idea of the empty range (∅) in Ree and Smith (2008). In the case of instantaneous extinction, a species can go extinct regardless of the number of regions where it is present. Although extinction by extirpation is appropriate when regions are small and each of them represents a single population (the extinction of a species takes place once the last population disappears), the scale at which ARE is normally conducted renders this type of extinction inappropriate (Polly & Sarwar, 2014). Furthermore, by using instantaneous extinction, we account for those events that involve a sudden decline in total population size that are not related to standard dynamics of region colonization/extirpation, so we can measure the contribution of each process independently. We therefore used instantaneous extinction in LEMAD, but extinction by extirpation could also be enabled. Our model assumes that lineages, including extinct lineages and ancestors, can be present in multiple regions, even if extant species are not. For instance, with three regions (A, B and C), LEMAD calculates the probability of the ancestors being present in A, B, C, AB, AC, BC or ABC (all possible combinations). By allowing this, we do not constrain the model to only consider region-endemic lineages, which could lead to underestimation of the importance of widespread historical lineages in shaping more narrow modern distributions. However, the model is flexible enough to set any restriction in the number of permitted regions per ancestral species. Note that LEMAD can handle any number of regions; however, computation time will exponentially increase with the number of regions. For instance, a phylogenetic tree with 66 species and 3 areas (yielding seven possible ancestral areas) can take around 10 min of computing time. With four areas (and 15 possible ancestral distributions), the calculation can take around 75 min. With six areas (and 63 possible states), the computing time can be as long as 100 h.

LEMAD enables two different sets of biogeographic assumptions (i.e. $LEMAD_{diva\_events}$ and $LEMAD_{dec\_events}$; we refer to both models under the term LEMAD) that match $DIVA_{events}$ and DEC_{events}. As they are different parameterizations of the same model, the comparison of their likelihoods is valid and straightforward.

2.4 | Accuracy assessment

In order to compare the accuracy of LEMAD and EF approaches under different extents of extinction, we modelled a number of scenarios in which we (i) simulated the evolutionary history of a clade along with the geographic evolution of its species, (ii) fit both models and (iii) compared their ancestral range estimations.

2.4.1 | Simulation procedure

The simulation started with one lineage in a random region (A, B and C) or combination of regions (AB, AC, BC or ABC); lineages undergo the following events: dispersal, extirpation, speciation and extinction. The simulation runs in continuous time where the waiting time between events is drawn from an exponential distribution (Gillespie algorithm; Doob, 1945; Gillespie, 1977). The duration of the simulation is chosen to ensure a final clade size of 150 species given the speciation rates (scenarios with high extinction were allowed to run longer, see below).

We kept track of the geographic distribution of lineages over time and of ancestor–descendant relationships and used this as a record to build a phylogenetic tree of the clade. As a result, the simulation produces a phylogenetic tree (without extinct branches, similar to standard reconstructed trees) and the geographic distribution of extant species. Notice that species (ancestors and extant lineages) could be in any of the seven states of the system (A, B, C, AB, AC, BC or ABC).

2.4.2 | Model fitting

We simulated two datasets that differed in modes of vicariance and in situ speciation, following the assumptions in $DIVA_{events}$ and DEC_{events}. For the simulations under $DIVA_{events}$, we fitted DIVA (from BioGeoBEARS) and $LEMAD_{diva\_events}$. Similarly, the simulations under DEC_{events} were fit with DEC (from BioGeoBEARS) and $LEMAD_{dec\_events}$. Next, we extracted the most likely ancestral distribution estimated by LEMAD and EF for every node in the phylogenetic reconstruction and compared to the record of ancestors directly from simulated datasets. This is, for a given ancestor/node, we took the distribution with the highest probability and compared to the distribution that was logged during the simulation. We defined a node successfully inferred when both distributions matched completely (if A is the simulated truth, only A would be a successful reconstruction. Neither AB nor ABC would be correct). We counted the number of nodes that were successfully recovered by both models in two sections of time during the history of the clade: recent and ancient time windows. We repeated the simulation-inference procedure under 18 different parameter combinations: rates of in situ speciation = 0.02, 0.03, 0.04 and vicariance = 0.02, 0.03, 0.04 to combine into three scenarios with overall speciation of 0.06; extinction = 0, 0.003, 0.03; dispersal/extirpation = 0.06, 0.6 (30 runs for each combination). To measure the accuracy in parameter estimation, we used the rates (geographic change, in situ speciation and vicariance) that are estimated during the analysis and compared them to the simulation generating rates. Lineage extinction was not estimated but was fixed to the generating rate as we were interested in the performance of the other (more informative) parameters.

We were also interested in measuring whether phylogenetic reconstructions and geographic data are informative about the modes of in situ and vicariant speciation, which constitute the main difference between $DIVA_{events}$ and DEC_{events}. Specifically, we measured the power of LEMAD to detect different sets of biogeographic assumptions. To this end, we simulated datasets under $DIVA_{events}$ and fitted $LEMAD_{diva\_events}$ and $LEMAD_{dec\_events}$ models and compared their likelihoods. It is expected that $LEMAD_{diva\_events}$ model should have higher likelihood than $LEMAD_{dec\_events}$ because the generating model was indeed, a $DIVA_{events}$ process. We counted the number of simulated datasets where this was the case. We also conducted the complementary analysis: we simulated datasets under DEC_{events} to fit and compare $LEMAD_{diva\_events}$ and $LEMAD_{dec\_events}$ models (30 runs for each case).

2.5 | An empirical example

The geographic origin of the American avian family Trochilidae (Hummingbirds) is still debated (McGuire et al., 2014). Previous ARE analyses have supported South America as the most likely area where the common ancestor of hummingbirds lived (22 million years ago; McGuire et al., 2007, 2014). Interestingly, the fossil record points to Eurasia as the source (Louchart et al., 2008; Mayr, 2004) from which the first hummingbird lineage spread via the Bering Strait 34–28; therefore, early diverging hummingbird lineages are expected to be found in North America. However, this is not the case, which leaves a time gap of several million years. To determine whether LEMAD could provide insights on this, we reconstructed the geographic distribution of a widespread and representative hummingbird clade (*Amazilia* sensu lato and closely related species) using both LEMAD and EF models. The phylogenetic tree was taken from McGuire et al. (2014) in combination with geographic information from Ornelas et al. (2014). Extant species and extinct lineages could be present in three regions: (A) South America, (B) Mesoamerica and (C) North America (West from the Isthmus of Tehuantepec) or a combination of them. We did not include Eurasia as a possible region as (1) no living species are present, and (2) the artificial inclusion of a Eurasian branch into the phylogenetic reconstruction would bias the analysis and model the distribution of recent ancestors in Eurasia which disagrees with the fossil record. As no information exists on how in situ and vicariant speciation occur in *Amazilia* (see first paragraph of Methods), we could not assume either $DIVA_{events}$ or DEC_{events} so we ran $LEMAD_{dec\_events}$ and $LEMAD_{diva\_events}$ and compared the fit using AIC weights. Additionally, the models were combined with three different assumptions for rates of lineage extinction: one in which extinction is the same as the estimate for speciation rate (using a standard birth–death model: 0.15), one in which extinction is 10 times less frequent than speciation (0.015) and one in which extinction is 10 times more frequent (1.5). Notice that by fixing extinction to a certain rate, the rates of in situ and vicariant speciation will adjust accordingly during the likelihood optimization. Phylogenetic reconstructions often do not include all species in a group (due to a lack of DNA samples for instance); LEMAD features functionality where the number of missing extant species is taken into account during the calculation (the so-called sampling fraction in diversification models; Fitzjohn et al., 2009). We included this completeness information for the *Amazilia* dataset.

3 | RESULTS

3.1 | Increase in accuracy by modelling extinct branches

Our simulations indicate that the reconstruction of the biogeographic history of a clade is notably improved when the set of branches that potentially existed and went extinct is incorporated into the analysis. The extent of the improvement depends on the relative rates of in situ and vicariance speciation, extinction and dispersal/extirpation (range evolution). For instance, LEMAD is more accurate than extinction free approaches (EF) when vicariance is higher than in situ speciation and there are high rates of range evolution. We find no parameter combination where EF outperforms LEMAD.

Although we find that low rates of range evolution led to few differences between EF and LEMAD, data simulated under the biogeographic assumptions of DEC_{events} show that ancient nodes are better estimated by LEMAD when lineage extinction is higher than zero. Under $DIVA_{events}$ and low rates of range evolution, neither ancient nor recent nodes are better estimated with LEMAD (Figures 2 and 3).

The scenarios with high rates of range evolution show increased accuracy in ancestral range estimation (ARE) when using LEMAD than when using EF. Under $DIVA_{events}$, the improvement is limited to recent nodes but also ancient ones when rates of vicariance are higher than in situ speciation. Datasets with DEC_{events} show that LEMAD outperforms EF in recent nodes in all scenarios; ancient nodes are also better estimated except when in situ speciation is dominant.

Even though the differences between LEMAD and EF are more important as extinction rate increases, simulations with zero extinction also suggest a better performance of LEMAD over EF approaches in most cases. However, recent ancestors are correctly recovered by both approaches at similar numbers when simulations featured low rates of range evolution. Finally, we find that the LEMAD estimates for dispersal/extirpation, in situ and vicariant speciation are accurate across all parameter combinations and, importantly, the model can correctly detect statistical differences in their relative contributions (Figures S1–S3). In summary, we recommend using LEMAD when rates of vicariant speciation are equal or higher than rates of in situ speciation, and when range expansion and contraction are highly dynamic (Table 1).

3.2 | Ability to distinguish the signal of $DIVA_{events}$ and DEC_{events}

For the simulations where in situ subset was not assumed ($DIVA_{events}$; see methods), we fitted LEMAD model in two versions:

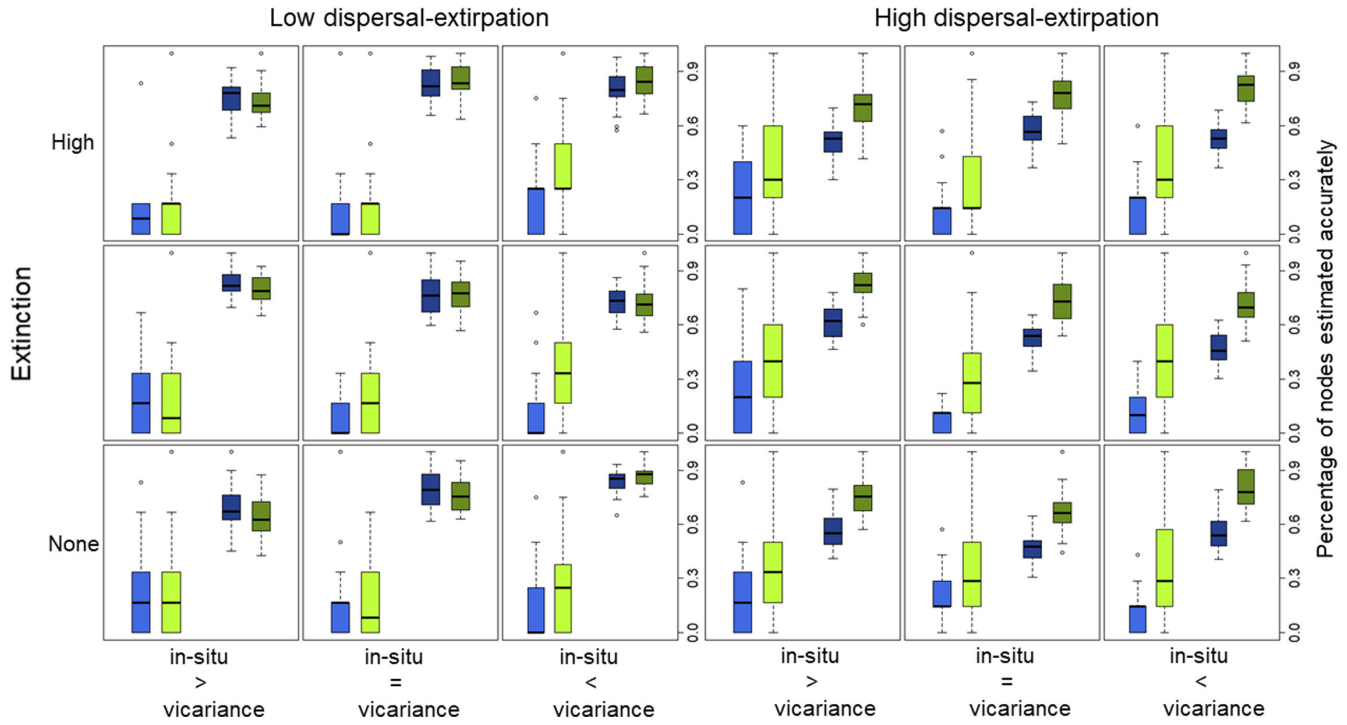

FIGURE 2 Accuracy in ancestral range estimation under DEC (in blue) and LEMAD (in green) models at recent (from half simulated time to present; dark shades) and ancient nodes (light shades). Eighteen scenarios were simulated with different rates of lineage extinction, range evolution (dispersal/extirpation) and relative contributions of in situ speciation and vicariance. For each panel, the *y*-axis shows the standardized number of ancestors whose distribution was correctly estimated by the models.

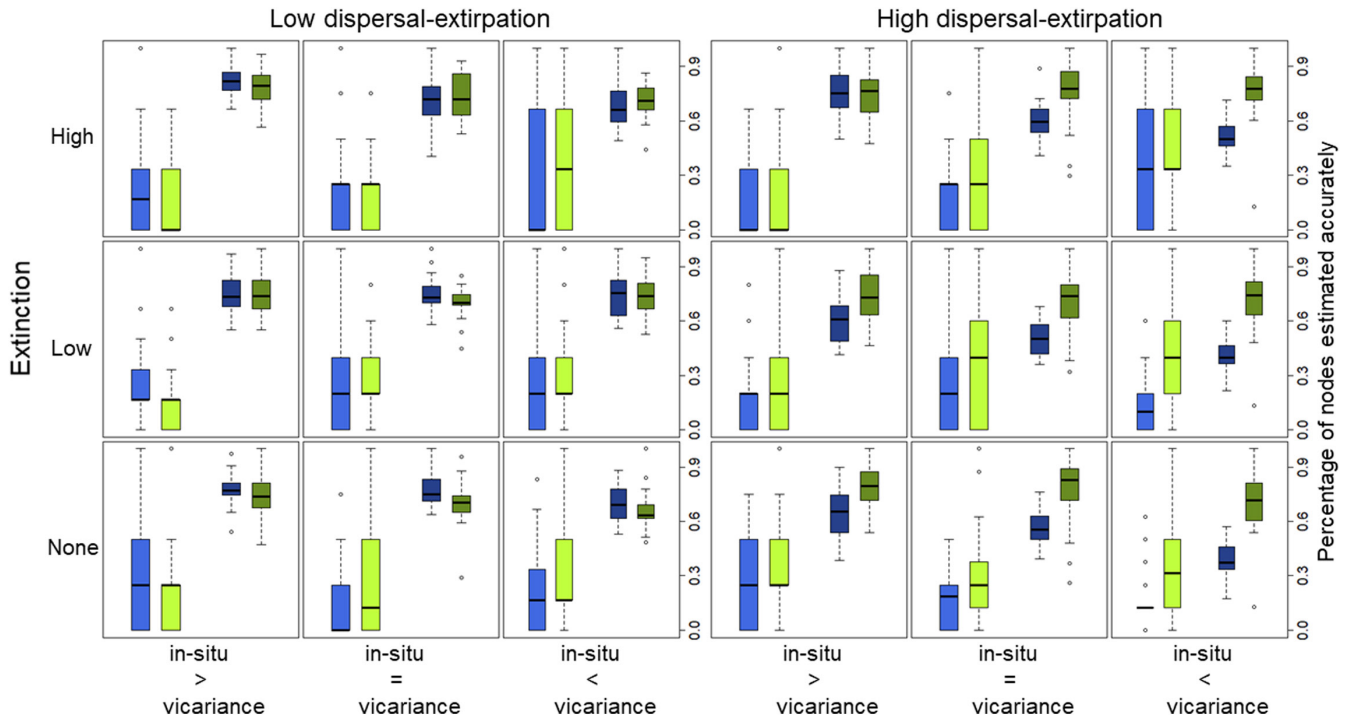

FIGURE 3 Accuracy in ancestral range estimation under DIVA (in blue) and LEMAD (in green) models at recent (from half simulated time to present; dark shades) and ancient nodes (light shades). Eighteen scenarios were simulated with different rates of lineage extinction, range evolution and relative contributions of in situ speciation and vicariance. For each panel, the *y*-axis shows the standardized number of ancestors whose distribution was correctly estimated by the models.

TABLE 1 Recommended scenarios to use LEMAD. Necessary condition is indicated with*

	Range evolution rate	Main speciation mode	Lineage extinction
When reconstructing nodes:			
Ancient	Any	Vicariance; equal contribution of both modes	Intermediate, High
Recent	High*	Any	Any

TABLE 2 Summary of LEMAD models fitted to 'Amazilia' hummingbird dataset under different assumptions on rates of extinction and modes of in situ and vicariant speciation

Biogeographic model	Assumed extinction	Log likelihood	Free parameters	AIC weights
$DIVA_{events}$	0.015	−289.49	3	0.81
$DIVA_{events}$	0.15	−291.10	3	0.16
DEC_{events}	0.015	−293.08	3	0.02
DEC_{events}	0.15	−295.54	3	<0.01
$DIVA_{events}$	1.5	−333.82	3	<0.01
DEC_{events}	1.5	−347.36	3	<0.01

$LEMAD_{diva\_events}$ and $LEMAD_{dec\_events}$. We find that in 86% of the simulated datasets, $LEMAD_{diva\_events}$ has the highest statistical support and for the remaining 14% of the simulations, $LEMAD_{dec\_events}$ was wrongly selected as the best model. It is important to note that in the datasets where LEMAD chose the right [generating] model (i.e. $DIVA_{events}$), the average difference in AIC weights is 0.43. In contrast, in simulations where the wrong model was selected, the difference was minimal (mean of AIC weights = 0.02). When DEC_{events} was the generating model (i.e. in situ subset, see methods), $LEMAD_{dec\_events}$ is correctly selected 76% of the times over $LEMAD_{diva\_events}$. A potential explanation on why LEMAD failed to select the correct model in some datasets is that in a three-area system like the one we are simulating, events of widespread vicariance are not possible which makes $DIVA_{events}$ and DEC_{events} less different from one another. Therefore, this analysis mainly explored the traces of the in situ subset assumption left in phylogenetic trees.

3.3 | Reconstruction of Amazilia biogeography

We find higher likelihood for LEMAD models with $DIVA_{events}$ than for LEMAD with DEC_{events} (difference in AIC weights = 0.95) which suggests that widespread species speciate by vicariance and not by in situ speciation. Within $LEMAD_{diva\_events}$, we find models with smaller extinction rate more likely; however, this result is not surprising as the estimated rate of extinction from a birth–death model was close to zero (Table 2). Instead of comparing across extinction rates and choosing the $DIVA_{events}$ model with the best AIC, we explore the parameter estimates and the reconstructed ancestral distributions for each model. Regardless of the assumed extinction rate, all reconstructions point to North America as the most likely region for the common ancestor of hummingbirds (Figure 4). In such a scenario, our simulation analysis finds that LEMAD is 50%–100% more effective than EF approaches in inferring the clade's common ancestor.

4 | DISCUSSION

We showed that ancestral range estimation can benefit from the -SSE framework by modelling lineage extinction, and that methods that neglect lineage extinction are less likely to accurately reconstruct true biogeographic histories of extant clades in a wide variety of scenarios. The parameterization of the model allows competing hypotheses for centres-of-origin and in situ versus vicariant speciation to be distinguished. With it, we found that North America is the most likely place of origin of *Amazilia* hummingbirds, which resolves a previous spatiotemporal disconnect between the hypothesized source region and the origin of first species divergence.

Empirical studies in island biogeography provide insights on how vicariance/in situ rates contribute to biodiversity patterns. Speciation after dispersal largely contributes to building species richness in small-sized islands and is responsible for the uniqueness of their assemblages (Losos & Schluter, 2000; Stuart et al., 2012). Archipelagos with small islands are expected to have high rates of vicariance, and therefore, LEMAD might be more appropriate for ancestral range estimation (ARE) than EF approaches. Nonetheless, in situ speciation becomes more frequent than vicariance as the size of the island increases which amounts to higher chances of geographic isolation and diversity of habitats (Kisel & Timothy, 2010); in fact, islands over a threshold size show evidence of rapid diversification (Algar & Losos, 2011; Losos & Schluter, 2000). Because the large geographic scale at which ARE is normally conducted (continents or large-sized islands), in situ speciation can be as frequent as vicariance. In this scenario and when DEC_{events} are assumed, the improvement provided by LEMAD is expected in recent and basal nodes. With $DIVA_{events}$, recent nodes are better estimated than EF methods whereas basal nodes show a non-significant improvement.

Similar to EF models, LEMAD assumes constant rates (extinction, vicariance and in situ speciation), which might not match empirical datasets in some cases. For instance, McGuire et al. (2014) report an important variation in richness across hummingbird subclades. This can be due to differential speciation (or extinction) rates among lineages (e.g. via diversity-dependent diversification; Etienne & Haegeman, 2012). McGuire et al. (2014) found that the difference in speciation rate between two subclades can be as large as 15-fold according to BAMM analysis. Heterogeneity in diversification rates which is independent from trait states or geographic distributions is likely to be ubiquitous across taxonomic groups besides hummingbirds and it is necessary to develop an ARE method that can handle this complexity. We argue that this should be the next methodological step forward. If the variation in speciation rates across lineages

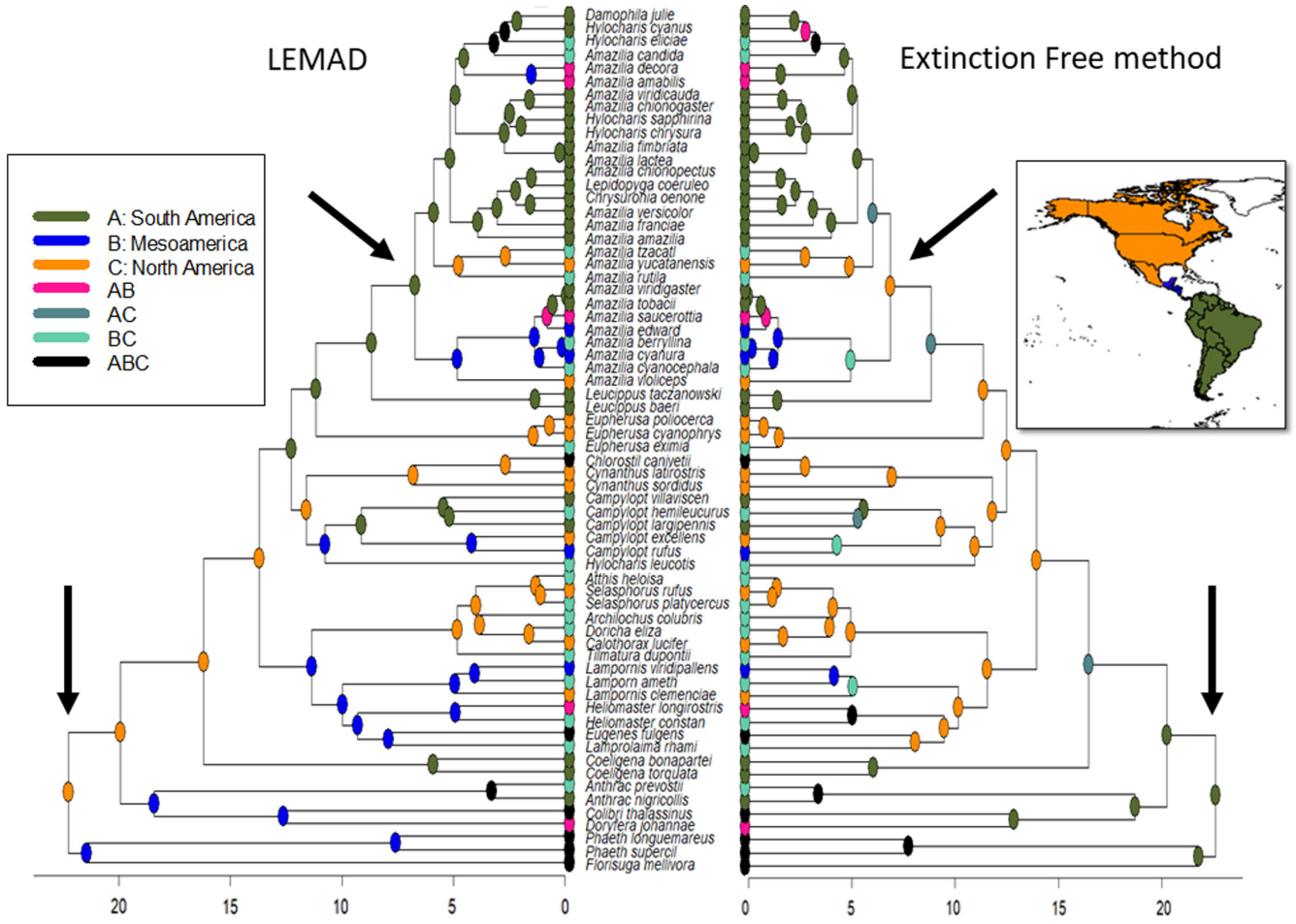

FIGURE 4 Estimated geographic distribution of 'Amazilia' hummingbirds' ancestors under two different approaches of state reconstruction. Extant and ancestral species (and extinct species in the case of LEMAD) could be present in (A) South America, (B) Mesoamerica and (C) North America (west from Tehuantepec) or a combination of them. Coloured circles show the most likely distribution. Arrows show some discrepancies between LEMAD and its extinction-free method counterpart on the ancestral range estimation of the entire hummingbird clade and 'Amazilia' group

is consistently a result of regional differences (i.e. lineages inhabiting a given area experience higher rates than in other regions), GeoSSE and GeoHiSSE (Caetano et al., 2018) are the proper tools to use. However, if more than three regions (or trait states) are to be analysed, SecSSE (Herrera-Alsina et al., 2019) can be used and with the right setup, it allows for character changes at cladogenetic events and not just along the branches extends (like in ClaSSE model; for an example with habitat preference, see Aduse-Poku et al., 2022). Any tool that would consider variable speciation rates across lineages should also incorporate variation in rates of expansion as the chances of vicariant events intrinsically depend on lineage dispersal (i.e. only multiregion lineages can undergo vicariance). Furthermore, opportunities for dispersal across regions can vary over time and assuming a single rate of range expansion/contraction might not be realistic in some cases (e.g. Buerki et al., 2011). Unlike DEC, the current implementation of LEMAD cannot handle this variation, but the framework could be adapted include it.

In previous studies, selecting DEC over DIVA was mostly based on the superior statistical properties (likelihood based) of DEC when compared to the parsimony method used in DIVA. After BioGeoBEARS was made available, researchers could confidently fit both methods to datasets and compare likelihoods, but surprisingly analysis is generally conducted with DEC, rather than DIVA, without justification. We recommend fitting both $LEMAD_{dec\_events}$ and $LEMAD_{diva\_events}$ to data and comparing likelihoods, instead of discarding either biogeographic model a priori. Our simulations show that LEMAD is capable of telling the two models apart, even though DEC_{events} are slightly less likely to be correctly detected than $DIVA_{events}$. We find that not only the relative contributions of in situ and vicariant speciation, lineage extinction and range evolution directly influence the precision of the ancestral reconstruction, but the set of biogeographic assumptions is also of paramount importance. For instance, unlike $DIVA_{events}$, DEC_{events} attribute some speciation events as in situ subset instead of vicariance followed by dispersal (Ree et al., 2005). This is reflected in the estimates for both processes in our analysis: even if the contributions of in situ and vicariant speciation are the same, we found high variability in vicariance estimates (higher than in situ events) when DEC_{events} underlie

simulations. Similarly, when using $DIVA_{events}$, the estimates for in situ speciation are likely to be more variable than those for vicariance. In both cases, high rates of lineage extinction increase the variability of rate estimates.

High rates of dispersal/extirpation have two main consequences on these biogeographic analyses. First, the impact of ignoring extinct branches in accurate ARE is higher than in the presence of low rates of range evolution. LEMAD is more likely to correctly track the change in geographic distribution of ancestors along the branches of a phylogenetic tree than EF methods, even with zero extinction. This could be due to how the likelihood at the root is handled by both approaches. In LEMAD, the probabilities of all the areas are multiplied by speciation rates whereas EF approaches do not consider speciation (Ree & Smith, 2008). This multiplication at the root (also called 'conditioning on extinction' because we are looking at a tree Nee et al., 1994) is used in all SSE diversification models. This may be responsible for its overall higher precision, which is magnified in systems with many range shifts. Second, with elevated rates of dispersal/extirpation, the uncertainty around speciation estimates is high. This is likely to occur because dispersal taking place right after in situ speciation (something expected with high rate of dispersal) looks similar to a vicariance event. In a similar way, an extirpation event following vicariance could be mistaken for in situ speciation. Importantly, although the estimates show important variation, the model can correctly detect statistical differences between rates of in situ and vicariant speciation.

LEMAD allows for the evaluation of contrasting models that make explicit assumptions regarding the rates of evolutionary events. Nonetheless, more meaningful hypotheses can be contrasted with fossils or other extinction estimates, which in turn would render a more accurate reconstruction of ancestral distributions (Mao et al., 2012). Alternatively, LEMAD can be modified to include extinct tree branches along with their last known distribution (Zhang et al., 2022; for an example of body size and extinct branches in a SSE implementation see Porto, 2022). The incorporation of known distributions of ancestors (i.e. constraining an internal node to have a certain distribution; see Meseguer et al., 2015) in LEMAD would be treated in a similar manner as the total likelihood is computed at the tree root, when giving different weights to the various regions. This feature, however, is not implemented yet. Dispersal could also be fixed to a specific rate; however, empirical evidence for rates of dispersal is challenging to collect. Unsurprisingly, the large geographical scale in ARE implies that regions are likely to be different from one another in both biotic and abiotic factors. Lineage dispersal in this context does not only entail the mobility to new localities but the successful arrival and further adaptation to potentially new conditions. It is likely that dispersal estimates from mark–release–recapture techniques (e.g. Hill et al., 1996), or other field-based measures would not be appropriate for ARE. One promising concept for testing with LEMAD is the taxon cycle, which posits that phases of range expansion and contraction occur along with habitat shifts (Ricklefs & Bermingham, 2002). The duration of these phases might offer a sensible starting point for developing hypotheses on rates of dispersal/extirpation. Finally, LEMAD enables the comparison of different assumptions on the distribution of the very first common ancestor to the entire clade, that is, the location of the centre-of-origin.

The biogeographic history for *Amazilia* hummingbirds reconstructed by LEMAD model showed clear differences with its EF counterpart. LEMAD found North America as the most likely region for the common ancestor of hummingbirds (Figure 4). This finding contrasts with previous studies where South America was found as the ancestral distribution. McGuire et al. (2014) proposed a northern arrival of hummingbirds to America with further expansion into South America immediately followed by extinction events that wiped out all hummingbird species from North America (recolonization of North America came at a later stage). However, their EF analysis could not prove this hypothesis. By considering extinction explicitly, our LEMAD analysis provides the missing piece of this puzzle, reconciling the South American distribution of the common ancestor of extant hummingbird species when ignoring extinction with North American distribution of the ancestor when extinction is considered.

5 | CONCLUSION

Lineage extinction seems less tangible than lineage formation; yet, we have shown that incorporating it into biogeographic models is crucial for a better reconstruction of ancestral areas. When using extinction-free methods, taxonomic groups can be inferred to have different centres of origin; however, this could be the result of dissimilar extinction rates across clades rather than actual differences in biogeographic histories. As a corollary, many taxa might have originated at the same place, we think that there are broad patterns which are yet to be discovered.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Ministry of Research and Technology, Indonesia, to Berry Juliandi (No. 2020/IT3.L1/PN/2021) and funding from NERC/NEWTON (NE/S006923/1 and NE/S006893/1) to Adam C. Algar, Greta Bocedi, Cecile Gubry-Rangin, Lesley Lancaster, Alexander S. T. Papadopulos and Justin M.J. Travis. The manuscript was improved by the constructive feedback from two anonymous reviewers. No permits were required to conduct this research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The R package lemad is available at https://github.com/leonel-halsina/lemad. And our code to reproduce simulations and reconstruct Amazilia's biogeographic history with LEMAD is available at https://doi.org/10.5061/dryad.34tmpg4p7. The DOI for LEMAD is: 10.5281/zenodo.7089334

ORCID

Leonel Herrera-Alsina https://orcid.org/0000-0003-0474-3592

Adam C. Algar https://orcid.org/0000-0001-8095-0097

REFERENCES

Aduse-Poku, K., van Bergen, E., Sáfián, S., Collins, S. C., Etienne, R. S., Herrera-Alsina, L., Brakefield, P. M., Brattström, O., Lohman, D. J., & Wahlberg, N. (2022). Miocene climate and habitat change drove diversification in Bicyclus, Africa's largest radiation of Satyrine butterflies. *Systematic Biology*, *71*(3), 570–588. https://doi.org/10.1093/sysbio/syab066

Algar, A. C., & Losos, J. B. (2011). Evolutionary assembly of Island faunas reverses the classic Island-mainland richness difference in Anolis lizards. *Journal of Biogeography*, *38*(6), 1125–1137. https://doi.org/10.1111/j.1365-2699.2010.02466.x

Barraclough, T. G., & Vogler, A. P. (2000). Detecting the geographical pattern of speciation from species-level phylogenies. *The American Naturalist*, *155*, 419–434.

Buerki, S., Forest, F., Alvarez, N., Nylander, J. A., Arrigo, N., & Sanmartín, I. (2011). An evaluation of new parsimony-based versus parametric inference methods in biogeography: A case study using the globally distributed plant family Sapindaceae. *Journal of Biogeography*, *38*(3), 531–550. https://doi.org/10.1111/j.1365-2699.2010.02432.x

Caetano, D. S., O'Meara, B. C., & Beaulieu, J. M. (2018). Hidden state models improve state-dependent diversification approaches, including biogeographical models. *Evolution*, *72*(11), 2308–2324. https://doi.org/10.1111/evo.13602

Crisp, M. D., Trewick, S. A., & Cook, L. G. (2011). Hypothesis testing in biogeography. *Trends in Ecology and Evolution*, *26*(2), 66–72. https://doi.org/10.1016/j.tree.2010.11.005

Ding, W.-N., Ree, R. H., Spicer, R. A., & Xing, Y.-W. (2020). Ancient orogenic and monsoon-driven assembly of the world's richest temperate alpine flora. *Science*, *369*, 578–581. https://www.science.org

Doob, J. L. (1945). Markoff Chains--Denumerable Case. *Transactions of the American Mathematical Society*, *58*(3), 455. https://doi.org/10.2307/1990339

Etienne, R. S., & Haegeman, B. (2012). A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence. *The American Naturalist*, *180*(4), E79–E85. https://doi.org/10.1086/667574

Fitzjohn, R. G. (2012). Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, *3*(6), 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 incompletely resolved phylogenies. *Systematic Biology*, *58*(6), 595–611. https://doi.org/10.1093/sysbio/syp067

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, *81*(25), 2340–2361. https://doi.org/10.1021/j100540a008

Goldberg, E. E., & Igić, B. (2012). Tempo and mode in plant breeding system evolution. *Evolution*, *66*(12), 3701–3709.

Goldberg, E. E., Lancaster, L. T., & Ree, R. H. (2011). Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Systematic Biology*, *60*(4), 451–465. https://doi.org/10.1093/sysbio/syr046

Herrera-Alsina, L., van Els, P., & Etienne, R. S. (2019). Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. *Systematic Biology*, *68*, 317–328. https://doi.org/10.1093/sysbio/syy057

Hill, J. K., Thomas, C. D., & Lewis, O. T. (1996). Effects of habitat patch size and isolation on dispersal by Hesperia comma butterflies: Implications for metapopulation structure. *The Journal of Animal Ecology*, *65*(6), 725. https://doi.org/10.2307/5671

Jablonski, D., & Sepkoski, J. J. (1996). Paleobiology, community ecology, and scales of ecological pattern. *Ecology*, *77*(5), 367–1378.

Kisel, Y., & Timothy, T. G. (2010). Speciation has a spatial scale that depends on levels of gene flow. *American Naturalist*, *175*(3), 316–334. https://doi.org/10.1086/650369

Lancaster, L. T., & Kay, K. M. (2013). Origin and diversification of the California flora: Re-examining classic hypotheses with molecular phylogenies. *Evolution*, *67*(4), 1041–1054. https://doi.org/10.1111/evo.12016

Losos, J. B., & Glor, R. E. (2003). Phylogenetic comparative methods and the geography of speciation. *Trends in Ecology and Evolution*, *18*(5), 220–227. https://doi.org/10.1016/S0169-5347(03)00037-5

Losos, J. B., & Schluter, D. (2000). Analysis of an evolutionary species-area relationship. *Nature*, *408*(6814), 847–850. https://doi.org/10.1038/35048558

Louchart, A., Tourment, N., Carrier, J., Roux, T., & Mourer-Chauviré, C. (2008). Hummingbird with modern feathering: An exceptionally well-preserved Oligocene fossil from southern France. *Naturwissenschaften*, *95*(2), 171–175. https://doi.org/10.1007/s00114-007-0309-0

Maddison, W. P., Midford, P. E., & Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, *56*(5), 701–710. https://doi.org/10.1080/10635150701607033

Mao, K., Milne, R. I., Zhang, L., Peng, Y., Liu, J., Thomas, P., Mill, R. R., & Renner, S. S. (2012). Distribution of living Cupressaceae reflects the breakup of Pangea. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(20), 7793–7798. https://doi.org/10.1073/pnas.1114319109

Matzke, N. J. (2013). Probabilistic historical biogeography: New models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. *Frontiers of Biogeography*, *5*(4), 242–248. https://doi.org/10.21425/F5FBG19694

Matzke, N. J. (2014). Model selection in historical biogeography reveals that founder-event speciation is a crucial process in Island clades. *Systematic Biology*, *63*(6), 951–970. https://doi.org/10.1093/sysbio/syu056

Mayr, G. (2004). Old World fossil record of modern-type hummingbirds. *Science*, *304*(5672), 861–864.

McGuire, J., Witt, C., Remsen, J., Corl, A., Rabosky, D., Altshuler, D., & Dudley, R. (2014). Molecular phylogenetics and the diverisification of hummingbirds. *Current Biology*, *24*, 910–196.

McGuire, J. A., Witt, C. C., Altshuler, D. L., & Remsen, J. V. (2007). Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. *Systematic Biology*, *56*(5), 837–856. https://doi.org/10.1080/10635150701656360

Meseguer, A. S., Antoine, P. O., Fouquet, A., Delsuc, F., & Condamine, F. L. (2020). The role of the neotropics as a source of world tetrapod biodiversity. *Global Ecology and Biogeography*, *29*(9), 1565–1578. https://doi.org/10.1111/geb.13141

Meseguer, A. S., Lobo, J. M., Ree, R., Beerling, D. J., & Sanmartín, I. (2015). Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: The case of hypericum (Hypericaceae). *Systematic Biology*, *64*(2), 215–232. https://doi.org/10.1093/sysbio/syu088

Nee, S., Holmes, E. C., May, R. M., Harvey, P. H., Harvey, P. H., Nee, S., Holmes, E. C., & May, R. M. (1994). Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions: Biological Sciences*, *344*(1307), 77–82.

Nguyen, L.-T. (2011). *An efficient algorithm for phylogeny reconstruction by maximum likelihood*. Technischen Universität Wien.

Ornelas, J. F., González, C., de los Monteros, A. E., Rodríguez-Gómez, F., & García-Feria, L. M. (2014). In and out of Mesoamerica: Temporal divergence of Amazilia hummingbirds pre-dates the orthodox account of the completion of the isthmus of Panama. *Journal of Biogeography*, *41*(1), 168–181. https://doi.org/10.1111/jbi.12184

Polly, P. D., & Sarwar, S. (2014). Extinction, extirpation, and exotics: Effects on the correlation between traits and environment at the continental level. *Annales Zoologici Fennici*, *51*(1–2), 209–226. https://doi.org/10.5735/086.051.0221

Porto, L. M. V. (2022). *Patterns of diversification and geographic distribution of Canidae over time*. PhD Thesis. University of Groningen, The Netherlands. https://doi.org/10.33612/diss.198160305

Ree, R. H., Moore, B. R., Webb, C. O., & Donoghue, M. J. (2005). A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution; International Journal of Organic Evolution*, *59*(11), 2299–2311. https://doi.org/10.1554/05-172.1

Ree, R. H., & Sanmartín, I. (2018). Conceptual and statistical problems with the DEC+J model of founder-event speciation and its comparison with DEC via model selection. *Journal of Biogeography*, *45*(4), 741–749. https://doi.org/10.1111/jbi.13173

Ree, R. H., & Smith, S. A. (2008). Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology*, *57*(1), 4–14. https://doi.org/10.1080/10635150701883881

Ricklefs, R. E., & Bermingham, E. (2002). The concept of the taxon cycle in biogeography. *Global Ecology and Biogeography*, *11*(5), 353–361. https://doi.org/10.1046/j.1466-822x.2002.00300.x

Ronquist, F. (1997). Dispersal-vicariance analysis: A new approach to the quantification of historical biogeography. *Systematic Biology*, *46*(1), 195–203. https://doi.org/10.1093/sysbio/46.1.195

Sanmartín, I., & Meseguer, A. S. (2016). Extinction in phylogenetics and biogeography: From timetrees to patterns of biotic assemblage. In *Frontiers in genetics* (Vol. 7, Issue 35). Frontiers Media S.A. https://doi.org/10.3389/fgene.2016.00035

Stuart, Y. E., Losos, J. B., & Algar, A. C. (2012). The Island-mainland species turnover relationship. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1744), 4071–4077. https://doi.org/10.1098/rspb.2012.0816

Zhang, Q., Ree, R. H., Salamin, N., Xing, Y., & Silvestro, D. (2022). Fossil-informed models reveal a Boreotropical origin and divergent evolutionary trajectories in the walnut family (Juglandaceae). *Systematic Biology*, *71*(1), 242–258. https://doi.org/10.1093/sysbio/syab030

BIOSKETCH

Leonel Herrera-Alsina focuses on how species diversity is spread across space and time. Rates of diversification and species coexistence are regulated by the distribution of standing diversity and geographic constraints which vary over time. To understand the interaction of these factors, he develops dynamic models of diversification which provide theoretical predictions or are applied to empirical datasets.

Author contributions: Conceptualization: LHA (lead), JMJT, ACA, LTL, ASTP, GB, CGR, OGO; funding acquisition: JMJT, BJ, ACA, LTL, ASTP, GB, CGR; data analysis: LHA; software preparation: LHA (lead), OGO (supporting); data curation: JFO; writing: LHA (lead), JMJT, ACA, LTL, ASTP, GB, OGO, CGR, JFO, BJ, PM (supporting), IMS (supporting), PL (supporting).

SUPPORTING INFORMATION

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

How to cite this article: Herrera-Alsina, L., Algar, A. C., Lancaster, L. T., Ornelas, J. F., Bocedi, G., Papadopulos, A. S. T., Gubry-Rangin, C., Osborne, O. G., Mynard, P., Sudiana, I. M., Lupiyaningdyah, P., Juliandi, B., & Travis, J. M. J. (2022). The missing link in biogeographic reconstruction: Accounting for lineage extinction rewrites history. *Journal of Biogeography*, *49*, 1941–1951. https://doi.org/10.1111/jbi.14489