Phylogenetic Inference for Biogeographic and Quantitative Trait Evolution

by

Michael Landis

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John P. Huelsenbeck, Chair
Professor Rasmus Nielsen
Associate Professor Ian H. Holmes

Fall 2015

Phylogenetic Inference for Biogeographic and Quantitative Trait Evolution

Abstract

Phylogenetic Inference for Biogeographic and Quantitative Trait Evolution

by

Michael Landis

Doctor of Philosophy in Integrative Biology

and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor John P. Huelsenbeck, Chair

A fundamental goal of evolutionary biology is to characterize the processes by which species traits evolve, and how those processes gave rise to the patterns of variation observed between species. Since interspecific variation often arises over millions of years, the tempo and mode of these processes cannot typically be observed directly or reproduced experimentally. Instead, they may be studied through a statistical framework called the phylogenetic comparative method. This dissertation focuses on phylogenetic models for two classes of traits: species geographical distributions (or biogeographic traits) and quantitative traits. The following contributions represent methodological advances that serve to render long-standing theoretical questions vulnerable to statistical analysis.

Chapter 1 develops an inference method to efficiently estimate historical biogeographic patterns using data augmented Markov chain Monte Carlo. This strategy increases the feasible number of areas per analysis from the tens to the thousands. Taking advantage of this increased resolution, the work introduces parameterizations for distance-dependent dispersal effects to greatly reduce model complexity. Analyzing Malesian *Vireya* (subgenus *Rhododendron*) biogeography, the method recovers Wallace's Line and Lydekker's Line as important geographical dispersal barriers, as well as ancestral range estimates for the clade.

Chapter 2 presents a technique called biogeographic dating that leverages paleogeographical information to estimate speciation events in absolute time. To achieve this, I construct a time-heterogeneous continuous-time Markov chain for the dispersal process, whose rate matrix takes values that are empirically informed by paleocontinental adjacencies. For biogeographic evolution, the time-heterogeneous process restores rate-time identifiability, thus enabling the estimation of absolute speciation times. Informed by the current paleogeographical literature, I construct an empirical dispersal graph using 25 areas and 26 epochs between the Cambrian (540 Ma) and the present (0 Ma). Applying biogeographic dating to

Testudines (crown turtles), I recover a root age concordant with fossil-based estimates ($\approx$ 205 Ma) to validate the efficacy of the method.

Chapter 3 introduces a class of models of continuous trait evolution that permit bursts of evolutionary change ("jumps"). Darwin's original conception of evolution proposed that species evolve gradually over time, which is typically modeled as a Brownian motion. However, many evolutionary mechanisms produce bursts in trait variation as punctuated change of large effect, such as rapid adaptation. I use Lévy processes to model these effects, which are a flexible class of stochastic processes that produce gradual and/or punctuational patterns of change. Applying a data augmented Bayesian method to primates, I show that body mass and endocranial volume measurements both bear the signature of evolution with jumps.

Contents

Acknowledgments

First of all, I thank my doctoral advisor, John Huelsenbeck, who has remained tremendously supportive throughout my studies. Of course, none of this work would have been possible without his sponsorship. Beyond that, he encouraged me to pursue my research interests whereever they led me. And when I inevitably found myself rather lost, he always made himself free to provide direction. In addition to John's mentorship, I thank Rasmus Nielsen, Ian Holmes, and Monty Slatkin for their guidance as advisors during the past several years. Their words helped me remain directed and motivated to complete my dissertation. I will always be indebted to Jeff Thorne, who sponsored me (then a stranger) for a fellowship with the National Evolutionary Synthesis Center, and gave me a home in his lab for several months. My thanks go to Deborah Dean at the Children's Hospital of Oakland for helping me take my first steps into the halls of science. Finally, I appreciate the support and encouragement I received from my unofficial mentors, who opened doors for me in my earlier years: Erick Matsen, Charles Marshall, Paul Fine, Nicolas Lartillot, and, above all, Brian Moore.

Apart from my mentors, I learned a great deal from my labmates and classmates. I shared the Huelsenbeck Lab with a cast of truly inspirational researchers, many of whom became good friends: Tracy Heath, Sebastian Höhna, Nick Matzke, Jeremy Brown, Chris Nasrallah, Dan Rabosky, and Bastien Boussau; Tracy and Sebastian deserve special recognition for always being free to share their advice, support, and insight. While there are too many Integrative Biology classmates to name in full, I would like to thank Lucy Chang, Darcy Kato Ernst, Sonal Singhal, Mason Liang, Ben Peter, Kelley Harris, Dave Armitage, and, especially, Josh Schraiber, Nick Matzke, and Melinda Yang, for making the community so memorable.

My friends and family outside UC Berkeley helped me stay grounded. First, I thank my friends who remained supportive, even when I would disappear into projects for months on end: Phil Roveto, Shawn Lewis, Elena Spitzer, Lee Bishop, Phil Morganelli, Casimir Blonski, and, the late and great, Andrew Bast. I am immensely grateful to my parents, Carol and David Landis, for raising me to value education and commitment; to my siblings, Sarah Gonzales and Matt Landis, for conditioning me to revel in rational argument; and to my nephews, Jacob and Lucas Gonzales, for reminding me that biology is amazing.

Last of all, Dharushana Muthulingam, my beloved wife, has been my bedrock throughout this entire endeavor. She has nurtured my first inklings to pursue research; she has endured all the late nights and lost weekends claimed in the name of this dissertation; but she has also been there to celebrate all my moments of elation. To Dee, I give my deepest heartfelt gratitude.

Chapter 1

Biogeography for large state spaces

1.1 Introduction

Historical biogeography—the study of the past geographic distribution of species and the processes that influence species distribution—remains a difficult problem in evolutionary biology. Inference of biogeographic history is made particularly challenging because of the many factors that influence species range, including various geological, climatic, ecological, and chance events. Both the diversity of factors influencing the geographic range of a species and the uncertainty regarding their relative importance motivates pursuit of biogeographic inference within a solid statistical framework. A statistical approach requires that the assumptions of an analysis be explicitly stated through the construction of probabilistic models that include parameters representing processes thought to impact the geographic distribution of species. This approach allows for the efficient estimation of model parameters and, perhaps more importantly, the rigorous comparison of alternative biogeographic models.

Over the past decade, several promising methods have been proposed that cast biogeographic inference in a statistical modeling framework. Lemmon and Lemmon (2008) and Lemey et al. (2009; 2010) proposed stochastic models that treat the distribution of species as continuous variables. A few years earlier, Ree et al. (2005) and Ree and Smith (2008) proposed stochastic models that treat the distribution of species as a discrete variable. For both approaches—those treating space as a continuous or a discrete variable—parameters are estimated using maximum likelihood or Bayesian inference.

The discrete-space model of Ree et al. (2005) is particularly intriguing because its basic statistical flexibility has the potential to profoundly change biogeographic inference, but is hampered by computational limitations. They modeled the colonization of and local extinction within a set of discrete areas as a continuous-time Markov process with a state space consisting of all possible geographic-range configurations. The machinery for computing the

likelihoods of discrete geographic ranges on phylogenetic trees is the same as that used to calculate the likelihood of discrete characters (*e.g.*, nucleotide sequences) on a tree; matrix exponentiation is used to calculate the probability of transitions among states/ranges along branches and the Felsenstein (1981) pruning algorithm (also see Gallager 1962) is used to account for different ancestral configurations at the interior nodes of the tree. Together, matrix exponentiation and the Felsenstein pruning algorithm allow the likelihood to account for all possible histories of area colonization and local extinction that could have given rise to the observed geographic distribution of species.

The conventional algorithms for calculating the likelihood, however, have practical limitations. Both matrix exponentiation and the pruning algorithm become computationally unmanageable when the number of areas becomes too large. Practically speaking, this means that inference under a discrete-space model, such as that proposed by Ree et al. (2005), is limited to about ten areas. With ten areas, there are a total of $2^{10} - 1 = 1023$ possible states (geographic ranges) and the rate matrix of the continuous-time Markov model is 1023×1023 in dimension. A recent implementation of the Ree et al. (2005) method allows up to 20 areas to be considered, but at the expense of making some restrictive assumptions about the number of areas that can be occupied concurrently per species (Webb and Ree 2012). The usual method for working around the limitations of the Ree et al. (2005) approach is to group areas together in such a way that the biologist considers no more than about ten areas. This solution, unfortunately, comes at a cost: hard earned species-distribution data are lumped, limiting the spatial resolution of the inferred biogeographic history; the inference of parameters suffers because fewer data are available for estimation; and the complexity of the models that can be distinguished is limited by the small number of areas that can be considered.

In this paper, we describe a computational method—referred to as 'data augmentation'—that allows the approach proposed by Ree et al. (2005) to be extended to hundreds or thousands of areas. The approach is inspired by the method described by Robinson et al. (2003) for the analysis of amino acid sequence data under complex models of non-independence, which relies on Markov chain Monte Carlo (MCMC; Metropolis et al. 1953; Hastings 1970) to carry out the tasks normally accomplished by means of matrix exponentiation and the Felsenstein pruning algorithm. The biogeographic model described by Ree et al. (2005) explicitly considers various scenarios by which ancestral ranges may become subdivided during speciation and inherited by daughter species. By contrast, the two biogeographic models that we describe here both assume that ancestral ranges are inherited identically: the first is a simple (null) model in which every area has an equal rate of colonization or extinction and a second model in which rates of colonization are distance dependent. We develop this approach in a Bayesian statistical framework in which model parameters are estimated using MCMC and candidate biogeographic models are compared using Bayes factors. We explore the statistical behavior of this approach by means of simulation, and demonstrate its

empirical application with an analysis of Malesian species within the flowering-plant clade, *Rhododendron* section *Vireya*.

1.2 Methods

Statistical Inference of Biogeographic History

We are interested in modeling the biogeographic distribution of M extant taxa over a geographic space that has been discretized into N areas, where each taxon occurs in at least a single area. The evolutionary relationships among the M taxa are described by a rooted, time-calibrated phylogenetic tree that in this paper is considered to be known without error. We label the tips of this tree to correspond to the observed species, $1, 2, \ldots, M$; the interior nodes of the tree are labeled in postorder sequence $M+1, M+2, \ldots, 2M$ (Figure 1.1). The ancestor of node i is denoted $\sigma(i)$. The most recent common ancestor of the M observed species (the 'root' node) is labeled $2M-1$. We also consider both the branch subtending the root node (the 'stem' branch) and its immediate ancestor (the 'stem' node), which is labeled $2M$. The times of the speciation events (nodes) on the tree are designated $t_1, t_2, \ldots, t_{2M}$. Typically, the species at the tips are contemporaneous and extant, such that $t_1 = t_2 = \ldots = t_M = 0$. The temporal duration of the branch below node i, typically in terms of millions of years, can be calculated as $T_i = t_{\sigma(i)} - t_i$.

Our use of 'geographic range' refers to the pattern of presence and absence of a lineage within the set of discrete geographic areas. For the models we will explore, all geographic ranges in which at least one area is occupied are admissible (*i.e.*, the case in which all areas are unoccupied is precluded). The occurrence of the ith species in the jth area is denoted $x_{i,j}$, where $x_{i,j}$ is equal to 0 or 1. Although we model geographic ranges as bit vectors, we represent them using bit strings (*i.e.*, a sequence of zeros and ones) to simplify our notation. For example, the bit string 101 corresponds to a geographic range for a species that is present in areas 1 and 3 and absent in area 2. The biogeographic state space, $\mathcal{S}$, includes the $2^N - 1$ geographic ranges for a model with N discrete areas. For example, all allowable geographic ranges, $\mathcal{S}$, for a model with $N = 3$ areas are

$$\mathcal{S} = \{001, 010, 100, 011, 101, 110, 111\},$$

and the number of distinct configurations for this state space is $n(\mathcal{S}) = 2^3 - 1 = 7$. We designate the observed geographic range for the ith species as $\mathbf{x}_i = (\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,N})$, where $\mathbf{X}_{obs} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$, and designate ancestral geographic ranges at interior nodes of the tree as $\mathbf{x}_{M+1}$, $\mathbf{x}_{M+2}, \ldots, \mathbf{x}_{2M}$.

The 'states' (geographic ranges) that we observe at the tips of the tree were generated through a potentially complicated history of colonization and local extinction. Figure 1.1B–D

depicts examples of biogeographic histories. A 'biogeographic history' is a specific sequence of colonization and/or local extinction events that could have given rise to the observed geographic ranges. An event of range expansion or contraction is denoted $x_{i,j,k}$; each event occurs on a specific branch (leading to node i) and involves a single area (j) at a point in time (k, indicating the relative time of the k^{th} event on branch i, $\tau_k^{(i)} \in \tau^{(\mathbf{i})}$, which we describe in more detail below). The history of range expansion or local extinction on the branch with index i involving area j is denoted $\mathbf{x}_{i,j} = (x_{i,j,1}, x_{i,j,2}, \ldots, x_{i,j,F})$, where events along branch i are ordered such that $x_{i,j,1}$ is the oldest and $x_{i,j,F}$ is the most recent. The collection of histories over all branches of the tree is denoted $\mathbf{X}_{aug} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{2M-1})$, representing the data augmented biogeographic history. For example, there are 6, 6, and 12 biogeographic events for the histories shown in Figure 1.1B, C, and D, respectively.

The probability of a particular biogeographic history can be calculated in a straightforward manner by assuming that the events of colonization and local extinction occur according to a continuous-time Markov chain (Ree et al. 2005). A continuous-time Markov chain is fully described by a matrix containing the instantaneous rates of change between all pairs of states (geographic ranges, in this case). This instantaneous-rate matrix, $\mathbf{Q}$, has off-diagonal elements that are all greater than or equal to zero and negative diagonal elements that are specified such that each row of the matrix sums to zero. The elements of $\mathbf{Q}$ are parameterized by functions of $\boldsymbol{\theta}$, the parameter vector, according to some dispersal model, $\mathcal{M}$. The probability of a biogeographic history is obtained using the information on the position of colonization/extinction events on the tree and information from the instantaneous-rate matrix. Consider, for example, a case in which the process starts with a geographic range of 001 at one end of a branch, with a subsequent colonization of area one at time t_1 (*i.e.*, changes from 001 $\rightarrow$ 101), and then remains in the geographic range 101 until the end of the branch at time t_2. The probability of this history is

$$\underbrace{-q_{001,001}e^{-(-q_{001,001}\,t_1)}}_{\text{Waiting time for colonization}} \times \underbrace{-\frac{q_{001,101}}{q_{001,001}}}_{\text{Probability of colonization event}} \times \underbrace{e^{-(-q_{101,101}\,(t_2-t_1))}}_{\text{Probability of no further events}}$$

There are an infinite number of biogeographic histories that can explain the observed geographic ranges. When calculating the probability of the observed geographic ranges at the tips of the phylogenetic tree, it is unreasonable to condition on a specific history of biogeographic change. After all, the past history of biogeographic change is not observable. Instead, the usual approach is to marginalize over all possible histories of biogeographic change that could give rise to the observed geographic ranges. The standard way to do this is to assume that events of colonization or local extinction occur according to a continuous-time Markov chain (Ree et al. 2005). Marginalizing over histories of biogeographic change is accomplished using two procedures. First, exponentiation of the instantaneous-rate matrix,

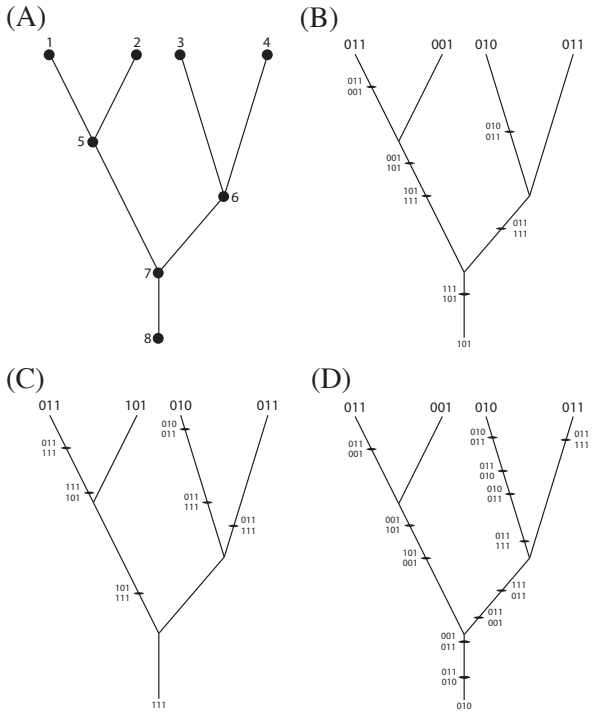


Figure 1.1: An example of a tree with $M = 4$ species. (A) Nodes on the tree are labeled such that the tips of the tree have the labels $1, 2, \ldots, M$ whereas the interior nodes of the tree are labeled $M + 1, M + 2, \ldots, 2M$. Note that in this paper we also consider the "stem" branch of the tree, which connects the root node (node 7) and its immediate common ancestor (node 8). (B–D) Several possible biogeographic histories—comprising 6, 6, and 12 events, respectively—that can explain the observed species ranges.

$\mathbf{Q}$, gives the probability density of all possible biogeographic changes along a branch

$$p(y \to z; t, \mathbf{Q}) = \left[e^{-\mathbf{Q}t}\right]_{yz},$$

where y is the ancestral geographic range, z is the current geographic range, and t is the duration of the branch on the tree. The geographic-range transition probabilities obtained in this way marginalize over all possible biogeographic histories along a single branch, but do not account for the possible combinations of geographic ranges that can occur at internal nodes of the phylogeny. The Felsenstein (1981) pruning algorithm is typically used to marginalize over the different combinations of 'states' (ancestral geographic ranges) at the interior nodes of the tree. Taken together, matrix exponentiation and the pruning algorithm comprise the

conventional approach for calculating the probability of observing the geographic ranges at the tips of the tree while accounting for all of the possible ways those observations could have been generated under the model.

The dimensions of the instantaneous-rate matrix, $\mathbf{Q}$, however, are $n(\mathcal{S}) \times n(\mathcal{S})$, where $n(\mathcal{S}) = 2^N - 1$, so the size of $\mathbf{Q}$ grows exponentially with respect to the number of geographic areas, N. Furthermore, computing the matrix exponential is of complexity $\mathcal{O}(n(\mathcal{S})^3)$ (Golub and Loan 1983). Thus, for values of $N \geq 20$, the number of computations required to exponentiate the rate matrix is quite large and computing the transition probabilities in this manner is intractable (Ree and Sanmartín 2009).

Statistical phylogenetic models encounter an analogous problem when modeling nucleotide evolution. As Felsenstein (1981) suggests, one might assume that each nucleotide site evolves under mutual independence to keep the state space small and amenable to matrix exponentiation. For biogeographic inference, however, the assumption of mutual independence would imply (implausibly) that the correlative effects between areas—such as geographic distance—are irrelevant to dispersal processes, which renders this assumption suitable only as a null model for testing the fitness of more plausible (*e.g.*, distance-dependent dispersal) biogeographic models.

Our primary motivation here is to remove the computational constraint that precludes the elaboration of more complex (and realistic) biogeographic models. As a result of this focus, we leave the rigorous comparison of inference across alternative models and methods as an open topic for future study.

A Distance-Dependent Biogeographic Model

The instantaneous-rate matrix, $\mathbf{Q}$, describes how the geographic range of a species can evolve through time. As with the formulation of Ree et al. (2005), we assume that in an instant of time only a single area can be gained or lost. In other words, each row of $\mathbf{Q}$ contains up to N positive, non-zero entries, which correspond to the rates at which any one of the N areas switches between absent and present (*i.e.*, the N $0 \rightarrow 1$ and $1 \rightarrow 0$ positive entries of the row). Additionally, each row contains a single element on the diagonal of the matrix, defined as $q_{i,i} = -\sum_{i \neq j} q_{i,j}$, which ensures that each row of $\mathbf{Q}$ sums to zero. The remaining entries in $\mathbf{Q}$ have a value of zero, as they entail an instantaneous change in geographic range involving two or more areas. This process corresponds to a dispersal-extinction (DE) model, which is somewhat simplified relative to the dispersal-extinction-cladogenesis (DEC) model (Ree et al. 2005), in that ancestral ranges are inherited identically. However, the current framework greatly expands the scope for the elaboration and inclusion of more diverse and realistic speciation scenarios.

We define a distance-dependent dispersal model, $\mathcal{M}_D$, where the rate of gaining a particular area ($0 \rightarrow 1$) depends on the relative proximity of available areas to those currently

occupied by a lineage. That is, the rate of colonizing a nearby area just outside the perimeter of the current geographic range should be greater than the rate of colonizing a relatively remote geographic area. The precise nature of the relationship between geographic distance and dispersal probability might be specified in numerous ways (see, *e.g.*, Wallace 1887; MacArthur and Wilson 1967; Hanski 1998). Our distance-dependent model specifies a simple relationship in which the probability of dispersal between two areas is inversely related to the geographic distance between them.

Let $q^{(a)}_{\mathbf{y},\mathbf{z}}$ be the rate of change from the geographic range $\mathbf{y}$ to the geographic range $\mathbf{z}$, where $\mathbf{y}$ and $\mathbf{z}$ differ only at the single area index a. Note that the rate function accepts any pair of bit vectors as arguments, allowing us to later assign configurations from $\mathbf{x}_{i,\bullet,k}$ to $\mathbf{y}$ and $\mathbf{z}$, $\mathbf{x}_{i,\bullet,k}$ being the geographic range of species i at time $\tau^{(i)}_k$. Also, let $\lambda_0 \in \boldsymbol{\theta}$ and $\lambda_1 \in \boldsymbol{\theta}$ be the respective rates at which an individual area is lost or gained within a geographic range, and $\eta(\mathbf{y}, \mathbf{z}, a, \beta)$ be a dispersal-rate modifier that accounts for correlative distance effects. We define the instantaneous dispersal rate as

$$q^{(a)}_{\mathbf{y},\mathbf{z}} = \begin{cases} \lambda_0 & \text{if } z_a = 0 \\ \lambda_1 \eta(\mathbf{y}, \mathbf{z}, a, \beta) & \text{if } z_a = 1 \\ 0 & \text{if } \mathbf{y} \text{ and } \mathbf{z} \text{ differ at more than one area} \\ 0 & \text{if } \mathbf{y} = 00\ldots0 \end{cases} \tag{1.1}$$

and the distance-dependent dispersal rate modifier as

$$\begin{aligned} \eta(\mathbf{y}, \mathbf{z}, a, \beta) = &\left(\sum_{n=1}^{N} \mathbb{1}_{\{y_n=1\}} d(G_n, G_a)^{-\beta} \right) \\ &\times \left(\frac{\sum_{m=1}^{N} \mathbb{1}_{\{z_m=0\}}}{\sum_{m=1}^{N} \mathbb{1}_{\{z_m=0\}} \left(\sum_{n=1}^{N} \mathbb{1}_{\{y_n=1\}} d(G_n, G_m)^{-\beta} \right)} \right) \end{aligned} \tag{1.2}$$

where we define $\mathbb{1}_{\{x=y\}}$ as the indicator function that equals one when both arguments are equal and zero otherwise, and $d(\cdot)$ as the Great Circle distance between two geographical coordinates on the surface of a sphere, known by

$$d(G_n, G_m) = 2r \sin^{-1} \left(\sqrt{\sin^2 \left(\frac{G_{m,\phi} - G_{n,\phi}}{2} \right) + \cos(G_{n,\phi}) \cos(G_{m,\phi}) \sin^2 \left(\frac{G_{m,\lambda} - G_{n,\lambda}}{2} \right)} \right),$$

where r is the radius of the sphere, and G_n is a vector with elements $G_{n,\phi}$ and $G_{n,\lambda}$ that correspond to the latitude and longitude of the the centroid of discrete area n. Here, we take a sphere with $r \approx 6.37 \times 10^6$ meters to approximate the size and shape of Earth.

Figure 1.2 will help develop intuition for how we model distance-dependent dispersal. In effect, the first term of $\eta(\cdot)$ computes the sum of inverse pairwise β-exponentiated geographic distances between the dispersal target, a, and all currently occupied areas of the geographic range. The second term normalizes the dispersal rate by the mean of all inverse pairwise geographic distances between all occupied-unoccupied area pairs. This normalization ensures that the sum of dispersal rates with or without the distance-dependence modifier are equal, which helps identify and interpret parameters λ_1 and β. If $\eta(\cdot) = 1$ or $\beta = 0$, then the rate of dispersal to area a equals the unmodified dispersal rate, λ_1. If $\beta > 0$, then the rate of dispersal to nearby areas is higher than that to more distant areas. Conversely, when $\beta < 0$, the rate of dispersal to more distant areas is higher than that to nearby areas. Finally, model $\mathcal{M}_D$ is equivalent to $\mathcal{M}_0$ when $\beta = 0$.

Note that the rate of gain depends on the distance-dependent correlation function $\eta(\cdot)$, but the rate of loss does not, so the distance-dependent dispersal model is not time reversible when $\beta \neq 0$. This fact has implications for evaluating the stationary frequency of geographic ranges at the root of the tree under this biogeographic model, which we detail below.

Sampling Biogeographic Histories

Our goal is to conduct inference under a dispersal model that captures the correlative effects of geographic distance between areas when N is large. For the computational reasons cited above, we cannot use matrix exponentiation to compute the likelihood under such a biogeographic model. Instead, we adapt a Bayesian data-augmentation approach that was introduced by Robinson et al. (2003) to model site-dependent protein evolution. Rather than analytically integrating over all possible biogeographic histories using matrix exponentiation, we numerically integrate over possible histories using data augmentation and Markov chain Monte Carlo.

We use the stochastic character-mapping algorithm described by Nielsen (2002) to sample biogeographic histories under the mutual-independence model, $\mathcal{M}_0$. This works by first sampling a set of geographic ranges for all internal nodes of the phylogeny and then sampling intermediate ranges over each of the branches connecting pairs of ancestor-descendant nodes. Upon completion, each branch is associated with a biogeographic history: the events comprising this history on each branch are ordered chronologically from past to present. Examples of such biogeographic histories are depicted in Figure 1.1B–D. We describe the process of sampling biogeographic histories in more detail below.

We first sample a set of geographic ranges for all M internal nodes from the joint posterior probability distribution of geographic-range configurations at the nodes. For tip nodes, we simply assign the observed species ranges. Next, we visit each individual branch in a preorder traversal (moving from the root to the tips) of the tree. For each branch, we simulate

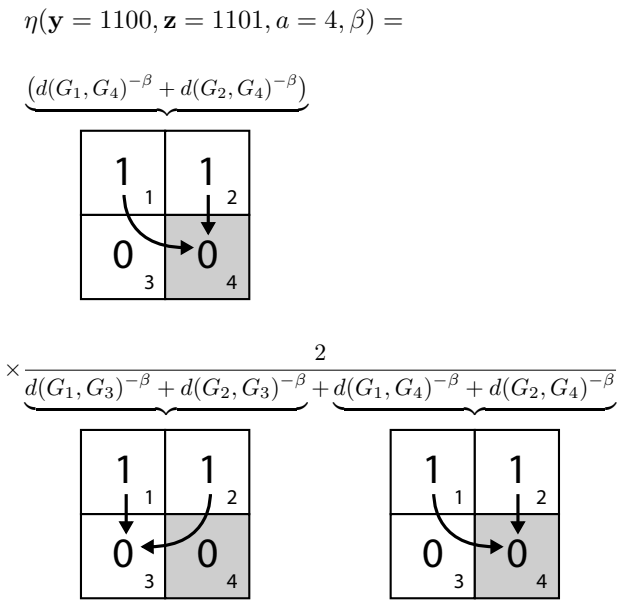


Figure 1.2: Cartoon of the computation of the distance-dependent dispersal-rate modifier, $\eta(\cdot)$. Here, we are interested in computing the rate of $\mathbf{y} = 1100$ transitioning to $\mathbf{z} = 1101$. The first term computes the sum of inverse distances raised to the power β between the area of interest (*i.e.*, 4) and all currently occupied areas (*i.e.*, areas 1 and 2). The second term then normalizes this quantity by dividing by the sum of inverse distances raised to the power β between all occupied-unoccupied area-pairs (*i.e.*, the denominator), then multiplying by number of currently unoccupied areas (*i.e.*, 2, the numerator).

a sequence of intermediate geographic ranges from the ancestral to the descendant node using rejection sampling; that is, the biogeographic history simulated along a branch must be consistent with the geographic ranges sampled/specified for the ancestor and descendant nodes of that branch. To do so, we first identify the initial geographic range at the the ancestral node, the final geographic range at the descendant node, and the duration of the branch separating these two nodes. We then sample a history of dispersal events for each area under the mutual-independence model, $\mathcal{M}_0$, under a simple instantaneous-rate matrix for a single area

$$\mathbf{Q}^* = \begin{pmatrix} -\lambda_0^* & \lambda_0^* \\ \lambda_1^* & -\lambda_1^* \end{pmatrix},$$

where λ_0^* and λ_1^* are the per-area rate of loss/local extinction ($1 \rightarrow 0$) and gain/colonization ($0 \rightarrow 1$), respectively. To iteratively sample the biogeographic history for each area, $j \in \{1, \ldots, N\}$, we initialize $\delta_0 = t_{\sigma(i)}$ and $k = 1$. Each iteration moves the process further along the branch by sampling a new event time δ from $\mathbf{Q}^*$, updating $\delta_0 = \delta_0 - \delta$, incrementing k, and inserting δ_0 into $\tau^{(i)}$ in sorted order as we go. Each event results in the state for area j changing to its complement (*i.e.*, $0 \rightarrow 1$, or $1 \rightarrow 0$), which we record in the branch history, $x_{i,j,k}$. We continue to sample dispersal events until the time of the next event is younger than the age of the end of branch, $\delta_0 < t_i$, whereupon we record the final event time as $\tau_F^{(i)} = t_i$. Since time is exponentially distributed, the probability that any two areas undergo dispersal events at precisely the same instant occurs with probability zero, which is consistent with the one-change-at-a-time assumption of the model.

When the biogeographic history for area j is sampled, we check to make sure it matches the geographic ranges sampled at the nodes. Inconsistent histories are rejected and resampled for each area. Additionally, we reject and resample events that induce the forbidden extinction configuration. For models in which the per-site (per-area) state space is large, rejection sampling path histories can be computationally inefficient (*c.f.*, Minin and Suchard 2007). This is not a concern in the present case, however, as the per-area state space is binary (*i.e.*, 1 or 0 for presence/absence of a species in an area), so we opt for the simpler algorithm.

We iterate this process of simulating branch-specific biogeographic histories for the remaining branches, which we visit in a pre-order sequence. This results in $\tau^{(i)}$ for each branch, an ordered vector of event times across all N areas, enabling us to compute the model likelihood given a sampled biogeographic history.

Computing the Likelihood of Biogeographic Histories

Since we can compute the rate at which any area is gained or lost given the current geographic range, we can compute the likelihood of a sampled biogeographic history by adopting a 'mechanistic' interpretation of the instantaneous-rate matrix, $\mathbf{Q}$. In general, waiting times between events in a continuous-time Markov process are exponentially distributed: when the process is in state i, the next event will occur with an exponentially distributed waiting time, where the rate of the exponential is equal to the overall rate of leaving state i: $q_{i,i} = -\sum_{j \neq i} q_{i,j}$. Moreover, the nature of the change at the next event is also specified by the instantaneous rate matrix: the relative probability that the next event entails a change from state i to state j is $p(i \rightarrow j) = \frac{q_{i,j}}{-\sum_{j \neq i} q_{i,j}}$. Accordingly, the probability that the next event entails a change from state i to state j at time t is simply equal to the probability of *any* event occurring at time t times the relative probability that the event is a change from i to j.

In the present case, we let $\mathbf{x}_{i,\bullet,k} = (x_{i,1,k}, x_{i,2,k}, \ldots, x_{i,N,k})$ be the state (sampled range) for lineage i at time $\tau_k^{(i)}$. Then, the probability that the next event is a the state change $\mathbf{y} \to \mathbf{z}$ at time t is the product of probability of the next sampled event occurring first among all possible events and the probability of any event occurring at time t, given as

$$\begin{aligned} p\left(\mathbf{y} \to \mathbf{z}; t, \boldsymbol{\theta}, \mathcal{M}\right) &= -\frac{q_{\mathbf{y},\mathbf{z}}}{q_{\mathbf{y},\mathbf{y}}} \left(-q_{\mathbf{y},\mathbf{y}}\right) e^{-(-q_{\mathbf{y},\mathbf{y}})t} \\ &= q_{\mathbf{y},\mathbf{z}} e^{q_{\mathbf{y},\mathbf{y}}t}, \end{aligned} \tag{1.3}$$

and the probability that no event occurs in time t is given as

$$p\left(\mathbf{y} \to \mathbf{y}; t, \boldsymbol{\theta}, \mathcal{M}\right) = e^{q_{\mathbf{y},\mathbf{y}}t}. \tag{1.4}$$

Note that the distance-dependent dispersal model defined in (1.1) depends on the superscript, (a), which indicates the single area that differs between ranges $\mathbf{y}$ and $\mathbf{z}$. Here, we suppress the superscript in the interest of simplifying the notation. Changes between between ranges that differ by more than one area have a transition rate of zero (they are prohibited under the one-change-at-a-time model), so this summation requires only N computations.

The likelihood of the biogeographic history over all branches of the phylogeny is then simply calculated as the product of all stepwise likelihoods (Figure 1.3),

$$\begin{aligned} L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}, \mathcal{M}) = \Bigg(\prod_i \Bigg(& \prod_{k=2}^{F_i - 1} \underbrace{p\left(\mathbf{x}_{i,\bullet,k-1} \to \mathbf{x}_{i,\bullet,k}; \Delta\tau_k^{(i)}, \boldsymbol{\theta}, \mathcal{M}\right)}_{\text{stepwise changes}} \Bigg) \\ & \times \underbrace{p\left(\mathbf{x}_{i,\bullet,F_i} \to \mathbf{x}_{i,\bullet,F_i}; \Delta\tau_{F_i}^{(i)}, \boldsymbol{\theta}, \mathcal{M}\right)}_{\text{no change}} \Bigg), \end{aligned}$$

where $F_i = n\left(\tau^{(i)}\right)$ is the number of events on branch i, $\Delta\tau_k^{(i)} = \left(\tau_{k-1}^{(i)} - \tau_k^{(i)}\right)$ is the temporal interval between events, and $\mathbf{X}_{obs}$ are the ranges observed at the tips.

Markov Chain Monte Carlo

We can compute the posterior probability of a single sampled biogeographic history as

$$p(\boldsymbol{\theta}, \mathbf{X}_{aug} \,|\, \mathbf{X}_{obs}, \mathcal{M}_D) \propto L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}, \mathcal{M}_D) p(\boldsymbol{\theta}).$$

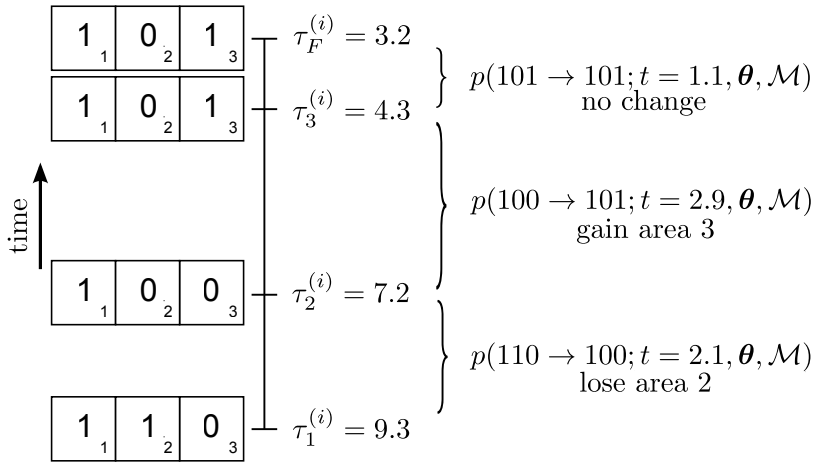


Figure 1.3: Cartoon of the likelihood terms. The biogeographic history for lineage i includes the lineage start at time $\tau_1^{(i)}$, an extinction event at area 2 at time $\tau_2^{(i)}$, a dispersal event into area 3 at time $\tau_3^{(i)}$, and the lineage end at time $\tau_F^{(i)}$, with all events laying within the time interval $(3.2, 9.3)$. The probability of a sampled geographic range at the start of the branch is conditioned on the previous (ancestral) geographic range and the time separating the geographic ranges, $\Delta\tau_k^{(i)} = \tau_{k-1}^{(i)} - \tau_k^{(i)}$. The likelihood is the product of the probabilities corresponding to each interval accounting for an area loss at time $\tau_2^{(i)}$, an area gain at time $\tau_3^{(i)}$, and no further changes occurring before the lineage terminates.

We approximate the joint posterior probability density of the biogeographic model parameters numerically using a Markov chain Monte Carlo (MCMC) algorithm. The general idea is to construct a Markov chain with a state space comprising the possible values for the model parameters and a stationary probability distribution that is the target distribution of interest (*i.e.*, the joint posterior probability distribution of the model parameters). Draws from the Markov chain at stationarity are valid, albeit dependent, samples from the posterior probability distribution of the biogeographic parameters (Tierney 1994). Accordingly, parameter estimates are based on the frequency of samples drawn from the stationary Markov chain.

By repeatedly sampling dispersal histories via MCMC, we numerically integrate over H,

$$p(\boldsymbol{\theta} \,|\, \mathbf{X}_{obs}, \mathcal{M}_D) \propto \int_{\mathbf{X}_{aug}} p(\boldsymbol{\theta}, \mathbf{X}_{aug} \,|\, \mathbf{X}_{obs}, \mathcal{M}_D).$$

To generate samples from this posterior, we rely on the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). Below, we describe our MCMC proposals for an audience whom we assume has some familiarity with MCMC.

Proposing parameters

Our method has two pairs of parameters for each area governing the rate at which it is added to or removed from the current biogeographic range: λ_0 and λ_1, which are used when computing the likelihood under the distance-dependent model; and λ_0^* and λ_1^*, which are used to sample biogeographic histories under the simpler mutual-independence model. All four rates must take on values greater than 0 and are distributed by half-Cauchy(0, 1) priors. We propose changes to the dispersal-rate parameters by first randomly selecting one of the four rates (uniformly with $P = 0.25$), then propose a new value for the selected rate parameter, $x' = xe^{\psi(u-0.5)}$, where x is the current dispersal rate, x' is the proposed dispersal rate, ψ is a tuning parameter, and $u \sim \text{Uniform}(0, 1)$. The probability of accepting a proposed change to the dispersal-rate parameters, λ_0 and λ_1, under the distance-dependent model, $\mathcal{M}_D$, is calculated using the Metropolis-Hastings ratio

$$R = min\left\{1, \frac{L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}', \mathcal{M}_D)}{L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}, \mathcal{M}_D)} \times \frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta})} \times \frac{\lambda'}{\lambda}\right\},$$

where first term is the ratio of the likelihoods of the proposed and current states, the second term is the ratio of the prior probabilities of the proposed and current states, and the final term is the simplified Hastings ratio that describes the ratio of the proposal probabilities for the proposed and current states.

To improve acceptance rates for proposed dispersal histories under the mutual-independence model, $\mathcal{M}_0$, we infer $(\lambda_0^*, \lambda_1^*) \in \boldsymbol{\theta}^*$ by conditioning the likelihood on $\mathcal{M}_0$ instead of $\mathcal{M}_D$, yielding the Metropolis-Hastings ratio

$$R = min\left\{1, \frac{L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}^{*\prime}, \mathcal{M}_0)}{L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}^*, \mathcal{M}_0)} \times \frac{p(\boldsymbol{\theta}^{*\prime})}{p(\boldsymbol{\theta}^*)} \times \frac{\lambda^{*\prime}}{\lambda^*}\right\}.$$

We specify a Cauchy(0, 1) prior for the distance-power parameter, β, and propose new values $\beta' = \mathcal{N}(\beta, \psi)$, where ψ is a tuning parameter. The Metropolis-Hastings ratio to update β is

$$R = min\left\{1, \frac{L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}', \mathcal{M}_D)}{L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}, \mathcal{M}_D)} \times \frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta})} \times 1\right\},$$

where the Hastings ratio simplifies to 1 owing to the symmetry of the normal distribution. We used the Cauchy and half-Cauchy distributions as priors because they are weakly informative and fat-tailed, causing our inference to prefer parameter values near zero while permitting parameters to take on large values should the data prove informative.

Proposing biogeographic histories

To update biogeographic histories, we sample an internal node uniformly at random and a set of areas, S, uniformly at random. We then propose a new biogeographic history by resampling the biogeographic histories for areas S for incident branches using the stochastic-mapping approach described earlier.

The Metropolis-Hastings ratio for this proposal is

$$R = min\left\{1, \frac{L(\mathbf{X}_{obs}, \mathbf{X}'_{aug}; \boldsymbol{\theta}, \mathcal{M}_D)}{L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}, \mathcal{M}_D)} \times \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \times \frac{L(\mathbf{X}_{obs}, \mathbf{X}_{aug}; \boldsymbol{\theta}^*, \mathcal{M}_0)}{L(\mathbf{X}_{obs}, \mathbf{X}'_{aug}; \boldsymbol{\theta}^*, \mathcal{M}_0)}\right\},$$

where the first term is the likelihood ratio under the full model, $\mathcal{M}_D$, and the second term is the proposal-density ratio that accounts for the probability of sampling the proposed biogeographic histories under the sampling model, $\mathcal{M}_0$, using the sampling parameters, $\boldsymbol{\theta}^*$. The parameters are not updated as part of this proposal, thus the ratio of prior probabilities may be safely omitted as it always equals 1.

Typically, the prior probability of each state (geographic range) at the root is equal to the corresponding stationary frequencies of the model. As mentioned above, our distance-dependent dispersal model is not time reversible, so we cannot approximate the stationary distribution by conventional means (*c.f.*, Robinson et al. 2003). Instead, we leverage the fact that the stationary frequencies of states (geographic ranges) of a model can be approximated by simulating the continuous-time Markov process over a sufficiently long branch. Accordingly, we append a long stem branch to the root node, sample an ancestral "consensus" configuration as the ancestral state at the stem node, then simulate a biogeographic history along the stem branch that is consistent with the states at the beginning (stem node) and end (root node) of the stem branch. Thus, we simulate into the stationary distribution of geographic ranges under the distance-dependent dispersal model along the stem branch, and then sample from the approximated stationary distribution at the root node using the same proposal machinery as is used for any internal node.

Model Selection

The mutual-independence model, $\mathcal{M}_0$, is equivalent to the distance-dependent dispersal model, $\mathcal{M}_D$, when $\beta = 0$. Since $\mathcal{M}_0 \subseteq \mathcal{M}_D$, we compute Bayes factors for these nested

models using the Savage-Dickey ratio (Dickey 1971; Verdinelli and Wasserman 1995), defined as

$$B_{D,0} = \frac{P_0(\beta = 0 \,|\, \mathcal{M}_D)}{P(\beta = 0 \,|\, \lambda_0, \lambda_1, \mathbf{x}_{obs}, \mathcal{M}_D)},$$

where $P_0(\beta = 0 \,|\, \mathcal{M}_D)$ is the prior probability and $P(\beta = 0 \,|\, \lambda_0, \lambda_1, \mathbf{x}_{obs}, \mathcal{M}_D)$ is the posterior probability under the more general distance-dependent dispersal model, $\mathcal{M}_D$, at the restriction point $\beta = 0$, where $\mathcal{M}_D$ is equivalent to the simpler mutual-independence model, $\mathcal{M}_0$. If the posterior probability under $\mathcal{M}_D$ at $\beta = 0$ is significantly greater than the corresponding prior probability, then the Bayes factor supports $\mathcal{M}_D$ (*i.e.*, $\mathcal{M}_D$ provides a better fit to the data). Since there is no analytical expression for the posterior probability, $P(\beta = 0 \,|\, \lambda_0, \lambda_1, \mathbf{x}_{obs}, \mathcal{M}_D)$, we approximate its distribution using the non-parametric Gaussian kernel density estimation method provided by default in `R` (R Core Team 2012).

Data Analysis

Simulation study

We simulated 50 dispersal datasets for each of eight values of β: 0, 0.25, 0.5, 1, 2, 3, 4, and 6. These data were simulated upon a geography with $20 \times 30 = 600$ uniformly spaced discrete areas positioned over the Bay Area, California. Phylogenies were simulated under a pure birth process with rate 1, then scaled to have a height comparable to our empirical study phylogeny. Dispersal and extinction rates were also chosen to resemble the rates inferred from the empirical analysis, but scaled to account for the increased number of areas. We then ran independent MCMC analyses for each dataset under the distance-dependent model. To identify the values of β that are indistinguishable from the mutual-independence model, we computed Bayes factors using the Savage-Dickey ratio for all posteriors inferred under the distance-dependent model.

We then quantified how well the posterior probabilities of dispersal histories correspond to the true biogeographic history known from the simulation. To do so, we compute the posterior probability of each area being occupied by each internal node for each analysis, then compute the sum of squared difference between each probability ($0 \leq p \leq 1$) and the corresponding true history ($p = 0$ or 1) recorded from the simulation. As this error term increases, the inferred ancestral ranges at nodes may be interpreted as less accurate.

Empirical study

We applied our method to 65 species of the plant clade *Rhododendron* section *Vireya*, which are distributed throughout the Malesian Archipelago. We used the species distributions and 20 discrete areas of endemism reported by Brown et al. (2006), and the time-calibrated

phylogeny reported by Webb and Ree (2012). To compute distances between areas, we used a single representative coordinate per area (depicted in Figure 1.8a). To simplify the analysis, we hold the geography to be constant throughout time.

Software configuration

Each MCMC analysis of the simulated data ran for 10^6 cycles, sampling parameters and node biogeographic histories every 10^3 cycles. For the empirical data, we ran five independent MCMC analyses, each set to run for 10^9 cycles, sampling every 10^4 cycles. To verify MCMC analyses converged to the same posterior distribution, we applied the Gelman diagnostic (Gelman and Rubin 1992) provided through the `coda` package (Plummer et al. 2006). Results from a single MCMC analysis are presented. The methods described here have been implemented in *BayArea*, for which `C++` source code is available for download at `http://code.google.com/p/bayarea`.

1.3 Results

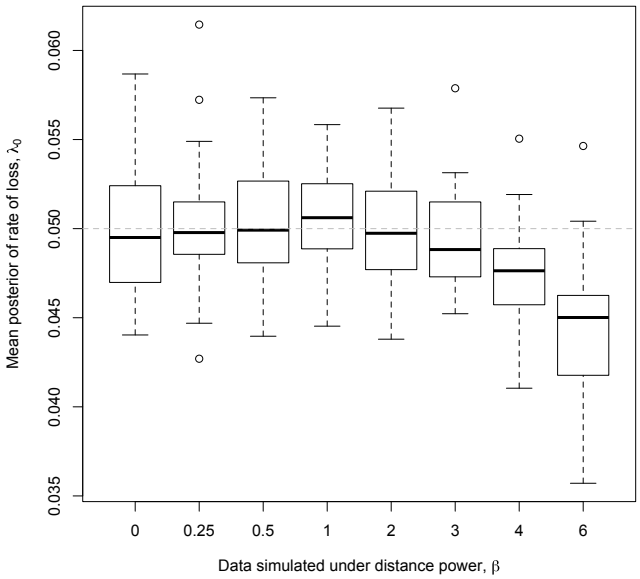
Simulation

For 50 phylogenies of 20 tips and a fixed geography of 600 areas (see Methods), we simulated 50 presence-absence data matrices for eight values of β: 0, 0.25, 0.5, 1, 2, 3, 4, and 6. Distributions of the mean posterior parameter values for the 8×50 MCMC analyses are shown in Figure 1.4. For $\beta \leq 3$, the model was able to retrieve the true simulation parameters accurately, but this accuracy degraded for $\beta \geq 4$ (see Discussion).

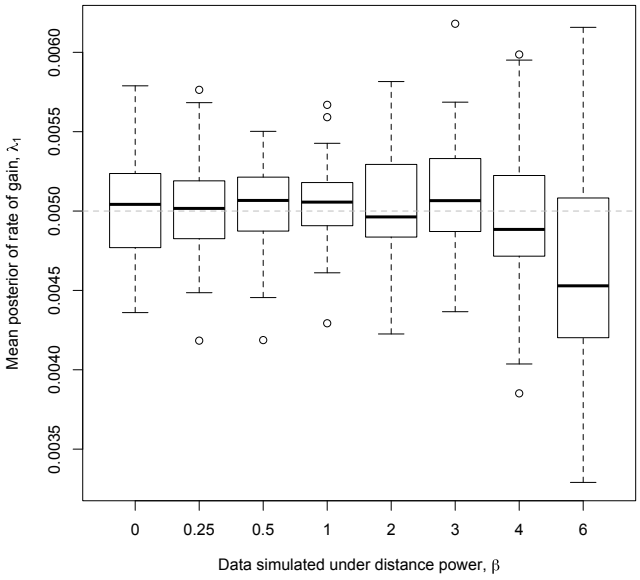
Figure 1.5 shows that Bayes factors consistently selected the correct model when data were simulated for $\beta \geq 1$ and for $\beta = 0$. For data simulated when $0 < \beta < 1$, we observed the greatest variance in the Bayes factor credible intervals. Data simulated under conditions in which distance had a weak effect on dispersal, *i.e.*, $\beta \leq 0.25$, were typically (and appropriately) indistinguishable from the mutual-independence model.

We then compared the true biogeographic history of each simulation to the corresponding posterior distribution of the sampled biogeographic histories. The sum of squared differences between posterior (estimated) and true (simulated) dispersal histories varied little for values of $\beta \leq 3$, with slight elevation in error for $\beta \geq 4$ (Figure 1.6). The elevated error for large values of the distance-power parameter, β, may be caused by the underestimated parameter values, or it may be an artifact of our error metric; it carries an independence assumption, so it over-penalizes distance-dependent dispersal histories that contain an excess of "near misses" relative to "wild misses".

A)


Mean posterior of rate of loss, λ_0
0.035
0.040
0.045
0.050
0.055
0.060
0
0.25
0.5
1
2
3
4
6
Data simulated under distance power, β

B)


Mean posterior of rate of gain, λ_1
0.0035
0.0040
0.0045
0.0050
0.0055
0.0060
0
0.25
0.5
1
2
3
4
6
Data simulated under distance power, β

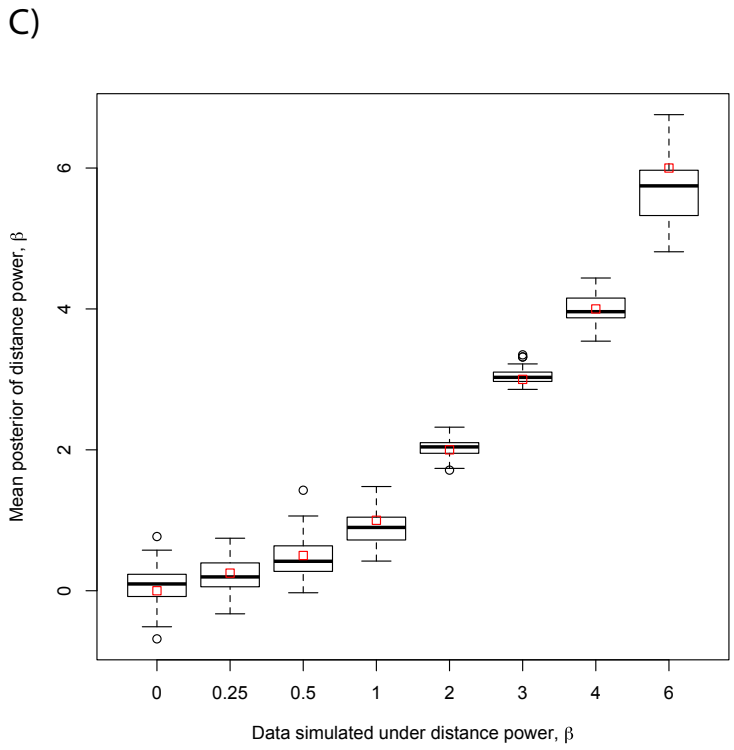


Figure 1.4: Distributions of means of posteriors of simulation study. Fifty datasets were simulated for each value of $\beta \in \{0, 0.25, 0.5, 1, 2, 3, 4, 6\}$ while $\lambda_0 = 0.05$ and $\lambda_1 = 0.005$ were held constant. For each set of 50 datasets, the mean of the posterior of each parameter was computed under the distance-dependent dispersal model. Distribution means are given by a bold line, while the 25th and 75th percentiles are given by the lower and upper edges of each box, called Q1 and Q3, respectively. The upper and lower whiskers indicate Q1 - IQR and Q3 + IQR, where IQR = 1.5 × (Q3 - Q1), and circles indicate outliers. The true parameter values are given by (A,B) the horizontal dashed line, and (C) the red squares.

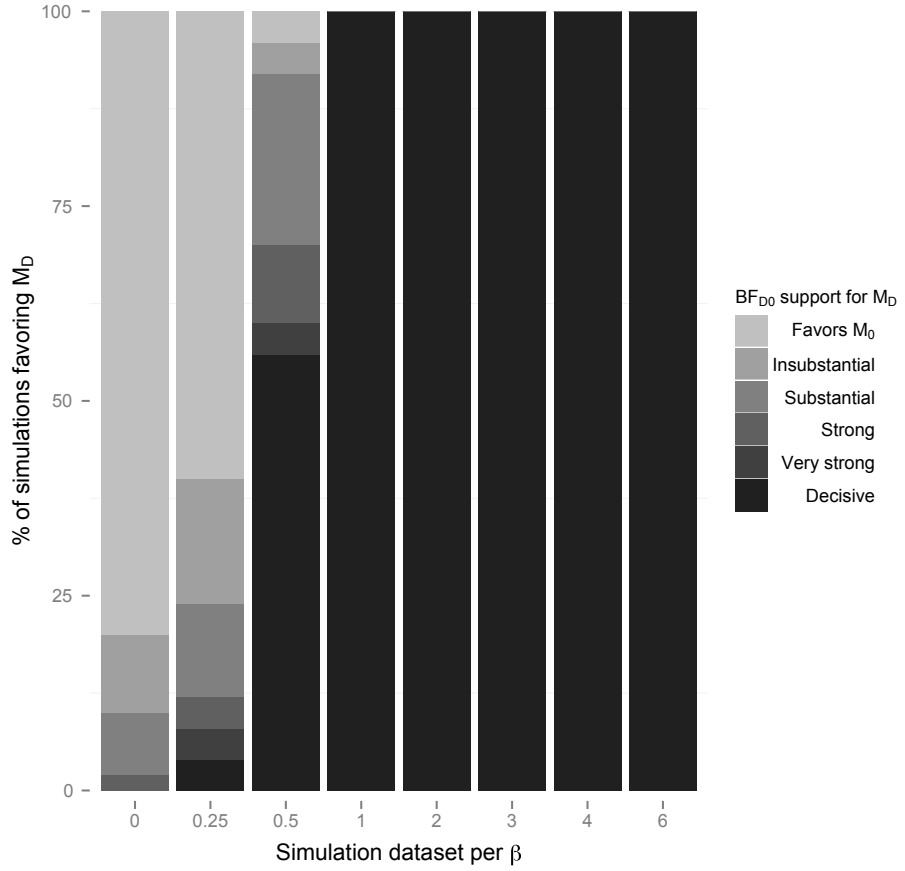


Figure 1.5: Distributions of Bayes factors for the simulation study. Fifty datasets were simulated for each value of $\beta \in \{0, 0.25, 0.5, 1, 2, 3, 4, 6\}$ while $\lambda_0 = 0.05$ and $\lambda_1 = 0.005$ were held constant. Columns display the frequencies of strengths of support in favor of the distance-dependent dispersal model, where strengths of support correspond to the intervals suggested by Jeffreys (1961): Favors M_0 on $(-\infty, 1)$; Insubstantial on $[1, 3)$; Substantial on $[3, 10)$; Strong on $[10, 30)$; Very strong on $[30, 100)$; Decisive on $[100, \infty)$. Each column corresponds to the strengths of support per 50 β-valued simulations. Bayes factors generally select the correct underlying model except for $\beta = 0.25$.

Empirical: *Vireya*

Bayes factors strongly favor the distance-dependent dispersal model over the mutual independence model to explain the biogeographic history of 65 rhododendron species in the

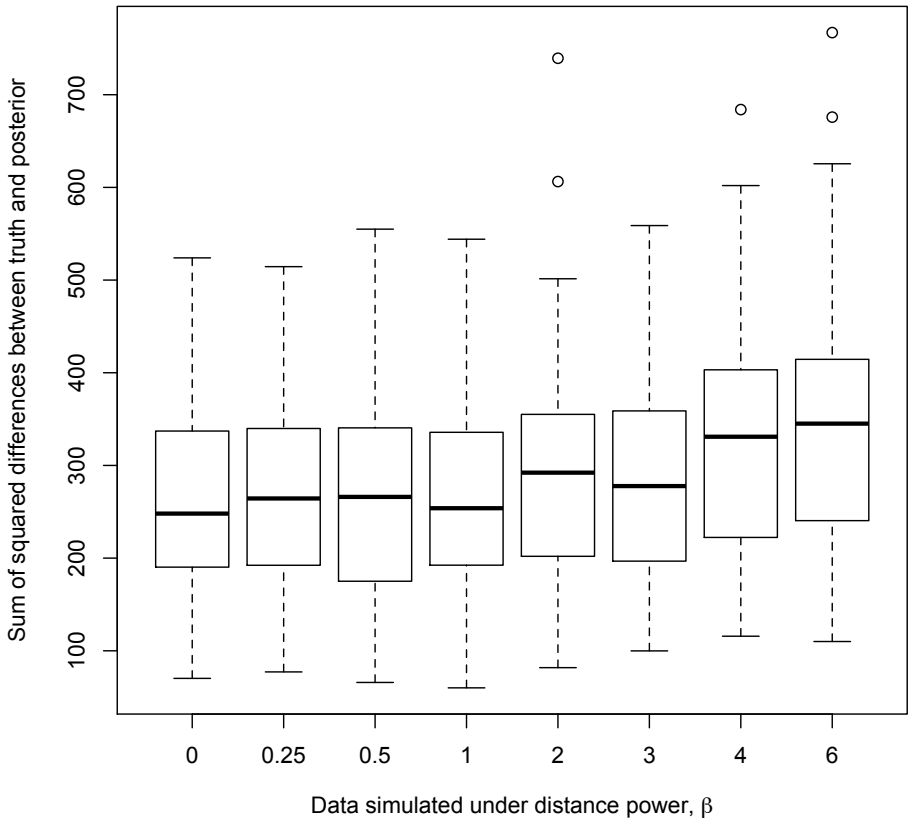


Figure 1.6: Errors for inferred dispersal histories of simulation study. The sum of squared differences between the posterior probability (*i.e.*, $0 < p < 1$) and the true history (*i.e.*, $p = 0$ or $p = 1$) for each area and each internal node were computed per simulated dataset. The box plots show the distribution of these sums for each batch of 50 simulated datasets per value of $\beta \in \{0, 0.25, 0.5, 1, 2, 3, 4, 6\}$. Distribution means are given by a bold line, while the 25th and 75th percentiles are given by the lower and upper edges of each box, called Q1 and Q3, respectively. The upper and lower whiskers indicate Q1 - IQR and Q3 + IQR, where IQR = 1.5 × (Q3 - Q1), and circles indicate outliers.

section *Vireya* over 20 biogeographical areas throughout Malesia. The estimated maximum *a posteriori* (MAP) value of the rate of area loss was $\lambda_0 = 0.13$, the rate of area gain was $\lambda_1 = 0.013$, and the distance power was $\beta = 2.65$ (Figure 1.7). Gelman-Rubin convergence

values for λ_0, λ_1, and β between all pairs of MCMC analyses were less than 1.1, which is consistent with all independent MCMC runs converging to the same posterior.

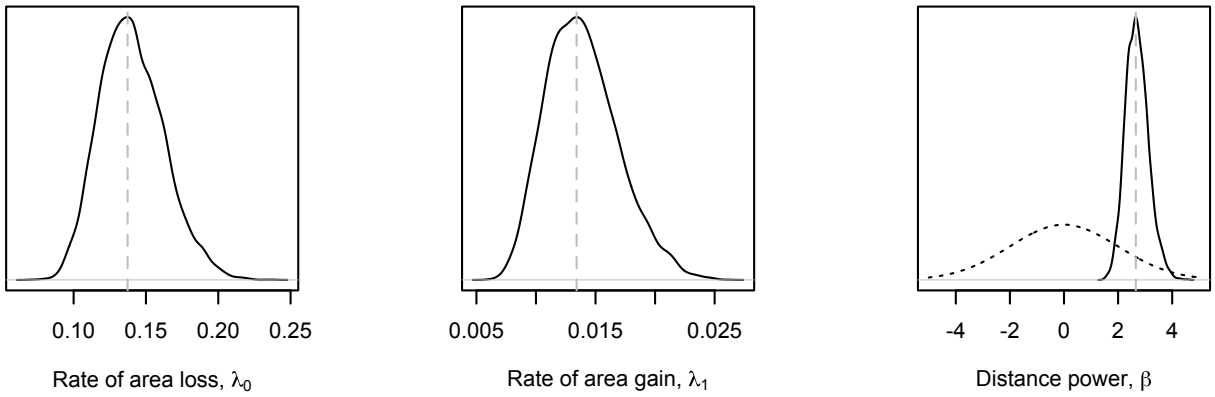


Figure 1.7: Marginal posterior densities for dispersal parameters from the Malesian *Rhododendron* dataset. Maximum *a posteriori* values (dashed gray line) for the distance-dependent dispersal model parameters are (A) $\lambda_0 = 0.13$, (B) $\lambda_1 = 0.013$, (C) $\beta = 2.65$. The dotted black line corresponds to the prior, $\beta \sim \text{Cauchy}(0, 1)$. Note that the posterior probability of $\beta = 0$ is approximately zero, resulting in "Decisive" support (*c.f.*, Jeffreys 1961) for the distance-dependent dispersal model over the mutual-independence model.

Figure 1.8 shows a summary of the inferred biogeographic history (Supplementary Figure 1 shows the full history and observed ranges). The per-area posterior probabilities of the ancestral ranges strongly favor migration eastward into the Malesian Archipelago originating from Southeast Asia. The inferred biogeographic scenario — multiple independent dispersal events from the Sunda Shelf across Wallace's Line into Wallacea — is favored over that of a single dispersal event followed by pervasive extinction events (Figure 1.8b). Lydekker's Line appears to be less permeable, with only a single lineage dispersing eastward from Wallacea across it onto the Sahul Shelf (Figure 1.8c). An interactive animation of the ancestral range reconstruction is hosted at `http://mlandis.github.com/phylowood/?url=examples/vireya.nhx`.

Readers might naturally wonder how inferences under the current method compare to those based on alternative statistical biogeographic methods, such as the DEC model of Ree et al. (2005). Despite their superficial similarities — both are likelihood-based methods that rely on continuous-time Markov models to describe the evolution of species geographic range — the methods differ to an extent that makes it difficult to draw any meaningful comparisons. Specifically, the two methods invoke models that differ in many respects (see

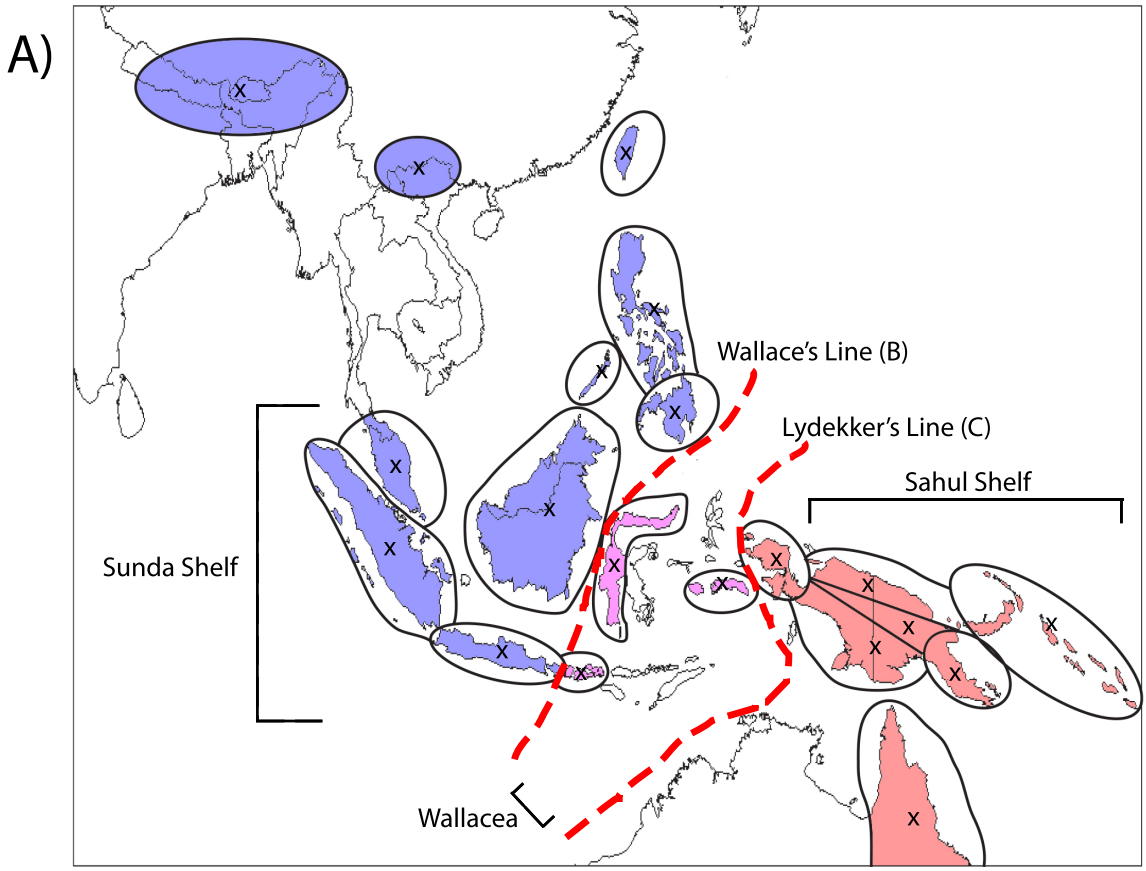
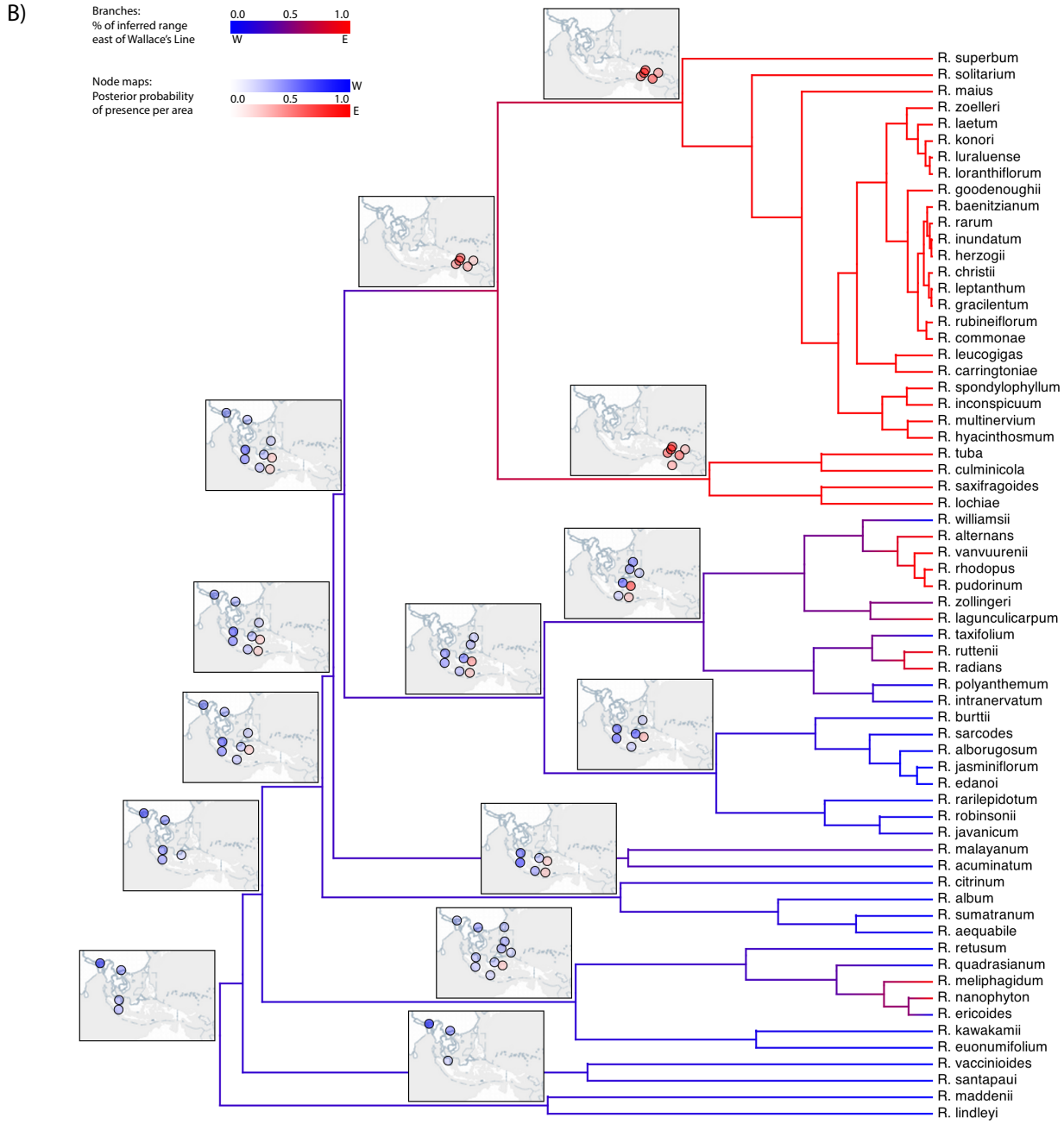
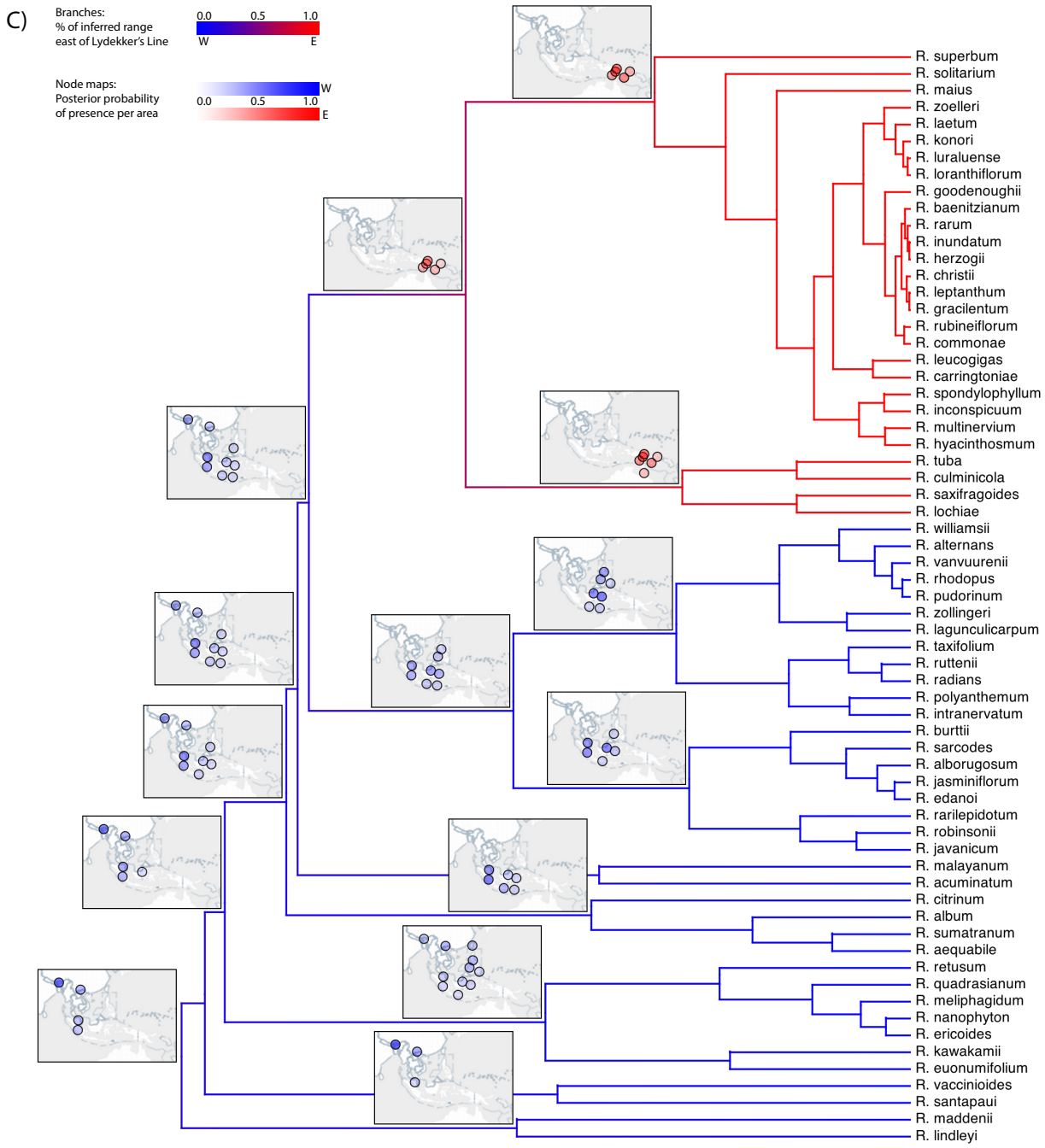


Figure 1.8: Biogeographic history of Malesian *Rhododendron*. (A) The region was parsed into 20 discrete geographic areas following Brown et al. (2006), which straddle two important biotic boundaries — Wallace's and Lydekker's Lines. Each circle corresponds to a discrete area. Distances between these areas are based on a single coordinate for each area, indicated by an 'x'. Posterior probability of being present in an area is proportional to the opacity of the circle. Occupied areas with posterior probabilities less than 0.12 are masked to ease interpretation. Circles are colored according to their position relative to Wallace's Line (B) or Lydekker's Line (C). Branches are colored by a gradient representing the sum of posterior probabilities of being present per area for descendant-ancestor pairs. We infer a continental Asian origin for Malesian rhododendrons with multiple dispersal events across Wallace's Line (B) and a single dispersal event across Lydekker's Line (C).

Discussion), and are implemented in different statistical frameworks (maximum likelihood *vs.* Bayesian inference).


B)
Branches:
% of inferred range
east of Wallace's Line
0.0
0.5
1.0
W
E
Node maps:
Posterior probability
of presence per area
W
0.0
0.5
1.0
E
R. superbum
R. solitarium
R. maius
R. zoelleri
R. laetum
R. konori
R. luraluense
R. loranthiflorum
R. goodenoughii
R. baenitzianum
R. rarum
R. inundatum
R. herzogii
R. christii
R. leptanthum
R. gracilentum
R. rubineiflorum
R. commonae
R. leucogigas
R. carringtoniae
R. spondylophyllum
R. inconspicuum
R. multinervium
R. hyacinthosmum
R. tuba
R. culminicola
R. saxifragoides
R. lochiae
R. williamsii
R. alternans
R. vanvuurenii
R. rhodopus
R. pudorinum
R. zollingeri
R. lagunculicarpum
R. taxifolium
R. ruttenii
R. radians
R. polyanthemum
R. intranervatum
R. burttii
R. sarcodes
R. alborugosum
R. jasminiflorum
R. edanoi
R. rarilepidotum
R. robinsonii
R. javanicum
R. malayanum
R. acuminatum
R. citrinum
R. album
R. sumatranum
R. aequabile
R. retusum
R. quadrasianum
R. meliphagidum
R. nanophyton
R. ericoides
R. kawakamii
R. euonumifolium
R. vaccinioides
R. santapaui
R. maddenii
R. lindleyi


C)
Branches:
% of inferred range
east of Lydekker's Line
0.0 0.5 1.0
W E
Node maps:
Posterior probability
of presence per area
W
0.0 0.5 1.0
E
R. superbum
R. solitarium
R. maius
R. zoelleri
R. laetum
R. konori
R. luraluense
R. loranthiflorum
R. goodenoughii
R. baenitzianum
R. rarum
R. inundatum
R. herzogii
R. christii
R. leptanthum
R. gracilentum
R. rubineiflorum
R. commonae
R. leucogigas
R. carringtoniae
R. spondylophyllum
R. inconspicuum
R. multinervium
R. hyacinthosmum
R. tuba
R. culminicola
R. saxifragoides
R. lochiae
R. williamsii
R. alternans
R. vanvuurenii
R. rhodopus
R. pudorinum
R. zollingeri
R. lagunculicarpum
R. taxifolium
R. ruttenii
R. radians
R. polyanthemum
R. intranervatum
R. burttii
R. sarcodes
R. alborugosum
R. jasminiflorum
R. edanoi
R. rarilepidotum
R. robinsonii
R. javanicum
R. malayanum
R. acuminatum
R. citrinum
R. album
R. sumatranum
R. aequabile
R. retusum
R. quadrasianum
R. meliphagidum
R. nanophyton
R. ericoides
R. kawakamii
R. euonumifolium
R. vaccinioides
R. santapaui
R. maddenii
R. lindleyi

1.4 Discussion

Historical biogeography has begun the transition to explicitly model-based statistical inference (Ree and Sanmartín 2009; Ronquist and Sanmartín 2011). These methods describe the biogeographic process by means of continuous-time Markov chain that models the colonization of—and extinction within—a set of discrete geographic areas, and calculate the likelihood of the observed species geographic ranges at the tips of the tree using matrix exponentiation (to integrate over possible biogeographic histories along branches) and Felsenstein's pruning algorithm (to integrate over possible ancestral ranges at the interior nodes of the tree). Although this is a vigorous area of research, reliance on matrix exponentiation ultimately entails serious computational constraints that limit both our ability to develop more elaborate and realistic biogeographic models and to apply these methods to more complex and typical empirical problems.

We offer a Bayesian solution to this constraint that relies on data augmentation and MCMC to numerically integrate over biogeographic histories to estimate the joint posterior probability of the parameters given the data. The primary implication of this approach is a substantial increase in the number of discrete areas that can be accommodated — by approximately two orders of magnitude. Moreover, we propose a simple distance-dependent dispersal model in which rates of area colonization are a function of geographic distance. The nature and strength of the distance effect on rates of colonization are governed by the distance-power parameter, β. When $\beta > 0$, dispersal events over long distances are penalized, whereas long-distance dispersal events are favored when $\beta < 0$. Importantly, when $\beta = 0$, the distance-dependent dispersal model collapses to the simpler mutual-independence model, and so $\mathcal{M}_0 \subseteq \mathcal{M}_D$. Because the models are nested, we can use the Savage-Dickey density ratio to compute Bayes factors for robust model selection.

In the remainder of this section, we attempt to develop an intuition regarding the behavior of this new biogeographic approach, describe some of the benefits and limitations of the current implementation, and consider how this approach might be profitably extended.

Exploring the behavior of the Bayesian biogeographic framework

We explored the statistical behavior of our biogeographic model and inference framework via analyses of simulated and empirical data. The simulation study comprised 50 dispersal datasets for 20 taxa and 600 areas that were simulated under each of eight strengths of distance effects, β: 0, 0.25, 0.5, 1, 2, 3, 4, and 6. For $\beta \leq 3$, we were generally able to infer the true parameters. However, estimation accuracy begins to suffer when $\beta \geq 4$, resulting in all parameters being slightly underestimated. Estimation accuracy is also high for inferences based on data simulated under large β values, so the poor accuracy appears to emerge from the phylogenetic structure underlying the data. Although values of $\beta \geq 4$ are

greater than those we have inferred from empirical data, we advise increased caution should one's inference lie in this range of parameters. Using the Savage-Dickey ratio to compute Bayes factors, we found our ability to select the correct model was largely determined by the strength of β (Figure 1.5). Future simulation studies should be extended to evaluate the effects of the phylogeny on inference (tree size, shape, uncertainty, etc.), the sensitivity of the model to various priors, and whether extreme parameter values introduce greater errors in ancestral geographic-range estimates.

As currently specified, the distance-dependent dispersal rate modifier, $\eta(\cdot)$, only changes the dispersal rate per area, but not the summed rates of colonization and extinction over the geographic range. Accordingly, the equilibrium number of occupied or unoccupied areas for the geographic range is largely determined by the ratio of λ_1 and λ_0 (the per-area rates of colonization and extinction, respectively). When the geographic range involves occupation of a relatively small fraction of available areas — as occurs when the number of areas increases — the area colonization/extinction rate ratio becomes small in order to explain the low observed frequencies of area occupancy at the tips of the tree. In such situations, these relatively simple parameters may fail to fit the data well. Moreover, the size of inferred ancestral geographic ranges (in terms of the number of occupied areas) tends to be larger than those observed at the tips of the tree. This phenomenon is also characteristic of other parsimony- and likelihood-based biogeographic methods (*e.g.*, Ronquist 1997; Ree et al. 2005; Clark et al. 2008; Buerki et al. 2011). One solution to both problems would be to favor sampled biogeographic histories with range sizes most similar to a carrying-capacity or range-size parameter.

We demonstrated the empirical application of our method with an analysis of the biogeographic history of 65 *Vireya* species distributed over 20 geographic areas across the Malesian Archipelago (Brown et al. 2006). Bayes factors strongly favored the distance-dependent model, with a maximum *a posteriori* estimate of $\beta = 2.65$ (see Figure 1.7). Brown et al. offered two hypotheses for the origin of *Rhododendron*: as an old genus that arose in Australia, or as a young genus that arose in Asia. Under our model, the posterior of sampled biogeographic histories at the root of the tree suggests that Asia is the most probable point from which the genus entered the Malesian archipelago (see Figure 1.8).

The inferred biogeographic history of *Vireya* involves several episodes of dispersal across Wallace's Line and a single episode of dispersal across Lydekker's Line (see Figure 1.8b,c). We note two points regarding these dispersal events. First, the earliest dispersal across Wallace's Line and the single dispersal across Lydekker's Line appear to have occurred at approximately the same time. Adopting 55Mya as the crown age of the *Rhododendron* phylogeny (Webb and Ree 2012) implies that these dispersal events occurred in the Late Eocene ($\sim$ 40Mya). At that time, many of the discrete areas in the western part of the Malesian Archipelago collectively formed a contiguous, emergent terrestrial region, Sundaland (Lohman et al. 2011), which may have facilitated the easterly dispersal of *Vireya* species

from their ancestral range in continental Asia across Sundaland. Moreover, the eastern border of Sundaland was not yet bounded by a contiguous deep oceanic trench, which may have facilitated the continued easterly dispersal from Sundaland into Wallacea (across Wallace's Line) and eastward out of Wallacea (across Lydekker's Line) into the eastern region of the Malesian Archipelago.

The second point pertains to the apparent prevalence of dispersal events across Wallace's line. The origin of *Vireya* in continental Asia may have permitted the accumulation of greater species diversity throughout Sundaland, west of Wallacea. This would have established a greater species-diversity gradient across Wallace's line than that for Lydekker's line. Consequently, there may have been more opportunity for species to disperse across the western boundary (Wallace's line) into Wallacea than there has been for species to disperse across the eastern boundary (Lydekker's line) out of Wallacea.

Advantages and limitations of the Bayesian biogeographic method

Increasing the number of areas offers several benefits. The most obvious, of course, is the ability to increase the geographic resolution of biogeographic inference. As we increase the number of areas, discrete biogeography better represents the continuous features of Earth. As an example, for a clade of terrestrial species that collectively share a global distribution, a statistical biogeographic analysis would want to discretize the (approximately) 1.5×10^8 km^2 of terrestrial space into a meaningful number of areas. With approximately 15 areas (the previous limit), the average area would be comparable in size to Canada ($\approx 10^7$ km^2); for approximately 1500 areas (manageable under the current approach), the average area would be comparable to the size of Ohio ($\approx 10^5$ km^2).

Second, biogeographic areas have traditionally been defined on the basis of empirical analysis. For systems that do not have well-defined biogeographic areas, our method allows the biogeographer to agnostically define areas according to a grid, as was done in our simulation study. By studying the congruence between posteriors of dispersal histories for alternatively discretized geographies, one could determine the optimal discretization for a particular system, including both the number and shapes of areas. For example, a researcher with intimate knowledge of a study system may derive a geographic discretization that produces radically different ancestral-range estimates than those based on a uniformly gridded discretization. Such a scenario suggests that one of the two discretizations does not properly "weight" the importance of certain geographic areas when inferring the biogeographic history.

Although it has benefits, the ability to increase the geographic resolution also raises new issues. At highly resolved spatial scales, for example, it may become more difficult to accurately specify the occupancy of species within individual cells of the geographic grid. Inference under our model conditions on the biogeographic ranges of species at the tips of the

tree, and errors in specifying these ranges are likely to lead inference astray. One solution to this issue would be to use species-distribution models to first predict the geographic ranges of species, and then treat these estimated ranges as the observed species' geographic ranges (analogous to the conventional practice of treating a multiple-sequence alignment—an inference predicted from the raw data—as the observations used to infer phylogeny).

Extending the Bayesian biogeographic method

The real benefit of the Bayesian framework is the tremendous extensibility that it affords. The current implementation makes various restrictive assumptions. For example, we assume a fixed (and known) tree, a static geological history, and a homogeneous environment. Below we touch briefly on three extensions that permit the approach to accommodate phylogenetic uncertainty, dynamic geological history, and environmental heterogeneity.

Our implementation assumes the phylogeny is known without error, a luxury that exists only under simulation. The most natural way to account for phylogenetic uncertainty would be to exploit a distribution of time-calibrated trees (estimated separately) as input for biogeographic inference. This approach is straightforward for methods that analytically integrate over biogeographic histories: simply define an MCMC proposal to draw a new tree from the marginal distribution of phylogenies. However, our model entails sampling biogeographic histories for a specific phylogeny. Accordingly, this extension will require the use of joint proposals for both biogeographic history and phylogeny that maintain good mixing of the MCMC (*i.e.*, that ensure reasonable acceptance probabilities). This will be a challenging task.

It is important to emphasize that our empirical analysis was conducted under the assumption of a static geological history: we explicitly ignore the substantial effects of tectonic drift, changes in sea level, the formation of islands, etc. This greatly simplifies the analysis, of course, since biogeographic likelihoods are computed by conditioning on a single, static set of geographic distances. Ideally, paleogeographic reconstructions would inform the changing proximity of areas through time, and biogeographic inference would be computed by conditioning on a temporally dynamic geography. For example, consider the scenario in which two continents drift apart as time advances, which may be characterized as a time-ordered vector of maps, each map corresponding to the geography appropriate to each interval of geological time. Since our phylogeny is also measured in units of absolute time, the rates of gain and loss could easily be modified to condition on the relevant set of geographical coordinates. In the above scenario, distances between areas between continents would increase with time, so dispersal events between continents would become increasingly unlikely.

By adopting a DEC-like approach wherein cladogenesis events differ in pattern from anagenic dispersal and extinction events, our model would have to define transition probabilities between larger numbers of configurations; it is trivial to compute the model likelihood with

a model that accounts for cladogenic events by conditioning on a single biogeographic history, but to numerically integrate over all possible cladogenic events via MCMC will require sophisticated proposal distributions.

Finally, we can incorporate other features of areas beyond their latitude and longitude—such as altitude, climate, and ecology—that may affect dispersal rates. Morphological evolution also has a noted role in biogeography—Bergmann's Rule (Freckleton et al. 2003), traits that effect long-distance dispersal ability (Carlquist 1966), etc.—and could be jointly inferred along with dispersal patterns (Lartillot and Poujol 2011). These factors could variously be incorporated as parameters to construct a suite of candidate biogeographic models. As we demonstrated for exploring the effect of geographic distance, marginal likelihoods under different biogeographic models could then be computed and Bayes factors used to identify biogeographically important model components.

Noting the simplicity of their biogeographic model, Ree et al. (2005) drew an analogy to the earliest work on probabilistic models of molecular evolution—the Jukes and Cantor (1969) model. Although it admittedly offered a rudimentary description of the process, this first model nevertheless provided a critical proof of concept that the problem could be profitably pursued in a statistical framework. To extend this analogy, we believe the current contribution resembles the subsequent paper by Felsenstein (1981), in which he proposed the pruning algorithm that—by virtue of conferring a tremendous increase in computational efficiency—heralded an era of progress in developing stochastic models for the analysis of DNA and amino acid sequence data that has been one of the great success stories in evolutionary biology. We are hopeful that the small steps made here will precipitate a similar era of productivity in the field of biogeographic inference that will enhance our ability to make progress on this important problem.

Chapter 2

Biogeographic dating of speciation times

2.1 Introduction

Time is a simple and fundamental axis of evolution. Knowing the order and timing of evolutionary events grants us insight into how vying evolutionary processes interact. With a perfectly accurate catalog of geologically-dated speciation times, many macroevolutionary questions would yield to simple interrogation, such as whether one clade exploded with diversity before or after a niche-analogous clade went extinct, or whether some number of contemporaneous biota were eradicated simultaneously by the same mass extinction event. Only rarely does the fossil record give audience to the exact history of evolutionary events: it is infamously irregular across time, space, and species, so biologists generally resort to inference to estimate when, where, and what happened to fill those gaps. That said, we have not yet found a perfect character or model to infer dates for divergence times, so advances in dating strategies are urgently needed. A brief survey of the field reveals why.

The molecular clock hypothesis of Zuckerkandl and Pauling (1962) states that if substitutions arise (i.e. alleles fix) at a constant rate, the expected number of substitutions is the product of the substitution rate and the time the substitution process has been operating. With data from extant taxa, we only observe the outcome of the evolutionary process for an unknown rate and an unknown amount of time. As such, rate and time are non-identifiable under standard models of molecular substitution, so inferred amounts of evolutionary change are often reported as a compound parameter, the product of rate and time, called length. If all species' shared the same substitution rate, a phylogeny with branches measured in lengths would give relative divergence times, i.e. proportional to absolute divergence times. While it is reasonable to say species' evolution shares a basis in time, substitution rates

differ between species and over macroevolutionary timescales (Wolfe et al. 1987; Martin and Palumbi 1993). Even when imposing a model of rate heterogeneity (Thorne et al. 1998; Drummond et al. 2006), only relative times may be estimated. Extrinsic information, i.e. a dated calibration point, is needed to establish an absolute time scaling, and typically takes form as a fossil occurrence or paleogeographical event.

When fossils are available, they currently provide the most accurate inroad to calibrate divergence events to geological timescales, largely because each fossil is associated with a geological occurrence time. Fossil ages may be used in several ways to calibrate divergence times. The simplest method is the fossil node calibration, whereby the fossil is associated with a clade and constrains its time of origin (Ho and Phillips 2009; Parham et al. 2011). Node calibrations are empirical priors, not data-dependent stochastic processes, so they depend entirely on experts' abilities to quantify the distribution of plausible ages for the given node. That is, node calibrations do not arise from a generative evolutionary process. Since molecular phylogenies cannot identify rate from time, then the time scaling is entirely determined by the prior, i.e. the posterior is perfectly prior-sensitive for rates and times. Rather than using prior node calibrations, fossil tip dating (Pyron 2011; Ronquist et al. 2012a) treats fossil occurrences as terminal taxa with morphological characters as part of any standard phylogenetic analysis. In this case, the model of character evolution, tree prior, and fossil ages generate the distribution of clade ages, relying on the fossil ages and a morphological clock to induce time calibrations. To provide a generative process of fossilization, Heath et al. (2014) introduced the fossilized birth-death process, by which lineages speciate, go extinct, or produce fossil observations. Using fossil tip dating with the fossilized birth-death process, Gavryushkina et al. (2015) demonstrated multiple calibration techniques may be used jointly in a theoretically consistent framework (i.e. without introducing model violation).

Of course, fossil calibrations require fossils, but many clades leave few to no known fossils due to taphonomic processes, which filter out species with too soft or too fragile of tissues, or with tissues that were buried in substrates that were too humid, too arid, or too rocky; or due to sampling biases, such as geographical or political biases imbalancing collection efforts (Behrensmeyer et al. 2000; Kidwell and Holland 2002). Although these biases do not prohibitively obscure the record for widespread species with robust mineralized skeletons—namely, large vertebrates and marine invertebrates—fossil-free calibration methods are desperately needed to date the remaining majority of nodes in the tree of life.

In this direction, analogous to fossil node dating, node dates may be calibrated using paleobiogeographic scenarios (Heads 2005; Renner 2005). For example, an ornithologist might reasonably argue that a bird known as endemic to a young island may have speciated only after the island was created, thus providing a maximum age of origination. Using this scenario as a calibration point excludes the possibility of alternative historical biogeographic explanations, e.g. the bird might have speciated off-island before the island surfaced and migrated there afterwards. See Heads (2005; 2011), Kodandaramaiah (2011), and Ho et al.

(2015) for discussion on the uses and pitfalls of biogeographic node calibrations. Like fossil node calibrations, biogeographic node calibrations fundamentally rely on some prior distribution of divergence times, opinions may vary from expert to expert, making results difficult to compare from a modeling perspective. Worsening matters, the time and context of biogeographic events are never directly observed, so asserting that a particular dispersal event into an island system resulted in a speciation event to calibrate a node fails to account for the uncertainty that the assumed evolutionary scenario took place at all. Ideally, all possible biogeographic and diversification scenarios would be considered, with each scenario given credence in proportion to its probability.

Inspired by advents in fossil dating models (Pyron 2011; Ronquist et al. 2012a; Heath et al. 2014), which have matured from phenomenological towards mechanistic approaches (Rodrigue and Philippe 2010), I present an explicitly data-dependent and process-based biogeographic method for divergence time dating to formalize the intuition underlying biogeographic node calibrations. Analogous to fossil tip dating, the goal is to allow the observed biogeographic states at the "tips" of the tree induce a posterior distribution of dated speciation times by way of an evolutionary process. By modeling dispersal rates between areas as subject to time-calibrated paleogeographical information, such as the merging and splitting of continental adjacencies due to tectonic drift, particular dispersal events between area-pairs are expected to occur with higher probability during certain geological time intervals than during others. For example, the dispersal rate between South America and Africa was likely higher when they were joined as West Gondwana (ca 120 Ma) than when separated as they are today. If the absolute timing of dispersal events on a phylogeny matters, then so must the absolute timing of divergence events. Unlike fossil tip dating, biogeographic dating should, in principle, be able to date speciation times only using extant taxa.

To illustrate how this is possible, I construct a toy biogeographic example to demonstrate when paleogeography may date divergence times, then follow with a more formal description of the model. By performing joint inference with molecular and biogeographic data, I demonstrate the effectiveness of biogeographic dating by applying it to simulated and empirical scenarios, showing rate and time are identifiable. While researchers have accounted for phylogenetic uncertainty in biogeographic analyses (Nylander et al. 2008; Lemey et al. 2009; Beaulieu et al. 2013), I am unaware of work demonstrating how paleogeographic calibrations may be leveraged to date divergence times via a biogeographic process. For the empirical analysis, I date the divergence times for *Testudines* using biogeographic dating, first without any fossils, then using a fossil root node calibration. Finally, I discuss the strengths and weaknesses of my method, and how it may be improved in future work.

2.2 Model

The anatomy of biogeographic dating

Briefly, I will introduce an example of how time-calibrated paleogeographical events may impart information through a biogeographic process to date speciation times, then later develop the details underlying the strategy, which I call biogeographic dating. Throughout the manuscript, I assume a rooted phylogeny, Ψ, with known topology but with unknown divergence times that I wish to estimate. Time is measured in geological units and as time until present, with $t = 0$ being the present, $t < 0$ being the past, and age being the negative amount of time until present. To keep the model of biogeographic evolution simple, the observed taxon occurrence matrix, $\mathbf{Z}$, is assumed to be generated by a discrete-valued dispersal process where each taxon is present in only a single area at a time (Sanmartín et al. 2008). For example, taxon $T1$ might be coded to be found in Area A or Area B but not both simultaneously. Although basic, this model is sufficient to make use of paleogeographical information, suggesting more realistic models will fare better.

Consider two areas, A and B, that drift into and out of contact over time. When in contact, dispersal is possible; when not, impossible. Represented as a graph, A and B are vertices, and the edge (A, B) exists only during time intervals when A and B are in contact. The addition and removal of dispersal routes demarcate time intervals, or *epochs*, each corresponding to some epoch index, $k \in \{1, \ldots, K\}$. To define how dispersal rates vary with k, I use an epoch-valued continuous-time Markov chain (CTMC) (Ree et al. 2005; Ree and Smith 2008; Bielejec et al. 2014). The adjacency matrix for the k^{th} time interval's graph is used to populate the elements of an instantaneous rate matrix for an epoch's CTMC such that the dispersal rate is equal to 1 when the indexed areas are adjacent and equals 0 otherwise. For a time-homogeneous CTMC, the transition probability matrix is typically written as $\mathbf{P}(t)$, which assumes the rate matrix, $\mathbf{Q}$, has been rescaled by some clock rate, μ, and applied to a branch of some length, t. For a time-heterogeneous CTMC, the value of the rate matrix changes as a function of the underlying time interval, $\mathbf{Q}(k)$. The transition probability matrix for the time-heterogeneous process, $\mathbf{P}(s, t)$, is the matrix-product of the constituent epochs' time-homogeneous transition probability matrices, and takes a value determined by the absolute time and order of paleogeographical events contained between the start time, s, and end time, t. Under this construction, certain types of dispersal events are more likely to occur during certain absolute (not relative) time intervals, which potentially influences probabilities of divergence times in absolute units.

Below, I give examples of when a key divergence time is likely to precede a split event (Figure 2.1) or to follow a merge event (Figure 2.2). To simplify the argument, I assume a single change must occur on a certain branch given the topology and tip states, though the logic holds in general.

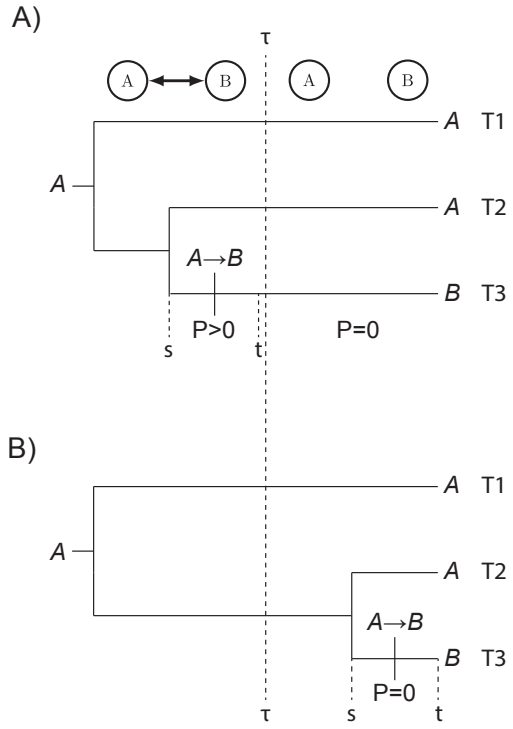


Figure 2.1: **Effects of a paleogeographical split on divergence times.** Area A splits from Area B at time τ. $T1$ and $T2$ have state A and the transition $A \to B$ most parsimoniously explains how $T3$ has state B. The transition probabilty for $P = [\mathbf{P}(s,t)]_{AB}$ is non-zero before the paleogeographical split event at time τ, and is zero afterwards. Two possible divergence and dispersal times are given: A) $T3$ originates before the split when the transition $A \to B$ has non-zero probability. B) $T3$ originates after the split when the transition $A \to B$ has probability zero.

In the first scenario (Figure 2.1), sister taxa $T2$ and $T3$ are found in Areas A and B, respectively. The divergence time, s, is a random variable to be inferred. At time τ, the dispersal route (A, B) is destroyed, inducing the transition probability $[\mathbf{P}(s,t)]_{AB} = 0$ between times τ and 0. Since $T2$ and $T3$ are found in different areas, at least one dispersal event must have occurred during an interval of non-zero dispersal probability. Then, the divergence event that gave rise to $T2$ an $T3$ must have also pre-dated τ, with at least one dispersal event occuring before the split event (Figure 2.1A). If $T2$ and $T3$ diverge after τ, a dispersal event from A to B is necessary to explain the observations (Figure 2.1B), but the model disfavors that divergence time because the required transition has probability zero.

In this case, the creation of a dispersal barrier informs the latest possible divergence time, a bound after which divergence between $T2$ and $T3$ is distinctly less probable if not impossible. It is also worth considering that a more complex process modeling vicariant speciation would provide tight bounds centered on τ (see Discussion).

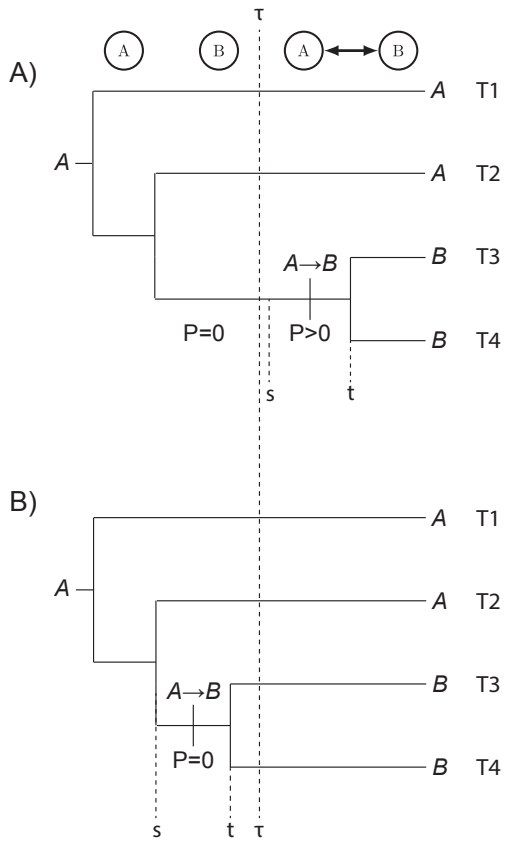


Figure 2.2: **Effects of paleogeographical merge on divergence times.** Area A merges with Area B at time τ. $T1$ and $T2$ have the state A and the transition $A \to B$ on the lineage leading to $(T3, T4)$ most parsimoniously explains how $T3$ and $T4$ have state B. The transition probabilty for $P = [\mathbf{P}(s,t)]_{AB}$ is zero before the paleogeographical merge event at time τ, and only non-zero afterwards. Two possible divergence and dispersal times are given: A) $T3$ and $T4$ originate after the merge when the transition $A \to B$ has non-zero probability. B) $T3$ and $T4$ originate before the merge when the transition $A \to B$ has probability zero.

In the second scenario (Figure 2.2), the removal of a dispersal barrier is capable of creating a maximum divergence time threshold, pushing divergence times towards the present. To

demonstrate this, say the ingroup sister taxa $T3$ and $T4$ both inhabit Area B and the root state is Area A. Before the areas merge, the rate of dispersal between A and B as zero, and non-zero afterwards. When speciation happens after the areas merge, then the ancestor of $(T3, T4)$ may disperse from A to B, allowing $T3$ and $T4$ to inherit state B (Figure 2.2A). Alternatively, if $T3$ and $T4$ originate before the areas merge, then the same dispersal event on the branch ancestral to $(T3, T4)$ has probability zero (Figure 2.2B).

Paleogeography, graphs, and Markov chains

How biogeography may date speciation times depends critically on the assumptions of the biogeographic model. The above examples depend on the notion of *reachability*, that two vertices (areas) are connected by some ordered set of edges (dispersal routes) of any length. In the adjacent-area dispersal model used here, one area might not be reachable from another area during some time interval, during which the corresponding transition probability is zero. That is, no path of any number of edges (series of dispersal events) may be constructed to connect the two areas. The concept of reachability may be extended to sets of partitioned areas: in graph theory, sets of vertices (areas) that are mutually reachable are called *(connected) components*. In terms of a graphically structured continuous time Markov chain, each component forms a *communicating class*: a set of states with positive transition probabilities only to other states in the set, and zero transition probabilities to other states (or communicating classes) in the state space. To avoid confusion with the "generic" biogeographical concept of components (Passalacqua 2015), and to emphasize the interaction of these partitioned states with respect to the underlying stochastic process, I hereafter refer to these sets of areas as communicating classes.

Taking terrestrial biogeography as an example, areas exclusive to Gondwana or Laurasia may each reasonably form communicating classes upon the break-up of Pangaea (Figure 2.3), meaning species are free to disperse within these paleocontinents, but not between them. For example, the set of communication classses is $S = \{\{\text{Afr}\}, \{\text{As}\}, \{\text{Ind}\}\}$ at $t = -100$, i.e. there are $|S| = 3$ communicating classes because no areas share edges (Figure 2.3C), while at $t = -10$ there is $|S| = 1$ communicating class since a path exists between all three pairs of areas (Figure 2.3E).

Specifying communicating classes is partly difficult because we do not know the ease of dispersal between areas for most species throughout space and time. Encoding zero-valued dispersal rates directly into the model should be avoided given the apparent prevalence of long distance dispersal, sweepstakes dispersal, etc. across dispersal barriers (Carlquist 1966). Moreover, zero-valued rates imply that dispersal events between certain areas are not simply improbable but completely impossible, creating troughs of zero likelihood in the likelihood surface for certain dated-phylogeny-character patterns (Buerki et al. 2011). In a biogeographic dating framework, this might unintentionally eliminate large numbers of

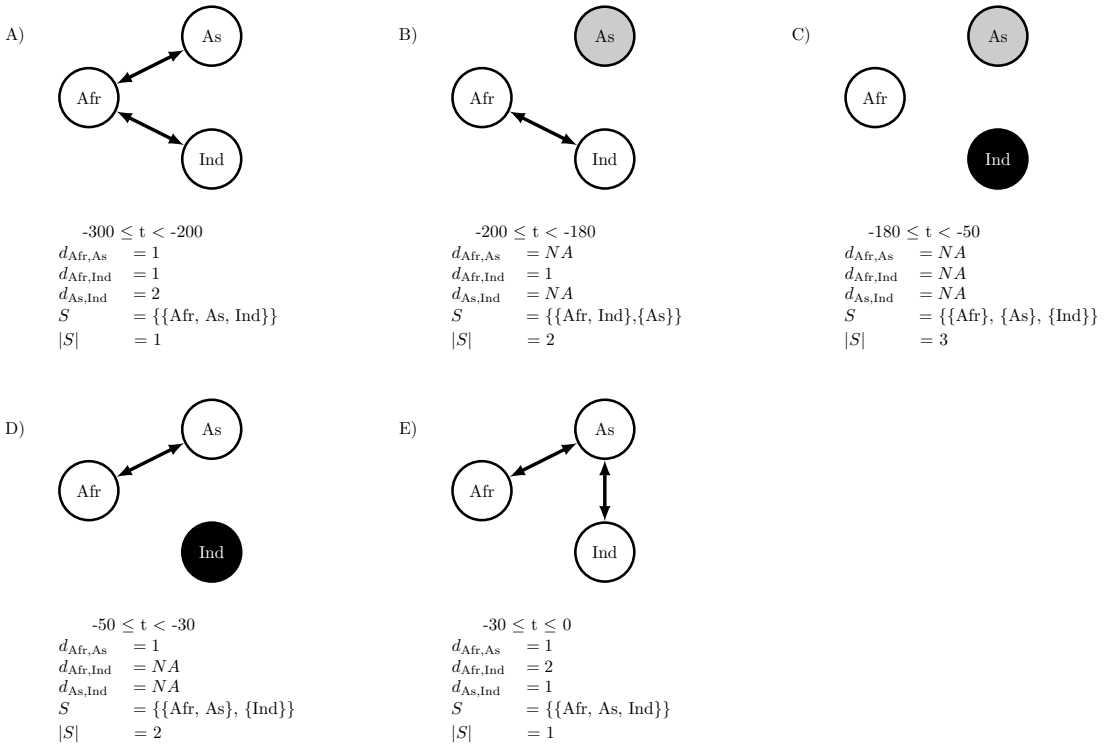


Figure 2.3: **Biogeographic communicating classes.** Dispersal routes shared by Africa (Afr), Asia (As), and India (Ind) are depicted for each time interval, t, over the past 300 Ma. Dispersal path lengths between areas i and j are given by $d_{i,j}$, with NA meaning there is no route between areas (areas i and j are mutually unreachable). communicating classes per interval are given by S and by the shared coloring of areas (vertices), with $|S|$ being the number of communicating classes.

speciation scenarios from the space of possible hypotheses, resulting in distorted estimates. To avoid these problems, I take the dispersal graph as the weighted average of three distinct dispersal graphs assuming short, medium, or long distance dispersal modes, each with their own set of communicating classes (see Section 2.2).

Fundamentally, biogeographic dating depends on how rapidly a species may disperse between two areas, and how that dispersal rate changes over time. In one extreme case, dispersals between mutually unreachable areas do not occur after infinite time, and hence have zero probability. At the other extreme, when dispersal may occur between any pair

of areas with equal probability over all time intervals, then paleogeography does not favor nor disfavor dispersal events (nor divergence events, implicitly) to occur during particular time intervals. In intermediate cases, so long as dispersal probabilities between areas vary across time intervals, the dispersal process informs when and what dispersal (and divergence) events occur. For instance, the transition probability of going from area i to j decreases as the average path length between i and j increases. During some time intervals, the average path length between two areas might be short, thus dispersal events occur more freely than when the average path is long. Comparing Figures 2.3A and 2.3E, the minimum number of events required to disperse from India to Africa is smaller during the Triassic ($t = -250$) than during the present ($t = 0$), and thus would have a relatively higher probability given the process operated for the same amount of time today (e.g. for a branch with the same length).

The concepts of adjacency, reachability, components, and communicating classes are not necessary to structure the rate matrix such that biogeographic events inform divergence times, though their simplicity is attractive. One could yield similar effects by parameterizing dispersal rates as functions of more complex area features, such as geographical distance between areas (Landis et al. 2013) or the size of areas (Tagliacollo et al. 2015). In this study, these concepts serve the practical purpose of summarizing perhaps the most salient feature of global paleogeography—that continents were not always configured as they were today—but also illuminate how time-heterogeneous dispersal rates produce transition probabilities that depend on geological time, which in turn inform the dates of speciation times.

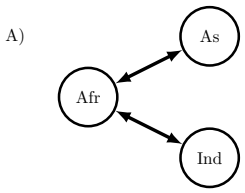
Time-heterogeneous dispersal process

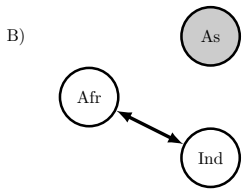
Let $\mathbf{Z}$ be a vector reporting biogeographic states for $M > 2$ taxa. The objective is to construct a time-heterogeneous CTMC where transition probabilities depend on time-calibrated paleogeographical features. For simplicity, species ranges are assumed to be endemic on the continental scale, so each taxon's range may be encoded as an integer in $Z_i \in \{1, 2, \ldots, N\}$, where N is the number of areas.

The paleogeographical features that determine the dispersal process rates are assumed to be a piecewise-constant model, sometimes called a stratified (Ree et al. 2005; Ree and Smith 2008) or epoch model (Bielejec et al. 2014), where $K - 1$ breakpoints are dated in geological time to create K time intervals. These breakpoint times populate the vector, $\tau = (\tau_0 = -\infty, \tau_1, \tau_2, \ldots, \tau_{K-1}, \tau_K = 0)$, with the oldest interval spanning deep into the past, and the youngest interval spanning to the present.

While a lineage exists during the k^{th} time interval, its biogeographic characters evolve according to that interval's rate matrix, $\mathbf{Q}(k)$, whose rates are informed by paleogeographical features present throughout time $\tau_{k-1} \leq t < \tau_k$. As a example of an paleogeographically-informed matrix structure, take $\mathbf{G}(k)$ to be a adjacency matrix indicating 1 when dispersal

may occur between two areas and 0 otherwise, during time interval k. This adjacency matrix is equivalent to an undirected graph where areas are vertices and edges are dispersal routes. Full examples of $\mathbf{G} = (\mathbf{G}(1), \mathbf{G}(2), \ldots, \mathbf{G}(K))$ describing Earth's paleocontinental adjacencies are given in detail later.





-300 ≤ t < -200

$\mu\mathbf{Q}(\mathrm{k}) =$

	Afr	As	Ind
Afr	-	0.5	0.5
As	0.5	-	0.0
Ind	0.5	0.0	-

$\mathbf{P}(-250, -249) =$

	Afr	As	Ind
Afr	0.48	0.26	0.26
As	0.26	0.67	0.07
Ind	0.26	0.07	0.67

-200 ≤ t < -180

$\mu\mathbf{Q}(\mathrm{k+1}) =$

	Afr	As	Ind
Afr	-	0.0	0.5
As	0.0	-	0.0
Ind	0.5	0.0	-

$\mathbf{P}(-190, -189) =$

	Afr	As	Ind
Afr	0.68	0.00	0.32
As	0.00	1.00	0.00
Ind	0.32	0.00	0.68

Figure 2.4: **Piecewise-constant dispersal rate matrices.** Dispersal routes shared by Africa (Afr), Asia (As), and India (Ind) are depicted for two time intervals, $-300 \leq t < -200$ and $-200 \leq t < 180$. Graphs and times correspond to those in Figure 2.3A,B. Transition probabilities are computed for a unit time during different epochs with a time-homogeneous biogeographic clock rate $\mu = 0.5$. A) The three areas are connected and all transitions have positive probability. B) As is unreachable from Afr and Ind, so transition probabilities into and out of As are zero.

With the paleogeographical vector $\mathbf{G}$, I define the transition rates of $\mathbf{Q}(k)$ as equal to $\mathbf{G}_z(k)$. Similar rate matrices are constructed for all K time intervals that contain possible supported root ages for the phylogeny, Ψ. Figure 2.4 gives a simple example for three areas, where Asia shares positive dispersal rate with Africa when they are merged and no dispersal while split.

For a piecewise-constant CTMC, the process' transition probability matrix is the product of transition probability matrices spanning m breakpoints. To simplify notation, let v be the vector marking important times of events, beginning with the start time of the branch, s, followed by the m breakpoints satisfying $s < \tau_k < t$, ending the the end time of the branch, t, such that $v = (s, \tau_k, \tau_{k+1}, \ldots, \tau_{k+m-1}, t)$, and let $u(v_i, \tau)$ be a "look-up" function that gives the index k that satisfies $\tau_{k-1} \leq v_i < \tau_k$. The transition probabilty matrix over the intervals

in v according to the piece-wise constant CTMC given by the vectors τ and $\mathbf{Q}$ is

$$\mathbf{P}_\tau(v, \mu; \tau, \mathbf{Q}) = \prod_{i=1}^{m+1} e^{\mu(v_{i+1}-v_i)\mathbf{Q}(u(v_i,\tau))}$$

The pruning algorithm (Felsenstein 1981) is agnostic as to how the transition probabilties are computed per branch, so introducing the piecewise-constant CTMC does not prohibit the efficient computation of phylogenetic model likelihoods. See Bielejec et al. (2014) for an excellent review of piecewise-constant CTMCs as applied to phylogenetics.

In the above case, the times s and t are generally identifiable from μ_z so long as $\mathbf{P}_\tau(v, \mu; \tau, \mathbf{Q}) \neq \mathbf{P}_\tau(v', \mu'; \tau, \mathbf{Q})$ for any supported values of v, μ, v', and μ'. Note, I include μ as an explicit parameter in the transition probability matrix function for clarity, though they are suppressed in standard CTMC notation when t equals the product of rate and time, then the process effectively runs for the time, $\mu(t - s)$. For example, assume that $\mathbf{Q}$ is a time-homogeneous Jukes-Cantor model with no paleogeographical constraints, i.e. all transition rates are equal independent of k. The transition probability matrix for this model is readily computed via matrix exponentiation

$$\mathbf{P}(s, t, \mu) = e^{\mu(t-s)\mathbf{Q}}.$$

Note that $\mathbf{P}(s, t, \mu) = \mathbf{P}_\tau(v, \mu; \tau, \mathbf{Q})$ when $v = (s, t)$ – i.e. the time-heterogeneous process spans no breakpoints when $m = 0$ and is equivalent to a time-homogeneous process for the interval (s, t).

For a time-homogeneous model, multiplying the rate and dividing the branch length by the same factor results in an identical transition probability matrix. In practice this means the simple model provides no information for the absolute value of μ and the tree height of Ψ, since all branch rates could likewise be multiplied by some constant while branch lengths were divided by the same constant, i.e. $\mathbf{P}(s, t, 1) = \mathbf{P}(s\mu^{-1}, t\mu^{-1}, \mu)$. Similarly, since $\mathbf{P}(s, t, \mu) = \mathbf{P}(s + c, t + c, \mu)$ for $c \geq 0$, the absolute time when the process begins does not matter, only the amount of time that has elapsed. Extending a branch length by a factor of c requires modifying other local branch lengths in kind to satisfy time tree constraints, so the identifiability of the absolute time interval (s, t) depends on how "relaxed" (Drummond et al. 2006) the assumed clock and divergence time priors are with respect to the magnutude of c, which together induce some (often unanticipated) joint prior distribution on divergence times and branch rates (Heled and Drummond 2012; Warnock et al. 2015). In either case, rate and time estimates under the time-homogeneous process result from the induced prior distributions rather than by informing the process directly.

Adjacent-area terrestrial dispersal graph

I identified $K = 26$ times and $N = 25$ areas to capture the general features of continental drift and its effects on terrestrial dispersal (Figure 2.5; for all graphs and a link to the animation, see Supplemental Figure S1). All adjacencies were constructed visually, referencing Blakey (2006) and GPlates (Boyden et al. 2011), then corroborated using various paleogeographical sources (Table S2). The paleogeographical state per time interval is summarized as an undirected graph, where areas are vertices and dispersal routes are edges.

To proceed, I treat the paleogeographical states over time as a vector of adjacency matrices, where $\mathbf{G}_\bullet(k)_{i,j} = 1$ if areas i and j share an edge at time interval k, and $\mathbf{G}_\bullet(k)_{i,j} = 0$ otherwise. Temporarily, I suppress the time index, k, for the rate matrix $\mathbf{Q}(k)$, since all time intervals' rate matrices are constructed in a similar manner. To mitigate the effects of model misspecification, $\mathbf{Q}$ is determined by a weighted average of three geological adjacency matrices

$$\mathbf{G}_z = b_s\mathbf{G}_s + b_m\mathbf{G}_m + b_l\mathbf{G}_l \tag{2.1}$$

where s, m, and l correspond to short distance, medium distance, and long distance mode parameters.

Short, medium, and long distance dispersal processes encode strong, weak, and no geographical constraint, respectively. As distance-constrained mode weights b_s and b_m increase, the dispersal process grows more informative of the process' previous state or communicating class (Figure 2.6). The vector of short distance dispersal graphs, $\mathbf{G}_s = (\mathbf{G}_s(1), \mathbf{G}_s(2), \ldots, \mathbf{G}_s(K))$, marks adjacencies for pairs of areas allowing terrestrial dispersal without travelling through intermediate areas (Figure 2.6A). Medium distance dispersal graphs, $\mathbf{G}_m$, include all adjacencies in $\mathbf{G}_s$ in addition to adjacencies for areas separated by lesser bodies of water, such as throughout the Malay Archipelago, while excluding transoceanic adjacencies, such as between South America and Africa (Figure 2.6B). Finally, long distance dispersal graphs, $\mathbf{G}_l$, allow dispersal events to occur between any pair of areas, regardless of potential barrier (Figure 2.6C).

To average over the three dispersal modes, b_s, b_m, and b_l are constrained to sum to 1, causing all elements in $\mathbf{G}_z$ to take values from 0 to 1 (Eqn 2.1). Importantly, adjacencies specified by $\mathbf{G}_s$ always equal 1, since those adjacencies are also found in $\mathbf{G}_m$ and $\mathbf{G}_l$. This means $\mathbf{Q}$ is a Jukes-Cantor rate matrix only when $b_l = 1$, but becomes increasingly paleogeographically-structured as $b_l \to 0$. Non-diagonal elements of $\mathbf{Q}$ equal those of $\mathbf{G}_z$, but are rescaled such that the average number of transitions per unit time equals 1, and diagonal elements of $\mathbf{Q}$ equal the negative sum of the remaining row elements. To compute transition probabilities, $\mathbf{Q}$ is later rescaled by a biogeographic clock rate, μ, prior to matrix exponentiation. The effects of the weights b_s, b_m, and b_l on dispersal rates between areas are shown in Figure 2.7.

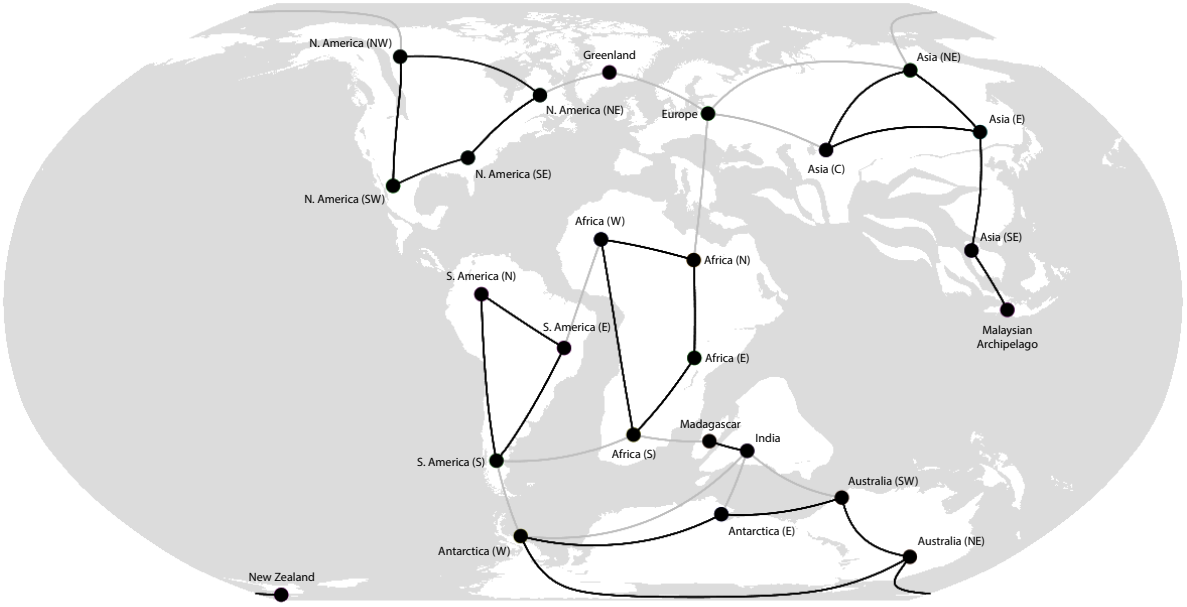


Figure 2.5: **Dispersal graph for Epoch 14, 110–100Ma: India and Madagascar separate from Australia and Antarctica.** A `gplates` (Gurnis et al. 2012) screenshot of Epoch 14 of 26 is displayed. Areas are marked by black vertices. Black edges indicate both short- and medium distance dispersal routes. Gray edges indicate exclusively medium distance dispersal routes. Long distance dispersal routes are not shown, but are implied to exist between all area-pairs. The short, medium, and long dispersal graphs have 8, 1, and 1 communicating classes, respectively. India and Madagascar each have only one short distance dispersal route, which they share. Both areas maintain medium distance dispersal routes with various Gondwanan continents during this epoch. The expansion of the Tethys Sea impedes dispersal into and out of Europe.

By the argument of that continental break-up (i.e. the creation of new communicating classes; Figure 2.1) introduces a bound on the minimum age of divergence, and that continental joining (i.e. unifying existing communicating classes; Figure 2.2) introduces a bound on the maximum age of divergence, then the paleogeographical model I constructed has the greatest potential to provide both upper and lower bounds on divergence times when the number of communicating classes is large, then small, then large again. This coincides with

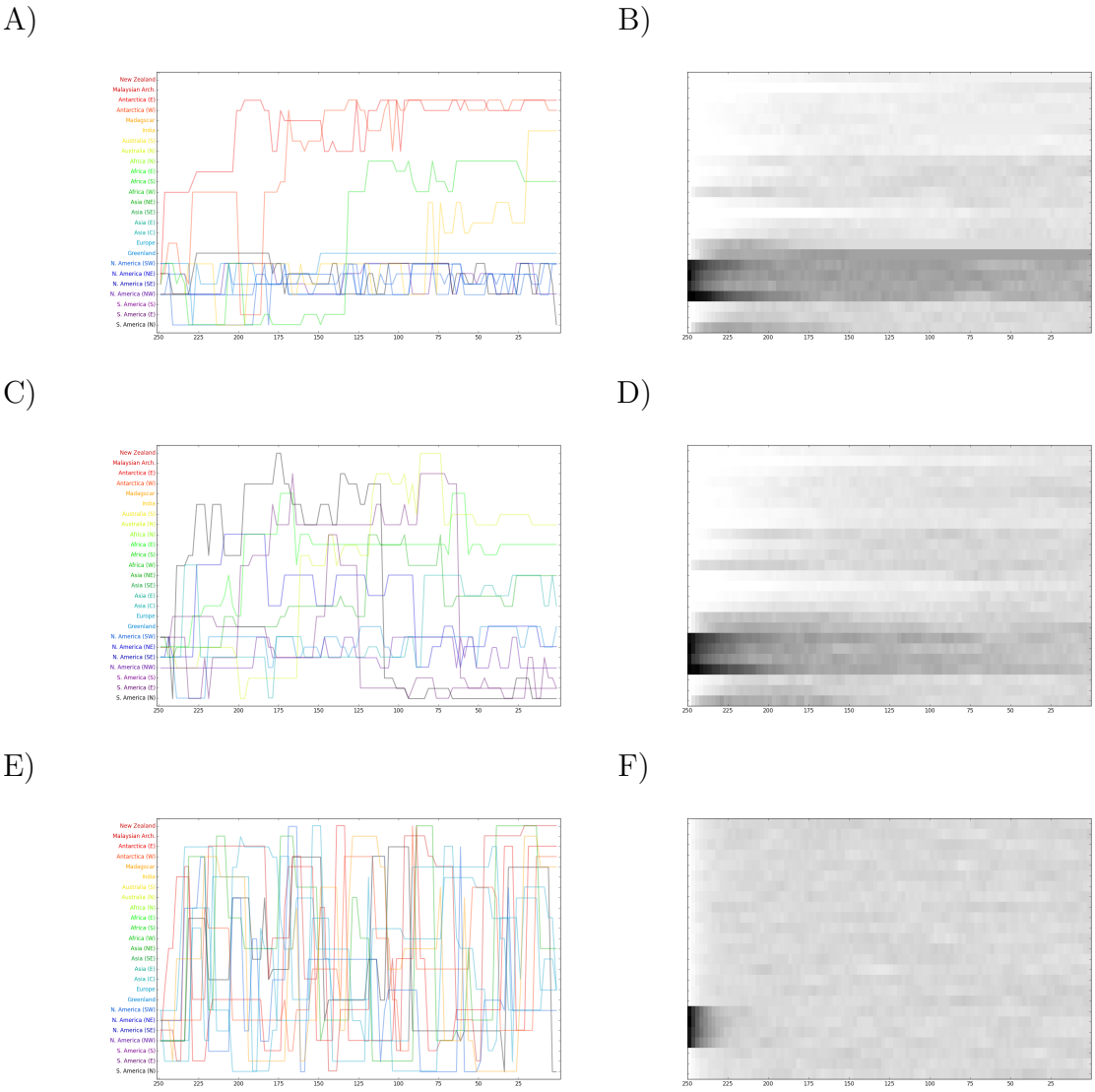
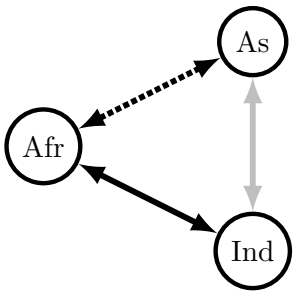


Figure 2.6: **Sample paths for paleogeographically informed biogeographic process.** The top, middle, and bottom panels show dispersal histories simulated by the pure short (A,B), medium (C,D), and long (E,F) distance process components. All processes originate in one of the four North American areas 250 Ma. The left column shows 10 of 2000 sample paths. Color indicates the area the lineage is found in the present (A,C,E). Colors for areas match those in Figure 2.8. The right column heatmap reports the sample frequencies for any of the 2000 dispersal process being in that state at that time (B,D,F).



$$b_s + b_m + b_l = 1.0$$

$$\begin{aligned} q_{\text{Afr,Ind}} &= 0.7 + 0.2 + 0.1 = 1.0 \\ q_{\text{Afr,As}} &= 0.2 + 0.1 = 0.3 \\ q_{\text{As,Ind}} &= 0.1 = 0.1 \end{aligned}$$

$$\mathbf{Q} = \begin{bmatrix} - & 0.3 & 1.0 \\ 0.3 & - & 0.1 \\ 1.0 & 0.1 & - \end{bmatrix}$$

Figure 2.7: **Example mode-weighted dispersal matrix.** Short, medium, and long distance dispersal edges are represented by solid black, dashed black, and solid gray lines, respectively. Short, medium, and long distance dispersal weights are $(b_s, b_m, b_l) = (0.7, 0.2, 0.1)$. The resulting mode-weighted dispersal matrix, $\mathbf{Q}$, is computed with areas (states) ordered as (Afr, As, Ind). Afr and Ind share a short distance dispersal edge, therefore the dispersal weight is $b_s + b_m + b_l = 1.0$. Afr and As share a medium distance edge with dispersal weight $b_m + b_l = 0.3$. Dispersal between As and Ind is only by long distance with weight $b_l = 0.1$.

the formation of Pangaea, dropping from 8 to 3 communicating classes at 280 Ma, followed by the fragmentation of Pangaea, increasing from 3 to 11 communicating classes between 170 Ma and 100 Ma (Figure 2.8). It is important to consider this bottleneck in the number of communicating classes will be informative of root age only for fortuitous combinations of species range and species phylogeny. Just as some clades lack a fossil record, others are bound to lack a biogeographic record that is informative of origination times.

2.3 Analysis

All posterior densities were estimated using Markov chain Monte Carlo (MCMC) as implemented in RevBayes, available at `revbayes.com` (Höhna et al. 2014). Data and analysis scripts are available at `github.com/mlandis/biogeo_dating`. Datasets are also available on Dryad at `datadryad.org/XXX`. Analyses were performed on the XSEDE supercomputing cluster (Towns et al. 2014).

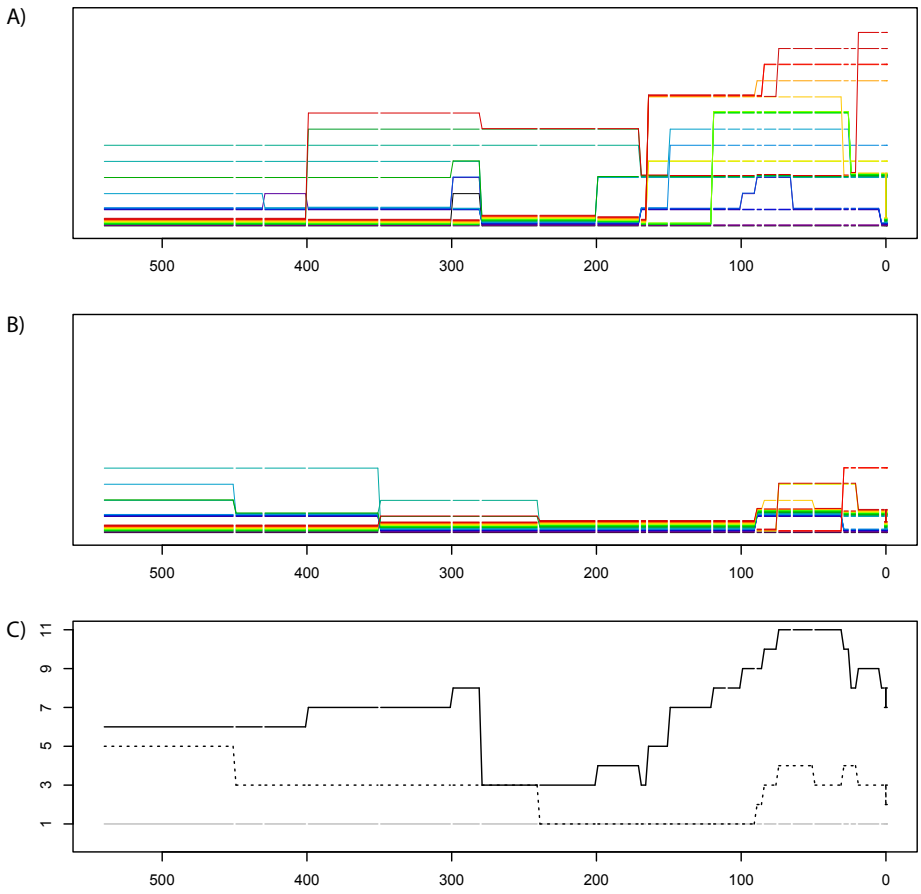


Figure 2.8: **Dispersal graph properties summarized over time.** communicating classes of short distance dispersal graph (A) and medium distance dispersal graph (B) are shown. Each of 25 areas is represented by one line. Colors of areas match those listed in Figure 2.6. Grouped lines indicate areas in one communicating class during an interval of time. Vertical lines indicate transitions of areas joining or leaving communication clases, i.e. due to paleogeographical events. When no transition event occurs for an area entering a new epoch, the line is interrupted with gap. (C) Number of communicating classes: the black line corresponds to the short distance dispersal graph (A), the dotted line corresponds to medium distance dispersal graph (B), and the gray line corresponds to the long distance dispersal graph, which always has one communicating class.

Simulation

Through simulation I tested whether biogeographic dating identifies rate from time. To do so, I designed the analysis so divergence times are informed solely from the molecular

and biogeographic data and their underlying processes (Table 2.1). As a convention, I use the subscript x to refer to molecular parameters and z to refer to biogeographic parameters. Specifically, I defined the molecular clock rate as $\mu_x = e/r$, where e gives the expected number of molecular substitutions per site and r gives the tree height. Both e and r are distributed independently by uniform priors on $(0, 1000)$. Biogeographic events occur with rate, $\mu_z = \mu_x 10^{s_z}$ where s_z has a uniform prior distribution on $(-3, 3)$. To further subdue effects from the prior on posterior parameter estimates, the tree prior assigns equal probability to all node age distributions. No node calibrations were used. Each dataset was analyzed with (+G) and without (–G) the paleogeographic-dependent dispersal process.

Two further assumptions were made to simplify the analyses. First, although divergence times were free to vary, the tree topology was assumed to be known. Second, molecular and biogeographic characters evolve by strict global clocks. In principle, inferring the topology or using relaxed clock models should increase the variance in posterior divergence time estimates, but not greatly distort the performance of –G relative to +G.

Phylogenies with $M = 50$ extant taxa were simulated using a birth-death process with birth rate, $\lambda = 0.25$, and death rate, $\mu = 0.15$, then rescaled so the root age equaled 250 Ma. Each dataset contained 500 nucleotides and 1 biogeographic character. Biogeographic characters were simulated under +G, where $\mathbf{G}_z$ is defined as piecewise-constant over 25 areas and 26 time intervals in the manner described in Section 2.2. In total, I simulated 100 datasets under the parameter values given in Table 2.1, where these values were chosen to reflect possible empirical estimates. Each dataset was analyzed under each of two models, then analyzed a second time to verify convergence (Gelman and Rubin 1992; Plummer et al. 2006). When summarizing posterior results, posterior mean-of-median and 95% highest posterior density (HPD95%) values were presented.

As expected, the results show the –G model extracts no information regarding the root age, so its posterior distribution equals its prior distribution, mean-of-median ≈ 499 (Figure 2.9A). In contrast, the +G model infers the mean-of-median root age 243 with a HPD95% interval width of 436, improving accuracy and precision in general.

Estimated divergence time accuracy was assessed with the statistic

$$d = \sum_i \frac{a_i - a_i^{(\text{true})}}{a_i^{(\text{true})}} \tag{2.2}$$

where a is a posterior sample of the node age vector and a_{true} is the true node age vector known through simulation. When a perfectly estimates $a^{(\text{true})}$ for all node ages, $d = 0$. When estimated node ages are too young (on average), $d < 0$, and when too old, $d > 0$. Inference under +G infers an mean $d = 0.19$ with a HPD95% interval width of ≈ 1.26, while –G performs substantially worse with $d = 0.92$ and width ≈ 2.75 (Figure 2.9B). Posterior estimates generally favored short over medium and long distance dispersal as

			Simulation		**Empirical**
	Parameter	X	$f(X)$	sim. value	$f(X)$
Tree	Root age	r	Uniform$(0, 1000)$	250	Uniform$(0, 540)$ or Uniform$(151.7, 251.4)$
	Time tree	Ψ	UniformTimeTree(r)	BD($\lambda = 0.25$,$\mu = 0.15$)	UniformTimeTree(r)
Molecular	Length	e	Uniform$(0, 1000)$	2.5	Uniform$(0, 1000)$
	Subst. rate	μ_x	e/r	determined (0.01)	e/r
	Exch. rates	r_x	Dirichlet(10)	from prior	Dirichlet(10)
	Stat. freqs	π_x	Dirichlet(10)	from prior	Dirichlet(10)
	Rate matrix	$\mathbf{Q}_x$	GTR(r_x, π_x)	determined	GTR(r_x, π_x)
	Branch rate mult.	$\rho_{x,i}$			Lognorm$(\ln\mu_x - \sigma_x^2/2, \sigma_x)$
	Branch rate var.	σ_x			Exponential(0.1)
	$+\Gamma 4$	Γ_x			Gamma(α, α)
	$+\Gamma 4$ hyperprior	α			Uniform$(0, 50)$
Biogeo.	Atlas-graph	$G(t)$	–G or +G	+G	+G
	Biogeo. rate	μ_z	$\mu_x 10^{s_z}$	determined (0.1)	$\mu_x 10^{s_z}$
	Biogeo. rate mod.	s_z	Uniform(-3, 3)	1.0	Uniform(-3, 3)
	Dispersal mode	(b_s, b_m, b_l)	Dirichlet(1)	(1000,10,1)/1011	Dirichlet$(1, 1, 1)$ or Dirichlet$(100, 10, 1)$
	Dispersal rates	$r_z(t)$	$\sum_{\{s,m,l\}} b_i G_i(t)$	determined	$\sum_{\{s,m,l\}} b_i G_i(t)$
	Stat. freqs	π_z	$(1, \ldots, 1)/25$	$(1, \ldots, 1)/25$	$(1, \ldots, 1)/25$
	Rate matrix	$\mathbf{Q}_z(t)$	GTR$(r_z(t), \pi_z)$	determined	GTR$(r_z(t), \pi_z)$

Table 2.1: **Model parameters.** Model parameter names and prior distributions are described in the manuscript body. All empirical priors were identical to simulated priors unless otherwise stated. Priors used for the empirical analyses but not simulated analyses are left blank. Determined means the parameter value was determined by other model parameters.

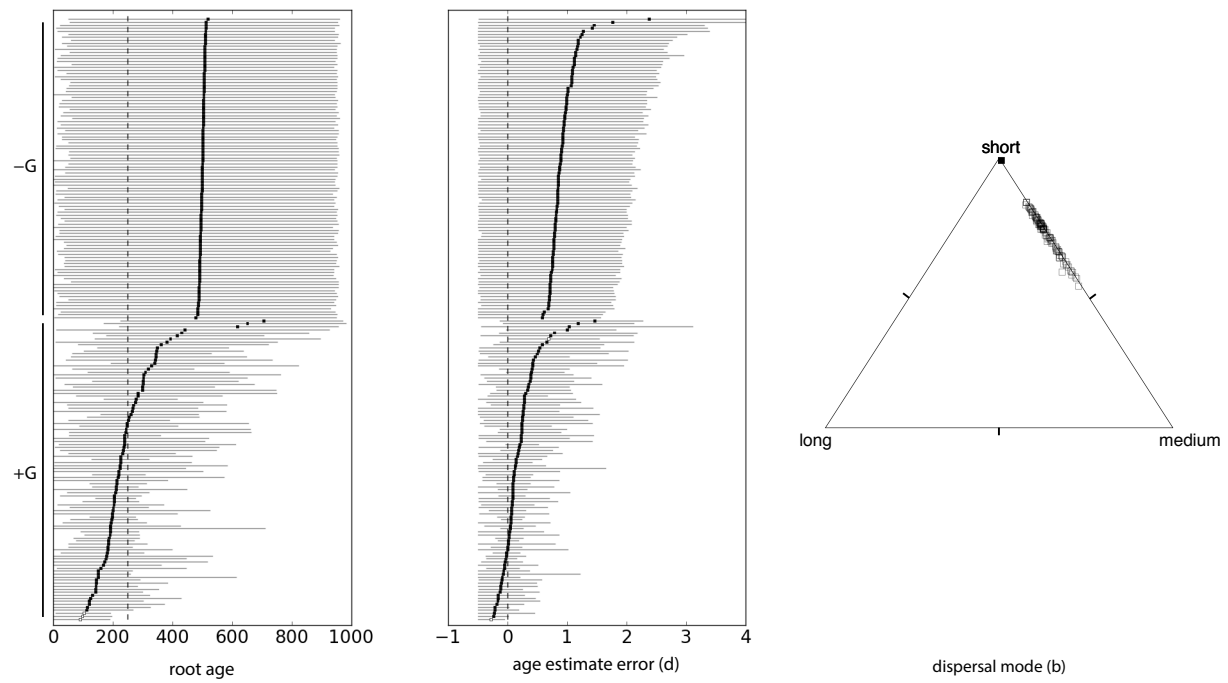


Figure 2.9: **Posterior estimates for simulated data.** A) Posterior estimates of root age. The true root age for all simulations is 250 Ma (dotted vertical line). B) Posterior estimates of relative node age error (Eqn 2.2). The true error term equals zero. Both A and B) –G analyses are on the top half, +G analyses are on the bottom. Each square marks the posterior mean root age estimate with the HPD95% credible interval. If the credible interval contains the true value, the square is filled. C) Posterior estimates of dispersal mode proportions for the +G simulations projected onto the unit 2-simplex. The filled circle gives the posterior median-of-medians, and the empty circles give posterior medians.

was assumed under simulation (Figure 2.9C). Dispersal mode parameter estimates were $(b_s, b_m, b_l) = (0.766, 0.229, 0.003)$, respectively, summarized as median-of-medians across simulated replicates.

Empirical: *Testudines*

To assess the accuracy of the method, I performed a biogeographic dating analysis on extant turtle species (*Testudines*). Extant turtles fall into two clades, *Pleurodira*, found in the Southern hemisphere, and *Cryptodira*, found predominantly in the Northern hemisphere. Their modern distribution shadows their biogeographic history, where *Testudines* are thought to be Gondwanan in origin with the ancestor to cryptodires dispersing into Laurasia during the Jurassic (Crawford et al. 2015). Since turtles preserve so readily in the fossil record, estimates of their phylogeny and divergence times have been profitably analyzed and re-analyzed by various researchers (Joyce 2007; Hugall et al. 2007; Danilov and Parham 2008; Alfaro et al. 2009; Dornburg et al. 2011; Joyce et al. 2013; Sterli et al. 2013; Warnock et al. 2015). This makes them ideal to assess the efficacy of biogeographic dating, which makes no use of their replete fossil record: if both biogeography-based and fossil-based methods generate similar results, they co-validate each others' correctness (assuming they are not both biased in the same manner).

To proceed, I assembled a moderately sized dataset. First, I aligned cytochrome B sequences for 185 turtle species (155 cryptodires, 30 pleurodires) using MUSCLE 3.8.31 (Edgar 2004) under the default settings. Assuming the 25-area model presented in Section 2.2, I consulted GBIF (`gbif.org`) and IUCN Red List (`iucnredlist.org`) to record the area(s) in which each species was found. Species occupying multiple areas were assigned ambiguous tip states for those areas. Missing data entries were assigned to the six sea turtle species used in this study to effectively eliminate their influence on the (terrestrial) biogeographic process. To simplify the analysis, I assumed the species tree topology was fixed according to Guillon et al. (2012), which was chosen for species coverage, pruning away unused taxa. All speciation times were considered random variables to be estimated. The tree topology and biogeographic states are shown in Supplemental Figure S2. All data are recorded on `datadryad.org/XXX`.

Like the simulation study, my aim is to show that the paleogeographically-aware +G model identifies the root age in units of absolute time. To reiterate, the posterior root age should be identical to the prior root age when the model cannot inform the root age. If the prior and posterior differ, then the data under the model are informative. The root age was constrained to Uniform(0, 540), forbidding the existence of Precambrian turtles. To improve biological realism, I further relaxed assumptions about rate variability for the molecular model of evolution, both among sites (Yang 1994) and across branches (Lepage et al. 2007; Drummond et al. 2006) (Table 2.1).

Biogeographic dating infers a posterior median root age of 201 with HPD95% credible interval of (115, 339) (Figure 2.10A). This is consistent current root age estimates informed from the fossil record (Figure 2.11). The posterior mode of dispersal mode is $(b_s, b_m, b_l) = (0.47, 0.51, 0.02)$, with short and medium distance dispersal occurring at approximately equal

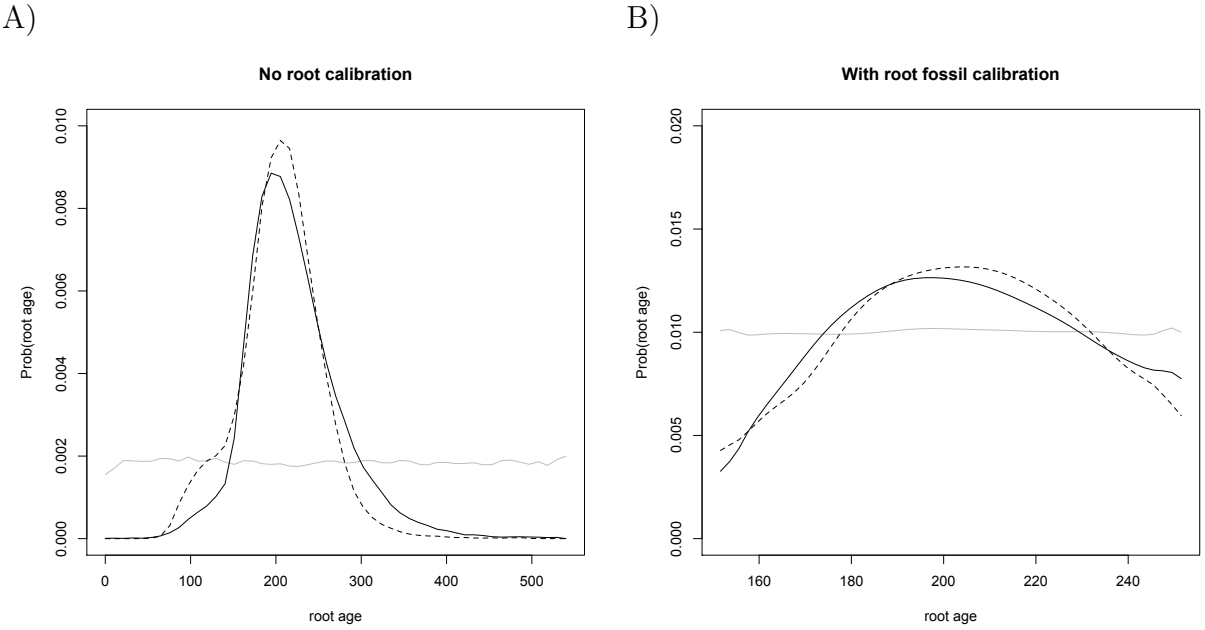


Figure 2.10: **Posterior root age of turtles by biogeographic dating.** Six root age posterior estimates were computed using biogeographic dating, each using variations on flat or short-biased priors for key parameters. Figure A assumes no knowledge of fossils with Uniform(0, 540) root age prior. Figure B follows Joyce et al. (2013) and assumes Uniform(151.7, 251.4) as a root node age calibration. The black solid posterior density assumes a flat prior on dispersal mode. The black dotted posterior density assumes an short-biased prior Dirichlet(100, 10, 1) on dispersal mode. The gray solid posterior density ignores paleogeography.

rates and long distance dispersal being rare by comparison. Biogeographic events occurred at a ratio of about 6:1 when compared to molecular events (posterior means: $\mu_x = 1.9\text{E}{-3}, \mu_z = 1.1\text{E}{-2}$). The posterior mode tree height measured in expected number of dispersal events is 2.3 with HPD95% (1.5, 3.0), i.e. as a treewide average, the current location of each taxon is the result of about two dispersal events.

The flat prior distribution for competing dispersal modes is Dirichlet(1, 1, 1) and does not capture the intuition that short distance dispersal should be far more common than long distance dispersal. I encoded this intuition in the dispersal mode prior, setting the distribution to Dirichlet(100, 10, 1), which induces expected proportion of 100:10:1 short:medium:long dispersal events. After re-analyzing the data with the short-biased dispersal prior, the posterior median and HPD95% credible interval were estimated to be, respectively, 204 $(96, 290)$ (Figure 2.10A).

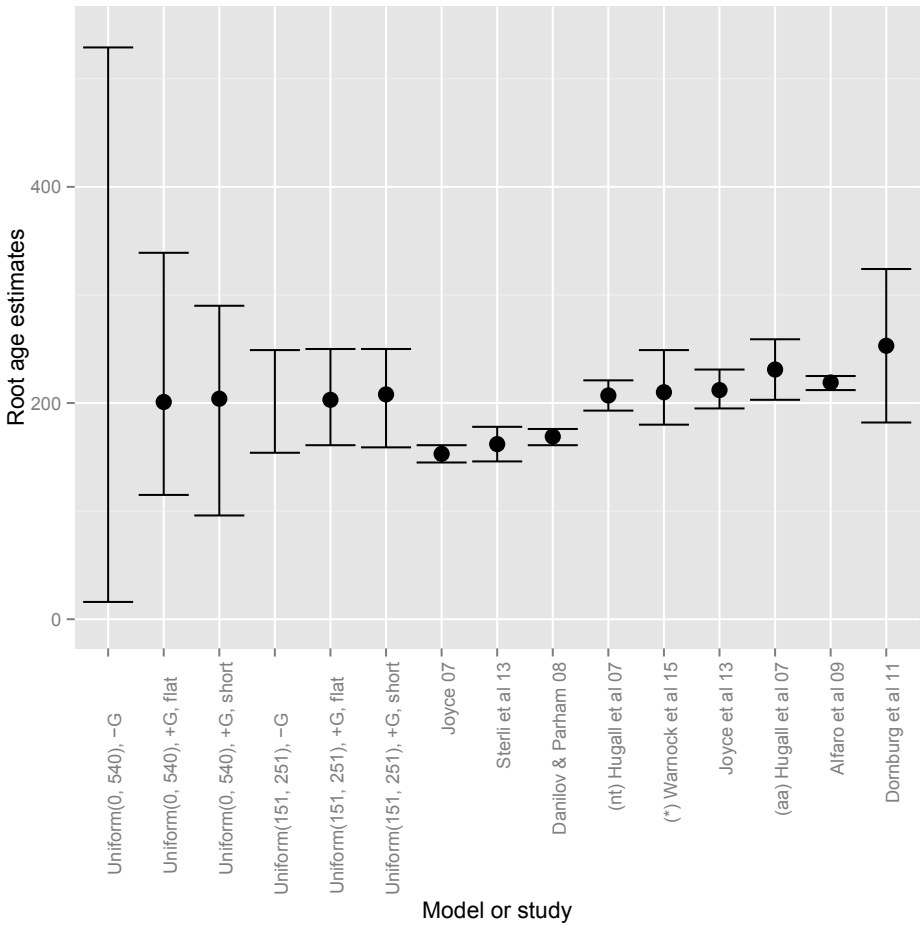


Figure 2.11: **Root age comparison.** Root age estimates are presented both for analysis conducted for this manuscript and as reported in existing publications. Existing estimates are as reported in Sterli et al. (2013) and supplemented recently reported results. Points and whiskers correspond to the point estimates and estimate confidence, which varies across analyses. The six left estimates were computed using biogeographic dating, each using variations on flat or short-biased priors for key parameters. Two of these analyses ignore paleogeography (–G) so the posterior root age is the uniform prior root age, whose mode (not shown) equals all values supported by the prior. Hugall et al. (2007) reports ages for analyses using amino acids (aa) and nucleotides (nt). Warnock et al. (2015) reports many estimates while exploring prior sensitivity, but only uniform prior results are shown here.

Biogeographic dating is compatible with fossil dating methods, so I repeated the analysis for both flat and informative prior dispersal modes while substituting the Uniform(0, 540)

prior on root age calibration for Uniform(151.7, 251.4) (Joyce et al. 2013). When taking biogeography into account, the model more strongly disfavors post-Pangaean origins for the clade than when biogeography is ignored, but the effect is mild. Posterior distributions of root age was relatively insensitive to the flat and short-biased dispersal mode priors, with posterior medians and credible intervals of 203 (161, 250) and 208 (159, 250), respectively.

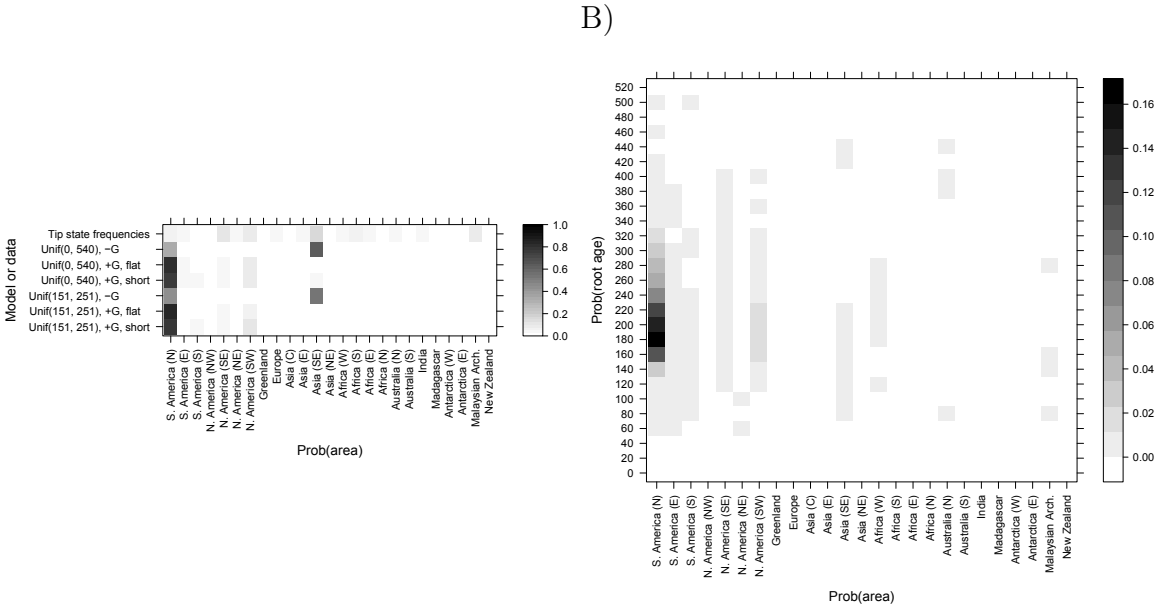


Figure 2.12: **Root state estimates.** A) Posterior probabilities of root state are given for the six empirical analyses. B) Joint-marginal posterior probabilities of root age and root state are given for the empirical analysis without a root calibration and with a flat dispersal mode prior. Root ages are binned into intervals of width 20.

All posterior root state estimates favored South America (N) for the paleogeographically-informed analyses (Figure 2.12A). Although this is in accord with the root node calibration adopted from Joyce et al. (2013)—*Caribemys oxfordiensis*, sampled from Cuba, and the oldest accepted crown group testudine—the fossil is described as a marine turtle, so the accordance may simply be coincidence. In contrast, the paleogeographically-naive models support Southeast Asian origin of *Testudines*, where, incidentally, Southeast Asia is the most frequently inhabited area among the 185 testudines. For the analysis with a flat dispersal mode prior and no root age calibration, all root states with high posterior probability appear to concur on the posterior root age density (Figure 2.12B), i.e. regardless of conditioning on South America (N), North America (SE), or North America (SW) as a root state, the posterior root age density is roughly equal.

2.4 Discussion

The major obstacle preventing the probabilistic union of paleogeographical knowledge, biogeographic inference, and divergence time estimation has been methodological, which I have attempted to redress in this manuscript. The intuition justifying prior-based fossil calibrations (Parham et al. 2011), i.e. that fossil occurrences should somehow inform divergence times, has recently been formalized into several models (Pyron 2011; Ronquist et al. 2012a; Heath et al. 2014). Here I present an analogous treatment for prior-based biogeographic calibrations, i.e. that biogeographic patterns of modern species echo time-calibrated paleobiogeographic events, by describing how epoch models (Ree et al. 2005; Ree and Smith 2008; Bielejec et al. 2014) are informative of absolute divergence times. Briefly, I accomplished this using a simple time-heterogeneous dispersal process (Sanmartín et al. 2008), where dispersal rates are piecewise-constant, and determined by a graph-based paleogeographical model (Section 2.2). The paleogeographical model itself was constructed by translating various published paleogeographical reconstructions (Figure 2.5) into a time-calibrated vector of dispersal graphs.

Through simulation, I showed biogeographic dating identifies tree height from the rates of molecular and biogeographic character change. This simulation framework could easily be extended to investigate for what phylogenetic, paleogeographic, and biogeographic conditions one is able to reliably extract information for the root age. For example, a clade with taxa invariant for some biogeographic state would contain little to no information about root age, provided the area has always existed and had a constant number of dispersal edges over time. At the other extreme, a clade with a very high dispersal rate or with a proclivity towards long distance dispersal might provide little due to signal saturation (Figure 2.6C). The breadth of applicability of biogeographic dating will depend critically on such factors, but because we do not expect to see closely related species uniformly distributed about Earth nor in complete sympatry, that breadth may not be so narrow, especially in comparison to the fossil record.

The majority of groups have poor fossil records, and biogeographic dating provides a second hope for dating divergence times. Since biogeographic dating does not rely on any fossilization process or data directly, it is readily compatible with existing fossil-based dating methods (Figure 2.10B). When fossils with geographical information are available, researchers have shown fossil taxa improve biogeographical analyses (Moore et al. 2008; Wood et al. 2012). In principle, the biogeographic process should guide placement of fossils on the phylogeny, and the age of the fossils should improve the certainty in estimates of ancestral biogeographic states (Slater et al. 2012), on which biogeographic dating relies. Joint inference of divergence times, biogeography, and fossilization stands to resolve recent paleobiogeographic conundrums that may arise when considering inferences separately (Beaulieu et al. 2013; Wilf and Escapa 2014).

Because time calibration through biogeographic inferences comes primarily from the paleogeographical record, not the fossil record, divergence times may be estimated from exclusively extant taxa under certain biogeographical and phylogenetic conditions. When fossils are available, however, biogeographic dating is compatible with other fossil-based dating methods (e.g. node calibrations, fossil tip dating, fossilized birth-death). As a proof of concept, I assumed a flat root age calibration prior for the origin time of turtles: the posterior root age was also flat when paleogeography was ignored, but Pangaean times of origin were strongly preferred when dispersal rates conditioned on paleogeography (Figure 2.10). Under the uninformative prior distributions on root age, biogeographic dating estimated turtles originated between the Mississipian (339 Ma) and Early Cretaceous (115 Ma) periods, with a median age of 201 Ma. Under an ignorance prior where short, medium, and long distance dispersal events have equal prior rates, short and medium distance dispersal modes are strongly favored over long distance dispersal. Posterior estimates changed little by informing the prior to strongly prefer short distance dispersal. Both with and without root age calibrations, and with flat and biased dispersal mode priors, biogeographic dating placed the posterior mode origin time of turtles at approximately 210–200 Ma, which is consistent with fossil-based estimates (Figure 2.11).

Model inadequacies and future extensions

The simulated and empirical studies demonstrate biogeographic dating improves divergence time estimates, with and without fossil calibrations, but many shortcomings in the model remain to be addressed. When any model is misspecified, inference is expected to produce uncertain, or worse, spurious results (Lemmon and Moriarty 2004), and biogeographic models are not exempted. I discuss some of the most apparent model misspecifications below.

Anagenetic range evolution models that properly allow species inhabit multiple areas should improve the informativeness of biogeographic data. Imagine taxa $T1$ and $T2$ inhabit areas $ABCDE$ and $FGHIJ$, respectively. Under the simple model assumed in this paper, the tip states are ambiguous with respect to their ranges, and for each ambiguous state only a single dispersal event is needed to reconcile their ranges. Under a pure anagenetic range evolution model (Ree et al. 2005), at least five dispersal events are needed for reconciliation. Additionally, some extant taxon ranges may span ancient barriers, such as a terrestrial species spanning both north and south of the Isthmus of Panama. This situation almost certainly requires a dispersal event to have occurred after the isthmus was formed when multiple-area ranges are used. For single-area species ranges coded as ambiguous states, the model is incapable of evaluating the likelihood that the species is found in both areas simultaneously, so additional information about the effects of the paleogeographical event on divergence times is potentially lost.

Any model where the diversification process and paleogeographical states (and events) are correlated will obviously improve divergence time estimates so long as that relationship is biogeographically realistic. Although the repertoire of cladogenetic models is expanding in terms of types of transition events, they do not yet account for geographical features, such as continental adjacency or geographical distance. Incorporating paleogeographical structure into cladogenetic models of geographically-isolated speciation, such as vicariance (Ronquist 1997), allopatric speciation (Ree et al. 2005; Goldberg et al. 2011), and jump dispersal (Matzke 2014), is crucial not only to generate information for biogeographic dating analyses, but also to improve the accuracy of ancestral range estimates. Ultimately, cladogenetic events are state-dependent speciation events, so the desired process would model range evolution jointly with the birth-death process (Maddison et al. 2007; Goldberg et al. 2011), but inference under these models for large state spaces is currently infeasible. Regardless, any cladogenetic range-division event requires a widespread range, which in turn implies it was preceeded by dispersal (range expansion) events. Thus, if we accept that paleogeography constrains the dispersal process, even a simple dispersal-only model will extract dating information when describing a far more complex evolutionary process.

That said, the simple paleogeographical model described herein (Section 2.2) has many shortcomings itself. It is only designed for terrestrial species originating in the last 540 Ma. Rates of dispersal between areas are classified into short, medium, and long distances, but with subjective criteria. The number of epochs and areas was limited by my ability to comb the literature for well-supported paleogeological events, while constrained by computational considerations. The timing of events was assumed to be known perfectly, despite the literature reporting ranges of estimates. Certainly factors such as global temperature, precipitation, ecoregion type, etc. affect dispersal rates between areas, but were ignored. All of these factors can and should be handled more rigorously in future studies by modeling these uncertainties as part of a joint Bayesian analysis (Höhna et al. 2014).

Despite these flaws, defining the paleogeographical model serves as an excercise to identify what features allow a biogeographic process to inform speciation times. Dispersal barriers are clearly clade-dependent, e.g. benthic marine species dispersal would be poorly modeled by the terrestrial graph. Since dispersal routes for the terrestrial graph might serve as dispersal barriers for a marine graph, there is potential for learning about mutually exclusive dispersal corridor use in a multi-clade analysis (Sanmartín et al. 2008). Classifying dispersal edges into dispersal mode classes may be made rigorous using clustering algorithms informed by paleogeographical features, or even abandoned in favor of modeling rates directly as functions of paleogeographical features like distance. Identifying significant areas and epochs remains challenging, where presumably more areas and epochs are better to approximate continuous space and time, but this is not without computational challenges (Ree and Sanmartín 2009; Webb and Ree 2012; Landis et al. 2013). Rather than fixing epoch event times to point estimates, one might assign empirical prior distributions based on collected estimates.

Ideally, paleogeographical event times and features would be estimated jointly with phylogenetic evidence, which would require interfacing phylogenetic inference with paleogeographical inference. This would be a profitable, but substantial, interdisciplinary undertaking.

Conclusion

Historical biogeography is undergoing a probabilistic renaissance, owing to the abundance of georeferenced biodiversity data now hosted online and the explosion of newly published biogeographic models and methods (Ree et al. 2005; Ree and Smith 2008; Sanmartín et al. 2008; Lemmon and Lemmon 2008; Lemey et al. 2010; Goldberg et al. 2011; Webb and Ree 2012; Landis et al. 2013; Matzke 2014; Tagliacollo et al. 2015). Making use of these advances, I have shown how patterns latent in biogeographic characters, when viewed with a paleogeographic perspective, provide information about the geological timing of speciation events. The method conditions directly on biogeographic observations to induce dated node age distributions, rather than imposing (potentially incorrect) beliefs about speciation times using node calibration densities, which are data-independent prior densities. Biogeographic dating may present new opportunities for dating phylogenies for fossil-poor clades since the technique requires no fossils. This establishes that historical biogeography has untapped practical use for statistical phylogenetic inference, and should not be considered of secondary interest, only to be analysed after the species tree is estimated.

Chapter 3

Lévy processes: evolution with jumps

3.1 Introduction

Morphological variation in continuous characters, such the body mass of therapods or the height of kelp, is one of the most visible examples of the diversity of life on Earth. A number of theoretical frameworks have been put forth to explain this variety of sizes and shapes seen in the natural world (Darwin 1859; Simpson 1953; Eldredge and Gould 1972; Stanley 1975). Gaussian processes – a class of stochastic processes which includes Brownian motion and the Ornstein-Uhlenbeck process – have been used extensively to model continuous trait evolution, e.g. body mass evolution (Freckleton et al. 2003) or gene expression level evolution (Brawand et al. 2011). These processes are a natural model for continuous character evolution because they are the continuum limit of a broad range of discrete-time character evolution models (Cavalli-Sforza and Edwards 1967; Lande 1976; Felsenstein 1985).

However, not all discrete-time models have a Gaussian process as their limit; many evolutionary processes may result in changes in a continuous character too abrupt to be accounted for by any Gaussian process. For example, rapid changes in population size can dramatically affect rates of allele fixation, and thus introduce abrupt changes in quantitative traits (Lande 1976). The ecological release of selective constraints may induce an adaptive radiation that increases disparity unevenly across a clade (Simpson 1953; Stanley 1975). Through cladogenesis under a punctuated equilibrium model of trait evolution, divergence events are paired with sudden trait change (Eldredge and Gould 1972). If cladogenetic evolutionary processes are present, continuous trait patterns seen in extant taxa may mislead inference due to speciation events "hidden" by extinction events (Bokma 2002).

Two main routes have been taken to account for the extra variation that these micro- and macro-evolutionary processes produce. One approach pioneered by O'Meara et al. (2006) is to allow for shifts in the rate of Brownian motion in different places on the phylogeny.

This method is similar in spirit to models of rate shifts in molecular evolution (Thorne et al. 1998; Huelsenbeck et al. 2000; Drummond and Suchard 2010). A number of refinements have since been proposed, such as the use of reversible jump Markov chain Monte Carlo (MCMC) to infer the timing and intensity of rate shifts (Eastman et al. 2011), which identified rate shifts in the evolution of primate body mass. Harmon et al. (2010) introduced an "early-burst" process to model rapid trait evolution following cladogenesis in which the rate of Brownian motion decreases exponentially along a branch, such that the rate of change is fastest immediately when a new lineage diverges and then decreases as the lineage grows older. For size and shape data across 49 clades of animals, they reported their early-burst model was favored in two datasets over Brownian motion and Ornstein-Uhlenbeck processes. While these models relax the time-homogeneity assumption of Gaussian process models, they remain fundamentally gradual, in the sense that the changes in traits cannot be too large in a short period of time. This results in the existence of intermediate forms, the hallmark of gradualism.

The other route explicitly models non-gradual evolution by augmenting Brownian motion with a process of "jumps". In a seminal work on models of continuous trait evolution, Hansen and Martins (1996) compared the covariance structure of models of punctuated equilibrium to other models of phenotypic trait evolution and found that one could not distinguish between punctuational models and Brownian motion models from covariance alone. Bokma (2008) described a method to identify punctuated evolution by modeling continuous trait evolution as the sum of Brownian motion and normally distributed jumps resulting from speciation events. The Bokma model accounts for hidden speciation events by first estimating the speciation and extinction rates, then conditioning on the rates as part of a Bayesian MCMC analysis. In a study on mammalian body mass evolution, this model inferred that cladogenetic, rather than anagenetic, processes produced the majority of trait diversity we see today (Mattila and Bokma 2008).

However, jumps in trait evolution may not be linked directly to cladogenesis. Using a pure-jump model, Uyeda et al. (2011) identified a once-per-million-year jump periodicity in vertebrate evolution by modeling trait evolution as the sum of white noise and normally distributed jumps drawn at the times of a Poisson process. Such pure-jump models may be appropriate for traits that are thought to have weak or no gradual evolution component, such as gene expression, which may depend only on the discrete events of transcription factor binding site recruitment and degradation. Khaitovich et al. (2005) introduced a pure-jump model of evolution in which gene expression levels evolve via jumps drawn from a skewed normal distribution at the times of a Poisson process. They reported evidence of skewness in primate gene expression evolution, a biologically interesting signal that could not have been explained by simple Brownian motion models (also see Chaix et al. (2008)).

This evidence of jumps motivates us to introduce a class of models to account for the wide range of modes of non-gradual evolution. Both Brownian motion and the compound Poisson

processes of Khaitovich et al. (2005); Uyeda et al. (2011) (but not the Ornstein-Uhlenbeck process) are members of a broader class of stochastic processes whose motion may be thought of as "drift and diffusion with jumps", viz. the class of Lévy processes. A Lévy process is the sum of three components: a directional drift (also called trend in the biology literature, not to be confused with genetic drift), a Brownian motion, and a pure-jump process. The last component allows Lévy processes to have jumps in their sample paths and, in the context of continuous trait evolution, account for abrupt shifts in continuous characters that pure diffusion models cannot easily explain. Qualitatively, these jumps give the distribution of trait change "fat tails", reflecting that there is a higher probability of larger amounts of trait change than under a Brownian motion. In the mathematical finance literature, Lévy processes have been successfully used to capture the "fat-tailed" behavior of stock prices (Li et al. 2008). We developed a Bayesian method that determines whether a Lévy process with jumps explains the data better than a single-rate Brownian motion and effectively infers the parameters of that Lévy process.

3.2 Model

Lévy processes

Stochastic processes with stationary and independent increments whose sample paths are right-continuous with left limits are called Lévy processes. We will highlight the key properties of this class of processes and state some important results. Kallenberg (2010; Ch. 15) provides a more detailed and technical exposition.

Let $\{X_t, t > 0\}$ be a Lévy process. There are two equivalent ways of characterizing X_t, by its transition density $\mathbb{P}(X_t = y \,|\, X_0 = x)$, or by its characteristic function, given by

$$\phi(k;t) = \mathbb{E}\left(e^{ikX_t} \,|\, X_0 = 0\right), \tag{3.1}$$

where $i = \sqrt{-1}$ is the imaginary unit and $\mathbb{E}(\cdot)$ is the expected value. Note that k is the variable on which the characteristic function acts. As an example, the transition density of a Brownian motion is

$$\mathbb{P}(X_t = y \,|\, X_0 = x) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{(y-x)^2}{2\sigma^2 t}}, \tag{3.2}$$

so the corresponding characteristic function is given by

$$\begin{aligned} \phi(k;t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{y^2}{2\sigma^2 t}} e^{iky} dy \\ &= e^{-t\frac{1}{2}\sigma^2 k}. \end{aligned} \tag{3.3}$$

A result known as the Lévy-Khinchine representation asserts that all Lévy processes have characteristic functions of the form

$$\phi(k;t) = \exp\left\{t\left(aik - \frac{1}{2}\sigma^2 k^2 + \int \left(e^{ikj} - 1 - ikj\mathbb{I}_{|j|<1}\right)\nu(dj)\right)\right\}, \tag{3.4}$$

where a and σ^2 are constants and $\nu(\cdot)$ is the so-called Lévy measure. Intuitively, the Lévy-Khinchine representation provides a mathematical decomposition of a Lévy process into its three constituent parts:

1. A constant directional drift (or trend) with rate a
2. A Brownian motion with rate σ^2
3. A pure-jump process that draws jumps from the Lévy measure $\nu(\cdot)$.

The processes we consider have no long-term directional trend, so $a = 0$. To get a better understanding of the Lévy measure, one can imagine that the process has probability $\nu(dj)dt$ of making a jump of size j during the time dt. If there are no jumps, then ν is identically 0 and (3.4) becomes (3.3). This shows that the only Lévy process with continuous sample paths is a single-rate Brownian motion.

Using the Lévy-Khinchine formula, it is possible to compute the moments of a Lévy process, assuming that they exist. Because we will only consider symmetric Lévy processes, we are only interested in the process' variance and excess kurtosis, the latter of which is a measure of the relative frequency of large evolutionary changes compared to a Brownian motion. These two moments are given by

$$V(t) = \mathbb{E}(X_t^2) = -\phi^{(2)}(0;t) \tag{3.5}$$

and

$$K(t) = \frac{\mathbb{E}(X_t^4)}{V(t)^2} - 3 = \frac{\phi^{(4)}(0;t)}{V(t)^2} - 3, \tag{3.6}$$

where $\phi^{(n)}$ is the nth derivative of ϕ.

Three examples of Lévy processes

In addition to a single-rate Brownian motion (BM), we implemented three different models that are representative of the range of behavior possible with Lévy processes. These models are a compound Poisson process with normally distributed jumps (jump normal, abbreviated JN), the variance gamma process (VG), and the α-stable process (AS). To gain an intuition for the behavior of each process, Figure 3.1 shows representative pure-jump sample paths

and the corresponding jump measures for JN, VG and AS, whose properties we examine in turn. Note, to accentuate the flavor of each jump measure under each parameterization in Figure 3.1, the Brownian motion rate was assigned to $\sigma = 0$. Parameters of each model are summarized in Table 3.1.

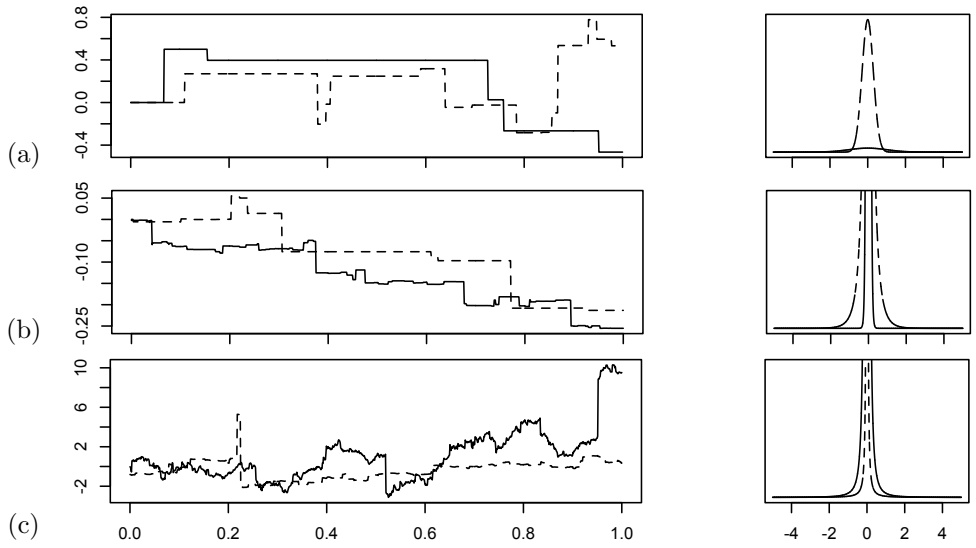


Figure 3.1: Sample paths of Lévy processes without Brownian motion (left panel) and their corresponding Lévy measures (right panel). Compound Poisson process with normally distributed jumps (JN; 1a) paths were sampled with parameters $\lambda = 2$, and $\delta = 1$ (solid line) and $\lambda = 20$, and $\delta = .3$ (dashed line). Variance gamma (VG; 1b) paths were sampled with parameters $\kappa = .1$, and $\tau = .2$ (solid line) and $\kappa = 1$, and $\tau = .6$ (dashed line). α-stable (AS; 1C) paths were sampled with parameters $\alpha = 1.5$, and $\beta = .1$ (solid line) and $\alpha = .9$, and $\beta = .005$ (dashed line).

The compound Poisson process

The JN model has Lévy measure

$$\nu(dj) = \lambda \frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{j^2}{2\delta^2}} dj.$$

With rate λ, the process makes jumps with values drawn from a centered normal distribution with standard deviation δ. As Figure 3.1a shows, the paths of the JN process are characterized by periods of stasis interrupted by bursts of rapid change. Looking at the Lévy measure, a process with more jumps will have a taller Lévy measure while a process with

Model	Parameter	Interpretation
Brownian motion (BM)	σ	rate of Brownian motion
Jump normal (JN)	σ	rate of Brownian motion
	λ	rate of jumps
	δ	standard deviation of jump size
Variance gamma (VG)	σ	rate of Brownian motion
	κ	relative rate of large jumps
	τ	size of jumps
α-stable (AS)	σ	rate of Brownian motion
	α	relative rate of small jumps
	β	size of jumps

Table 3.1: Model parameters and interpretations for all implemented models

larger jumps will have a fatter Lévy measure. The transition density of the JN process with no Brownian motion is known and is given by

$$\mathbb{P}(J_t = j \mid J_0 = 0, \delta, \lambda) = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \frac{1}{\sqrt{2\pi n \delta^2}} e^{-\frac{j^2}{2n\delta^2}}. \tag{3.7}$$

The variance and excess kurtosis of a process with both BM, with rate σ^2, and JN motion are

$$V(t) = (\sigma^2 + \lambda\delta^2)t \tag{3.8}$$

and

$$K(t) = \frac{3\lambda\delta^4}{(\sigma^2 + \lambda\delta^2)^2 t}, \tag{3.9}$$

respectively.

The variance gamma process

The VG model has Lévy measure

$$\nu(dj) = \frac{1}{\kappa|j|} e^{-\sqrt{\frac{2}{\kappa\tau^2}}|j|} dj.$$

Here, τ controls the size of jumps while κ controls the relative probability of large versus small jumps. The Lévy measure has infinite mass, and thus the VG process is infinitely active, meaning that in any finite period of time, the process makes infinitely many jumps.

However, as can be seen in Figure 3.1b most of those jumps are arbitrarily small. When κ is large, the VG process only makes very large or very small jumps.

Like the JN process, the transition density of the VG process with no Brownian motion is known analytically,

$$\mathbb{P}(J_t = j \mid J_0 = 0, \tau, \kappa) = \frac{2^{\frac{2t-3\kappa}{4\kappa}} \kappa^{-\frac{2t+\kappa}{4\kappa}}}{\Gamma(t/\kappa)\sqrt{\pi\tau^2}} \left(\frac{\tau^2}{j^2}\right)^{\frac{-2t+\kappa}{4\kappa}} K_{|t/\kappa-1/2|}\left(\sqrt{\frac{2j^2}{\kappa\tau^2}}\right), \tag{3.10}$$

where $\Gamma(\cdot)$ is the gamma function and $K_\epsilon(\cdot)$ is the modified Bessel function of the second kind with index ϵ (Abramowitz and Stegun 1964; Ch. 9,10).

The variance and excess kurtosis of a process with both BM and VG motion are

$$V(t) = (\sigma^2 + \tau^2)t \tag{3.11}$$

and

$$K(t) = \frac{3\kappa\tau^4}{(\sigma^2 + \tau^2)^2 t} \tag{3.12}$$

respectively.

The α-stable process

The AS model has Lévy measure

$$\nu(dj) = \frac{\beta^\alpha}{|j|^{1+\alpha}} dj,$$

where β is a scale parameter, controlling the magnitude of jumps taken and $0 \leq \alpha \leq 2$ is the so-called stability parameter. For every $\alpha < 2$, the Lévy measure has infinite mass, so the AS process is infinitely active. However, Figure 3.1c shows that the behavior of the AS process is quite different from the VG process. In particular, the AS process does not experience as strong a trade-off between small and large jumps as the VG process does. As $\alpha \to 0$, the tails of the Lévy measure become heavier and heavier, but the relative proportion of probability for medium sized jumps remains nearly constant, as opposed to the VG process. This is manifested in the fact that the AS process has infinite p^{th} moment for $p > \alpha$ when $\alpha < 2$; thus, the variance and the excess kurtosis of the process do not exist for $\alpha < 2$. In addition, unlike the JN and VG processes, the transition density is not known in closed form. However, the characteristic function of the AS process without Brownian motion is known to be

$$\phi(k; t) = e^{t|\beta k|^\alpha}, \tag{3.13}$$

and so we can make use of the Fourier inversion theorem to numerically compute the transition density of the AS process without Brownian motion,

$$\begin{aligned}\mathbb{P}(J_t = j \,|\, J_0 = 0, \beta, \alpha) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ikj}\phi(k;t)dk \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \cos(kj)\phi(k;t)dk, \end{aligned} \tag{3.14}$$

where the second equality follows because $\phi(k;t)$ is real and even.

3.3 Methods

Inference of Lévy processes

We use a Bayesian framework to analyze Lévy processes evolving on a phylogeny. Let $p(\theta)$ be the prior density for the parameters of the Lévy process model and $L(D \,|\, \theta)$ be the likelihood of the observed data given the parameters. We want to compute the posterior density,

$$p(\theta \,|\, D) \propto L(D \,|\, \theta)p(\theta). \tag{3.15}$$

To compute the likelihood of a Lévy process on a phylogeny, we use Felsenstein's pruning algorithm (Felsenstein 1981). To calculate $L_i(y_i)$, the likelihood of the data observed in all species that are descended from node i, given that the trait value at node i equals y_i, we use the likelihood at the descendent nodes j and k. Letting $\{X_t, t > 0\}$ be the Lévy process under consideration,

$$L_i(y_i \,|\, \theta) = \left(\int \mathbb{P}(X_{t_j} = y_j \,|\, X_0 = y_i)L_j(y_j \,|\, \theta)dy_j\right)\left(\int \mathbb{P}(X_{t_k} = y_k \,|\, X_0 = y_i)L_k(y_k \,|\, \theta)dy_k\right), \tag{3.16}$$

where t_j and t_k are the branch lengths leading to nodes j and k, respectively. At the root (node 0), we assume an improper uniform prior for the trait value y_0, and we integrate over all possible values of the root node to obtain

$$L(D \,|\, \theta) = \int_{-\infty}^{\infty} L_0(y_0 \,|\, \theta)dy_0. \tag{3.17}$$

However, the integrals in (3.16) and (3.17) are intractable for most Lévy processes. To get around this, we exploit the fact that if X is a Lévy process consisting of a Brownian motion with no directional drift and diffusion rate σ^2, and a pure-jump process, the Lèvy-Khinchine representation guarantees that $X = B + J$, where B is a Brownian motion and J is the

pure-jump process, and B and J are independent. Then conditional on $J = j$, the transition density of X is given by

$$\mathbb{P}(X_t = y \mid X_0 = x, J = j) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{((y-j)-x)^2}{2\sigma^2 t}}. \tag{3.18}$$

This follows because the Brownian motion has to get to $y - j$ and then the jump process will do the rest. Thus, conditioned on all the jumps on the branch leading up to a node, $\mathbf{J} = \{J^{(n)}, \ldots, J^{(1)}\}$ for a tree with n non-root nodes, $L(D \mid \theta, \mathbf{J})$ is the likelihood of the data under Brownian motion where branch i has branch specific offset $J^{(i)}$. Then,

$$p(\theta, \mathbf{J} \mid D) \propto L(D \mid \theta, \mathbf{J}) p(\mathbf{J} \mid \theta) p(\theta), \tag{3.19}$$

where

$$p(\mathbf{J} \mid \theta) = \prod_i \mathbb{P}(J^{(i)}_{t_i} = j^{(i)} \mid J_0 = 0, \theta)$$

is the joint probability of the jumps along each branch (determined by the specific jump model adopted). We want to integrate over the jumps to get

$$p(\theta \mid D) = \int p(\theta, \mathbf{J} \mid D) d\mathbf{J}, \tag{3.20}$$

but this integral remains intractable. Instead, we approximate the integral by using MCMC to obtain samples from the joint posterior distribution of the parameters and the jumps. Marginalizing over the sampled jumps approximates the integral in the right-hand side of (3.20).

To obtain posterior samples of the jumps, we serially update each branch in a post-order traversal of tree by proposing a new value $J^{(i)\prime}$ from a normal distribution centered at the current sampled $J^{(i)}$ and with variance 0.5. This variance lead to good mixing for the data we considered, but should be specified by the user as appropriate. We then accept or reject the proposed jump update using the Metropolis-Hastings ratio,

$$\mathbb{P}(\text{Accept } J^{(i)\prime}) = \frac{L(D \mid \theta, \mathbf{J}')}{L(D \mid \theta, \mathbf{J})} \frac{p(\mathbf{J}' \mid \theta)}{p(\mathbf{J} \mid \theta)},$$

where $\mathbf{J}' = \{J^{(n)}, \ldots, J^{(i+1)}, J^{(i)\prime}, J^{(i-1)}, \ldots, J^{(1)}\}$ is the vector of jumps with only one branch updated. Note that the proposal ratio is equal to 1 because of the symmetry of the normal distribution and the prior ratio is equal to 1 because no parameters are updated. This method is similar to the path sampling method of Robinson et al. (2003), in that we use MCMC to sample and integrate over hidden states (the unobserved jumps).

During the MCMC run, we randomly choose to update either the jumps or the model parameters. When we choose to update a model parameter, we randomly choose a model

parameter to update. All parameters except for α from the AS process are positive and real and so were assigned scaling proposal distributions. Because $0 < \alpha < 2$, we use a truncated normal proposal distribution to update α. Parameter updates are accepted or rejected according to the Metroplis-Hastings ratio,

$$\mathbb{P}(\text{Accept } \theta') = \frac{L(D \mid \theta', \mathbf{J})}{L(D \mid \theta, \mathbf{J})} \frac{p(\mathbf{J} \mid \theta')}{p(\mathbf{J} \mid \theta')} \frac{p(\theta')}{p(\theta)} \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)},$$

where θ is the randomly selected parameter, θ' is the proposed update, and $q(\cdot \mid \cdot)$ is the proposal distribution.

Data

We log-transformed the male-female means of body mass, endocranial volume (ECV), and mass-to-ECV ratio data reported in Isler et al. (2008). The branches of the phylogeny provided by Isler et al. (2008) were measured in increments of half-million years. In favor of higher resolution of branch lengths, we substituted the Isler et al. phylogeny with the Redding et al. (2010) primate phylogeny included in the R package `auteur` (Eastman et al. 2011). We intersected the Isler et al. dataset with the Redding et al. phylogeny, which resulted in 126 taxa with data present in the phylogeny. This phylogeny has 1267 myr of total branch and a root height of 65 myr. The resulting phylogeny was used for all analyses and simulations reported in this paper.

Software configuration

The software used in this study was programmed in C++, borrowing code from GNU Scientific Library (Contributors 2010) and MrBayes (Ronquist et al. 2012b). The source code may be found at `http://github.com/mlandis/creepy-jerk`. With one exception, all parameters were assigned half-Cauchy distributions with scale parameters of 1 as prior densities. Under the AS processs, $0 \leq \alpha \leq 2$, so we used a uniform distribution on $[0, 2]$ as its prior. Each posterior distribution was computed by running MCMC for 2×10^6 cycles, sampling every 10^3 cycles, where the first 10^5 cycles were discarded as part of the burn-in. The R package `coda` (Plummer et al. 2006) was used to verify MCMC convergence. For the BM, JN, VG, and AS models, one MCMC run took 0.5, 8, 6, and 48 hours, respectively. This discrepancy results from the fact that, while the JN and VG models have analytical solutions for their jump densities, we had to approximate the AS jump density using time-consuming numerical integration.

Analysis

We characterized how Lévy processes perform in the context of phylogenetic inference for both simulated data and real data. We used simulated data to test the accuracy of parameter inference and quantify the power to reject BM when the true model is a Lévy process with jumps. We then analyzed the primate data set to both estimate parameters and determine if a BM model is rejected in favor of a Lévy process with jumps in biological data. Our analysis examines the aforementioned four Lévy processes: BM, JN, VG, and AS.

To test the BM model, we performed a 3-step procedure similar to a parametric bootstrap. First, data was analyzed under a pure BM model, resulting in an estimate of the BM rate, σ^2. Then, 20 "jump-absent" datasets were simulated under BM with the inferred rate. Finally, each simulated dataset was analyzed using a "jump-present" model, and the average posterior distributions of either the variance and excess kurtosis (for JN and VG) or the parameter α (for AS) were compared to the posterior distribution of those parameters inferred from the original data. Note that the variance, excess kurtosis and α calculated here do *not* describe the data observed at the tips, but rather their expected values as a function of time, see equations (3.8), (3.9), (3.11), and (3.12).

By inspecting the posterior distribution of the variance and excess kurtosis for JN and VG between data and the Brownian motion simulations, we determined whether there was evidence for non-Gaussian evolution. Under the BM model, the expected excess kurtosis is 0, so if the posterior of the excess kurtosis placed significant mass away from 0, we interpreted that as strong evidence for non-Gaussian evolution. For the AS model, these moments are not defined; however, when $\alpha = 2$ the AS process is equivalent to BM, so if the posterior distribution of α placed significant mass away from 2, we took that as evidence for non-Gaussian evolution.

When we were able to reject BM in favor of a Lévy process with jumps, we characterized the amount of trait change attributable to those jumps by computing the signal-to-noise ratio, defined as the ratio of the mean to standard deviation, of the posterior distributions of the sampled jumps for each branch in the phylogeny. To normalize for branch-length effects, we further divided all signal-to-noise ratios by their respective branch lengths. When the signal-to-noise ratio equals zero, the Brownian motion component of the model alone is capable of producing the observed trait changes along that branch. A non-zero signal-to-noise ratio was interpreted as evidence that traits along the branch evolved faster than could be explained by the model's Brownian motion component.

3.4 Results

Simulated data

We simulated 20 datasets for each model (JN, VG, and AS) on the primate phylogeny (see Methods) and computed posterior distributions under the true model for each simulated dataset. Figure 3.2 presents boxplots of the maximum a posteriori estimates for each simulation, with the horizontal line indicating the true parameter value. Inference under the JN and AS model recovered the true parameters with minor error. Inference under the VG model recovered σ and τ reasonably well while underestimating κ by an order of magnitude. The mean and root mean square errors of the posteriors are recorded in Supplemental Table 1.

We then applied our method to test for the presence of jumps to the simulated datasets. The results are shown in Figure 3.3. When the true model is either JN or VG, the inferred variance was approximately equal between the jump-present and BM simulations, but the excess kurtosis was different. For the AS model, the inferred α deviated significantly from 2 only in the jump-present data. The maximum a posteriori estimates and 95% highest posterior density intervals may be found in Supplemental Table 2.

Empirical: Primates

Next, we computed the posterior distributions for body mass, ECV, and mass-to-ECV ratios for the BM, JN, VG, and AS models. The maximum a posteriori estimates and 95% highest posterior density intervals for each dataset are provided in Supplemental Table 3. We applied our test to detect evolution that cannot be explained by BM to each data set. For the sake of brevity, we only present results for body mass under the JN model (Figure 3.4) and ECV under the AS model (Figure 3.5), although several models showed evidence of non-Gaussianity in the evolution of these traits. For the mass-to-ECV data, no Lévy process with jumps was preferred over BM (Figure 3.6). Supplemental Table 4 has more detailed numbers, including parameter estimates for each model.

For body mass under the JN model, the estimates of both the jump rate λ and the jump size δ are non-zero. This is seen when comparing the posterior estimates of the excess kurtosis, which are qualitatively different between the BM simulations and the real data (Figure 3.4). In addition, the posterior estimates of the variance of the process were nearly identical between the BM simulations and the real data. Together, these provided evidence that the evolution of primate body mass is not well-explained by BM alone.

For ECV under the AS model, the posterior density of α inferred from the ECV data placed extremely little mass on 2, while the BM simulations consistently resulted in maximum

a posteriori estimates of $\alpha = 2.0$ (Figure 3.5), evidence of non-Gaussian evolution of primate ECV.

For the mass-to-ECV ratio, we found no remarkable deviation from Brownian motion (Figure 3.6). This is reflected in the fact that the posteriors of the kurtosis for the JN and VG models, as well as the posterior of α for the AS model, were extremely close to the posteriors inferred from the BM simulations.

Figure 7 shows the primate phylogeny with branches colored according to their branch-normalized signal-to-noise ratios. Since we rejected the BM in favor of a Lévy process with jumps in the body mass and ECV data, non-zero signal-to-noise ratios are possibly explained by jumps in trait evolution.

3.5 Discussion

Darwin (1859) first proposed that what is now called continuous character evolution occurs gradually, with species changing very little over short time periods. Since then, some (Simpson 1953; Eldredge and Gould 1972; Stanley 1975) have suggested that evolution occasionally happens more quickly, with rapid changes in characters occurring over short periods of evolutionary time. However, most studies of continuous trait evolution that use comparative data rely on a Brownian motion model. Because the path of a Brownian motion is continuous; i.e, the value of the trait at the next moment in time is necessarily very close to the value of the trait at the current moment, the most natural interpretation of these models excludes the possibility of saltational change. Moreover, even though some saltational processes can produce the same distribution of tip-data as a Brownian motion, these are highly restricted–for example, if jumps occur only at nodes in the tree that lead to extant taxa.

A natural generalization of Brownian motion that allows for paths that are not strictly continuous is the class of Lévy processes. The discontinuities in the path can be thought of as "jumps", in which the character changes instantly without any intermediate forms. These jumps approximate rapid changes in character value over a short time-scale and result in distributions of character change that have "fat tails"; in statistical literature, distributions with fat tails are said to be leptokurtic.

We examined three specific Lévy processes: a compound Poisson with normally distributed jumps (JN), variance gamma (VG), and α-stable (AS). All processes also include a Brownian motion component, and hence can be interpreted as modeling gradual evolution punctuated by large, sudden changes in trait value. The JN process waits an exponentially distributed amount of time with rate λ before making a jump whose size is drawn from a normal distribution with standard deviation δ. The VG and AS processes are so-called infinitely active processes that jump infinitely often. However, most of the jumps are arbitrarily small, and so the processes are well-behaved. An important difference between the

VG and AS processes is that the AS process is much more likely to take extremely large jumps, compared to the VG process (as reflected in the fact that the variance of the AS process is infinite). For the VG process, the parameter κ corresponds to the rate of very large jumps and the parameter τ controls the variance of the jumps that are taken. In the AS process, the parameter α is confined between 0 and 2. As α approaches 2, the process converges in distribution to a Brownian motion, while as it approaches 0 the process makes larger jumps more frequently. The parameter β controls the scale of jumps that are taken.

These processes can be interpreted in a biological context. The JN process reflects the classic idea of stasis punctuated by rapid character change, and has some history in the literature (Hansen and Martins 1996; Bokma 2008; Uyeda et al. 2011). VG and AS are more exotic models; however, they may capture certain aspects of evolution that would otherwise be impossible to model. For example, in the Lande (1976) description of the impact of genetic drift on quantitative traits, trait evolution is a Brownian motion on a time scale determined by the effective population size: evolution works more slowly in large population and more quickly in small populations. Because the VG process arises as a time-change of a Brownian motion (Madan and Carr 1998), it can capture the impact of fluctuating population size on continuous character evolution. The AS process, on the other hand, is a natural generalization of BM that has many of the same features, but allows for fatter tails and erratic sample paths.

Because analytic computation of the likelihood using Felsenstein's pruning algorithm is not possible for the Lévy processes that we considered, we developed a MCMC method to estimate the parameters of a Lévy process. The MCMC algorithm samples possible jump histories along each branch of the phylogeny. Using data augmentation for ancestral states, similar in spirit to that of Robinson et al. (2003), we numerically integrate over the history of jumps. Because any Lévy process can be split into a Brownian motion and pure-jump components, our method is applicable to any Lévy process outside of the examples we considered here.

To determine whether a phylogeny contains sufficient information to reject single-rate Brownian motion in favor of a more general Lévy process, we conducted simulation studies using each of the models that we implemented. Figure 3.2 shows that we were able to recover the parameters of the JN and AS processes with high accuracy. However, for the VG process, τ, which controls the variance of the jumps, was well estimated, but the rate of large jumps, κ, is underestimated. We are uncertain why κ is consistently underestimated but suspect that tree shape plays an important role.

We then made use of the fact that non-Brownian Lévy processes have more frequent large deviations in short time periods than BM. This large deviation is manifested as excess kurtosis. Using the characteristic function of a Lévy process (i.e. the Lévy-Khinchine formula), we calculated the posterior distribution of the variance and excess kurtosis per unit time. Because the Gaussian distribution has zero excess kurtosis, this posterior estimate should be

close to zero when BM is a good model for trait evolution and have significant mass away from zero when the trait evolution is non-Gaussian. In the case of the α-stable process, the excess kurtosis is not defined and so we focused our attention on the parameter α. As $\alpha \to 2$, the α-stable process becomes a BM; thus, if the posterior distribution of α was not very close to 2, the evolution of the continuous character was inferred to be non-Brownian.

We applied our MCMC method to data from 126 primate species (Isler et al. 2008; Eastman et al. 2011) to uncover evidence of non-Gaussian evolution in a large group of mammals. For each species, we obtained measurements of body mass, endocranial volume (ECV) and also examined the ratio of mass-to-ECV. For the body mass and ECV data, we found evidence supporting Lévy process with jumps over BM and highlighted results under the JN and AS models, respectively, while the mass-to-ECV ratio appeared to evolve as a Brownian motion. The parameters inferred for body mass suggest that there is a burst of body size evolution equivalent to 5 to 6 million years of gradual evolution approximately once every 4 million years, which is within the same order of magnitude of jump periodicity as reported by Uyeda et al. (2011). ECV evolution was fit by an α-stable process, with an intermediate value of $\alpha = 1.7$, consistent with a mode of evolution in which character changes are mostly gradual but punctuated by infrequent, extremely large jumps.

We also obtained a posterior distribution on the amount of trait change in excess to the Brownian motion component of the Lévy process on every branch of the phylogeny. Using this data, we identified branches of the primate phylogeny that showed evidence for evolution that was faster than the Brownian motion component of the Lévy process could explain. In Figure 7, we colored the branches of the primate phylogeny according to the signal-to-noise ratio of the jump size on that branch, normalized by the branch length. Because the jumps account for the "extra" distance that the Brownian motion component of the model cannot explain, large magnitudes of this ratio correspond to branches where there is relatively strong evidence for trait evolution faster than the average BM rate on the tree. This signal weakens deep in the tree, as well as for long branches, although it is interesting to note that some deep branches show excess evolution relative to their branch length (e.g. body mass in the common ancestor of old world monkeys and apes). We identified several clades that showed strong evidence of unexpectedly rapid evolution prior to diversification. For example, the ancestor of the great apes (indicated by an arrow) shows evidence of unexpectedly rapid evolution in both body mass and ECV, while evolution in the ancestors of the Old World and New World monkeys is well explained by the average rate of Brownian motion.

While our method is able to discriminate between Gaussian and non-Gaussian evolutionary models, we were not able to find a test statistic that could discriminate between the different jump processes. Bayesian methods of model testing, such as Bayes factors, require computing the marginal likelihood. However, because of the stochastic nature of our method for integrating over the large number of possible jump histories using MCMC, many methods for estimating the marginal likelihood of a model are unstable or require an unfeasibly large

number of MCMC cycles. Moreover, since the method we present does not compute the marginal likelihood of the parameters alone (with the jumps integrated out), we cannot use information criteria such as the Akaike Information Criterion to conduct model fitness tests. In future work, we plan to implement a Bayesian reversible-jump MCMC method to distinguish between different jump models. This will help to identify how much signal the data contains to single out any particular Lévy process model of evolution. While the method presented in this paper conducts inference under time-homogeneous Lévy processes, nothing prevents the model from being implemented in a rate-shifting framework (see O'Meara et al. (2006); Eastman et al. (2011)). This will further help to distinguish jump events from rate-shifting events.

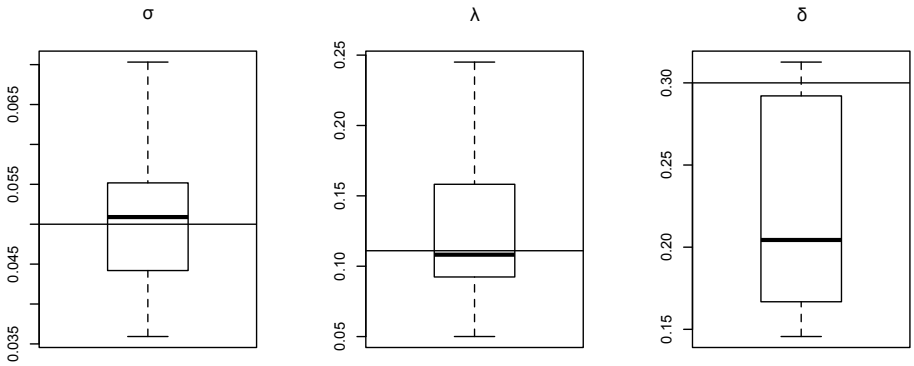
Previous methods describing inference of Lévy processes in the mathematical finance literature have shown that it is possible to precisely infer parameters and accurately choose models with time-series data. However, the correlation structure of a phylogeny complicates inference. As noted by Ané (2008), phylogenetic inference of trait evolution is strongly affected by tree shape, and proposed an effective sample size to gauge how powerful a given topology is for the inference of model parameters. Boettiger et al. (2012) explored the impact of tree shape on the ability for model tests to distinguish Brownian motion models from Ornstein-Uhlenbeck models of continuous trait evolution. Further examination of how tree shape affects inference will become particularly important as increasingly complex models of continuous character evolution are put forward (Khaitovich et al. 2005; Bokma 2008; Harmon et al. 2010; Eastman et al. 2011).

We face two other problems owing to the nature of the phylogeny and the data being analyzed. To illustrate these problems, consider that the clearest signal of excess kurtosis that our model captures lies in terminal sister nodes, where one lineage has evolved as expected under Brownian motion, but the other lineage has experienced an abnormally large jump in trait change. First, assigning data with measurement or sampling error to the tips could introduce (or mask) an excess of trait change, and lead to the false inference of the presence of jumps in trait evolution for the phylogeny. If this is a concern, tips may be modeled with noise at the potential price of losing power to reject Brownian motion in favor of a Lévy process with jumps. Second, the phylogeny is assumed to be fully resolved with errorless branch lengths. If a trait truly evolved by single-rate Brownian motion but exhibits an excess of trait change for the specified branch length, it is possible that that branch's trait evolution is simply an outlier among realizable evolution histories, but it is also possible that the true branch length is longer than indicated. A potential solution is to include posterior samples of the branch length from a Bayesian phylogenetic analysis.

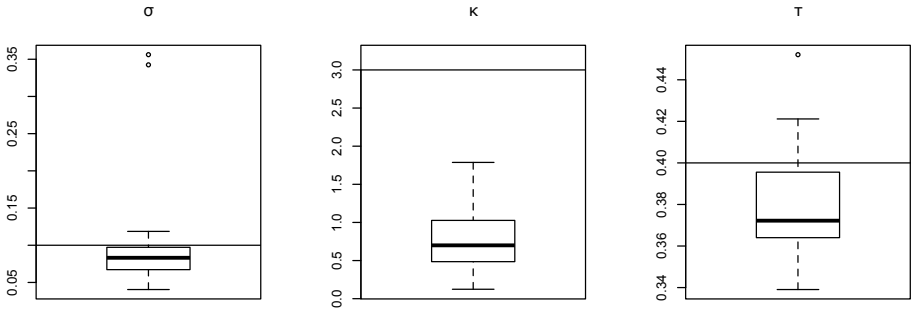
When fitting models of evolution to comparative data, it is important to keep in mind the distinction between the model of evolutionary change and the joint distribution of trait values at the tips that such a model produces. This mapping is not one-to-one; many different models can result in the same joint distribution at the tips and are therefore indistinguishable

from the data alone. To choose between these otherwise equivalent models, scientists must look beyond comparative data, e.g. to the fossil record and mechanistic biological models. Here, we have used BM as a representative process that results in a multivariate normal distribution with a particular covariance structure. Other processes that produce this same joint distribution exist. Similarly, though we fit models with jumps, there are many gradual processes that can produce the exact same distribution at the tips as a jump model, such as models which use Brownian motion with random rate shifts (although these models may not have straight-forward or desirable biological interpretations).

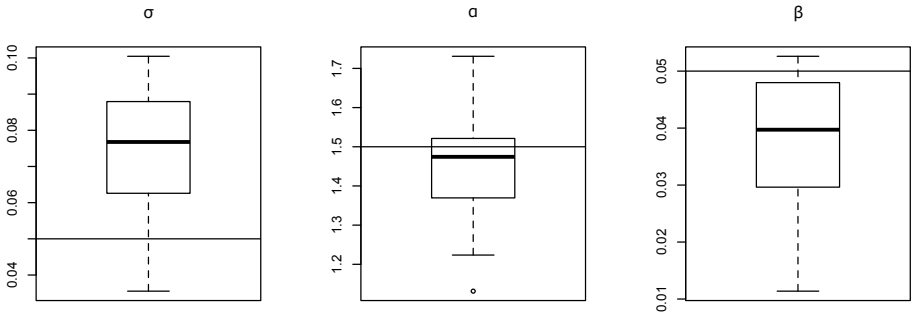
Many other Lévy processes exist. We have only showcased a few, but our method can be applied to any Lévy process with a known characteristic function. It will be interesting to see whether different evolutionary processes, different clades, or different traits are best modeled by certain types of Lévy processes, be it Brownian motion or the α-stable process.



(a) JN

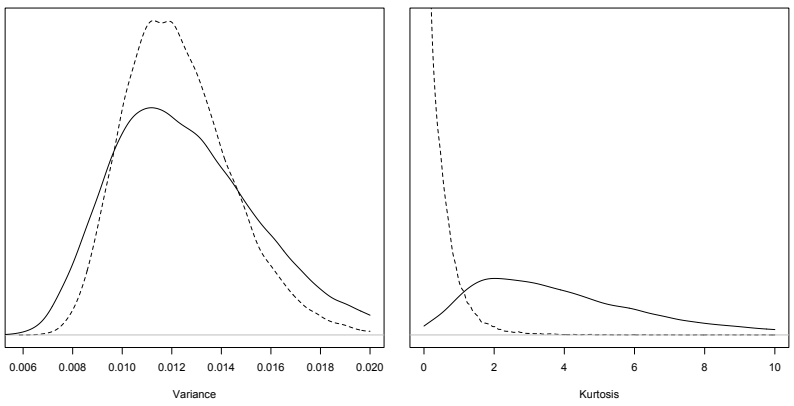


(b) VG

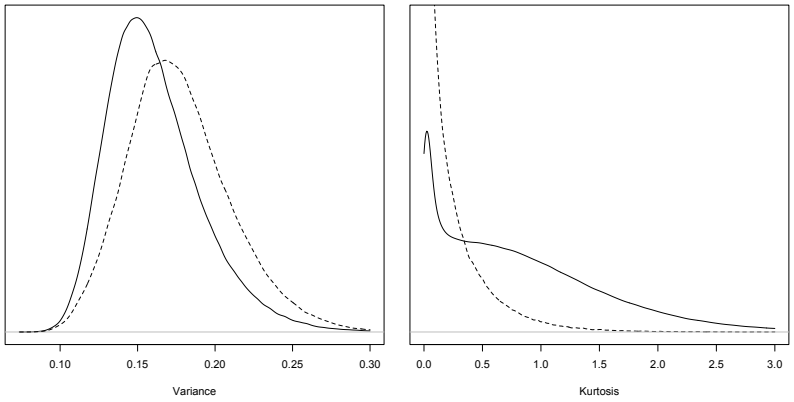


(c) AS

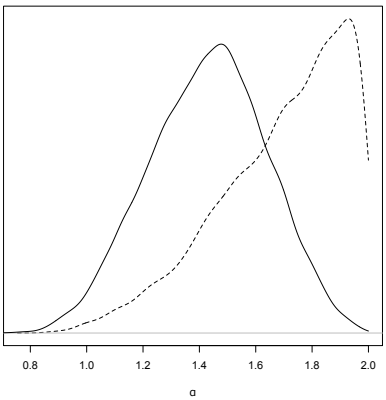
Figure 3.2: Box plots of maximum a posteriori model parameter estimates under JN (2a), VG (2b), and AS (2c) for 20 replicates of jump-present data simulated under each model. The horizontal line shows the true parameter value underlying the simulated data. The true parameters for JN are $\sigma = .05$, $\lambda = .111$, and $\delta = .3$. The true parameters for VG are $\sigma = .1$, $\kappa = 3$, and $\tau = .4$. The true parameters for AS are $\sigma = .05$, $\alpha = 1.5$, and $\beta = .05$.



(a) JN (b) JN

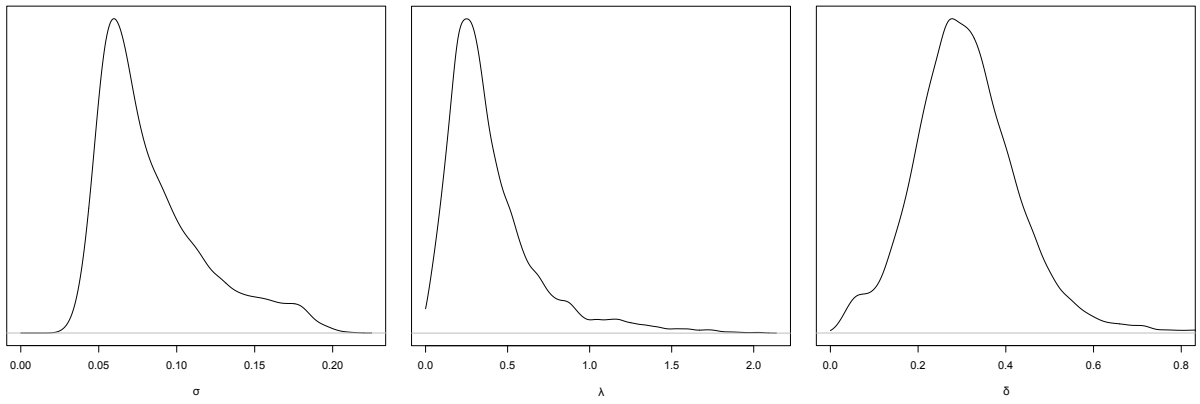


(c) VG (d) VG

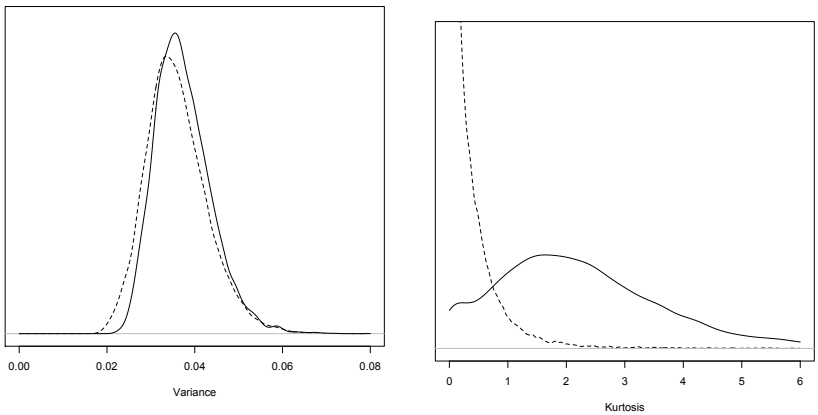


(e) AS

Figure 3.3: Average posteriors of model summary statistics under JN (3a), VG (3b), and AS (3c) upon simulated data. Solid lines indicate average posteriors from 20 replicates of jump-present data simulated under the same model. Dashed lines indicate average posteriors from 20 replicates of jump-absent data simulated under pure Brownian motion parameterized with equivalent variance per unit time ($\sigma = .1118$, $\sigma = .4050$, and $\sigma = .2389$ for analysis by JN, VG, and AS, respectively).

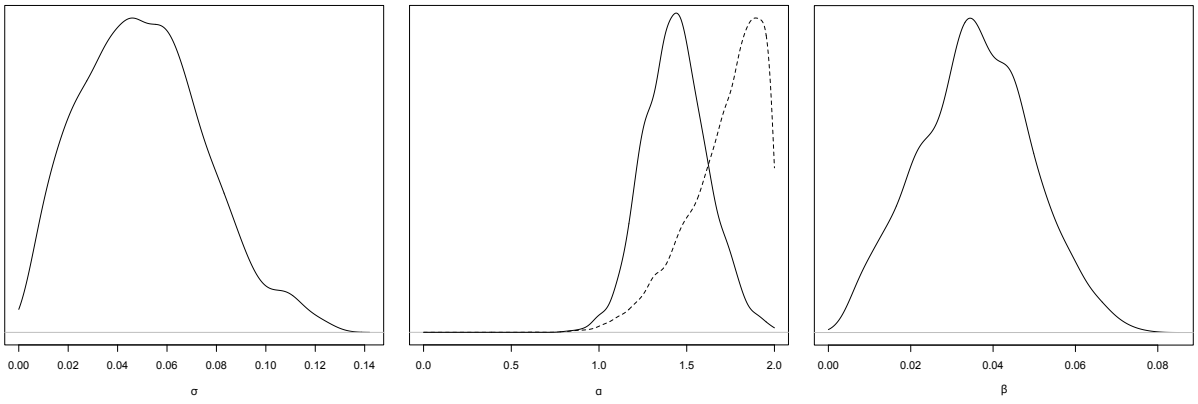


(a) (b) (c)



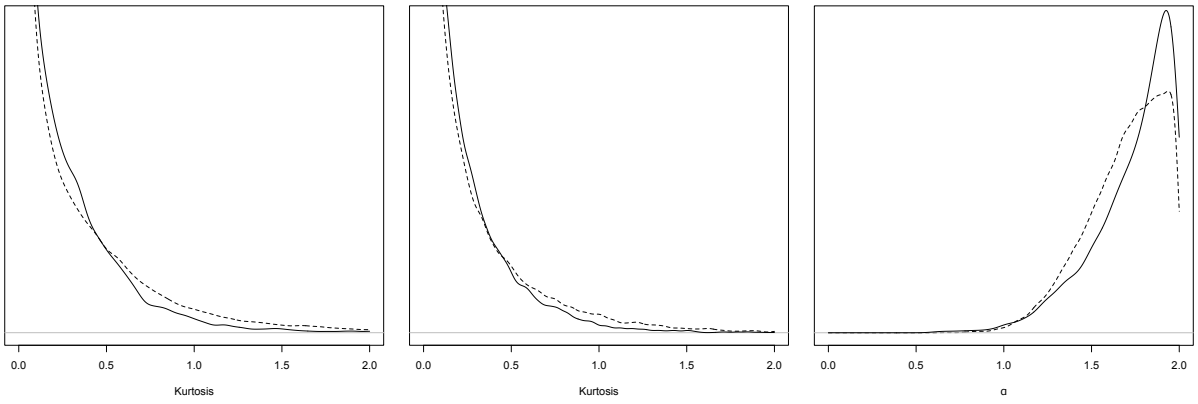
(d) (e)

Figure 3.4: Posteriors of model summary statistics under JN upon primate body mass data. Figures 4a, 4b, and 4c are the model parameters with maximum a posteriori estimates $\hat{\sigma} = .0596$, $\hat{\lambda} = .2497$, and $\hat{\delta} = .2929$, respectively. Figures 4d and 4e are the model variance and kurtosis per unit time. Solid lines indicate posteriors from the empirical data. Dashed lines indicate average posteriors from 20 replicates of jump-absent data simulated under pure Brownian motion parameterized with equivalent variance per unit time ($\sigma = .18$).



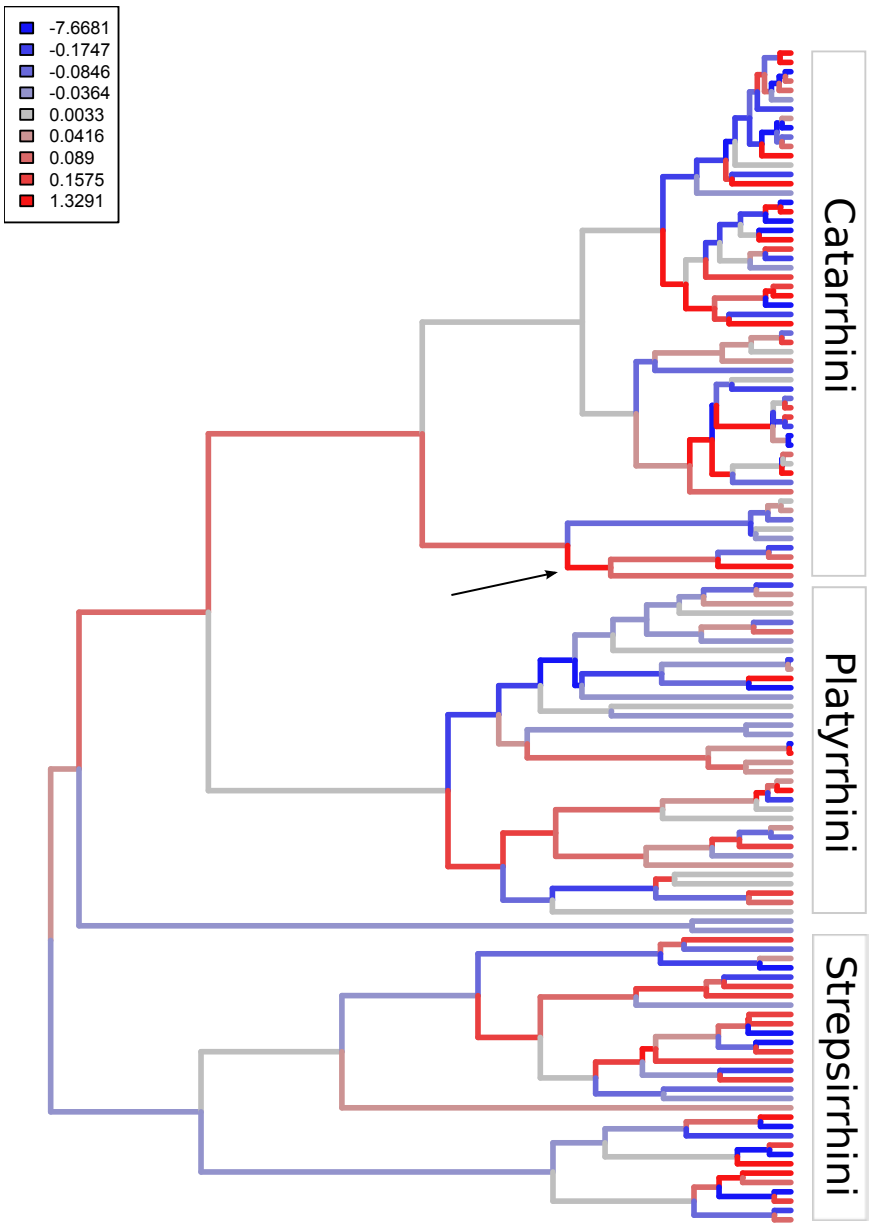
(a) (b) (c)

Figure 3.5: Posteriors of model summary statistics under AS upon primate endocranial volume data. Figures 5a, 5b, and 5c are the model parameters with maximum a posteriori estimates $\hat{\sigma} = .1541$, $\hat{\alpha} = 1.670$, and $\hat{\beta} = .0698$, respectively. Solid lines indicate posteriors from the empirical data. Dashed lines indicate average posteriors from 20 replicates of jump-absent data simulated under pure Brownian motion parameterized with equivalent variance per unit time ($\sigma = .12$).

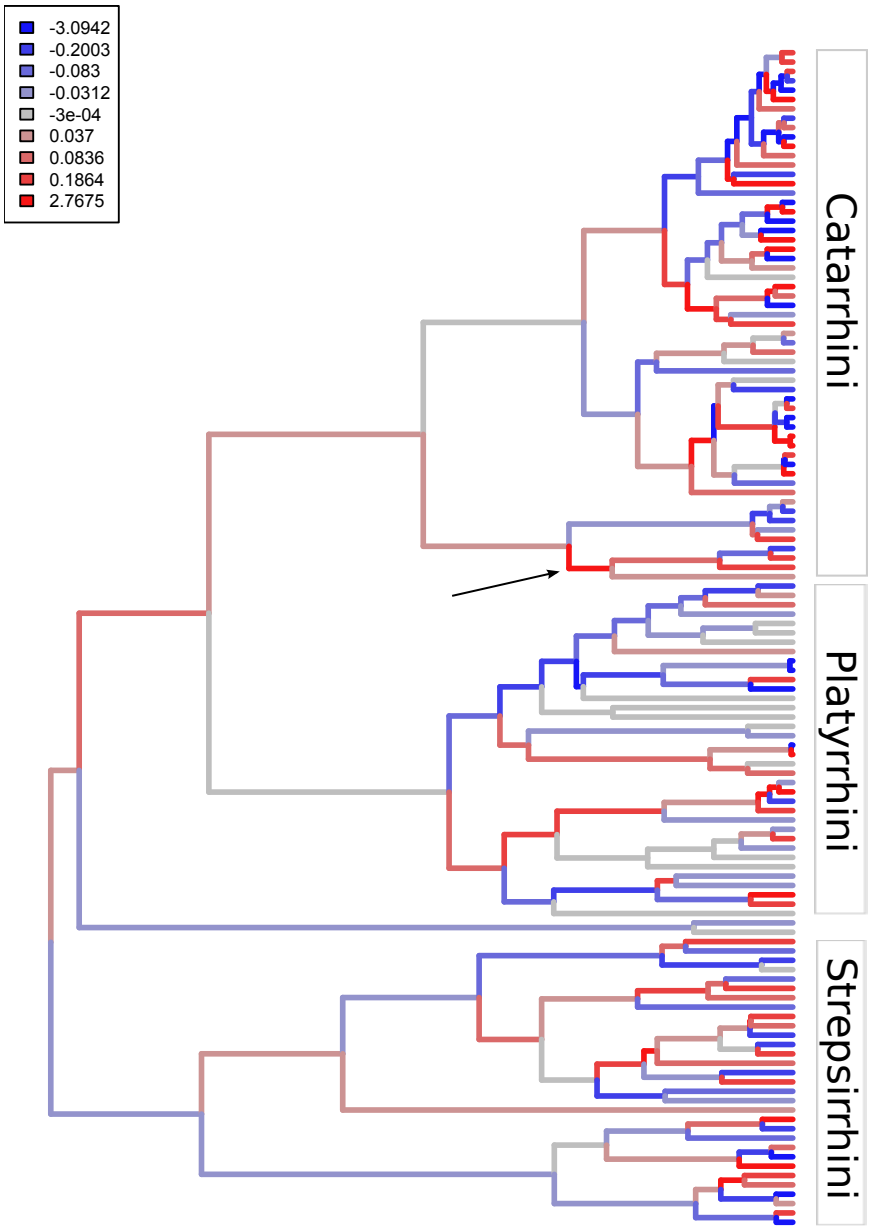


(a) (b) (c)

Figure 3.6: Posteriors of model summary statistics under JN (6a), VG (6b), and AS (6c) upon primate body mass-to-endocranial volume ratio data. Solid lines indicate posteriors from the empirical data. Dashed lines indicate average posteriors from 20 replicates of jump-absent data simulated under pure Brownian motion parameterized with equivalent variance per unit time ($\sigma = .096$).



(a) Body mass under JN



(b) ECV under AS

Figure 3.7: Branch-normalized signal-to-noise ratios of posterior jump distributions. The primate phylogeny with inferred evolutionary histories for body mass under JN (Fig. 7A) and for endocranial volume under AS (Fig. 7B) are shown. Branches are colored according to the quantile containing their branch length-normalized signal-to-noise ratios. A value of approximately zero indicates trait evolution explained predominantly by the Brownian motion component of the fitted model. Uncolored branches in light gray indicate the tendency for the model to explain trait evolution with jumps valued according to the figure legend. The arrow points to the most recent common ancestor of great apes.

Appendix A

Appendix: Biogeographic dating

A.1 Paleogeographical dispersal graphs from Cambrian to present (540–0 Ma)

To construct the instantaneous rate matrix for the epochal dispersal process described in Landis (2015), I defined three dispersal graphs for short, medium, and long-distance dispersal (Figure S2.5). To determine the timing and nature of epochs structuring the paleogeographical model, I first surveyed paleotectonic reconstructions published by Seton et al. (2012) and Wright et al. (2013) available through `gplates` (Gurnis et al. 2012) and paleogeographic maps published by Blakey (2008), then corroborated those findings with various independent sources in the literature (Dietz and Holden 1970; Ziegler et al. 1979; Duque-Caro 1990; Elias et al. 1996; Veevers 2004; Schettino and Scotese 2005; Algeo et al. 2007; Ali and Aitchison 2008; Fiorillo 2008; Lohman et al. 2011; McQuarrie and van Hinsbergen 2013; White et al. 2013; Nance et al. 2014). From this survey I identified twenty-five areas (Table A.1) and twenty-six epochs (Table A.2) as sufficient to coarsely model Earth's paleogeographic history from the Cambrian until the present. Note, each epoch does does not correspond to a single event, i.e. the creation or destruction of a *single* dispersal edge, but rather some group of events that are roughly coincident in time. This is done because the time required to compute the epoch model likelihood increases approximately linearly with respect to the number of epochs: the fewer the number of epochs, the faster the analysis. In practice, I binned events into intervals of 5 Myr, then, since two epochs with identical rate matrices is equivalent to a single epoch with the same rate matrix, I thinned the number of bins by concatenating eventless bins to the next youngest eventful bin. Since the three youngest epochs occur in the last 5 Myr, they were excluded from the binning-thinning procedure. Event times are rounded point estimates without error: they serve only as a consensus across sources to limit when dispersal events of different types might or might not plausibly occur.

Classes of dispersal edges are defined as follows. Each short-distance dispersal edge required areas be connected by land and immediately adjacent. Medium-distance dispersal edges required areas be immediately adjacent and be connected by land or by short water barriers, using the distance between Madagascar and Africa or throughout the Indoaustralian archipelago as a rough measure. In addition, because major paleotectonic events typically happen over tens of millions of years, medium-distance dispersal edges often preceded the establishment of short-distance dispersal edges (i.e. when continents begin to merge) or remained following the destruction of short-distance dispersal edges (i.e. when continents begin to split). All pairs of areas per time interval shared edges for the long-distance dispersal graph, i.e. the graph was fully connected. Strict criteria for presence or absence of an edge were not used, so the presented model should be taken as summary of various sources rather than a quantitative reconstruction.

The dispersal graph animation was generated using the `gplates` Markup Language (Qin et al. 2012). The animation may be viewed here:

`http://figshare.com/s/2a8329e06c6d11e587bd06ec4b8d1f61`.

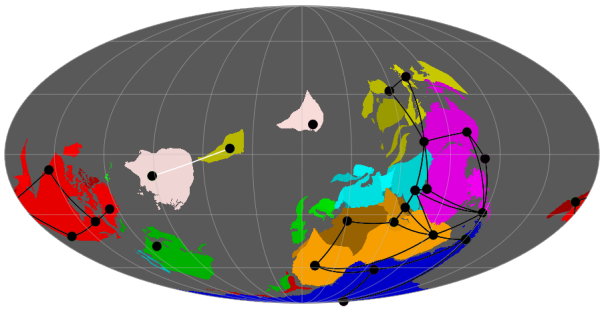
State	Abbrev.	Name
0	SAmN	South America (N)
1	SAmE	South America (E)
2	SAmS	South America (S)
3	NAmNW	North America (NW)
4	NAmNW	North America (SE)
5	NAmNW	North America (NE)
6	NAmNW	North America (SW)
7	Grn	Greenland
8	Eur	Europe
9	AsC	Asia (C)
A	AsE	Asia (E)
B	AsSE	Asia (SE)
C	AsNE	Asia (NE)
D	AfrW	Africa (W)
E	AfrS	Africa (S)
F	AfrE	Africa (E)
G	AfrN	Africa (N)
H	AusNW	Australia (NW)
I	AusSE	Australia (SE)
J	Ind	India
K	Mdg	Madgascar
L	AntW	Antarctica (W)
M	AntE	Antarctica (E)
N	Mly	Malaysian Archipelago
O	NZ	New Zealand

Table A.1: **List of areas.** Each row gives the full and abbreviated area names, plus the corresponding state value used for the analysis.

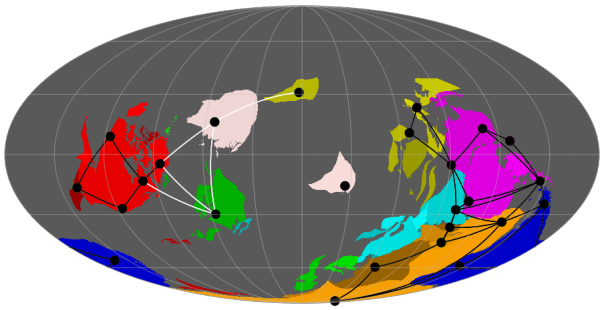
Index	Interval (approx.)	Start	End	Key events	Add'l Refs.
1	Cambrian	540	450		Z79, V04
2	Ordovician	450	430	NAm, As, Eur approach each other	Z79, V04
3	Silurian	430	400	NAm merges with Eur	Z79, V04
4	Devonian	400	350	Mly, AsSE separates from Aus	Z79, V04
5	Carboniferous	350	300	Mly, AsSE fully splits from Aus	Z79, V04
				NAm approaches SAm	Z79, V04
6	Early–Mid Permian	300	280	AsC, AsE, AsNE assembles	Z79, V04
				NAm, SAm floods	RR85, V04
7	Mid–Late Permian	280	240	Pangaea forms	V04, SS05
8	Triassic	240	200	AfrN approaches Eur	V04, SS05
				AsE positions between AsC, AsNE and AsSE, Mly	V04, SS05
9	Early Jurassic	200	170	Tethys Sea spreads between Laurasia and Gondwana	V04, SS05
10	Mid Jurassic	170	160	Laurasia separates from Gondwana	V04. SS05
				Asia forms	V04. SS05
11	Late Jurassic	160	150	Laurasia fully splits from Gondwana	SS05, AA08
				E.Gondwana separates from W.Gondwana	SS05
12	Early Cretaceous	150	120	Laurasia fragments	SS05
				Tethys Sea forms	SS05
13	Early Cretaceous	120	110	Afr separates from SAm	DH70
				NAmNW and AsNE near each other	G03
14	Early Cretaceous	110	100	Ind, Mdg separates from Aus, Ant	SS05
15	Late Cretaceous	100	90	W. Interior Seaway forms	A07
				Ind separates from Aus, Ant	AA08
16	Late Cretaceous	90	85	Afr fully splits from SAm	SS05
				Ind fully splits from Aus, Ant	AA08
				Bering Land Bridge forms	F08
17	Late Cretaceous	85	75	Aus separates from Ant	SS05, W13
18	Late Cretaceous	75	65	Ind fully splits from Mdg	AA08
				Aus fully splits from Ant	SS05, W13
19	Paleocene–Eocene	65	50	W. Interior Seaway removed	A07
				Bering Land Bridge removed	F08
20	Eocene–Oligocene	50	30	Ind approaches with AsC, AsSE	AA08
21	Oligocene	30	25	SAm separates from Ant	W13
				SAm approaches NAm	D90
				Ind merges with AsC, AsSE	AA08
22	Oligocene–Miocene	25	20	Tethys Sea removed	MH13
23	Miocene	20	13	Indoaustralian archipelago forms	L11
24	Pliocene–Pleistocene	13	0.1	SAm merges with NAm	D90, M15
25	Pleistocene	0.1	0.02	Bering Land Bridge forms	E96
26	Pleistocene–Holocene	0.02	0	Modern Earth	

Table A.2: **List of events.** Each row identifies one epoch assumed in the dispersal model, including the time interval, the key events, and supplemental references. All events and times were first established using Seton et al. (2012) and Wright et al. (2013) using `gplates` (Gurnis et al. 2012) and Blakey (2008). When applicable, events and times were supported by Supplemental References, given by the following abbreviations: DH70 = Dietz and Holden (1970), Z79 = Ziegler et al. (1979), RR85 = Ross and Ross (1985), D90 = Duque-Caro (1990), E96 = Elias et al. (1996), G03 = Golonka et al. (2003), V04 = Veevers (2004), SS05 = Schettino and Scotese (2005), A07 = Algeo et al. (2007), AA08 = Ali and Aitchison (2008), F08 = Fiorillo (2008), L11 = Lohman et al. (2011), MH13 = McQuarrie and van Hinsbergen (2013), W13 = White et al. (2013), N14 = Nance et al. (2014), M15 = Montes et al. (2015). Area abbreviations are given in Table A.1.

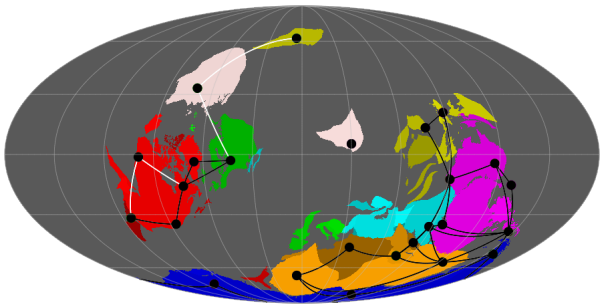
Figure A.1: **Paleogeographical dispersal graph for Earth from Cambrian until present.** Dispersal graphs for 26 epochs and 25 areas are shown. Areas are marked by black vertices. Modern continents share colors. Black edges indicate short- and medium-distance dispersal routes. White edges indicate exclusively medium-distance dispersal routes. Long-distance dispersal routes are not shown, but implied to exist between all area-pairs within an epoch. All images were produced using `gplates` (Gurnis et al. 2012).



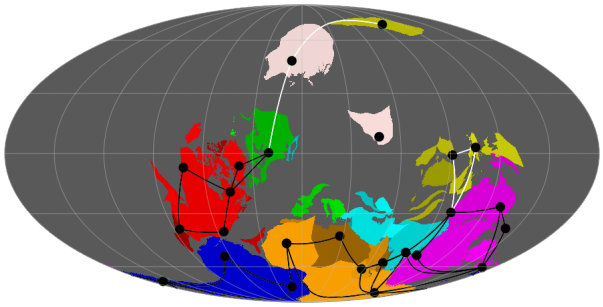
Epoch 1 of 26, 540–450Ma.



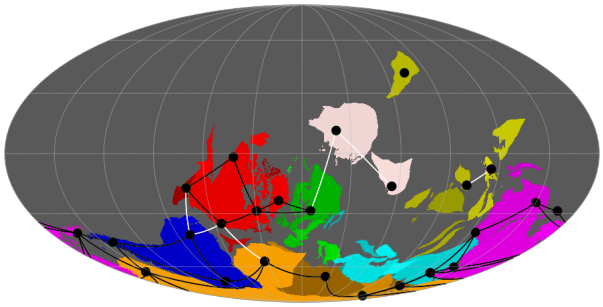
Epoch 2 of 26, 450–430Ma.



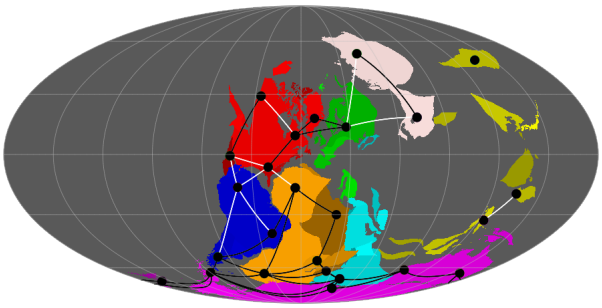
Epoch 3 of 26, 430–400Ma.



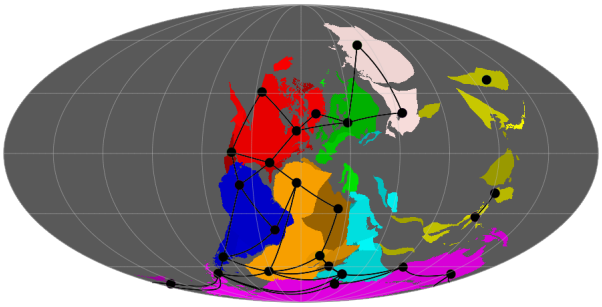
Epoch 4 of 26, 400–350Ma.



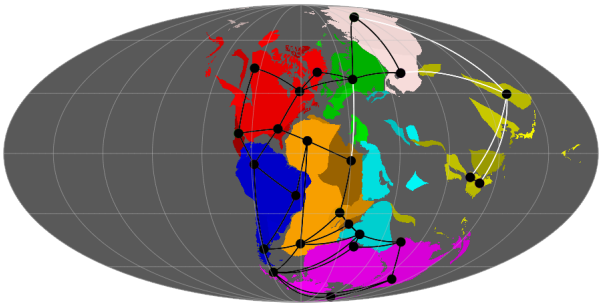
Epoch 5 of 26, 350–300Ma.



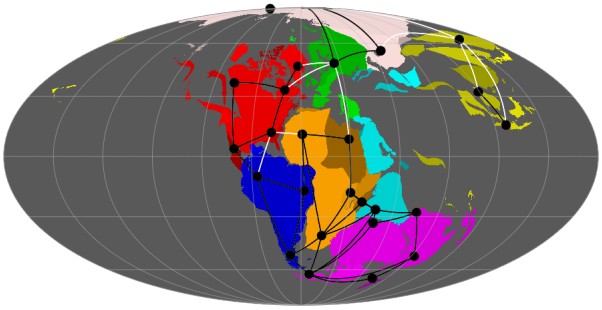
Epoch 6 of 26, 300–280Ma.



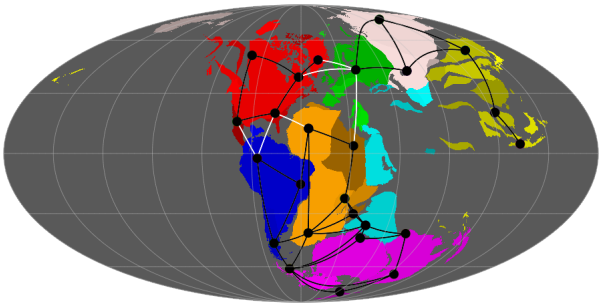
Epoch 7 of 26, 280–240Ma.



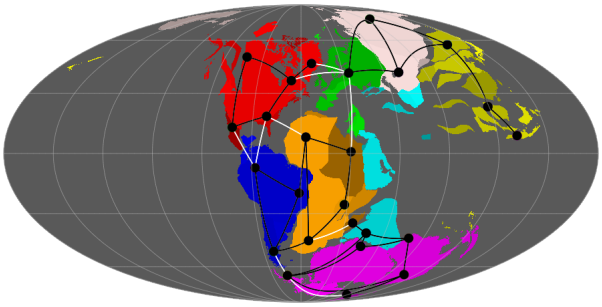
Epoch 8 of 26, 240–200Ma.



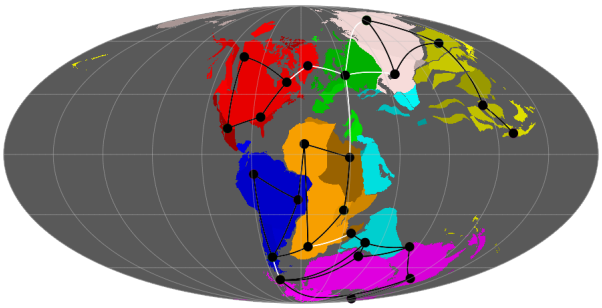
Epoch 9 of 26, 200–170Ma.



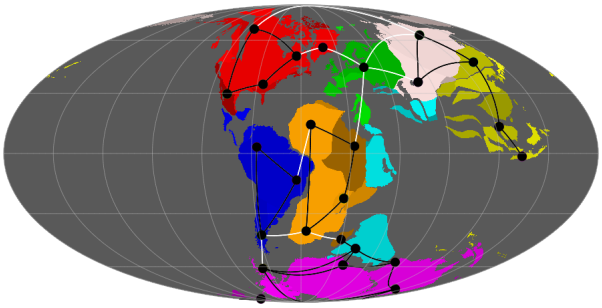
Epoch 10 of 26, 170–160Ma.



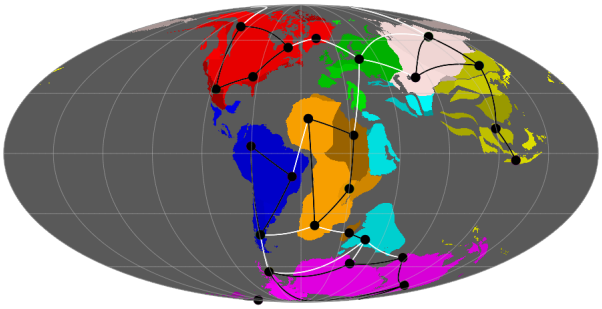
Epoch 11 of 26, 160–150Ma.



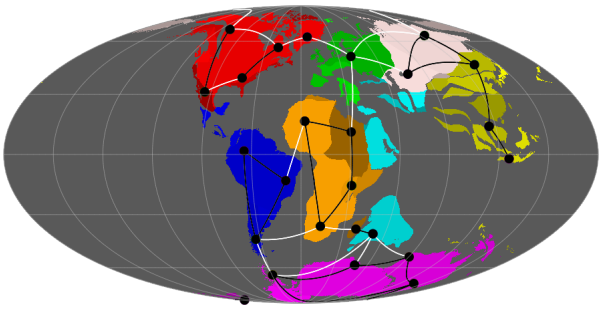
Epoch 12 of 26, 150–120Ma.



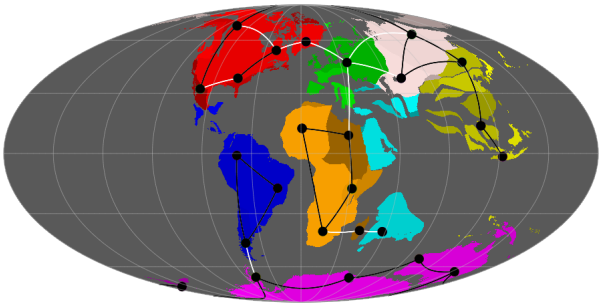
Epoch 13 of 26, 120–110Ma.



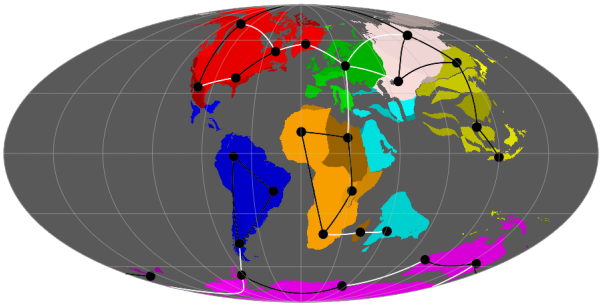
Epoch 14 of 26, 110–100Ma.



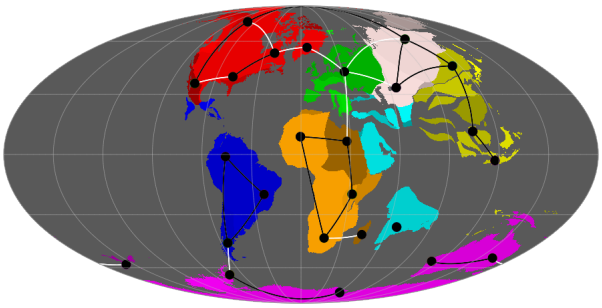
Epoch 15 of 26, 100–90Ma.



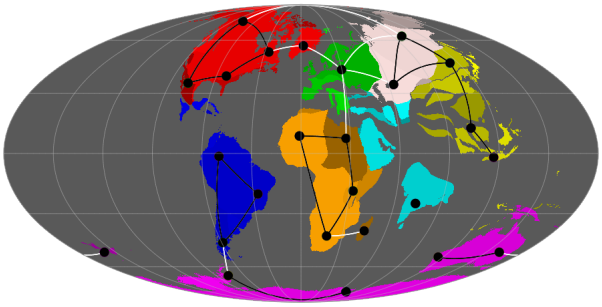
Epoch 16 of 26, 90–85Ma.



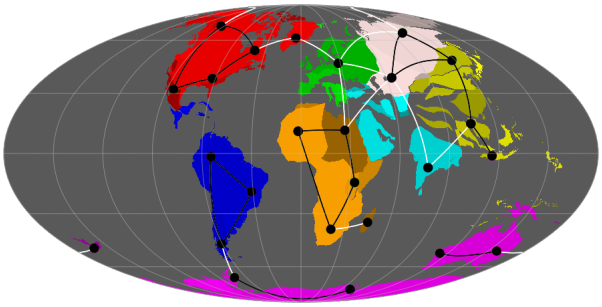
Epoch 17 of 26, 85–75Ma.



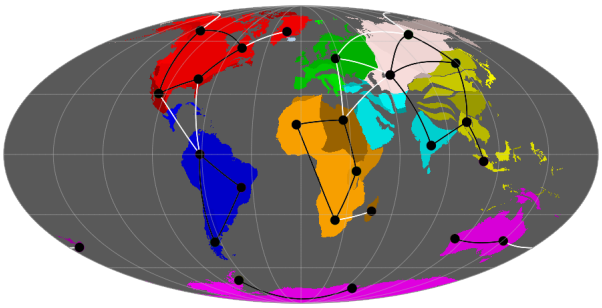
Epoch 18 of 26, 75–65Ma.



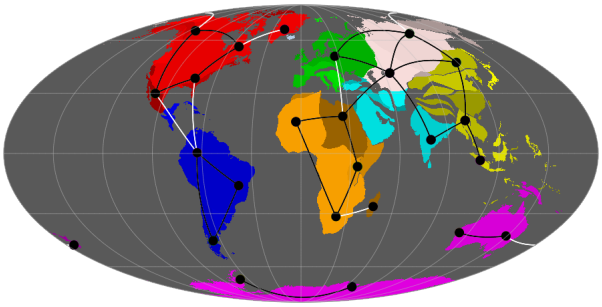
Epoch 19 of 26, 65–50Ma.



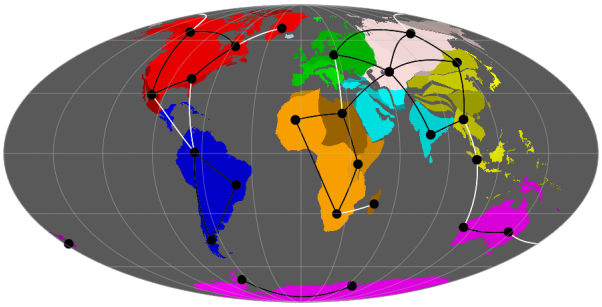
Epoch 20 of 26, 50–30Ma.



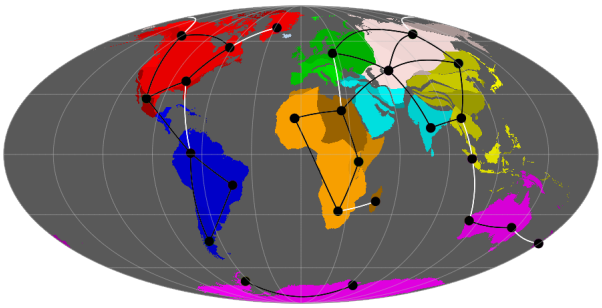
Epoch 21 of 26, 30–25Ma.



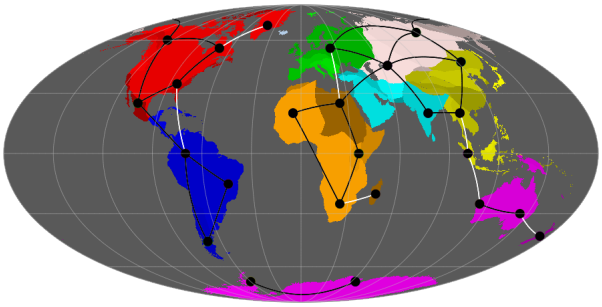
Epoch 22 of 26, 25–20Ma.



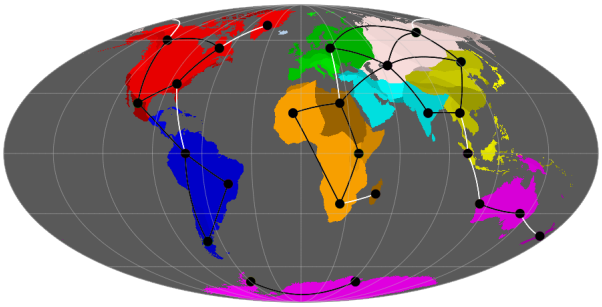
Epoch 23 of 26, 20–13Ma.



Epoch 24 of 26, 13–0.1Ma.



Epoch 25 of 26, 0.1–0.02Ma.



Epoch 26 of 26, 0.02–present.

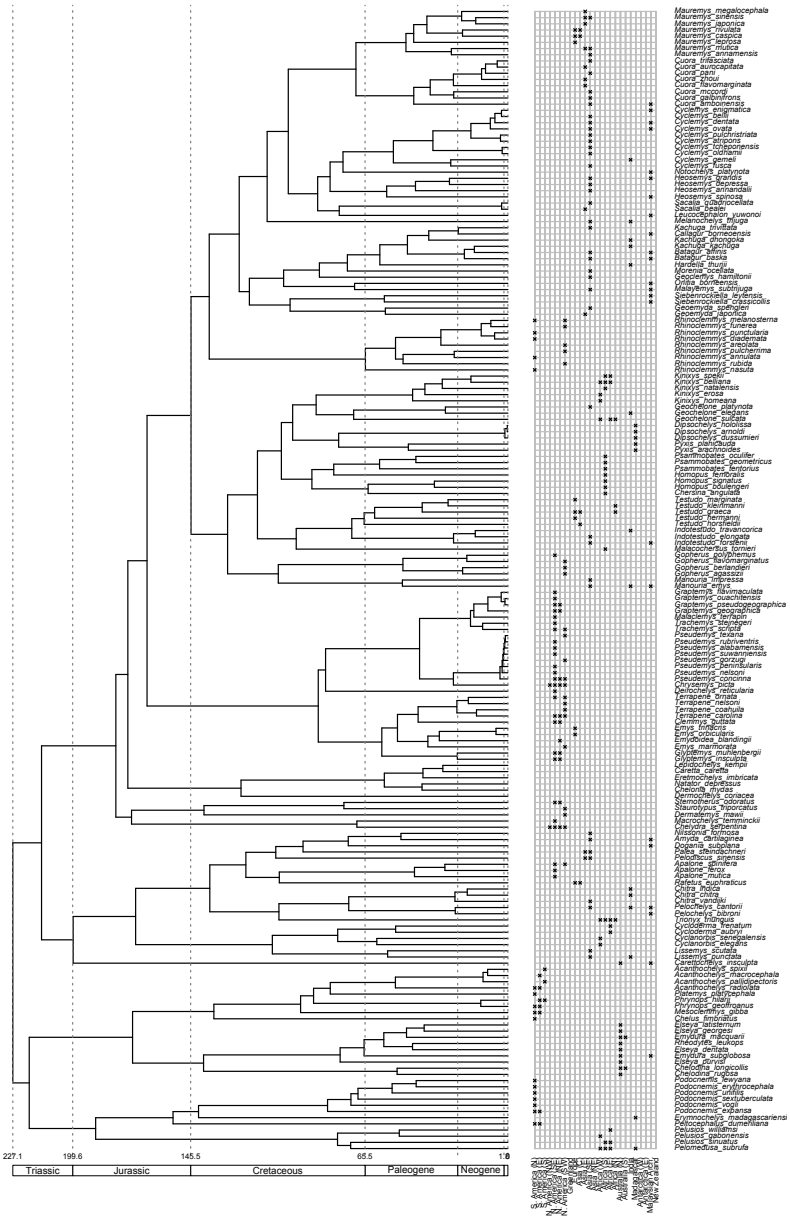


Figure A.2: **Maximum clade credibility tree for +G without root contraints.** Each taxon is marked with an 'X' for the area it is found in. Node ages are summarized using "Common Ancestor heights" using TreeAnnotator v2.3.0. Note, sea turtles are not assigned any areas (see main text).

Appendix B

Appendix: Lévy processess

Parameter	σ
True	0.1118
Mean	0.1085
RMSE	0.0068

(a) BM

Parameter	σ	λ	δ
True	.0500	.1110	.3000
Mean	.0506	.1271	.2194
RMSE	.0081	.0527	.1004

(b) JN

Parameter	σ	κ	τ
True	.1000	3.000	.4000
Mean	.0937	.7634	.3773
RMSE	.0698	2.269	.0347

(c) VG

Parameter	σ	α	β
True	.0500	1.500	.0500
Mean	.0684	1.430	.0392
RMSE	.0220	.1277	.0131

(d) AS

Table B.1: Mean and root-mean-square error (RMSE) for inference under Brownian motion (BM; 1A), compound Poisson with normally distributed jumps (JN; 1B), variance gamma (VG; 1C), and α-stable (AS; 1D) upon data simulated under the same model.

	Simulation (JN)	Simulation (BM)
V	.0120	.0119
	(.0085, .0176)	(.0092, .0158)
K	2.960	.0828
	(.7132, 9.067)	(2.112e-05, 1.184)

(a) JN

	Simulation (VG)	Simulation (BM)
V	.1549	.1713
	(.1189, .2077)	(.1337, .2243)
K	.3451	.0124
	(.0100, 1.953)	(1.909e-07, .7089)

(b) VG

	Simulation (AS)	Simulation (BM)
α	1.454	1.822
	(1.065, 1.748)	(1.292, 1.994)

(c) AS

Table B.2: MAP (maximum a priori) and 95% HPD (highest posterior density; below MAP in parentheses) for inference under JN (2A), VG (2B), and AS (2C) upon simulated data. The first column lists results for inference upon jump-present data simulated under the same model. The second column lists results for inference upon jump-absent data simulated under pure Brownian motion with equivalent variance ($\sigma = .1118$, $\sigma = .4050$, and $\sigma = .2389$ for analysis by JN, VG, and AS, respectively).

	BM	JN	AS	VG
σ	.1797	.0596	.1541	.0476
	(.1592, .2088)	(.0372, .1604)	(.0224, .1862)	(.0095, .1683)
λ	-	.2497	-	-
	-	(.0079, 1.188)	-	-
δ	-	.2929	-	-
	-	(.0758, .5389)	-	-
α	-	-	1.670	-
	-	-	(1.213, 1.998)	-
β	-	-	.0698	-
	-	-	(.0028, 0.1085)	-
κ	-	-	-	.6394
	-	-	-	(.0051, 2.990)
τ	-	-	-	.1767
	-	-	-	(.0689, .2155)

(a) Body mass

	BM	JN	AS	VG
σ	.1167	.0754	.0460	.0219
	(.1029, .1344)	(.0309, .1153)	(.0050, .0927)	(.0064, .1077)
λ	-	.0569	-	-
	-	(.0057, 1.122)	-	-
δ	-	.1313	-	-
	-	(.0244, .7425)	-	-
α	-	-	1.440	-
	-	-	(1.098, 1.797)	-
β	-	-	.0344	-
	-	-	(.0079, .0602)	-
κ	-	-	-	.4100
	-	-	-	(.0035, 3.095)
τ	-	-	-	.1145
	-	-	-	(.0481, .1427)

(b) ECV

	BM	JN	AS	VG
σ	.0965	.0871	.0912	.0770
	(.0853, .1111)	(.0485, .1071)	(.0304, .1070)	(.0186, .1009)
λ	-	.2209	-	-
	-	(.0001, 1.621)	-	-
δ	-	.0757	-	-
	-	(.0107, .1420)	-	-
α	-	-	1.926	-
	-	-	(1.269, 2.000)	-
β	-	-	.0076	-
	-	-	(2e-06, .0624)	-
κ	-	-	-	.1842
	-	-	-	(.0005, 1.570)
τ	-	-	-	.0818
	-	-	-	(.0261, .1041)

(c) Mass-to-ECV ratio

Table B.3: MAP and 95% HPD (below MAP in parentheses) of parameters inferred for evolution of primate body mass (3A), endocranial volume (ECV; 3B), and mass-to-ratio (3C). Results for BM, JN, VG, and AS listed per dataset.

	Body mass (JN)	Simulation (BM)
V	.0345	.0346
	(.0260, .0499)	(.0268, .0457)
K	1.401	.0133
	(1.030e-05, 4.796)	(2.966e-07, 1.041)

(a) Body mass, JN

	ECV (AS)	Simulation (BM)
α	1.440	1.859
	(1.098, 1.797)	(1.278, 2.000)

(b) ECV, AS

	Mass-to-ECV (JN)	Simulation (BM)
V	.0091	.0091
	(.0072, .0123)	(.0071, .0121)
K	.1061	.0900
	(2.166e-07, .8429)	(2.535e-05, 1.275)

(c) Mass-to-ECV ratio, JN

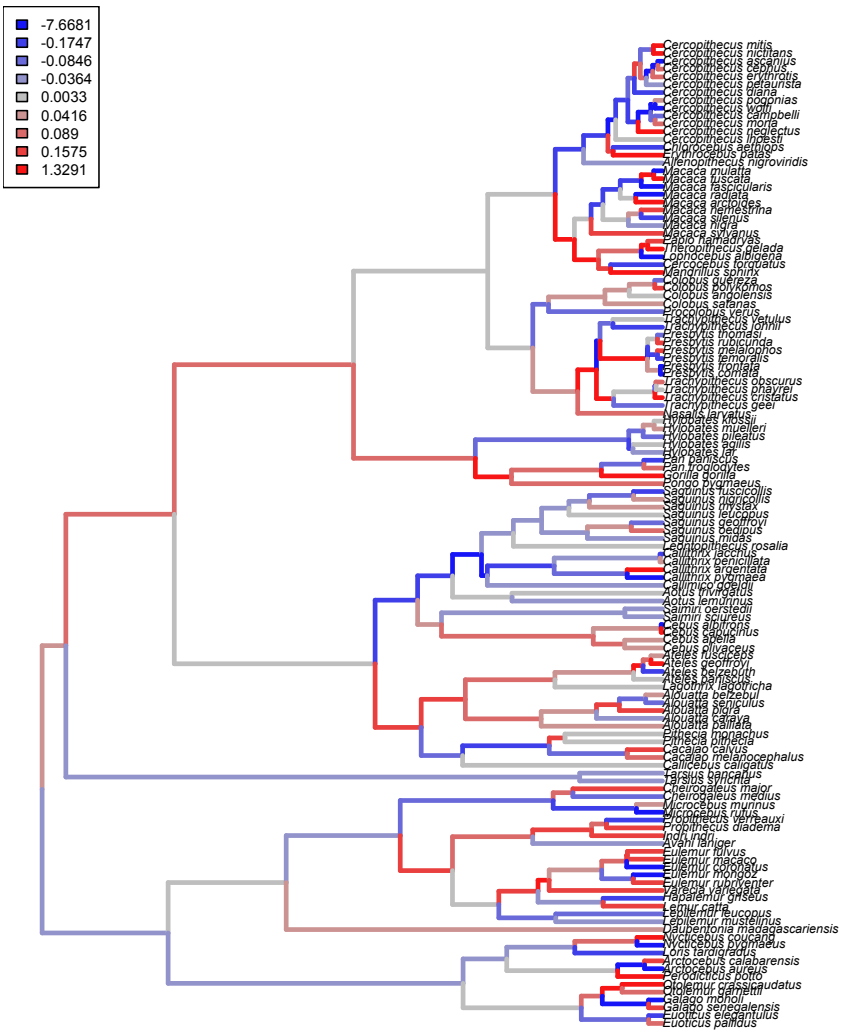
	Mass-to-ECV (VG)	Simulation (BM)
V	.0091	.0090
	(.0071, .0122)	(.0070, .0120)
K	.0193	.0360
	(5.552e-07, .6841)	(3.643e-05, .8430)

(d) Mass-to-ECV ratio, VG

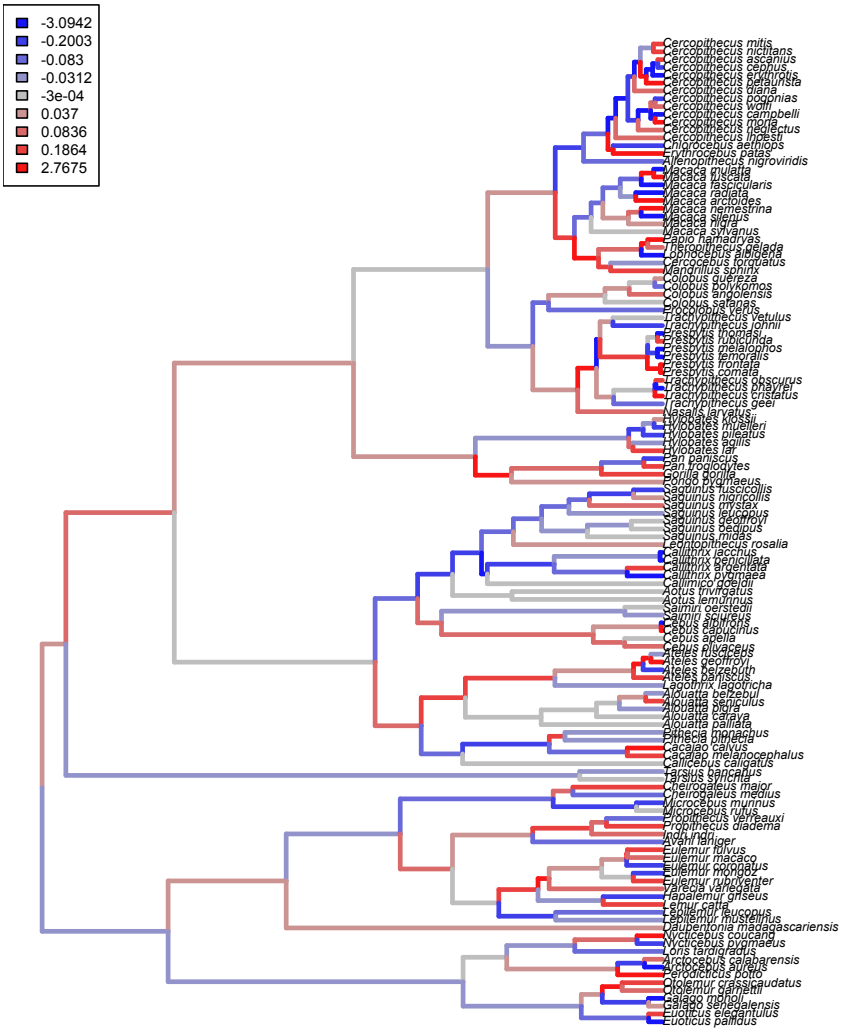
	Mass-to-ECV (AS)	Simulation (BM)
α	1.890	1.833
	(1.175, 1.998)	(1.280, 1.998)

(e) Mass-to-ECV-ratio, AS

Table B.4: MAP and 95% HPD (below MAP in parentheses) for inference under JN upon body mass data (4A), AS upon ECV data (4B), JN upon ratio data (4C), VG upon ratio data (4D), and AS upon ratio data (4E). The first column lists results for inference upon the empirical data. The second column lists results for inference upon jump-absent data simulated under pure Brownian motion with equivalent variance ($\sigma = .18$, $\sigma = .12$, and $\sigma = .096$ for body mass, ECV, and ratio data, respectively).



(a) Body mass under JN



(b) ECV under AS

Figure B.1: Branch-normalized signal-to-noise ratios (SNR) of posterior jump distributions. The primate phylogeny with inferred evolutionary histories for body mass under JN (Supp. Fig. 1A) and for ECV under AS (Supp. Fig. 1B) are shown. Branches are colored according to the quantile containing their branch length-normalized SNR ratios. A value of approximately zero indicates trait evolution explained predominantly by pure Brownian motion. Red and blue indicate the tendency for the model to explain trait evolution with positive and negative valued jumps, respectively. Supplemental Figure 1 is identical to Figure 7 except the tips are labeled with species names and the most recent common ancestor of great apes is not marked.

Bibliography

Abramowitz, M. and I. . Stegun. 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables vol. 55. Dover Publications.

Alfaro, M. E., F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, and L. J. Harmon. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proceedings of the National Academy of Sciences 106:13410–13414.

Algeo, T. J., T. W. Lyons, R. C. Blakey, and J. D. Over. 2007. Hydrographic conditions of the devono–carboniferous north american seaway inferred from sedimentary mo–toc relationships. Palaeogeography, Palaeoclimatology, Palaeoecology 256:204–230.

Ali, J. R. and J. C. Aitchison. 2008. Gondwana to asia: Plate tectonics, paleogeography and the biological connectivity of the indian sub-continent from the middle jurassic through latest eocene (166–35 ma). Earth-Science Reviews 88:145–166.

Ané, C. 2008. Analysis of comparative data with hierarchical autocorrelation. Annals of Applied Statistics 2:1078–1102.

Beaulieu, J. M., D. C. Tank, and M. J. Donoghue. 2013. A southern hemisphere origin for campanulid angiosperms, with traces of the break-up of gondwana. BMC Evolutionary Biology 13:80.

Behrensmeyer, A. K., S. M. Kidwell, and R. A. Gastaldo. 2000. Taphonomy and paleobiology. Paleobiology 26:103–147.

Bielejec, F., P. Lemey, G. Baele, A. Rambaut, and M. A. Suchard. 2014. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. Systematic Biology 63:493–504.

Blakey, R. 2006. Global paleogeographic views of earth history: Late Precambrian to Recent. R. Blakey.

Blakey, R. C. 2008. Gondwana paleogeography from assembly to breakup - a 500 my odyssey. Geological Society of America Special Papers 441:1–28.

Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? measuring the power of comparative methods. Evolution 66:2240–2251.

Bokma, F. 2002. Detection of punctuated equilibrium from molecular phylogenies. Journal of Evolutionary Biology 15:1048–1056.

Bokma, F. 2008. Detection of "punctuated equilibrium" by Bayesian estimation of speciation and extinction rates, ancestral character states, and rates of anagenetic and cladogenetic evolution on a molecular phylogeny. Evolution 62:2718–2726.

Boyden, J. A., R. D. Müller, M. Gurnis, T. H. Torsvik, J. A. Clark, M. Turner, H. Ivey-Law, R. J. Watson, and J. S. Cannon. 2011. Next-generation plate-tectonic reconstructions using gplates. Geoinformatics: cyberinfrastructure for the solid earth sciences Pages 95–114.

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343–348.

Brown, G., G. Nelson, and P. Y. Ladiges. 2006. Historical biogeography of rhododendron section vireya and the malesian archipelago. Journal of biogeography 33:1929–1944.

Buerki, S., F. Forest, N. Alvarez, J. A. A. Nylander, N. Arrigo, and I. Sanmartín. 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 38:531–550.

Carlquist, S. 1966. The biota of long-distance dispersal. i. principles of dispersal and evolution. Quarterly Review of Biology Pages 247–270.

Cavalli-Sforza, L. L. and A. W. F. Edwards. 1967. Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics 19:233–257.

Chaix, R., M. Somel, D. P. Kreil, P. Khaitovich, and G. A. Lunter. 2008. Evolution of primate gene expression: drift and corrective sweeps? Genetics 180:1379–1389.

Clark, J. R., R. H. Ree, M. E. Alfaro, M. G. King, W. L. Wagner, and E. H. Roalson. 2008. A comparative study in ancestral range reconstruction methods: retracing the uncertain histories of insular lineages. Systematic Biology 57:693–707.

Contributors, G. P. 2010. GSL - GNU Scientific Library - GNU Project - Free Software Foundation (FSF). The GNU Operating System.

Crawford, N. G., J. F. Parham, A. B. Sellas, B. C. Faircloth, T. C. Glenn, T. J. Papenfuss, J. B. Henderson, M. H. Hansen, and B. W. Simison. 2015. A phylogenomic analysis of turtles. Molecular Phylogenetics and Evolution 83:250–257.

Danilov, I. G. and J. F. Parham. 2008. A reassessment of some poorly known turtles from the middle jurassic of china, with comments on the antiquity of extant turtles. Journal of Vertebrate Paleontology 28:306–318.

Darwin, C. 1859. On the Origin of Species. J. Murray, London.

Dickey, J. 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. The Annals of Statistics 42:204–223.

Dietz, R. S. and J. C. Holden. 1970. Reconstruction of pangaea: breakup and dispersion of continents, permian to present. Journal of Geophysical Research 75:4939–4956.

Dornburg, A., J. M. Beaulieu, J. C. Oliver, and T. J. Near. 2011. Integrating fossil preservation biases in the selection of calibrations for molecular divergence time estimation. Systematic Biology 60:519–527.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4:e88.

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

Duque-Caro, H. 1990. Neogene stratigraphy, paleoceanography and paleobiogeography in northwest south america and the evolution of the panama seaway. Palaeogeography, Palaeoclimatology, Palaeoecology 77:203–234.

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

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792–1797.

Eldredge, N. and S. J. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradualism. Models in Paleobiology 82:115.

Elias, S. A., S. K. Short, C. H. Nelson, and H. H. Birks. 1996. Life and times of the bering land bridge. Nature 382:60–63.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Felsenstein, J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1–15.

Fiorillo, A. R. 2008. Dinosaurs of alaska: implications for the cretaceous origin of beringia. Geological Society of America Special Papers 442:313–326.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2003. Bergmann's rule and body size in mammals. The American Naturalist 161:821–825.

Gallager, R. G. 1962. Low-density parity-check codes. IRE Trans. Inform. Theory 8:21–28.

Gavryushkina, A., T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch, and A. J. Drummond. 2015. Bayesian total evidence dating reveals the recent crown radiation of penguins. arXiv preprint arXiv:1506.04797 .

Gelman, A. and D. B. Rubin. 1992. Inferences from iterative simulation using multiple sequences. Statistical Science 7:457–511.

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

Golonka, J., N. Y. Bocharova, D. Ford, M. E. Edrich, J. Bednarczyk, and J. Wildharber. 2003. Paleogeographic reconstructions and basins development of the Arctic. Marine and Petroleum Geology 20:211–248.

Golub, G. H. and C. F. V. Loan. 1983. Matrix Computations. Johns Hopkins University Press, Baltimore, Maryland.

Guillon, J.-M., L. Guéry, V. Hulin, and M. Girondot. 2012. A large phylogeny of turtles (testudines) using molecular data. Contributions to Zoology 81:147–158.

Gurnis, M., M. Turner, S. Zahirovic, L. DiCaprio, S. Spasojevic, R. D. Müller, J. Boyden, M. Seton, V. C. Manea, and D. J. Bower. 2012. Plate tectonic reconstructions with continuously closing plates. Computers & Geosciences 38:35–42.

Hansen, T. F. and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution Pages 1404–1417.

Hanski, I. 1998. Metapopulation dynamics. Nature 396:41–49.

Harmon, L. J., J. B. Losos, T. J. Davies, R. G. Gillespie, J. L. Gittleman, W. B. Jennings, K. H. Kozak, M. A. McPeek, F. Moreno-Roark, T. J. Near, et al. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Heads, M. 2005. Dating nodes on molecular phylogenies: a critique of molecular biogeography. Cladistics 21:62–78.

Heads, M. 2011. Old taxa on young islands: a critique of the use of island age to date island-endemic clades and calibrate phylogenies. Systematic Biology 60:204–218.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:E2957–E2966.

Heled, J. and A. J. Drummond. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. Systematic Biology 61:138–149.

Ho, S. Y., K. J. Tong, C. S. Foster, A. M. Ritchie, N. Lo, and M. D. Crisp. 2015. Biogeographic calibrations for the molecular clock. Biology letters 11:20150194.

Ho, S. Y. W. and M. J. Phillips. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Systematic Biology 58:367–380.

Höhna, S., T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P. Huelsenbeck. 2014. Probabilistic graphical model representation in phylogenetics. Systematic Biology 63:753–771.

Huelsenbeck, J. P., B. Larget, and D. L. Swofford. 2000. A compound poisson process for relaxing the molecular clock. Genetics 154:1879–1892.

Hugall, A. F., R. Foster, and M. S. Y. Lee. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene rag-1. Systematic Biology 56:543–563.

Isler, K., E. C. Kirk, J. Miller, G. A. Albrecht, B. R. Gelvin, and R. D. Martin. 2008. Endocranial volumes of primate species: scaling analyses using a comprehensive and reliable data set. Journal of Human Evolution 55:967–978.

Jeffreys, H. 1961. Theory of Probability. Oxford University Press, Oxford.

Joyce, W. G. 2007. Phylogenetic relationships of mesozoic turtles. Bulletin of the Peabody Museum of Natural History 48:3–102.

Joyce, W. G., J. F. Parham, T. R. Lyson, R. C. M. Warnock, and P. C. J. Donoghue. 2013. A divergence dating analysis of turtles using fossil calibrations: an example of best practices. Journal of Paleontology 87:612–634.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian Protein Metabolism (H. N. Munro, ed.) Academic Press.

Kallenberg, O. 2010. Foundations of Modern Probability. Springer.

Khaitovich, P., S. Pääbo, and G. Weiss. 2005. Toward a neutral evolutionary model of gene expression. Genetics 170:929–939.

Kidwell, S. M. and S. M. Holland. 2002. The quality of the fossil record: implications for evolutionary analyses. Annual Review of Ecology and Systematics Pages 561–588.

Kodandaramaiah, U. 2011. Tectonic calibrations in molecular dating. Current Zoology 57:116–124.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution Pages 314–334.

Landis, M. J. 2015. Biogeographic dating of speciation times using paleogeographically informed processes. Systematic Biology In review.

Landis, M. J., N. J. Matzke, B. R. Moore, and J. P. Huelsenbeck. 2013. Bayesian analysis of biogeography when the number of areas is large. Systematic Biology 62:789–804.

Lartillot, N. and R. Poujol. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Molecular Biology and Evolution Pages 729–744.

Lemey, P., A. Rambaut, A. J. Drummond, and M. A. Suchard. 2009. Bayesian phylogeography finds its roots. PLoS Computational Biology 5:e1000520.

Lemey, P., A. Rambaut, J. J. Welch, and M. A. Suchard. 2010. Phylogeography takes a relaxed random walk in continuous space and time. Moleculate Biology and Evolution 27:1877–1885.

Lemmon, A. A. and E. M. Lemmon. 2008. A likelihood framework for estimating phylogeographic history on a continuous landscape. Systematic Biology 57:544–561.

Lemmon, A. R. and E. C. Moriarty. 2004. The importance of proper model assumption in bayesian phylogenetics. Systematic Biology 53:265–277.

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Molecular Biology and Evolution 24:2669–2680.

Li, H., M. T. Wells, and C. L. Yu. 2008. A Bayesian analysis of return dynamics with Lévy jumps. Review of Financial Studies 21:2345–2378.

Lohman, D. J., M. de Bruyn, T. Page, K. von Rintelen, R. Hall, P. K. L. Ng, H.-T. Shih, G. R. Carvalho, and T. von Rintelen. 2011. Biogeography of the Indo-Australian archipelago. Annual Review of Ecology, Evolution, and Systematics 42:205–226.

MacArthur, R. H. and E. O. Wilson. 1967. The theory of island biogeography. Princeton University Press, New Jersey.

Madan, D. B. and P. P. Carr. 1998. The variance gamma process and option pricing. European Finance Review Pages 79–105.

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Systematic Biology 56:701–710.

Martin, A. P. and S. R. Palumbi. 1993. Body size, metabolic rate, generation time, and the molecular clock. Proceedings of the National Academy of Sciences 90:4087–4091.

Mattila, T. M. and F. Bokma. 2008. Extant mammal body masses suggest punctuated equilibrium. Proceedings of Royal Society of London [Biol] 275:2195–2199.

Matzke, N. J. 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. Systematic Biology Page syu056.

McQuarrie, N. and D. J. J. van Hinsbergen. 2013. Retrodeforming the arabia-eurasia collision zone: Age of collision versus magnitude of continental subduction. Geology 41:315–318.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21:1087–1092.

Minin, V. N. and M. A. Suchard. 2007. Counting labeled transitions in continuous-time markov models of evolution. Journal of Mathematical Biology .

Montes, C., A. Cardona, C. Jaramillo, A. Pardo, J. C. Silva, V. Valencia, C. Ayala, L. C. Pérez-Angel, L. A. Rodriguez-Parra, V. Ramirez, and H. N. no. 2015. Middle miocene closure of the central american seaway. Science 348:226–229.

Moore, B. R., S. A. Smith, R. H. Ree, and M. J. Donoghue. 2008. Incorporating fossil data in biogeographic inference: A likelihood approach. Evolution .

Nance, R. D., J. B. Murphy, and M. Santosh. 2014. The supercontinent cycle: A retrospective essay. Gondwana Research 25:4–29.

Nielsen, R. 2002. Mapping mutations on phylogenies. Systematic Biology 51:729–739.

Nylander, J. A., U. Olsson, P. Alström, and I. Sanmartín. 2008. Accounting for phylogenetic uncertainty in biogeography: a bayesian approach to dispersal-vicariance analysis of the thrushes (aves: Turdus). Systematic Biology 57:257–268.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–33.

Parham, J. F., P. C. J. Donoghue, C. J. Bell, T. D. Calway, J. J. Head, P. A. Holroyd, J. G. Inoue, R. B. Irmis, W. G. Joyce, D. T. Ksepka, J. S. L. Patané, N. D. Smith, J. E. Tarver, M. van Tuinen, Z. Yang, K. D. Angielczyk, J. M. Greenwood, C. A. Hipsley, L. Jacobs, P. J. Makovicky, J. Müller, K. T. Smith, J. M. Theodor, and R. C. M. Warnock. 2011. Best practices for justifying fossil calibrations. Systematic Biology Page syr107.

Passalacqua, N. G. 2015. On the definition of element, chorotype and component in biogeography. Journal of Biogeography 42:611–618.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News Pages 7–11.

Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of lissamphibia. Systematic Biology 60:466–481.

Qin, X., R. Müller, J. Cannon, T. Landgrebe, C. Heine, R. Watson, and M. Turner. 2012. The gplates geological information model and markup language. Geosci. Instrum. Methods Data Syst. Discuss 2:365–428.

R Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria ISBN 3-900051-07-0.

Redding, D. W., C. DeWolff, and A. Ø. Mooers. 2010. Evolutionary distinctiveness, threat status, and ecological oddity in primates. Conservation Biology 24:1052–1058.

Ree, R. H., B. R. Moore, C. O. Webb, and M. J. Donoghue. 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. Evolution 59:2299–2311.

Ree, R. H. and I. Sanmartín. 2009. Prospects and challenges for parametric models in historical biogeographical inference. Journal of Biogeography 36:1211–1220.

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

Renner, S. S. 2005. Relaxed molecular clocks for dating historical plant dispersal events. Trends in Plant Science 10:550–558.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Molecular Biology and Evolution 20:1692–1704.

Rodrigue, N. and H. Philippe. 2010. Mechanistic revisions of phenomenological modeling strategies in molecular evolution. Trends in Genetics 26:248–252.

Ronquist, F. 1997. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. Systematic Biology 46:195–203.

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. L. Murray, and A. P. Rasnitsyn. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. Systematic Biology 61:973–999.

Ronquist, F. and I. Sanmartín. 2011. Phylogenetic methods in biogeography. Annual Review of Ecology, Evolution, and Systematics 42:441–464.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012b. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Systematic biology 61:539–542.

Ross, C. A. and J. R. P. Ross. 1985. Late paleozoic depositional sequences are synchronous and worldwide. Geology 13:194–197.

Sanmartín, I., P. V. D. Mark, and F. Ronquist. 2008. Inferring dispersal: a bayesian approach to phylogeny-based island biogeography, with special reference to the canary islands. Journal of Biogeography 35:428–449.

Schettino, A. and C. R. Scotese. 2005. Apparent polar wander paths for the major continents (200 ma to the present day): a palaeomagnetic reference frame for global plate tectonic reconstructions. Geophysical Journal International 163:727–759.

Seton, M., R. D. Müller, S. Zahirovic, C. Gaina, T. Torsvik, G. Shephard, A. Talsma, M. Gurnis, M. Turner, and S. Maus. 2012. Global continental and ocean basin reconstructions since 200ma. Earth-Science Reviews 113:212–270.

Simpson, G. G. 1953. The Major Features of Evolution. Columbia University Press., New York.

Slater, G. J., L. J. Harmon, and M. E. Alfaro. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944.

Stanley, S. M. 1975. A theory of evolution above the species level. Proceedings of the National Academy of Sciences, USA 72:646.

Sterli, J., D. Pol, and M. Laurin. 2013. Incorporating phylogenetic uncertainty on phylogeny-based palaeontological dating and the timing of turtle diversification. Cladistics 29:233–246.

Tagliacollo, V. A., S. M. Duke-Sylvester, W. A. Matamoros, P. Chakrabarty, and J. S. Albert. 2015. Coordinated dispersal and pre-isthmian assembly of the central american ichthyofauna. Systematic Biology Page syv064.

Thorne, J., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution 15:1647–1657.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). Annals of Statistics 22:1701–1762.

Towns, J., T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, et al. 2014. Xsede: Accelerating scientific discovery. Computing in Science & Engineering 16:62–74.

Uyeda, J. C., T. F. Hansen, S. J. Arnold, and J. Pienaar. 2011. The million-year wait for macroevolutionary bursts. Proceedings of the National Academy of Sciences 108:15908–15913.

Veevers, J. J. 2004. Gondwanaland from 650–500 ma assembly through 320 ma merger in pangea to 185–100 ma breakup: supercontinental tectonics via stratigraphy and radiometric dating. Earth-Science Reviews 68:1–132.

Verdinelli, I. and L. Wasserman. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. Journal of the American Statistical Association 90:614–618.

Wallace, A. R. 1887. Oceanic islands: Their physical and biological relations. Bulletin of the American Geographical Society 19:1–21.

Warnock, R. C., J. F. Parham, W. G. Joyce, T. R. Lyson, and P. C. J. Donoghue. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proceedings of the Royal Society of London B: Biological Sciences 282:20141013.

Webb, C. O. and R. H. Ree. 2012. Historical biogeography inference in Malesia. Pages 191–215 *in* Biotic evolution and environmental change in Southeast Asia (D. Gower, K. Johnson, J. Richardson, B. Rosen, L. Ruber, and S. Williams, eds.) Cambridge University Press.

White, L. T., G. M. Gibson, and G. S. Lister. 2013. A reassessment of paleogeographic reconstructions of eastern gondwana: Bringing geology back into the equation. Gondwana Research 24:984–998.

Wilf, P. and I. H. Escapa. 2014. Green web or megabiased clock? plant fossils from gondwanan patagonia speak on evolutionary radiations. New Phytologist 207:283–290.

Wolfe, K. H., W.-H. Li, and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear dnas. Proceedings of the National Academy of Sciences 84:9054–9058.

Wood, H. M., N. J. Matzke, R. G. Gillespie, and C. E. Griswold. 2012. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Systematic Biology 62:264–284.

Wright, N., S. Zahirovic, R. Müller, and M. Seton. 2013. Towards community-driven paleogeographic reconstructions: integrating open-access paleogeographic and paleobiology data with plate tectonics. Biogeosciences 10:1529–1541.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution 39:306–314.

Ziegler, A. M., C. R. Scotese, W. S. McKerrow, M. E. Johnson, and R. K. Bambach. 1979. Paleozoic Paleogeography. Annual Review of Earth and Planetary Sciences 7:473–502.

Zuckerkandl, E. and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pages 189–225 *in* Horizons in Biochemistry (M. Kasha and B. Pullman, eds.) Academic Press, New York.