Appendix from A. Skeels and M. Cardillo, "Reconstructing the Geography of Speciation from Contemporary Biodiversity Data" (Am. Nat., vol. 193, no. 2, p. 000)

Expanded Methods

Model Overview

We develop a DREaD model that simulates the diversification of a clade and the evolution of geographic ranges between speciation events. The model is spatially explicit and takes place against a background of a gridded, heterogeneous landscape, in which each grid cell contains a value for a single, continuously varying hypothetical environmental variable. Each species in the model is defined by three key attributes: geographic range, niche position, and niche breadth. The geographic range is defined by the species occupancy of grid cells, and species are able to occupy space and disperse through the landscape as they track cells that fit their niche (niche position ± niche breadth) through space and time. Our model builds on previous simulation studies of the evolution of geographic ranges (e.g., Rangel et al. 2007; Colwell and Rangel 2010; Qiao et al. 2016) but differs in that it explicitly models range movement in dynamic environmental space under different initial range overlap conditions at the point of speciation (geographic speciation modes). The R code to perform the simulation and an accompanying vignette can be found in a supplemental file[1] and is deposited in the Dryad Digital Repository: https://dx.doi.org/10.5061/dryad.d9j09bm (Skeels and Cardillo 2019), and the functions used to perform different parts of the simulation model are referenced throughout this appendix.

Simulating Landscape and Seed Species

Our simulation (DREaD function) begins by generating a background environment. Each simulation generates a new landscape with a single hypothetical environmental layer on a grid of 100 × 100 cells ($n = 10{,}000$), using unconditional Gaussian simulation in the gstat package in R (Pebesma 2004). The layer is spatially autocorrelated with kriging to represent a heterogeneous environmental layer with no defined direction in the spatial gradient, and values of the cells are scaled in a range of 0–20 for consistency (generateEnv function). The degree of spatial autocorrelation is held constant between simulations, but the environmental values of each grid cell differs for each simulation replicate.

The simulation proceeds by seeding an initial species (the ancestor to all species for that simulation replicate), which has an initial niche position, niche breadth, and occupied range (seedSpecies function). The initial species is seeded by randomly selecting a grid cell in the domain. A boundary is drawn around this cell by sampling the distance from the selected cell to each range boundary (north, south, east, and west) from a uniform distribution between 1 and 10 cells. The species niche position is given as the environmental value of the selected cell, and the species niche breadth is drawn from a uniform distribution between 1 and 10. All cells within the range boundaries that fall within this breadth define the initial species range. The simulation then progresses in discrete time steps.

Dispersal and Environmental Change

At each time step each species is able to expand its range via dispersal into new grid cells that lie within its inherent dispersal capacity (D, a clade-wide value shared by all species within each simulation replicate and drawn from a uniform distribution; table A1) and have environmental values that fall within the species niche (niche position ± niche breadth). If no cells within the dispersal distance contain suitable habitat, the species will not expand its range (rangeDispersal function).

Concurrently, the environment changes at each time step by changing the value of the environmental variable in each grid cell according one of two models of environmental change: (1) a cyclical model where environmental change is modeled as a sine wave with parameters for amplitude, ENV_A, and frequency, ENV_F (Rangel et al. 2007), and (2) a directional model, with environmental change modeled as a linear increase with a parameter for the slope, ENV_S. Each

[1] Code that appears in the *American Naturalist* is provided as a convenience to the readers. It has not necessarily been tested as part of the peer review.

environmental-change model is either spatially homogenous, where each grid cell changes the same amount at each time step, or spatially heterogeneous, where the degree of change in each cell is a linear function of its latitude, so that the magnitude of environmental change follows a latitudinal gradient. Whether a simulation uses a directional or a cyclical model of environmental change, as well as whether or not environmental change varies spatially, is defined by two binary parameters ($\mathrm{ENV}_{\mathrm{mode}}$ and $\mathrm{ENV}_{\mathrm{hetero}}$, respectively). In total, we model four different environmental-change scenarios: (1) cyclical homogenous, (2) cyclical heterogeneous, (3) directional homogenous, and (4) directional heterogeneous (environmentalChange function). In response to environmental change, at each time step a species will reposition its geographic range. If environmental change causes some occupied cells to contain values that are no longer within the species niche breadth, these cells are no longer part of the species range. This may lead to range contraction or range fragmentation, and at its extreme, if environmental change removes all cells with suitable habitat from the species range, then the species is considered extinct.

Event Selection

After dispersal and environmental change occur, for each species at each time step, one of six events can occur: vicariant, sympatric, parapatric, or dispersal speciation; extinction; or no event. Each speciation and extinction event is assigned a probability (described below), and the probability of no event is 1 minus the sum of these probabilities. At each time step, one event is sampled in proportion to these probabilities. In the case of no event, niche evolution proceeds. Probabilities for each event are defined in the following subsections.

Speciation

We explicitly model four different geographic modes of speciation: vicariant, sympatric, parapatric, and founder. Although the geographic mode of speciation is now considered to be a continuum from complete vicariant isolation to complete sympatric overlap (Fitzpatrick et al. 2009), here we treat the geography of speciation as taking place within discrete modes in order to simplify the process of speciation and more easily pull apart complex interactions. We believe that these speciation modes, though not exhaustive, cover a majority of the overlap conditions likely at the point of speciation. In DREaD, speciation is a stochastic process that is a peaked function of range size. The probability of speciation for each species at each time step is modeled by the function

$$\lambda * e^{-(r-B^2)/2C^2},$$

where r is the range size of the species, λ is the probability of speciation when $r = B$, and C defines the peak of the curve. The value of B was set to 5,000, which is the range size of a species that occupies half of the total domain, and λ was set so that the sum of speciation probabilities is equal to 0.0415 when a species range size is equal to B.

We simulated five different models, changing the predominant mode of speciation: one each where a given speciation mode (vicariant, founder, sympatric, or parapatric) is predominant and a mixed model where each of these four modes had equal probability of occurring. For a mixed model of speciation, λ was fixed at 0.010375 equally for each speciation mode. When a predominant mode was enforced, λ was set to 0.04 for the predominant mode, and for all other speciation modes λ was set to 0.0005. The per-lineage speciation rate, therefore, is dynamic and greater in intermediate-ranged species (fig. A1).

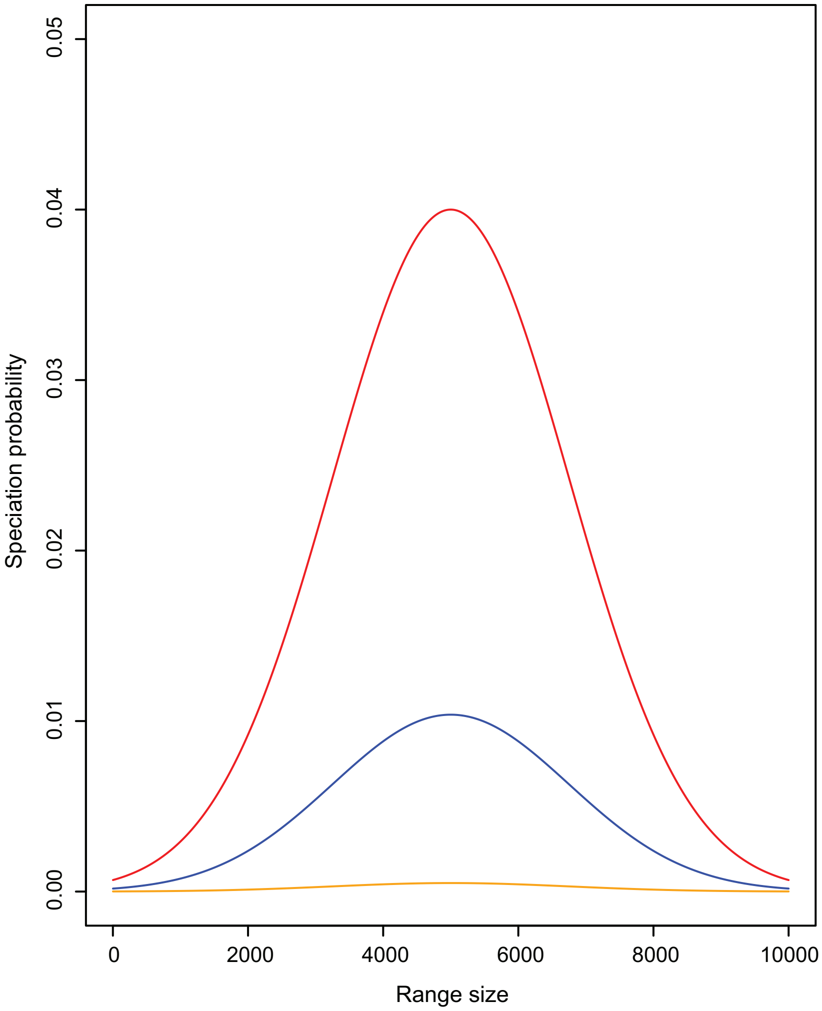

Figure A1: Relationship between speciation rate and range size. Speciation rate is a peaked function of range size. For a predominant speciation mode (red), the speciation rate is greatest when range size = 5,000 and the probability of speciation with a nonpredominant speciation mode is low (yellow). When the speciation mode model is mixed, all speciation modes have equal probability (blue), which sum to 0.0415 when range size = 5,000.

Each of the four modes is modeled as follows.

Sympatric. Under the sympatric model, one daughter species maintains the range of the parent and the other occupies a range that lies completely within the boundaries of the parent species range. This is chosen by randomly drawing four coordinates from within the parent species range, which form the boundaries of the daughter species range (speciateSympatric function).

Parapatric. Parapatric speciation occurs via budding at the range periphery. The new species is formed by creating an abutting range that may partially overlap the parent species range. This is done by selecting a cell within dispersal distance from the parent species range boundary and drawing four distances from a uniform distribution from this point to be the range boundaries of a new quadrant (speciateParapatric function).

Founder. Founder speciation follows a founder-event model, where founder events can found a new species in non-contiguous geographic space. Founder speciation proceeds by selecting a cell within the domain to be colonized with a probability inversely related to the shortest distance from the parent species range. The range boundaries are drawn by selecting 4 distance values from a uniform distribution (from 1 to D) from the colonized cell (speciateFounder function).

Vicariant. Vicariant speciation is modeled in two different ways, depending on the geometry of the species range. First, if the species range is a single contiguous area, vicariant speciation will occur via bisection of this range, whereby a line is drawn randomly through the species range, dividing it into two. The bisection is ambiguous with respect to the range size asymmetry of the daughter species ranges. The second method of vicariant division is used if a species range is fragmented. In this case, the parent species range is split so that each daughter species range is composed of a cluster of range fragments that are in closer spatial proximity to each other than to the sister species range. Clustering is performed with a k-means method on the X-Y coordinates of the range fragments (speciateVicariant function).

Extinction

Extinction is modeled in two ways in our simulation. First, there is deterministic extinction, which occurs if a species range collapses entirely. Second, there is a probabilistic rate of extinction that is a function of range size. Extinction probability is modeled by the function (following Rangel et al. 2007)

$$-\log \frac{r/m}{(1/\mu)^2},$$

where r is the species range size, m is the extinction range size threshold, and μ adjusts the magnitude of the function, which was fixed at 0.2. Here, extinction from stochastic processes is possible only in species with $r < m$, such that there is a point at which a species range is large enough that it can no longer become extinct from stochastic processes and is able to speciate (fig. A2).

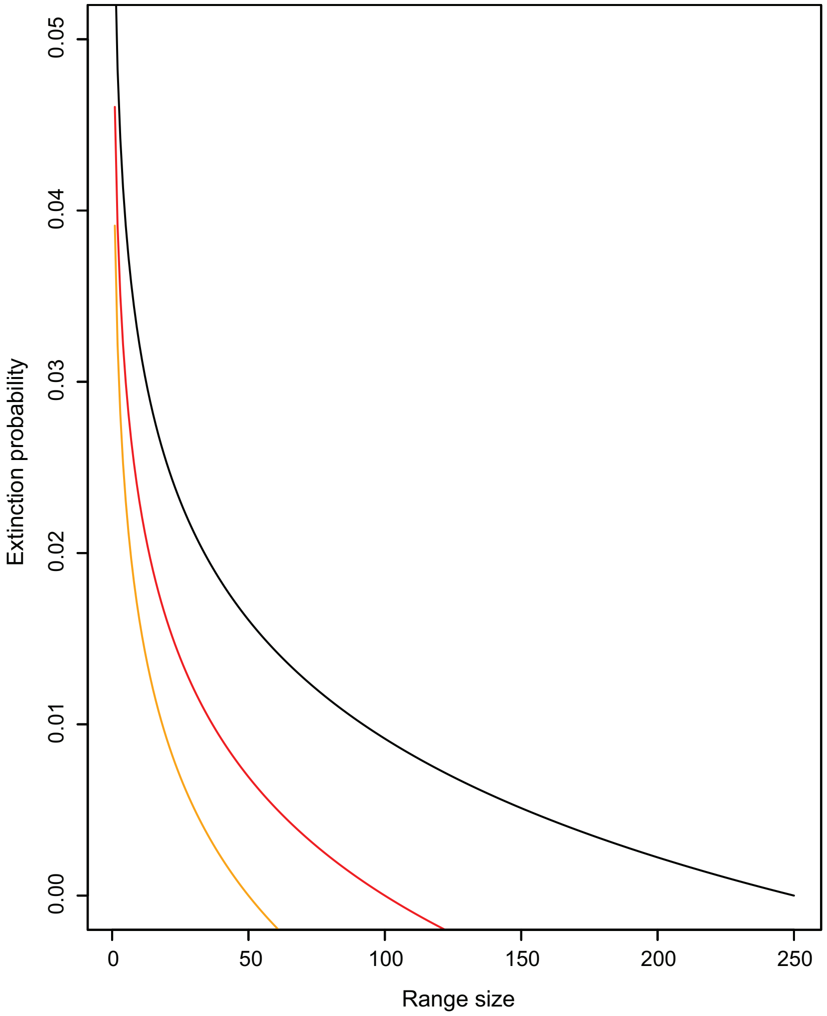

Figure A2: Relationship between range size and extinction rate. Extinction probability is highest in small-range species and becomes 0 when a species range size = m. The value of m was drawn from a uniform distribution between 50 and 250 in our simulation study. This figure shows the relationship when m = 50 (orange), m = 100 (red), or m = 250 (black).

Niche Evolution and Phylogenetic Signal in the Niche

We model phylogenetic signal of the niche at speciation as follows. Under vicariant speciation, immediately following speciation both daughter species niche positions are recentered toward the mean environmental value of the species new range. Under founder, parapatric, or sympatric speciation, the niche position of the budded daughter species (the smaller-ranged daughter species) is recentered. The degree to which the species shifts its inherited niche position value toward this new value is modeled with the parameter PS, the proportion of the step between the current niche position and the mean environmental value that the new species will take. A PS value of 1 means that the species will move completely toward the mean, while a value of 0 means that the new species will inherit the same niche position as the parent. In this way, the PS parameter controls the strength of "punctuational" evolution of the niche at speciation events (nicheRecenter function). We also model evolution of the niche along the branches of the phylogeny by allowing a species niche position and niche breadth to drift independently under a modified random-walk model of trait evolution (nicheEvolution function), controlled by the rate parameters NE_P (for niche position) and NE_B (for niche breadth). Species niche positions and breadths drift randomly under the single condition that the environmental value of at least one grid cell within the species range must remain within the species niche (niche position ± niche breadth), which ensures that species do not evolve an environmental niche that causes their immediate extinction (nicheEvolution).

Model Parameters

We simulated range evolution and speciation under different scenarios of environmental change, niche evolution, dispersal rate, clade size, and geographic speciation modes by sampling parameters from prior distributions. Priors were informed with preliminary simulations to qualitatively determine a match to broadly plausible evolutionary scenarios. For example, the rate of niche evolution was drawn from a uniform distribution between 0.005 and 2; these extreme values represent cases of strong environmental-niche conservatism and strong environmental-niche lability, respectively, that would be rare in many real clades (e.g., if the environmental variable is considered to represent temperature, then values of niche evolution of ≈2 could lead to shifts between tropical and temperate niche positions in relatively few time steps). Environmental-change parameters, on the other hand, were constrained by the mechanics of the simulation model, as high values of the slope of the directional model and the amplitude of the cyclical model led to repeated mass extinction, because species could not keep pace with the changing climate. This set an upper limit on the bounds of environmental-change parameters. Because of the computational cost of running the simulation model and the potentially inefficient method of sampling parameter space in a Monte Carlo simulation framework, we sampled parameter space with Sobol sequences (Burhenne et al. 2011), a type of quasi-random, low-discrepancy sequence that aims to prevent any one region of parameter space from being disproportionately over- or undersampled (table A1).

We ran the simulation 36,000 times, until a clade of size *n* was generated (table A1), discarding 269 simulations that could not be completed because the parameter combination led to total clade extinction more than five times successively. This led to roughly 7,200 replicates for each speciation scenario, which is considered adequate for model selection in an ABC framework (Pudlo et al. 2016). For each simulation we recorded the phylogenetic relationships of taxa, the polygons of each tip species range, the environmental grid, and a data frame containing information on the final niche position, niche breadth, range size, and mode of speciation for all tip species and internal nodes. Our simulations were written and performed in R, version 3.4.2, using the packages ape (Paradis et al. 2004), raster (Hijmans 2016), sp (Bivand et al. 2013), rgeos (Bivand and Rundel 2016), and fpc (Hennig 2015).

Table A1: Sampling range of each parameter in the simulation model

Parameter	Description	Sampling range
D	Distance of cells from occupied cells available during dispersal; dispersal kernel	1–10
NE_B	Niche breadth evolution rate	.0025–1
NE_P	Niche position evolution rate	.005–2
PS	Phylogenetic signal	.25–1
ENV_A	Amplitude of the environmental-change sine wave	.25–2
ENV_F	Frequency of the environmental-change sine wave	.25–2
ENV_S	Slope of linear environmental change	.001–.5
n	Simulated clade size	10–150
m	Range size at which the risk of stochastic extinction = 0	50–250
ENV_{hetero}	Binary parameter controlling whether environmental change is spatially heterogeneous	0–1
ENV_{mode}	Binary parameter controlling the model of environmental change	0–1
Speciation mode	Discrete parameter controlling whether speciation model is vicariant (1), founder (2), sympatric (3), parapatric (4), or mixed (5)	1–5

Note: A range of parameter values were sampled to explore parameter space and observe the effect on the patterns displayed by each clade. Dispersal capacity is modeled as a dispersal distance (*D*). Environmental change is modeled as both the frequency (ENV_F) and the amplitude (ENV_A) of a sine wave under the cyclical model of environmental change and as a slope (ENV_S) in the directional model of environmental change. The model of environmental change is set by a binary parameter (ENV_{mode}), and whether the model of environmental change is spatially homogenous or varies across a spatial gradient is defined by another binary parameter (ENV_{hetero}). Niche evolution is modeled as both niche position and niche breadth evolution rate (NE_P and NE_B, respectively), and phylogenetic signal in the niche is controlled by a single parameter (PS). The simulation proceeds until the phylogeny of the clade reaches a specified number of tips, with clade size (*n*). We modeled five different geographic speciation scenarios (speciation mode): one each where a particular mode was predominant (vicariant, sympatric, parapatric, and founder) and one mixed speciation scenario.

Data Analysis

At the completion of each simulation, we calculated 30 summary metrics that have been used, or might be used, to help reconstruct the predominant mode of speciation. These fall into five main categories: (1) range overlap metrics, (2) range asymmetry metrics, (3) range size and position metrics, (4) range isolation metrics, and (5) phylogenetic tree shape metrics. More details on each metric are provided in table A2. All summary statistics were generated in R, version 3.4.2. (generateSummaryStatistics function).

We compared the distributions of each summary statistic for simulated clades generated under different geographic speciation modes, using pairwise KSts. Our simulations generated large sample sizes (~7,200 replicates for each speciation scenario), which may return significant *P* values even when the effect sizes are small. To reduce the chance of misinterpreting significant differences, we randomly subsampled 500 simulation replicates 100 times and took the mean values of the test statistic and *P* value.

Next, we asked whether the signal of speciation is stronger than the signal of geographic-range evolution in present-day (simulated) data, by exploring which model parameters explained the greatest amount of variation in the summary statistics, independent of all other model parameters. To do this, we used a hierarchical partitioning protocol (Chevan and Sutherland 2017) that assesses all possible combinations of independent variables (model parameters) on the response (summary statistics) in a generalized linear modeling framework and partitions the variance according to a goodness-of-fit statistic (R^2), allowing for the independent assessment of each parameter's contribution while removing the effects of multicollinearity (MacNally 2000). Hierarchical partitioning was implemented in the R package hier.part (Walsh and MacNally 2013).

Table A2: The 30 summary metrics used for the study of the geographic mode of speciation with description and supporting references

Summary metric	Abbreviation	Description
Age-range correlation (ARC)	$\text{ARC}_{\text{slope}}$, $\text{ARC}_{\text{intercept}}$	Slope and intercept of regression between phylogenetic node age and geographic-range overlap (RO) among nodal descendants (Fitzpatrick and Turelli 2006)
Sister species RO × divergence times	RO_{slope}, $\text{RO}_{\text{intercept}}$	Slope and intercept of regression between sister species RO and divergence times; similar to ARC but uses only sister species
Mean sister species RO	RO_{mean}	RO = proportion of range of the smaller-ranged sister species found within that of the larger-ranged sister species
Sister species sympatry proportions	RO_{50}, RO_{75}, RO_{90}, RO_{100}	Proportion of species with RO ≥ 0, .5, .75, or .9 or RO = 1.0 (Cardillo and Warren 2016)
Sister species RO skew	RO_{skew}	Degree of bias in the distribution of sister species RO toward higher or lower values
Sister species RO kurtosis	RO_{kurt}	Degree of clustering or dispersion of sister species RO values
Difference between sister species RO and sister species–outgroup RO	TO_{mean}, TO_{SD}	Mean and SD of the difference in the RO between sister species and the RO of each sister species with an outgroup species; scaled between −1, where sister species ranges overlap completely with outgroup ranges and not at all with each other's, and 1, where sister species ranges completely overlap with each other and not at all with outgroup ranges (Cardillo 2015)
Bimodality of sister species RO	Bimod_{50}, Bimod_{75}, Bimod_{90}, Bimod_{100}	Degree to which sister species RO distributions are either unimodally sympatric, unimodally vicariant, or evenly distributed between vicariant and sympatric (Phillimore et al. 2008); sympatry ≥ .5, .75, or .9 or sympatry = 1
Mean sister species range size asymmetry	$\text{Asym}_{\text{mean}}$	Mean ratio of range sizes between sister species
Range asymmetry × divergence times	$\text{Asym}_{\text{slope}}$, $\text{Asym}_{\text{intercept}}$	Slope and intercept of regression between range asymmetry and divergence times (Grossenbacher et al. 2014)
Mean and SD of standardized range size	RS_{mean}, RS_{SD}	Standardized range size for each species = (range size)/(largest range size in clade); measured across all tip species in the phylogeny
Skew of range size distribution across all tips	RS_{skew}	Degree to which clades show bias in the distribution of range sizes toward higher or lower values (Pigot et al. 2010)
Mean standardized distance between sister species ranges	RD_{mean}	Standardized range distance = (minimum distance between sister species ranges)/(maximum distance between two species in the clade)
Range isolation × divergence times	RD_{slope}, $\text{RD}_{\text{intercept}}$	Slope and intercept of regression between standardized range distance and divergence times
β tree-splitting parameter	β	Phylogenetic tree imbalance = measure of the uneven distribution of species between clades descended from nodes across a phylogeny (Aldous 1996; Blum and François 2006; Pigot et al. 2010)
Colless's index	CI	Phylogenetic tree imbalance (Blum and François 2005)
Sackin's index	SI	Phylogenetic tree imbalance (Blum and François 2005)
γ	γ	Distribution of node heights through time, indicating degree to which phylogeny conforms to temporally constant diversification rates (Pybus and Harvey 2000)

Empirical Data Collection

We collected spatial and phylogenetic data for 30 empirical case studies (six plant, two fish, one invertebrate, four amphibian, four reptile, five mammal, and eight bird clades), selected to cover a range of taxonomic groups and levels, clade sizes, and geographic regions (table A3). We selected monophyletic clades on the basis of availability of well-sampled phylogenetic data, with all clades having >80% of known species included and most having >90%. Phylogenies were obtained, where possible, from the public databases TreeBase (www.treebase.org) and Data Dryad (datadryad.org), while one tree was obtained from an unpublished source—the South American *Liolaemus* (Esquerré et al., forthcoming). For bird and mammal clades, we used subsets of large composite supertrees constructed from multiple smaller phylogenies (Fritz 2009; Jetz et al. 2012). Phylogenies for two clades (*Amphiprion* fish and *Stenodactylus* lizards) were not time calibrated and were made ultrametric (with node heights scaled to a relative timescale) with the chronos function in the ape package in R. When phylogenies contained more than one representative from each species, these were pruned to species level by randomly selecting one subspecific lineage to represent the species.

Spatial data for some clades were obtained as spatial polygons from the IUCN (http://www.iucnredlist.org) or BirdLife (BirdLife 2016), which depict species range extents based on both occurrence data and expert assessment of species contemporary distributions. For other clades, we used point occurrence records from the GBIF (http://www.gbif.org) cleaned of obvious outliers, and in five cases we obtained spatial data supplied from the supplemental materials of the associated article presenting the phylogeny (*Pyrgus* butterflies: Pitteloud et al. 2017; *Mimulus* plants: Grossenbacher et al. 2014) or relating to the clade (*Banksia*, *Hakea*, and *Protea* plants: Skeels and Cardillo 2017). Spatial polygons were estimated from occurrence points with a fixed-*k* convex-hull method (following Getz and Wilmers 2004; Cardillo and Warren 2016). Finally, for consistency with our simulated data, all spatial polygons were transformed into a gridded raster format with a grid resolution of 10 arcsec to perform data analysis. From the spatial raster and phylogenetic data we obtained the same summary statistics as from the simulated data.

Table A3: Details of clades used in empirical analysis

Higher taxon, clade	Total species	Species in phylogeny	Species with occurrence records	Species with polygons	Reference
Plant:					
Banksia	170	157	157	0	Cardillo and Pratt 2013
Hakea	149	136	136	0	Cardillo et al. 2017
Protea	110	90	85	0	Valente et al. 2009
Mimulus	120	114	90	0	Grossenbacher et al. 2014
Sidalcea	25	24	24	0	Sabath et al. 2016
Bursera	100	85	83	0	Sabath et al. 2016
Fish:					
Sebastes	110	95	92	0	Ingram and Kai 2014
Amphiprion	30	27	17	10	Litsios et al. 2012
Invertebrate:					
Pyrgus	37	36	35	0	Pitteloud et al. 2017
Amphibian:					
Myobatrachidae	129	117	39	76	Vidal-García et al. 2014
Pseudacris	18	17	1	15	Pyron and Wiens 2013
Litoria	65	64	0	63	Rosauer et al. 2009
Plethodon	55	45	6	36	Wiens et al. 2006
Reptile:					
Anolis	119	100	81	17	Mahler et al. 2010
Pygopodidae	172	155	125	18	Brennan and Oliver 2017
Liolaemus	225	189	16	125	Esquerré et al., forthcoming
Stenodactylus	12	12	9	3	Metallinou et al. 2012
Mammal:					
Lemuridae	21	17	0	17	Fritz 2009
Bovidae	143	143	0	137	Fritz 2009

Table A3 (*Continued*)

Higher taxon, clade	Total species	Species in phylogeny	Species with occurrence records	Species with polygons	Reference
Diprotodontia	146	146	0	136	Fritz 2009
Geomyidae	39	39	0	39	Fritz 2009
Viverridae	33	33	0	33	Fritz 2009
Bird:					
Vidua	19	19	0	19	Jetz et al. 2012
Cuculidae	126	126	0	126	Jetz et al. 2012
Petroicidae	45	45	0	45	Jetz et al. 2012
Paradisaeidae	42	42	0	42	Jetz et al. 2012
Bucerotidae	51	51	0	51	Jetz et al. 2012
Rhinocryptidae	54	52	0	52	Jetz et al. 2012
Paridae	53	53	0	53	Jetz et al. 2012
Cacatuidae	21	21	0	21	Jetz et al. 2012

Note: Columns indicate the major taxonomic group each clade belongs to, the total number of described species, the number of species included in the phylogeny, the number of species with spatial-occurrence record data from the Global Biodiversity Information Facility (http://www.gbif.org), the number of species with spatial polygon data from the International Union for the Conservation of Nature (IUCN; http://www.iucnredlist.org) or Birdlife (2016), and the source of each phylogeny.

Inferring the Mode of Speciation in Empirical Data Sets Using Model Selection

For many complex biogeographic models, determining a likelihood function for the purpose of model selection becomes intractable. However, there are multiple likelihood-free model selection and model classification techniques available. Here we use both a machine-learning LDA and two ABC approaches, mnL and NN. Unlike traditional Bayesian or maximum likelihood inference, ABC avoids the calculation of a likelihood function by simulating data from prior distributions of the simulation parameters and summarizing these data with well-informed summary metrics (Csilléry et al. 2010; Blum et al. 2017). A simple ABC algorithm rejects or accepts sampled parameters on the basis of the distance between the simulated and observed summary statistics (Csilléry et al. 2010). However, to account for the discrepancy between accepted and observed summary metrics, local linear regression techniques or nonlinear neural-network machine learning methods can be used to better approximate the true posterior (Blum and François 2010; Csilléry et al. 2010). Model selection can then be performed by estimating the proportion of each model in the posterior.

To perform model selection, we first removed highly correlated variables (Pearson correlation coefficient; $r > 0.9$) and applied a variable selection procedure to reduce the number of summary statistics used in model selection, using the stepclass function in R package klaR (Weihs et al. 2005). We then tested the discriminatory ability of our candidate summary statistics to distinguish speciation modes for each method by using a leave-one-out cross-validation procedure, calculating the rate of model misspecification for each geographic speciation mode (the reclassification accuracy). After cross-validating each method's ability to reclassify the simulated data, given the summary statistics, we then inferred the geographic mode of speciation in the 30 empirical data sets, using LDA, implemented with the caret package in R (Kuhn 2016), and two ABC methods (mnL and NN), implemented with the abc package (Csilléry et al. 2012). To examine model adequacy, we plotted each empirical data set in two dimensions along the first two axes of an LDA on the simulated data set (fig. 6). A schematic of the workflow for simulation-based model selection can be found in figure 2.