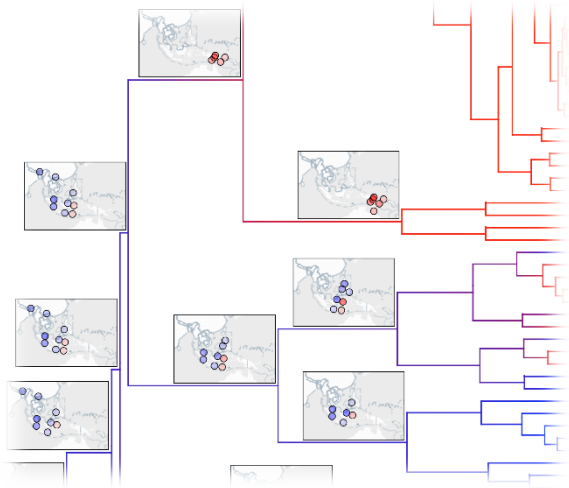
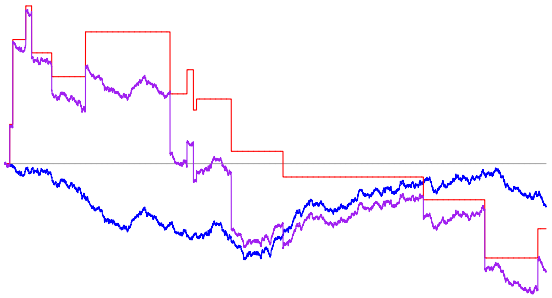
Phylogeny & Biogeography

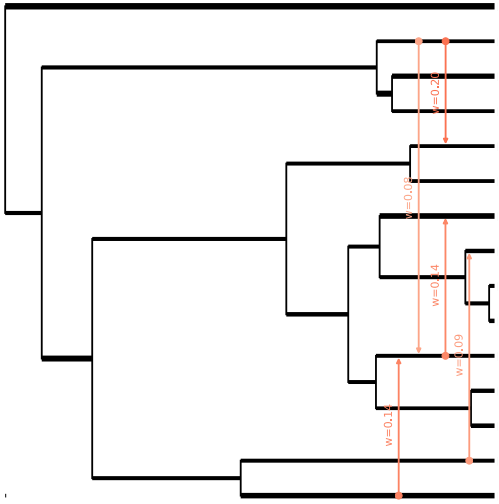
Michael Landis
mlandis@berkeley.edu
11 March 2014



Range evolution



Continuous trait evolution



Admixture graphs

Biogeography

How life is distributed in space and time

Watercolor, Joanna Barnum

"Every species has come into existence coincident both in space and time with a pre-existing closely allied species."

AR Wallace, 1855

Outline

Definitions and examples

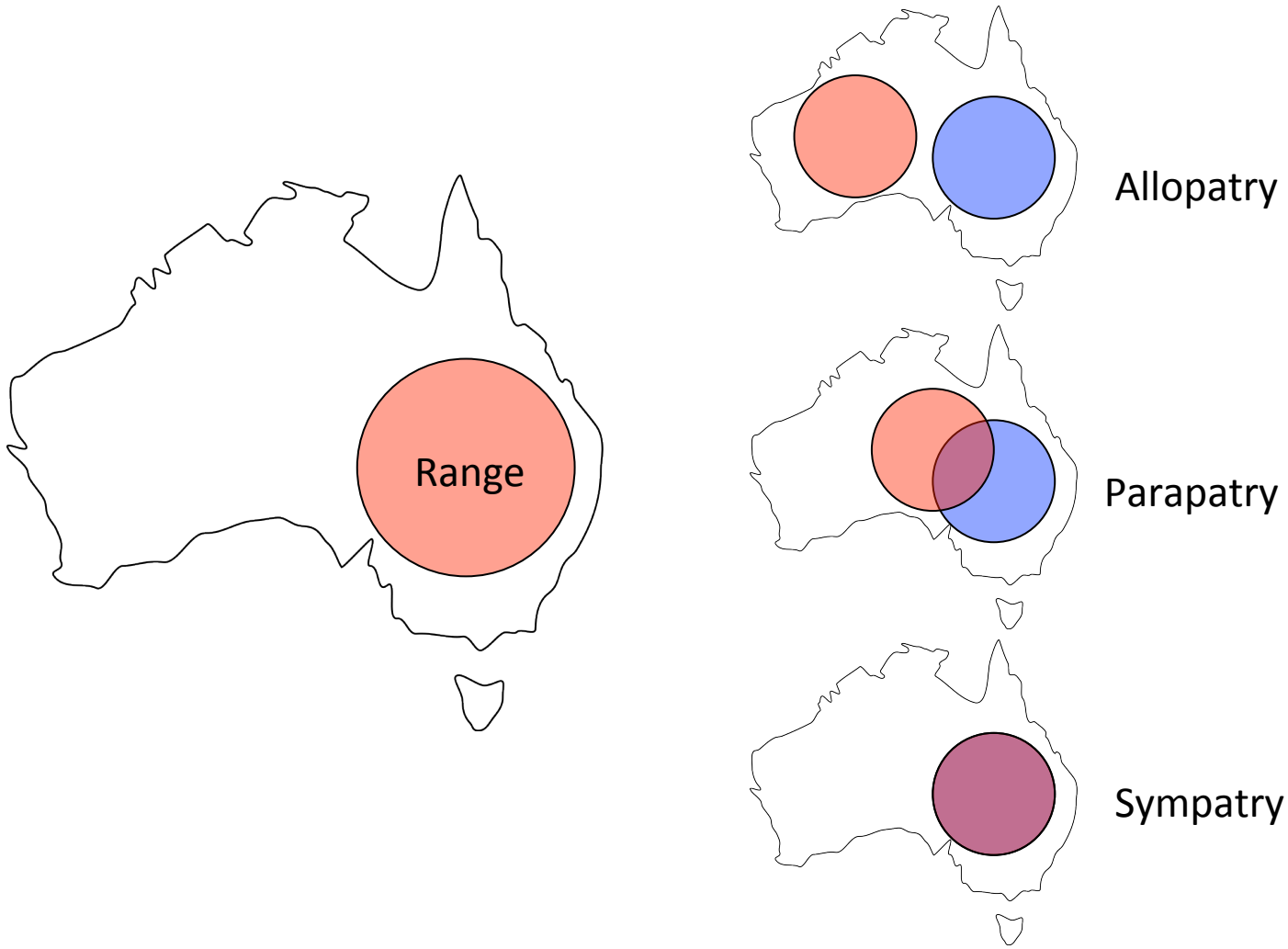
Phylogenetic inference

Discrete models

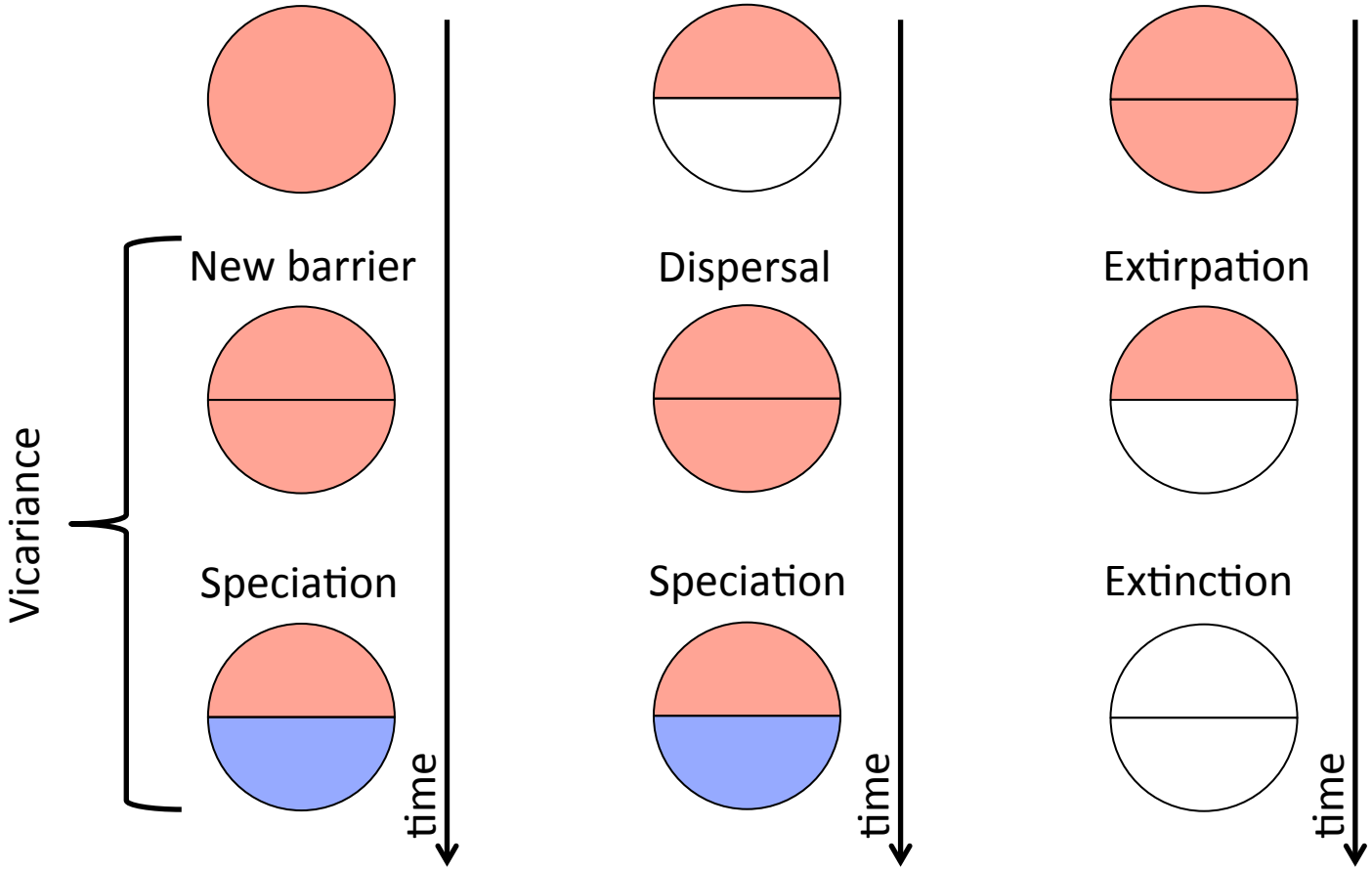
Continuous models

Biogeography lab

Biogeographic patterns



Biogeographic processes



Some of the big puzzles

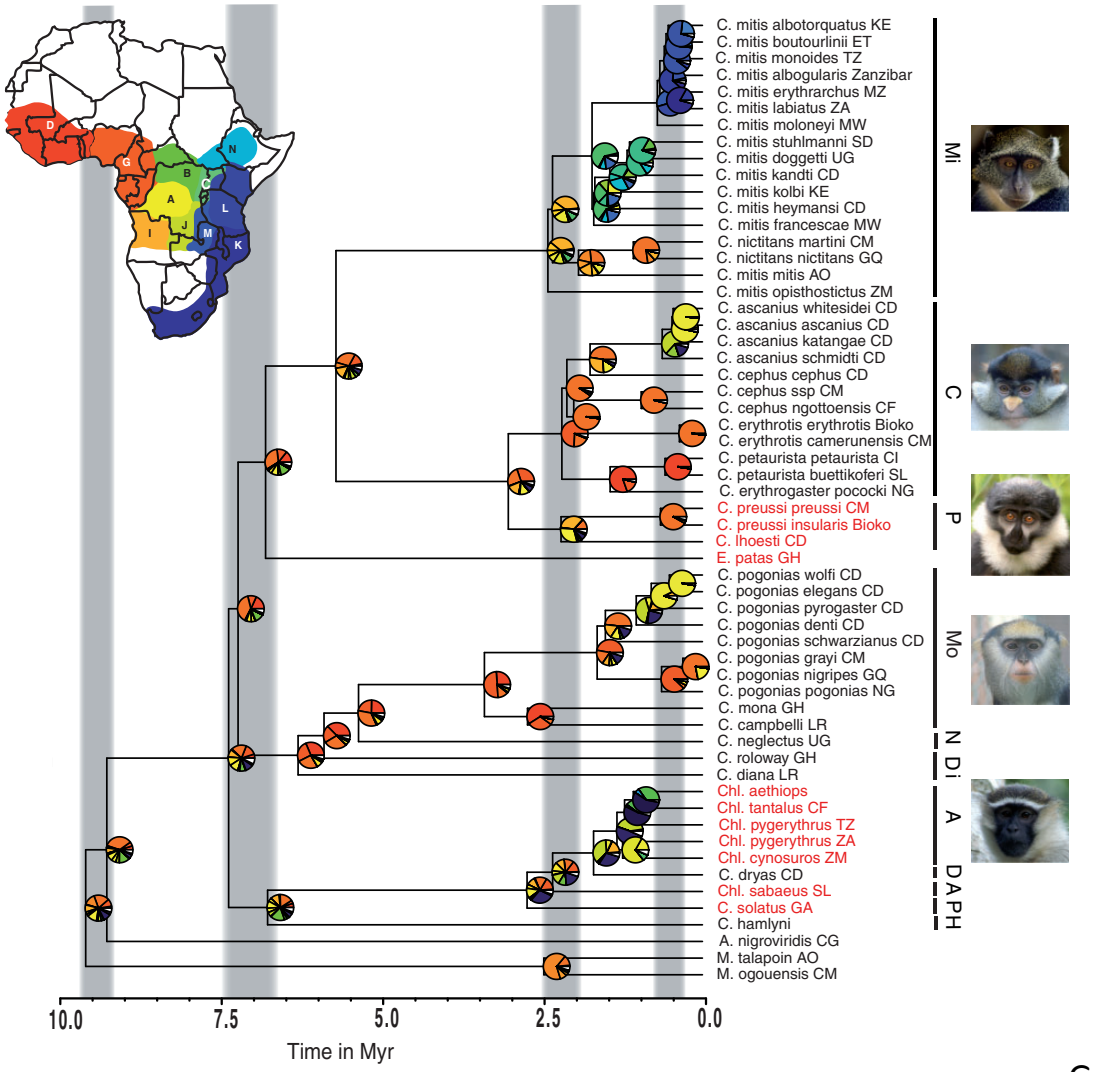
Where did ancestral species live?

How does range size affect speciation/extinction?

What species traits help/harm colonization?

How does geography affect range?

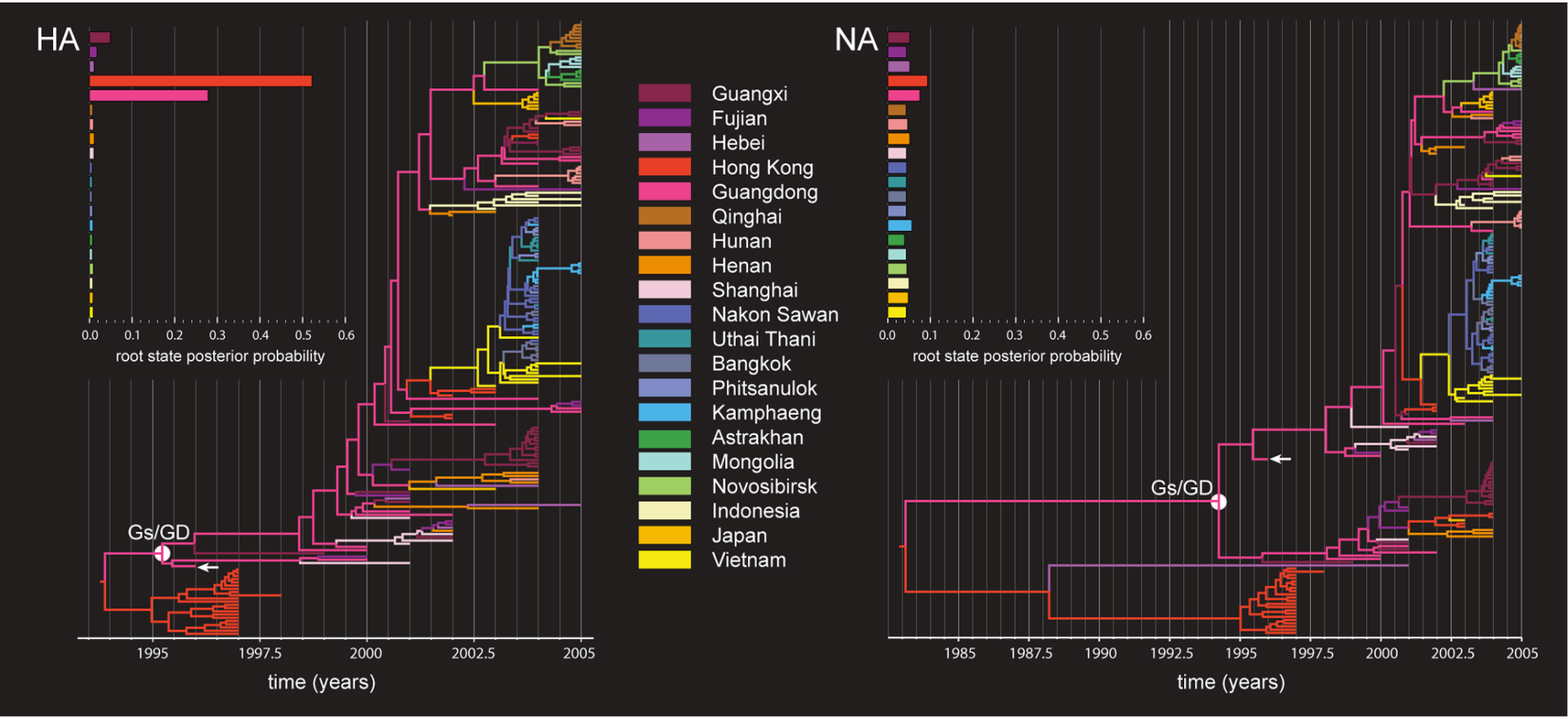
Biodiversity



Cercopithecidae (Primates)

Guschanski *et al.,* 2013 (Syst Biol)

Epidemiology



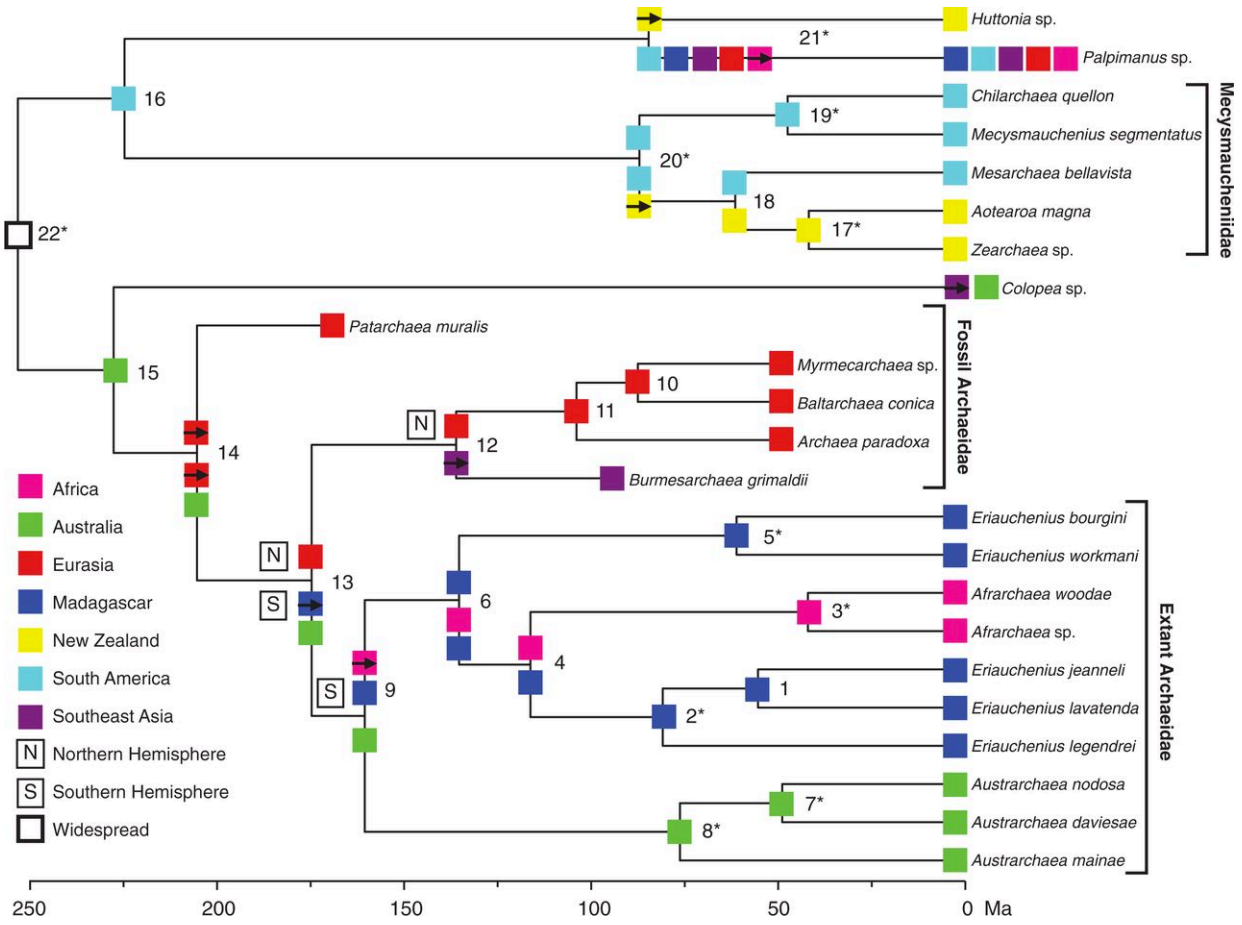
Hemagglutin (HA)

H5N1
(Avian Flu)

Neuraminidase (NA)

Lemey *et al.,* 2008 (PLoS Comp Biol)

Divergence time estimation



Archaeidae
(Assassin spiders)

Wood *et al.,* 2012 (Syst Biol)

Statistical phylogenetics

Familiar strategy:

- Data matrix (homology)
- Time-calibrated phylogeny
- Transition probability of change along branches
- Integrate over ancestral characters

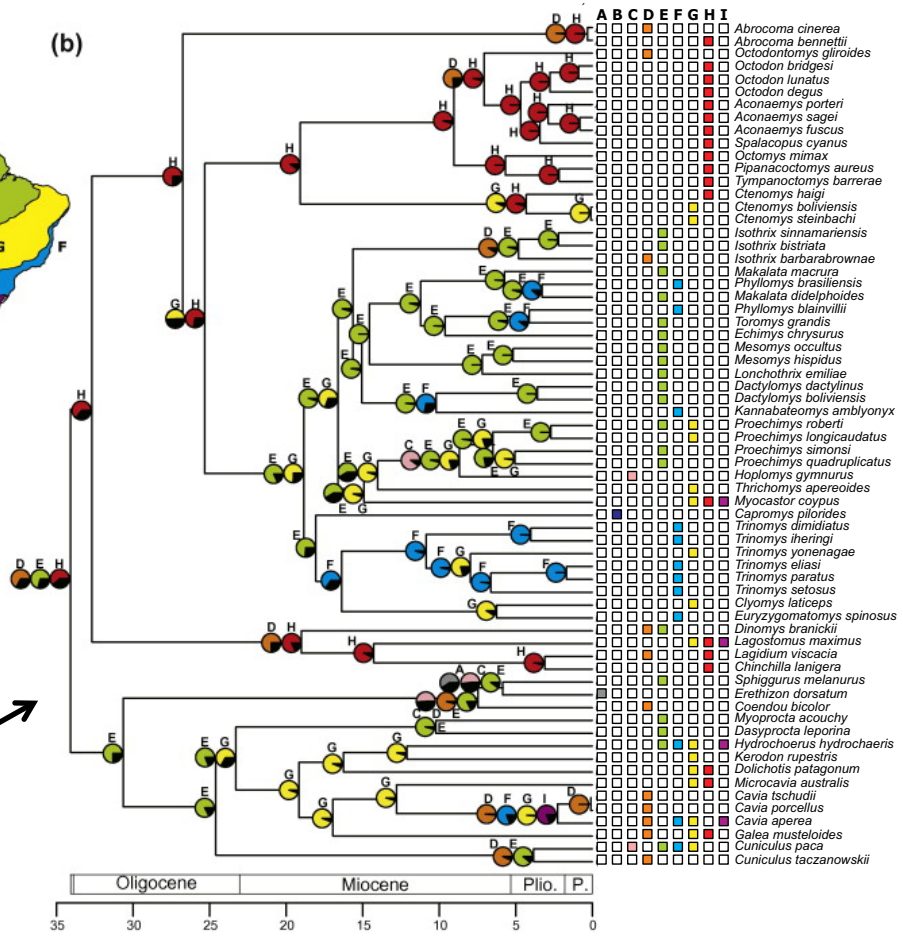
Gives us $\mathcal{L}(X; \theta, T, M)$

Character states

Time-calibrated phylogeny

Data matrix

Octodon degus



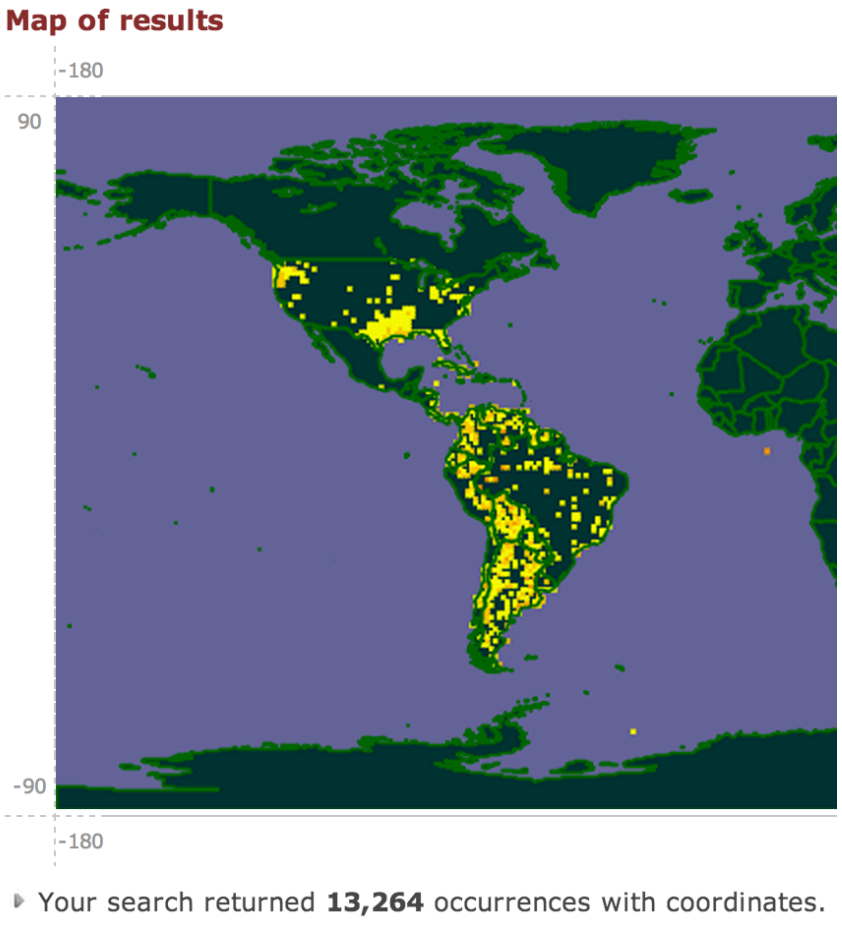
Ancestral state reconstructions

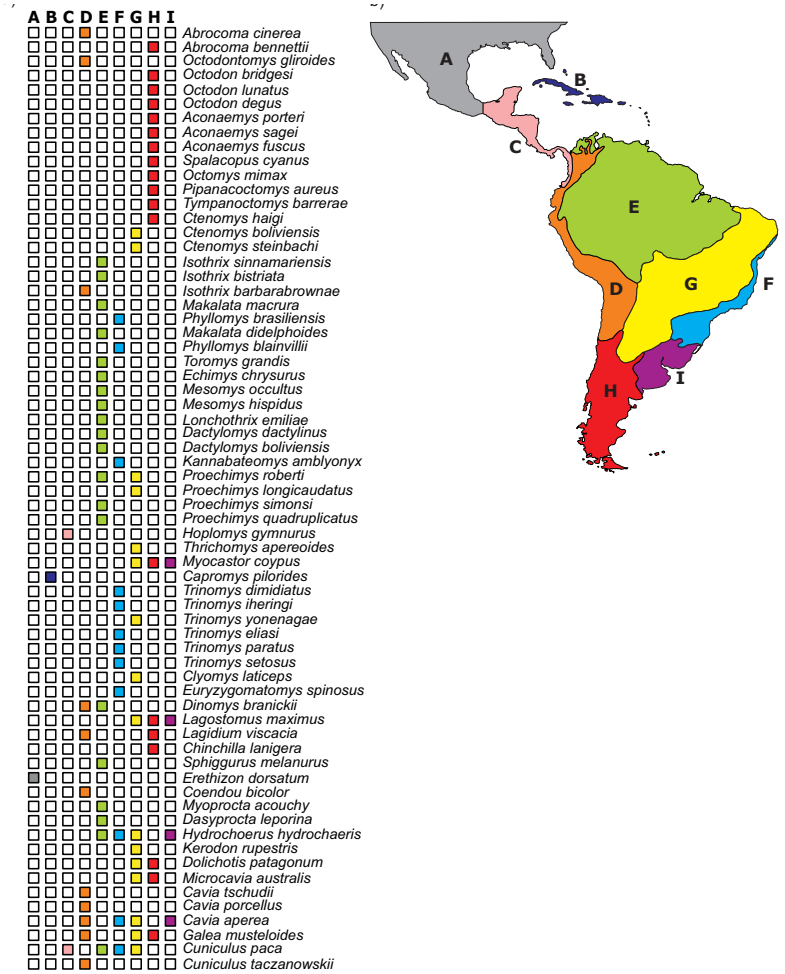
Photo, José Cañas

Upham & Patterson, 2012 (Mol Phylo Evol)

Species occurrence data
(gbif.org, 2013)

Discrete presence-absence
(Upham & Patterson, 2012)





Individual or range

The individuals in a taxon share a range.

	Discrete	**Continuous**
Individual or Endemic	occupied area	geographical point
Range	set of occupied areas	set of geographical points

Data matrix

X_{ij} taxon *i*, character *j*

Continuous

e.g. latitude-longitude

$$X_i = (\phi, \lambda) = (38.54^\circ\mathrm{N}, 121.75^\circ\mathrm{W})$$

Discrete

e.g. single area

$$X_i = \mathrm{Africa}$$

presence-absence (range)

$$X_i = (0, 0, 1, 0, 0, 1, 1, 1)$$

Models

Continuous

e.g. Brownian motion (Gaussian)

$$P(x \to y; t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-y)^2}{2\sigma^2}\right\}$$

Discrete

e.g. continuous-time Markov chain

$$P(x \to y; t) = \left[e^{Qt}\right]_{x,y}$$

Island Model

Dispersal-only model

One area per taxon (endemic/individual)

Learn favored dispersal routes

Work by:

- Sanmartín *et al.*, 2008 (Syst Biol)
- Lemey *et al.,* 2009 (PLoS Comp Biol)

Island Model

I. Sanmartín, P. van der Mark and F. Ronquist

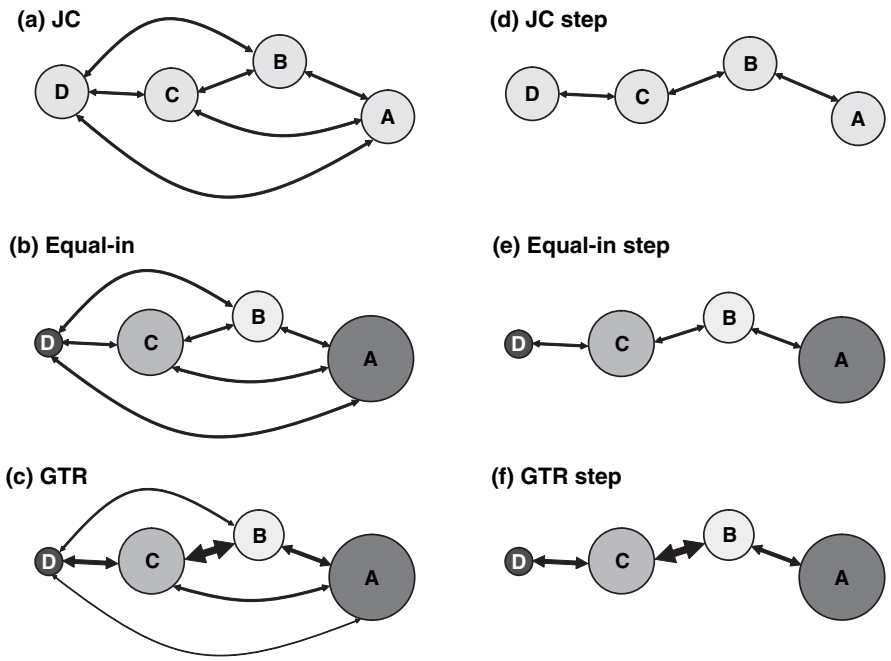


Figure 2 Bayesian Island Models: Each circle represents an island; circle size represents the relative carrying capacity of the island (expected number of lineages at equilibrium); arrow width represents the relative dispersal rate between two single islands. (a) Jukes–Cantor (JC) model: all carrying capacities equal, all dispersal rates equal. (b) Equal-in model: unequal carrying capacities, equal dispersal rates. (c) General Time Reversible (GTR) model: unequal carrying capacities, unequal dispersal rates. (d–f) Stepping-stone variant of each model. (d) JC step: all carrying capacities equal, dispersal rates equal between adjacent islands, zero between non-adjacent islands. (e) Equal-in step: unequal carrying capacities, all dispersal rates equal between adjacent islands, zero between non-adjacent islands. (f) GTR step: all carrying capacities unequal, all dispersal rates unequal between adjacent islands, zero between non-adjacent islands.

Embedding the graph in Q

General Time Reversible (a – c)

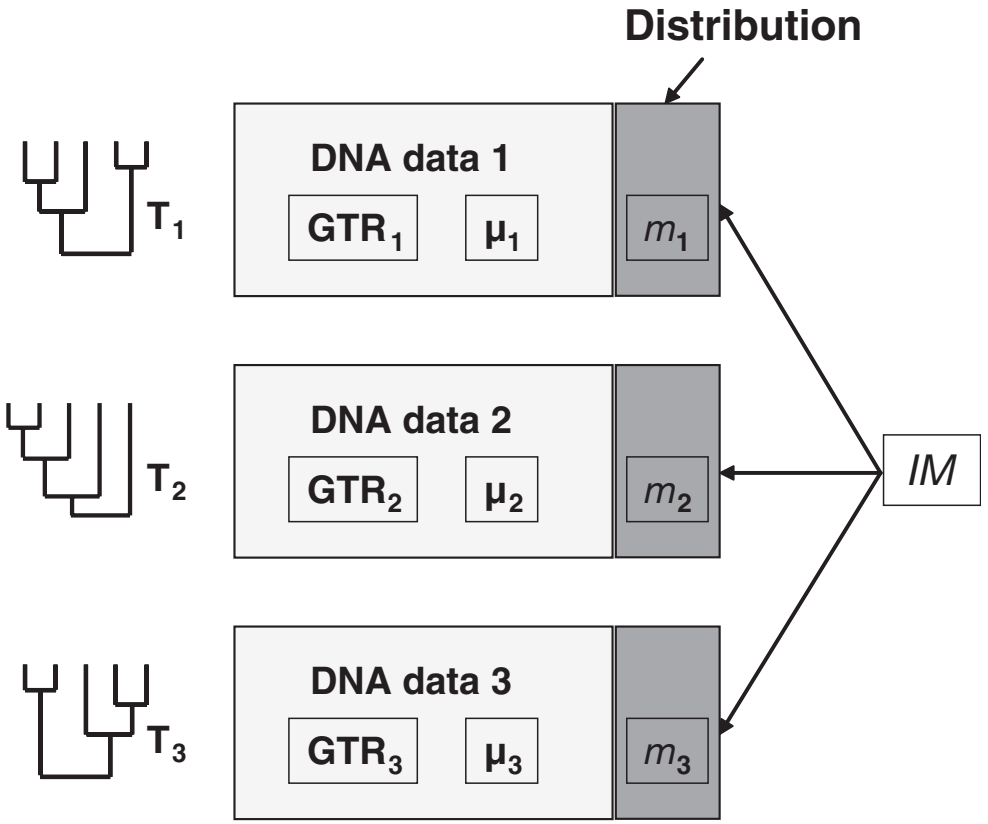
$$Q = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} - & \pi_B r_{AB} & \pi_C r_{AC} & \pi_D r_{AD} \\ \pi_A r_{AB} & - & \pi_C r_{BC} & \pi_D r_{BD} \\ \pi_A r_{AC} & \pi_B r_{BC} & - & \pi_D r_{CD} \\ \pi_A r_{AD} & \pi_B r_{BD} & \pi_C r_{CD} & - \end{pmatrix} \end{array}$$

Stepping Stone (b – f)

$$Q = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} - & \pi_B r_{AB} & 0 & 0 \\ \pi_A r_{AB} & - & \pi_C r_{BC} & 0 \\ 0 & \pi_B r_{BC} & - & \pi_D r_{CD} \\ 0 & 0 & \pi_C r_{CD} & - \end{pmatrix} \end{array}$$

Shared:
Dispersal process

Independent:
molecular process,
molecular speed,
dispersal speed,
clock tree



Bayes Factors (harmonic mean)

13 groups, 393 species, 954 taxa

Island model	Ln model likelihood
JC step	−101704.09
Equal-in	−101667.87
JC	−101649.92
Equal-in step	−101628.31
GTR	−101624.19
GTR step	−101618.94 ←
	*(−101642.9)

*Model likelihood for the 'long analysis' (30 million generations, four runs); see text.

LAGRANGE

Dispersal-(Local) Extinction-Cladogenesis (DEC)

Many areas per taxon (range)

D,E as parameterized event classes

Work by:

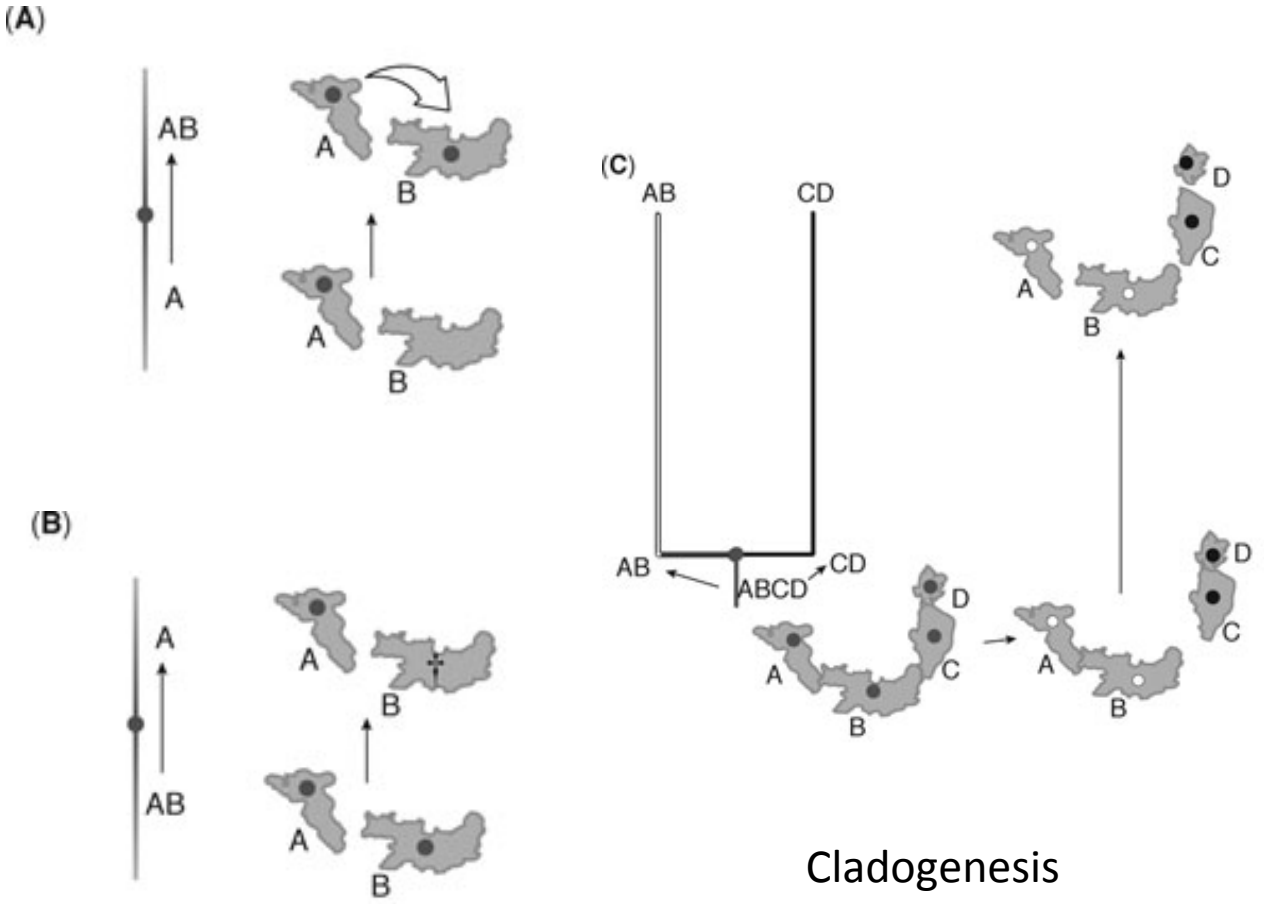
Ree *et al.*, 2005 (Syst Biol)

Ree & Smith, 2008 (Syst Biol)

DEC event types

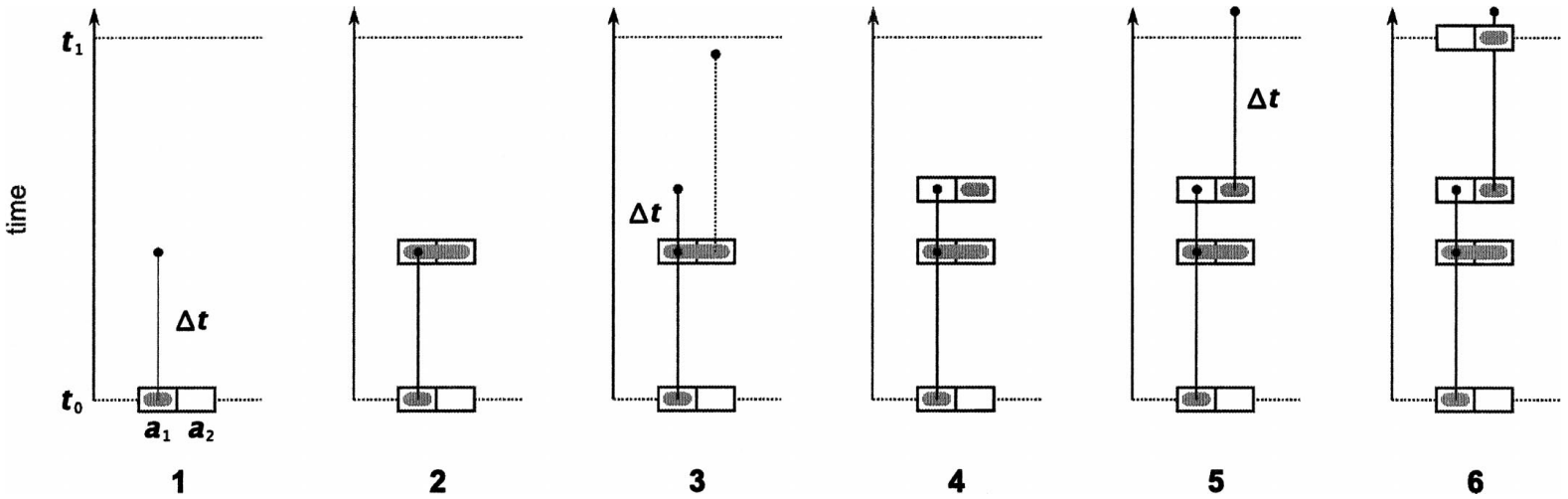
Dispersal

Extirpation



Cladogenesis

Dispersal & Extinction



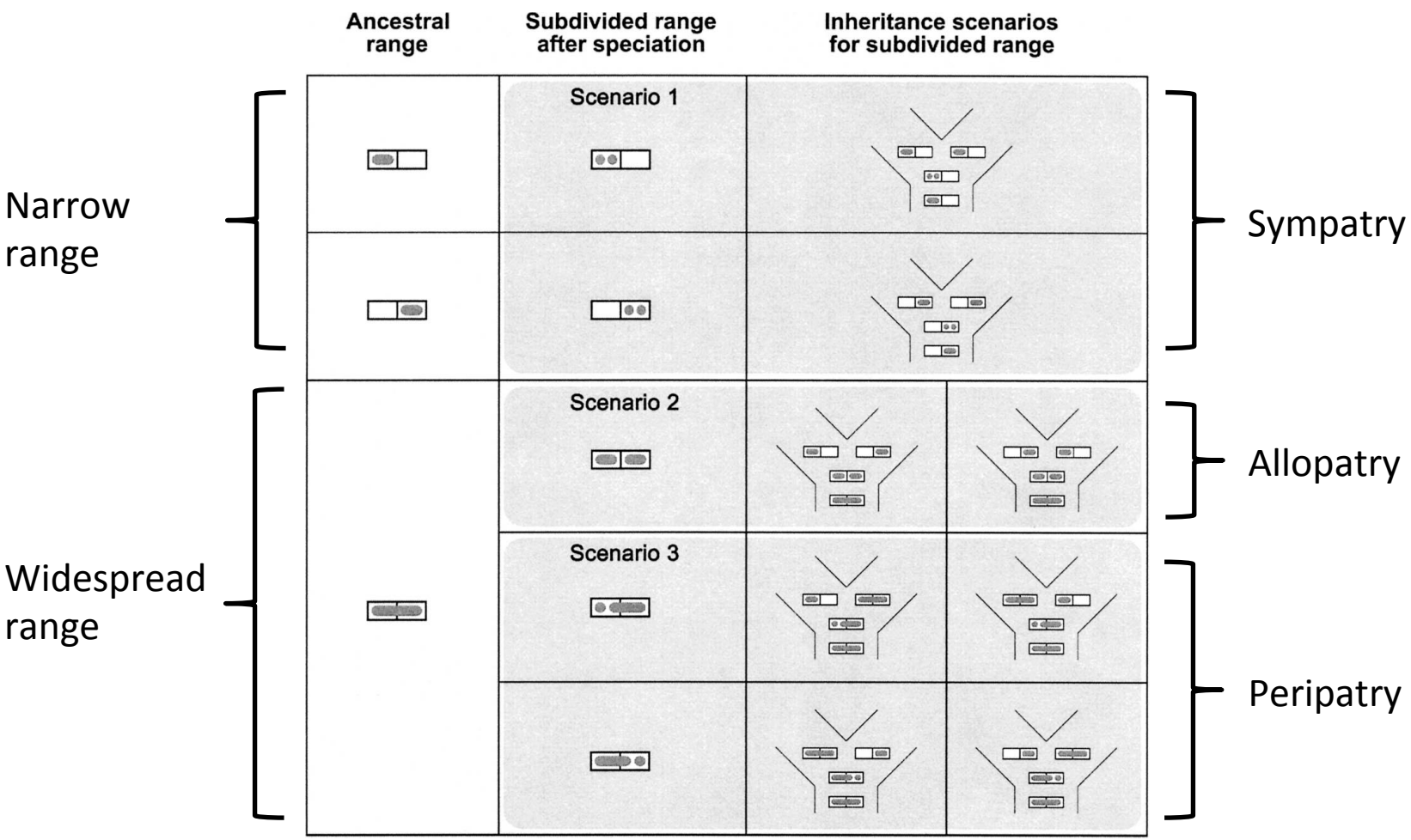
Exponentially-distributed times between events

Rate matrix for anagenesis

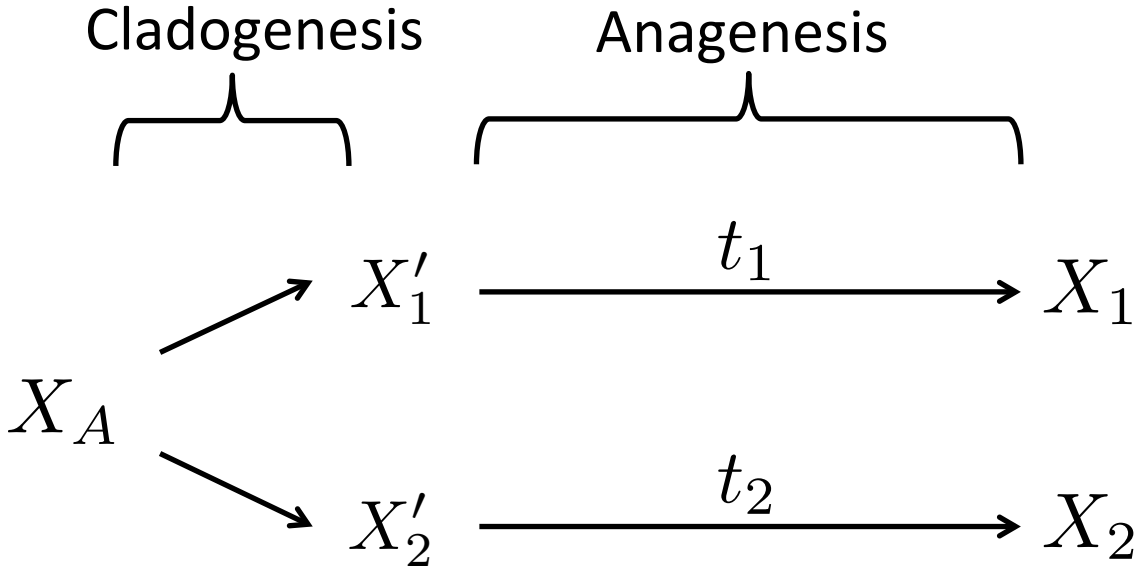
$$Q = \left[\begin{array}{c|cccccccc} & \emptyset & 1 & 2 & 3 & 12 & 13 & 23 & 123 \\ \hline \emptyset & — & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & E_1 & — & 0 & 0 & D_{12} & D_{13} & 0 & 0 \\ 2 & E_2 & 0 & — & 0 & D_{21} & 0 & D_{23} & 0 \\ 3 & E_3 & 0 & 0 & — & 0 & D_{31} & D_{32} & 0 \\ 12 & 0 & E_2 & E_1 & 0 & — & 0 & 0 & D_{13} + D_{23} \\ 13 & 0 & E_3 & 0 & E_1 & 0 & — & 0 & D_{12} + D_{32} \\ 23 & 0 & 0 & E_3 & E_2 & 0 & 0 & — & D_{21} + D_{31} \\ 123 & 0 & 0 & 0 & 0 & E_3 & E_2 & E_1 & — \end{array}\right]$$

$$\mathbf{P}_{ij}(t) = \left[\exp\left\{\mathbf{Q}t\right\}\right]_{ij}$$

Cladogenesis

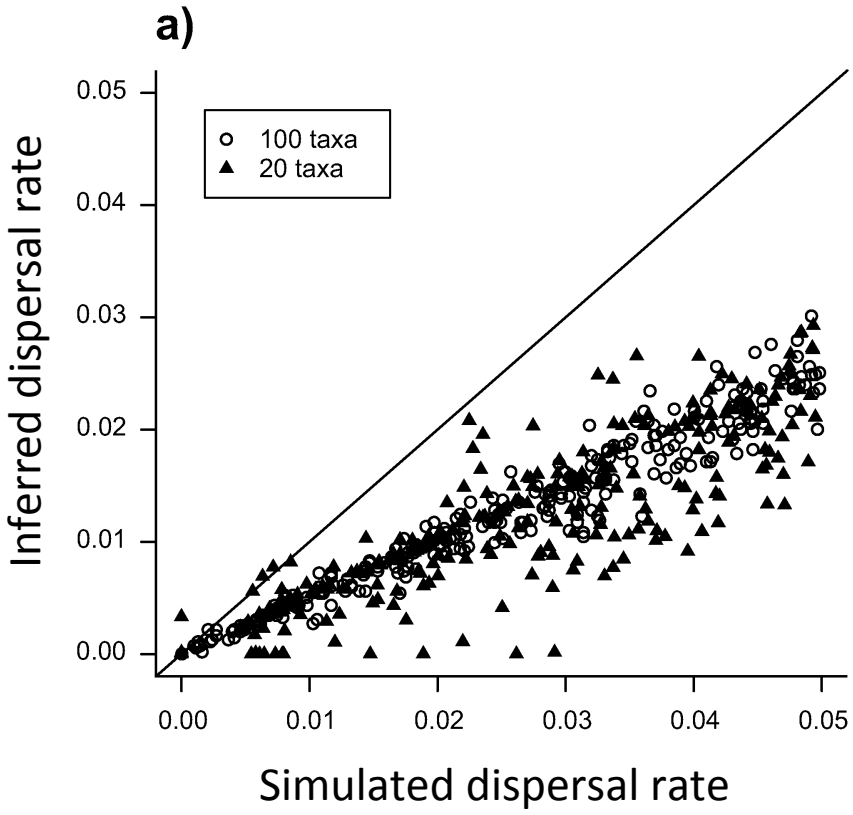


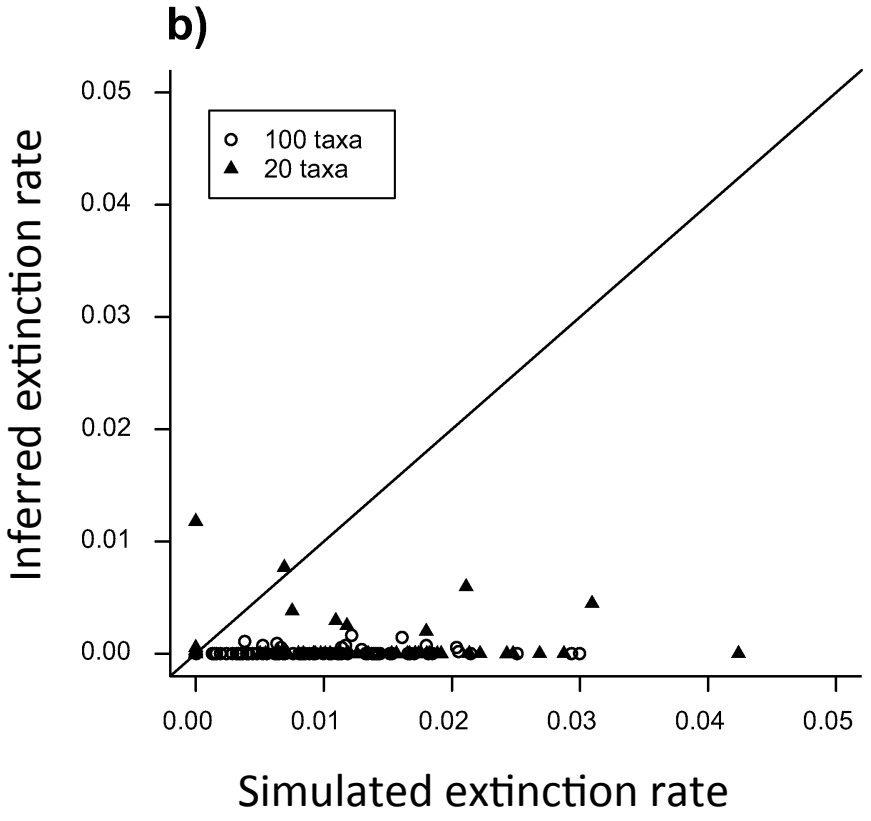
DEC likelihood



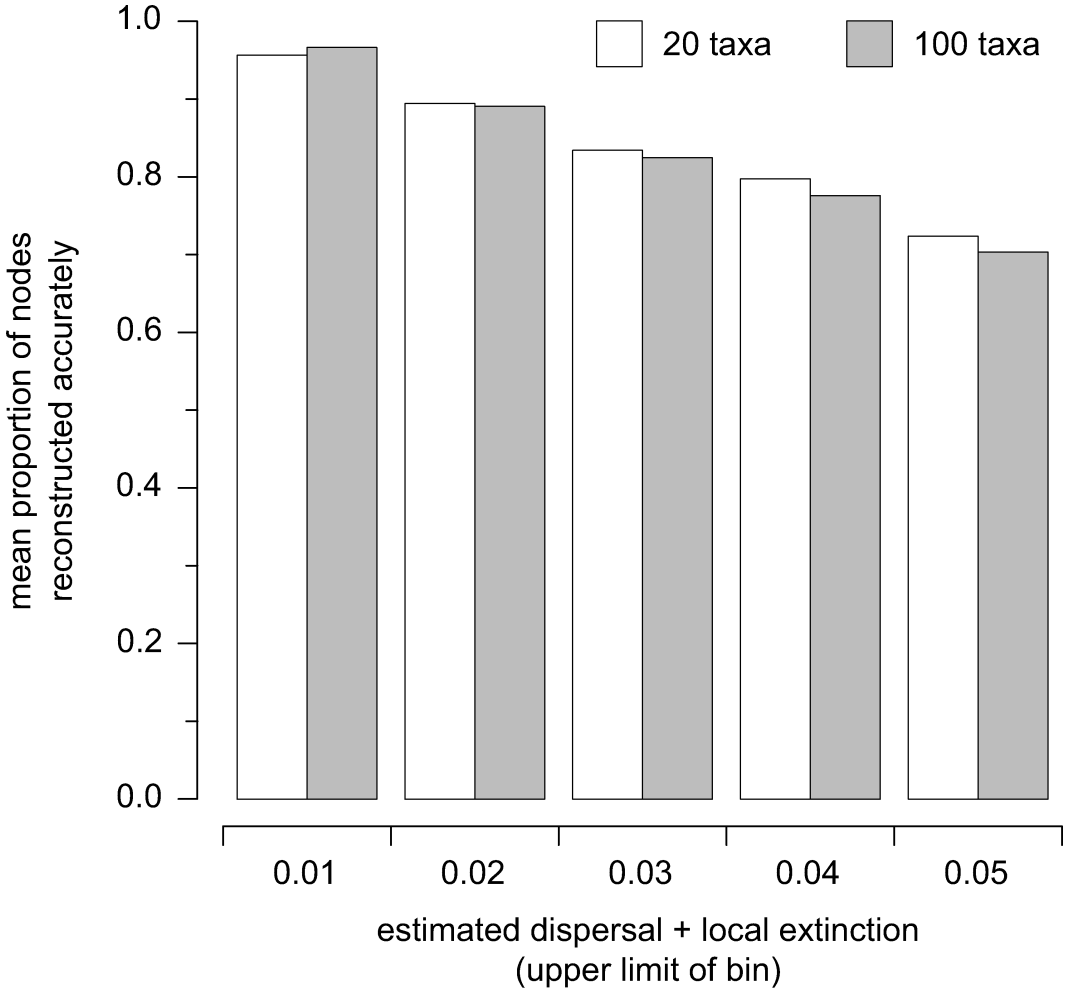
$$p(X_1', X_2' \mid X_A) \times p(X_1 \mid X_1', t_1, \theta) \times p(X_2 \mid X_2', t_2, \theta)$$

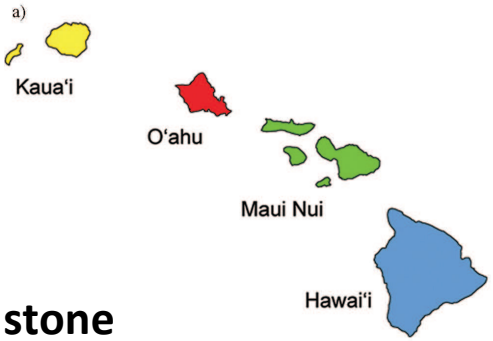
MLE inference vs truth

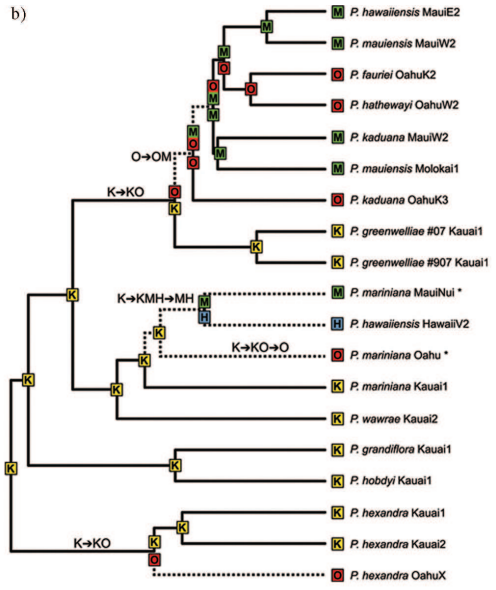




MLE range reconstructions

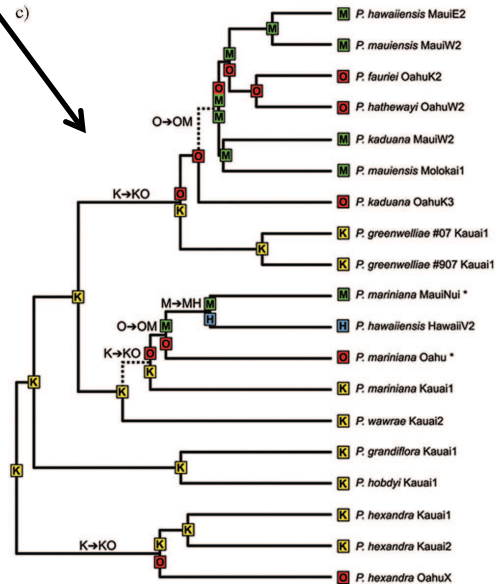


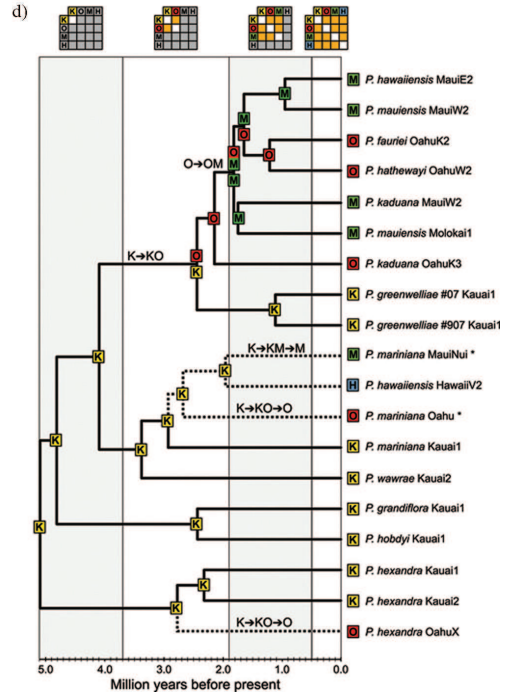




Unconstrained
(GTR)

Stepping stone
(Small adjacent ranges:
K, O, M, H, KO, OM, MH)





Psychotria mariniana

Stratified
(time-dependent GTR)

BioGeoBEARS

Generalized DEC model

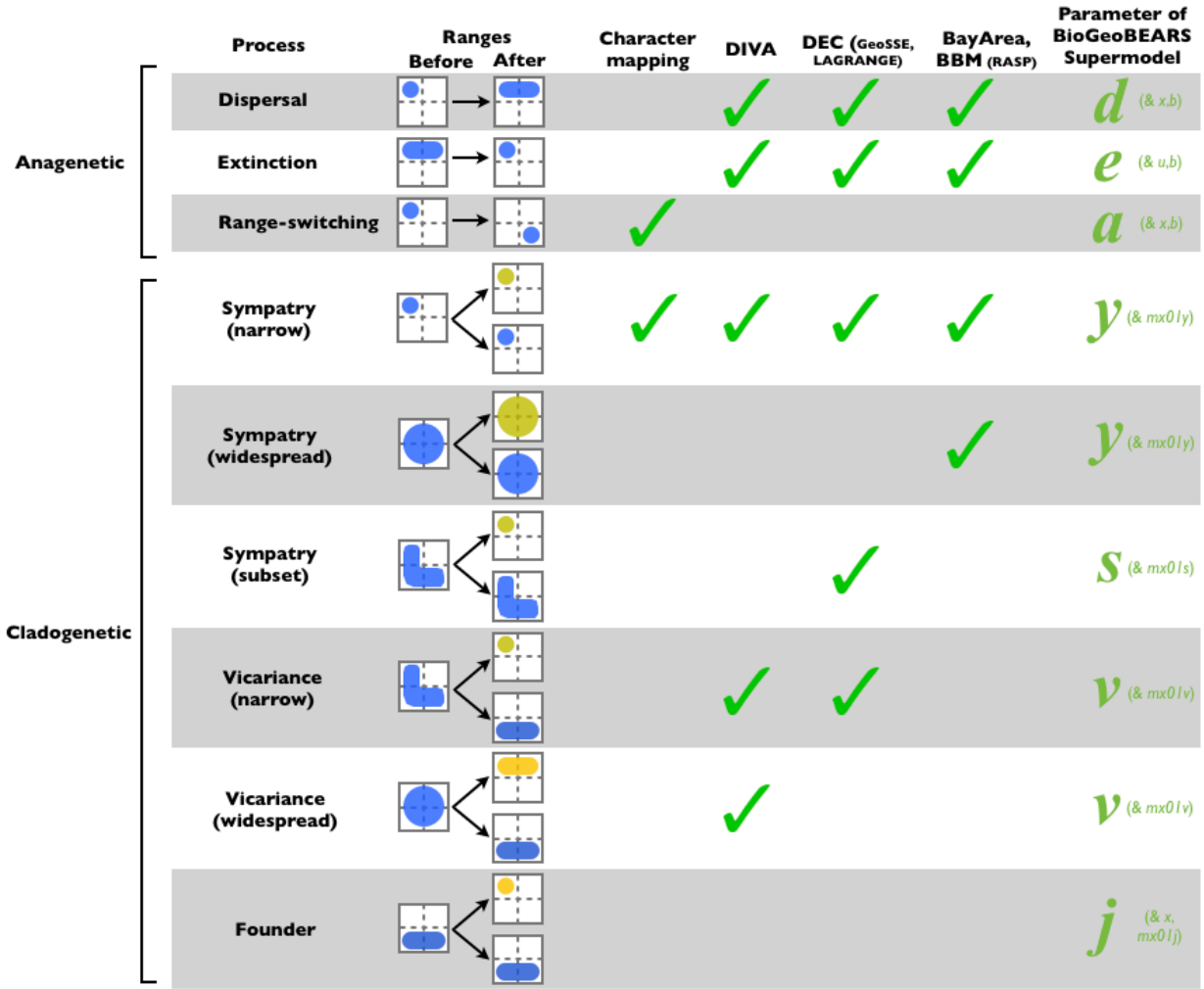
Many areas per taxon (range)

Seven parameterized event classes

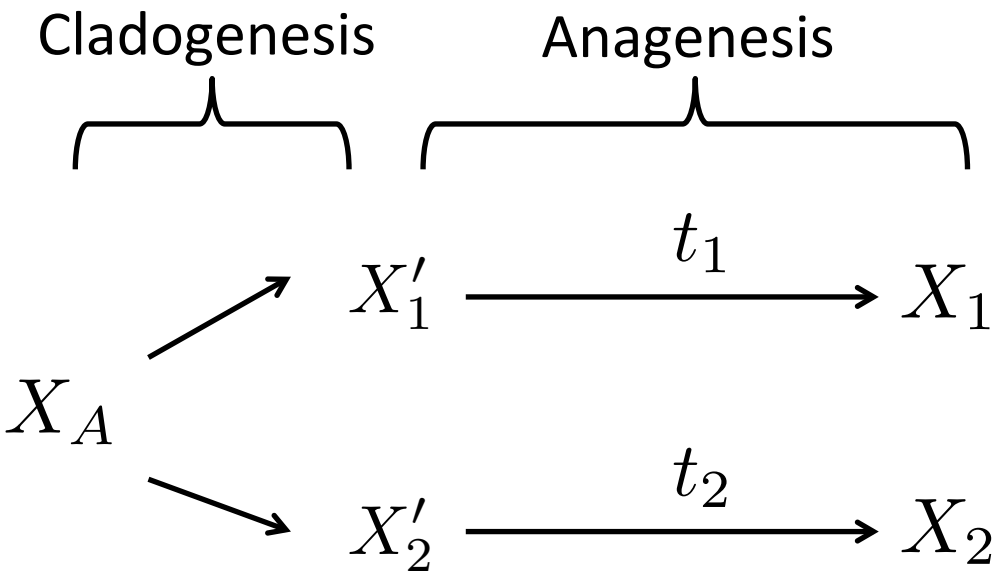
Work by:

Matzke, 2013 (Frontiers Biogeo)

Generalized DEC model

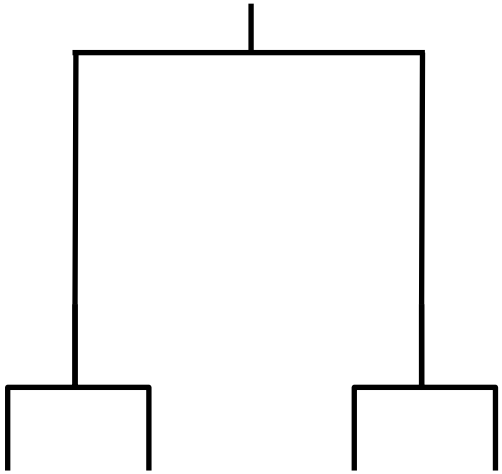


BiogeoBEARS likelihood

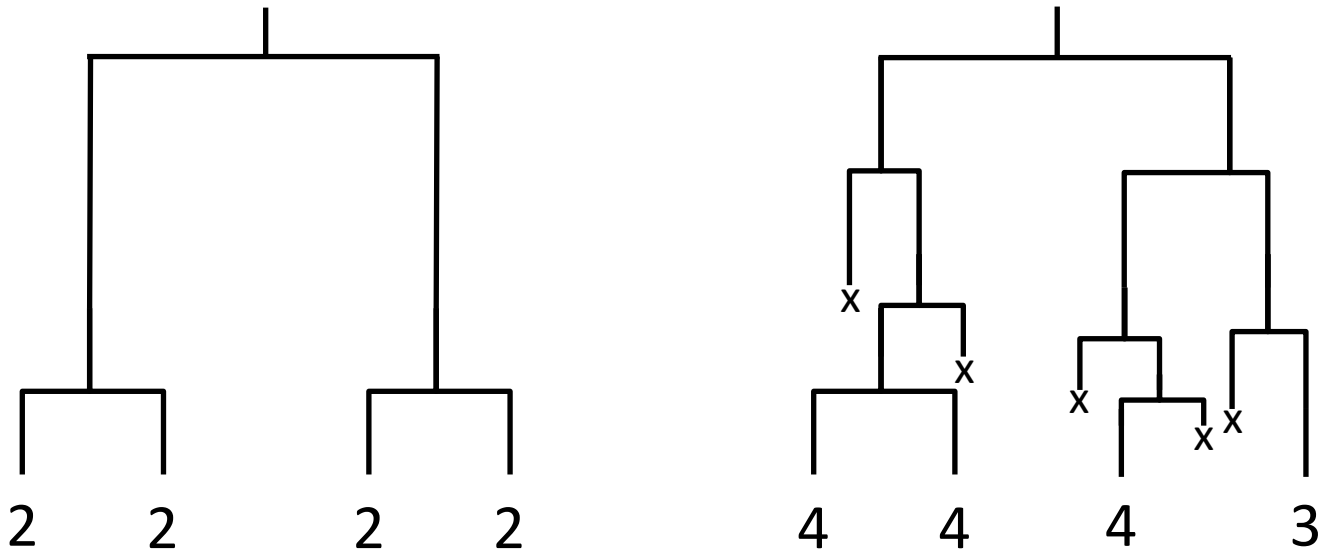


$$p(X_1', X_2' \mid X_A, \theta) \times p(X_1 \mid X_1', t_1, \theta) \times p(X_2 \mid X_2', t_2, \theta)$$

Parameterized cladogenesis



Speciation hidden by extinction



Geographic State Speciation Extinction (GeoSSE)

DEC model

Joint birth-death process & range evolution

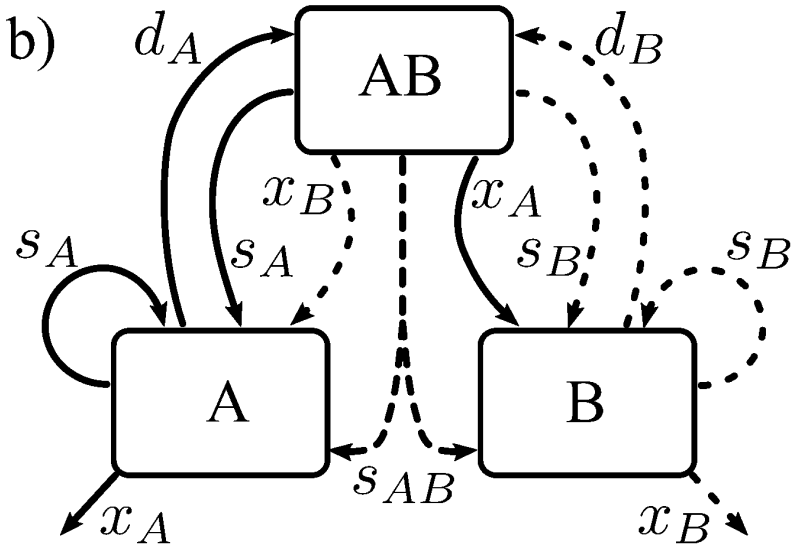
Accounts for “hidden” speciation

Range evolution, speciation, extinction intertwined

Work by:

- Goldberg *et al.*, 2011 (Syst. Biol.)
- Goldberg & Igić, 2012 (Evolution)

	Parameter	Areas	Event
Speciation	s_A	A	New lineage in area A
	s_B	B	New lineage in area B
	s_{AB}	AB	New lineage in area A or B
Dispersal	d_A, d_B	A or B	This lineage gains an area
Extinction	x_A, x_B	AB	This lineage loses an area
		A or B	This lineage goes extinct



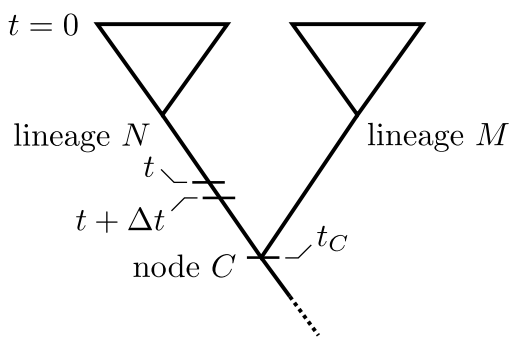
GeoSSE likelihood

Likelihood of tree and character states

$$\frac{\mathrm{d}D_{NA}}{\mathrm{d}t} = -(s_A + d_A + x_A)D_{NA}(t) + d_A D_{NAB}(t) + 2s_A D_{NA}(t)E_A(t),$$

$$\frac{\mathrm{d}D_{NB}}{\mathrm{d}t} = -(s_B + d_B + x_B)D_{NB}(t) + d_B D_{NAB}(t) + 2s_B D_{NB}(t)E_B(t),$$

$$\begin{aligned}\frac{\mathrm{d}D_{NAB}}{\mathrm{d}t} = &-(s_A + s_B + s_{AB} + x_A + x_B)D_{NAB}(t) \\ &+ x_A D_{NB}(t) + x_B D_{NA}(t) \\ &+ s_A[E_A(t)D_{NAB}(t) + E_{AB}(t)D_{NA}(t)] \\ &+ s_B[E_B(t)D_{NAB}(t) + E_{AB}(t)D_{NB}(t)] \\ &+ s_{AB}[E_A(t)D_{NB}(t) + E_B(t)D_{NA}(t)],\end{aligned}$$



Likelihood of extinction

$$\frac{\mathrm{d}E_A}{\mathrm{d}t} = -(s_A + d_A + x_A)E_A(t) + x_A + d_A E_{AB}(t) + s_A E_A(t)^2, \quad (3a)$$

$$\frac{\mathrm{d}E_B}{\mathrm{d}t} = -(s_B + d_B + x_B)E_B(t) + x_B + d_B E_{AB}(t) + s_B E_B(t)^2, \quad (3b)$$

$$\begin{aligned}\frac{\mathrm{d}E_{AB}}{\mathrm{d}t} = &-(s_A + s_B + s_{AB} + x_A + x_B)E_{AB}(t) + x_A E_B(t) \\ &+ x_B E_A(t) + s_A E_{AB}(t)E_A(t) + s_B E_{AB}(t)E_B(t) \\ &+ s_{AB}E_A(t)E_B(t). \quad (3c)\end{aligned}$$

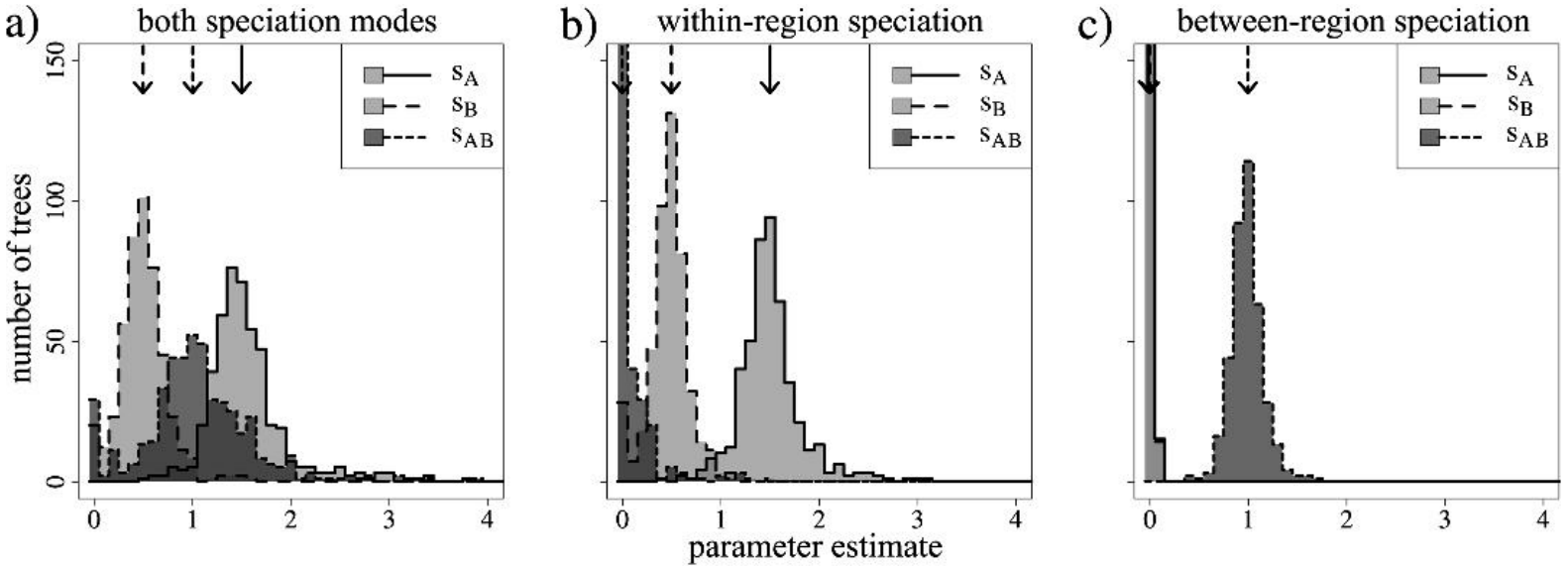
Likelihood of "observed" speciation

$$D_{CA}(t_C) = D_{NA}(t_C)D_{MA}(t_C)s_A, \quad (2a)$$

$$D_{CB}(t_C) = D_{NB}(t_C)D_{MB}(t_C)s_B, \quad (2b)$$

$$\begin{aligned}D_{CAB}(t_C) = &\frac{1}{2}[D_{NAB}(t_C)D_{MA}(t_C) + D_{NA}(t_C)D_{MAB}(t_C)]s_A \\ &+ \frac{1}{2}[D_{NAB}(t_C)D_{MB}(t_C) + D_{NB}(t_C)D_{MAB}(t_C)]s_B \\ &+ \frac{1}{2}[D_{NA}(t_C)D_{MB}(t_C) + D_{NB}(t_C)D_{MA}(t_C)]s_{AB}. \quad (2c)\end{aligned}$$

MLE inference vs truth



$s_A > s_B > s_{AB}$

$s_A > s_B$

$s_{AB} = 0$

$s_A = s_B = 0$

$s_{AB} > 0$

BayArea

Dispersal-Extirpation model

Distance effects as a free parameter

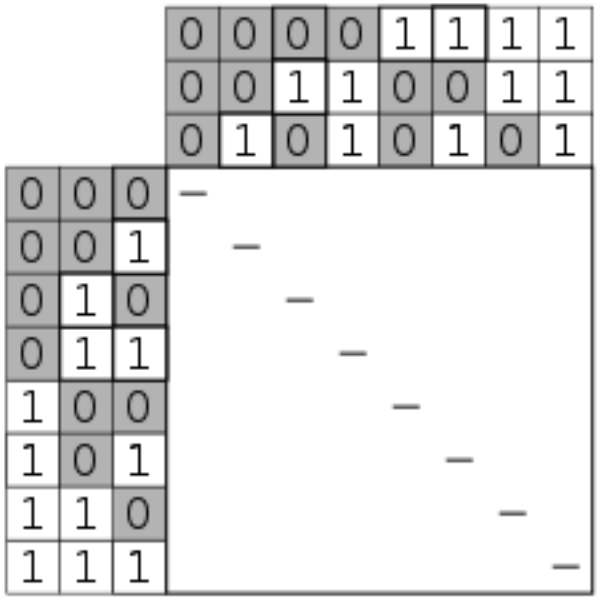
Scales for many areas

Work by:

Landis *et al.*, 2013 (Syst Biol)

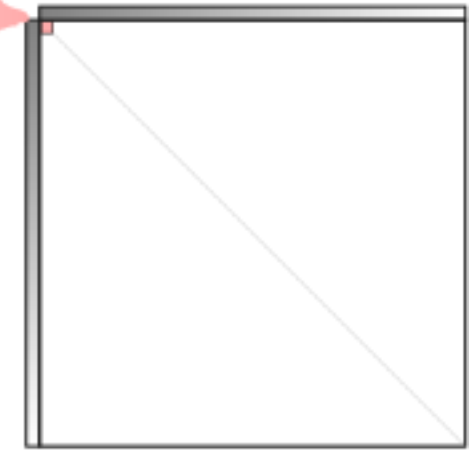
For more areas, Q explodes

3 areas



$2^3 \times 2^3 = 8 \times 8$

10 areas



$2^{10} \times 2^{10} = 1024 \times 1024$

Intractable for more than ten areas

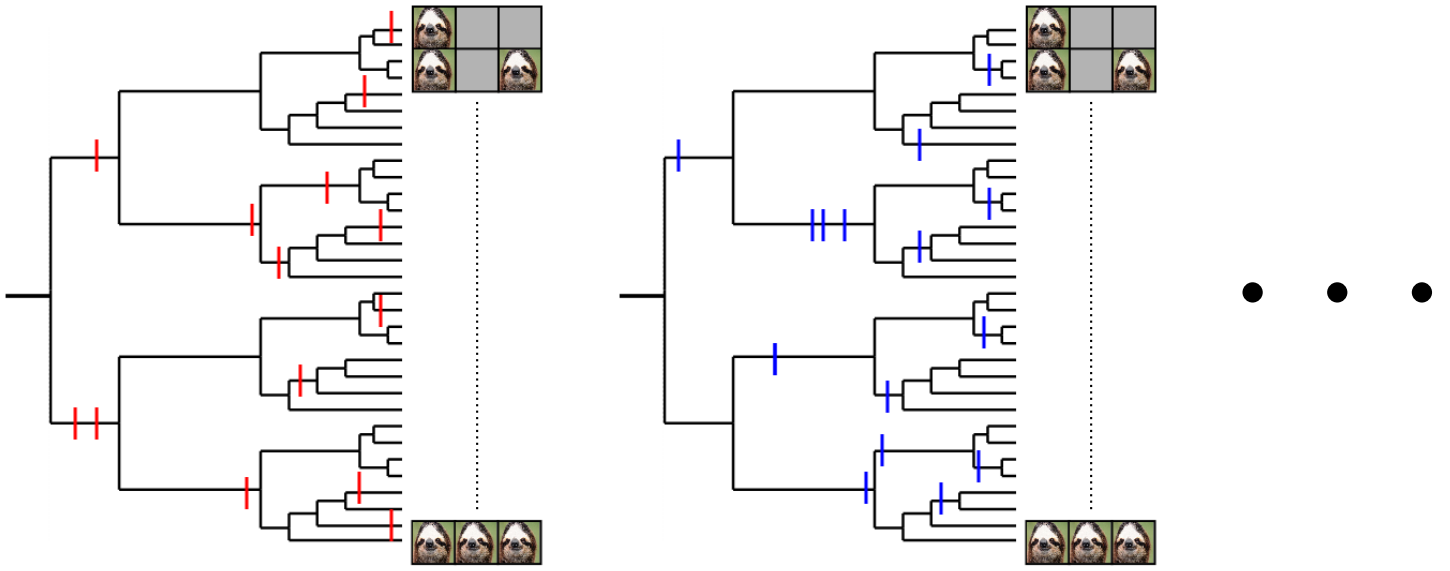
How to infer large Q?

Inspired by Robinson *et al*., 2003 (*Mol Biol Evol*)

Key concepts

1. Propose biogeographic histories, H
2. Compute likelihood, $\mathcal{L}_{\theta,H}$
3. Approximate $P(\theta, H \mid D)$ using Markov chain Monte Carlo (MCMC)

1. Propose biogeographic histories, H



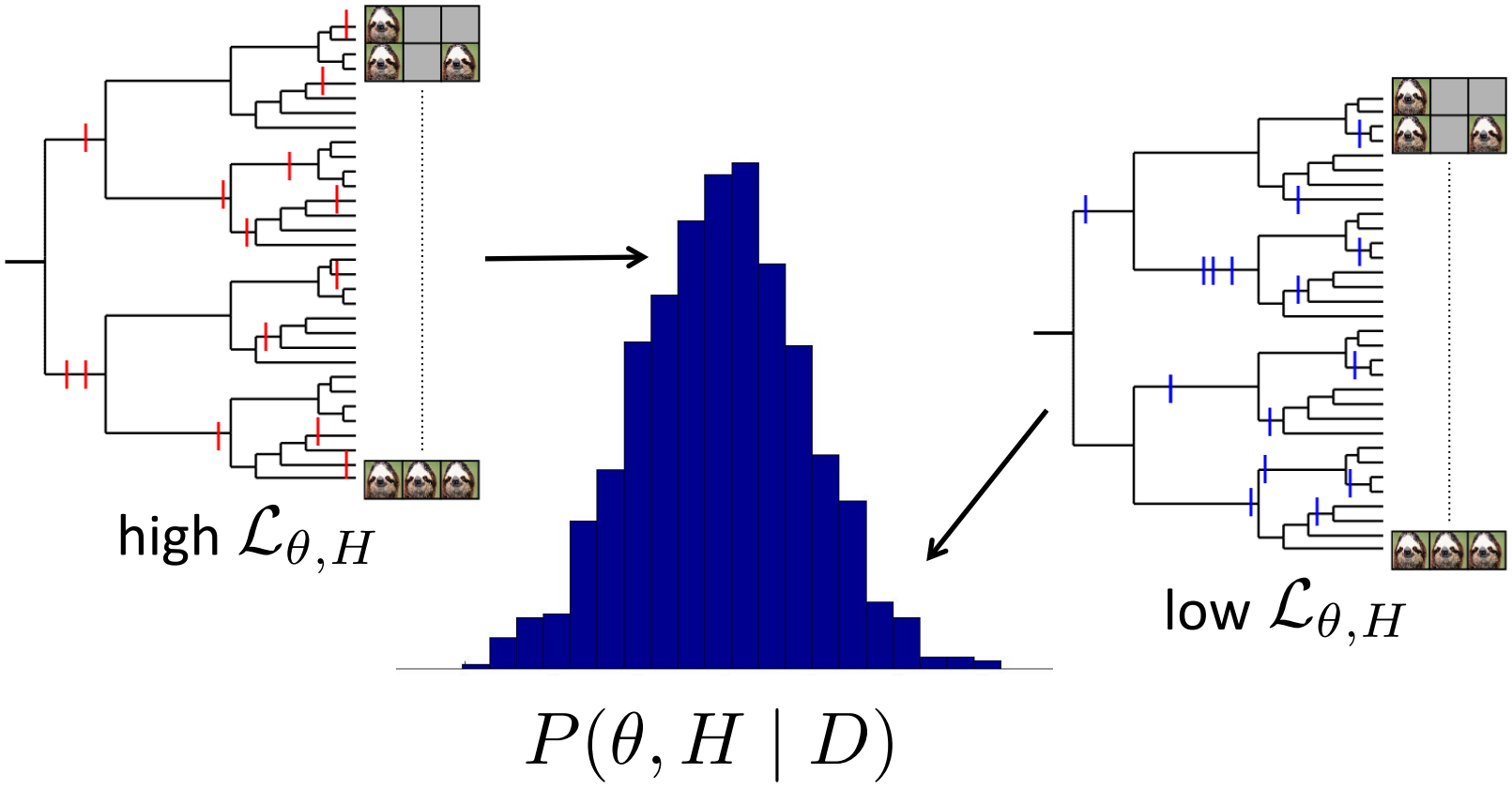
2. Compute likelihood, $\mathcal{L}_{\theta,H}$

Range evolution events from range $j \to i$:

$r = \sum r_j$	sum of rates leaving j
re^{-rt}	prob any event at time t
r_i/r	prob next event is $j \to i$

$\mathcal{L}_{\theta,H}$ = product of event types & times over tree

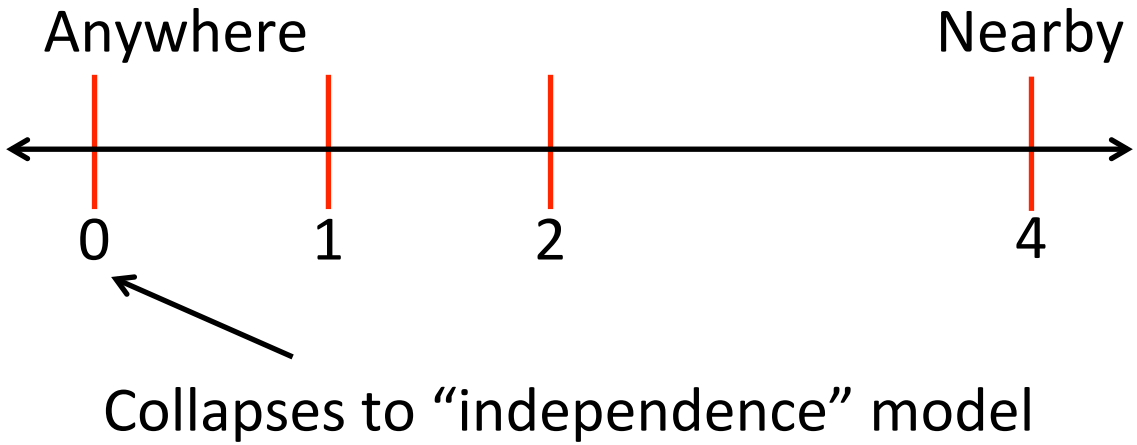
3. Approximate $P(\theta, H \mid D)$ using MCMC



Distance-dependent dispersal model

Infer distance effect parameter

Where is next dispersal event given current range?



Distance dependent rate matrix

$$R^{(a)}_{Y_i,Y_j} = \begin{cases} \lambda_0 & \text{if } Y_{j,a} = 0 \\ \lambda_1 \eta(Y_i, Y_j, a, \beta) & \text{if } Y_{j,a} = 1 \\ 0 & \text{if } Y_i \text{ and } Y_j \text{ differ at more than one area} \\ 0 & \text{if } Y_j = (0, 0, \ldots, 0) \end{cases}$$

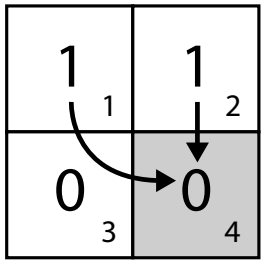
Extirpation	Uniform at random
Dispersal	Modified by distance
Extinction	Forbidden

Distance-dependent rate modifier

$$\eta(Y_i = (1,1,0,0) \rightarrow Y_j = (1,1,0,1), a = 4, \beta) =$$

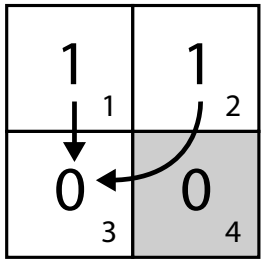
Rate-modifier

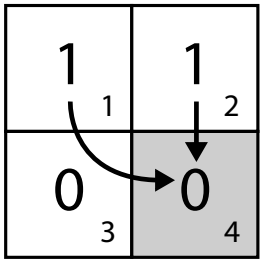
$$\underbrace{d(G_1, G_4)^{-\beta} + d(G_2, G_4)^{-\beta}}$$



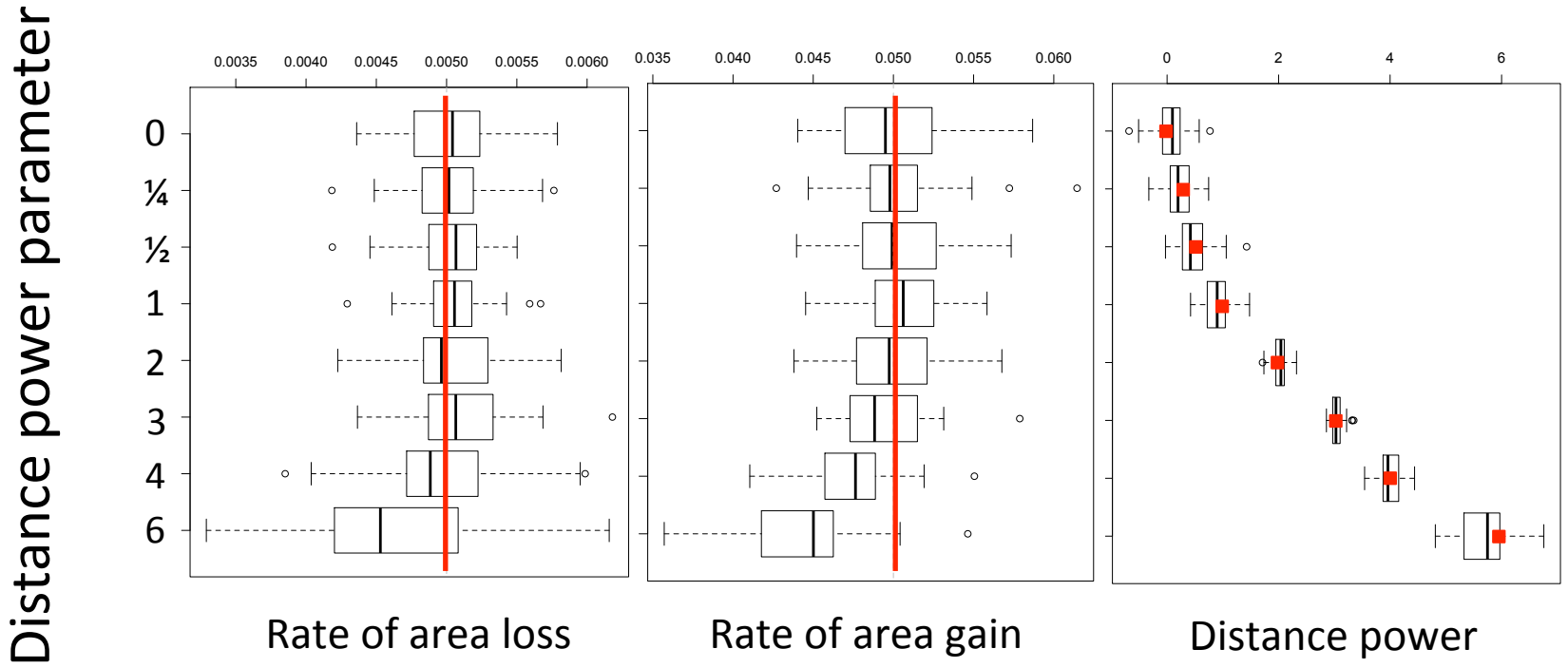
Normalization

$$\times \frac{2}{\underbrace{d(G_1, G_3)^{-\beta} + d(G_2, G_3)^{-\beta}} + \underbrace{d(G_1, G_4)^{-\beta} + d(G_2, G_4)^{-\beta}}}$$





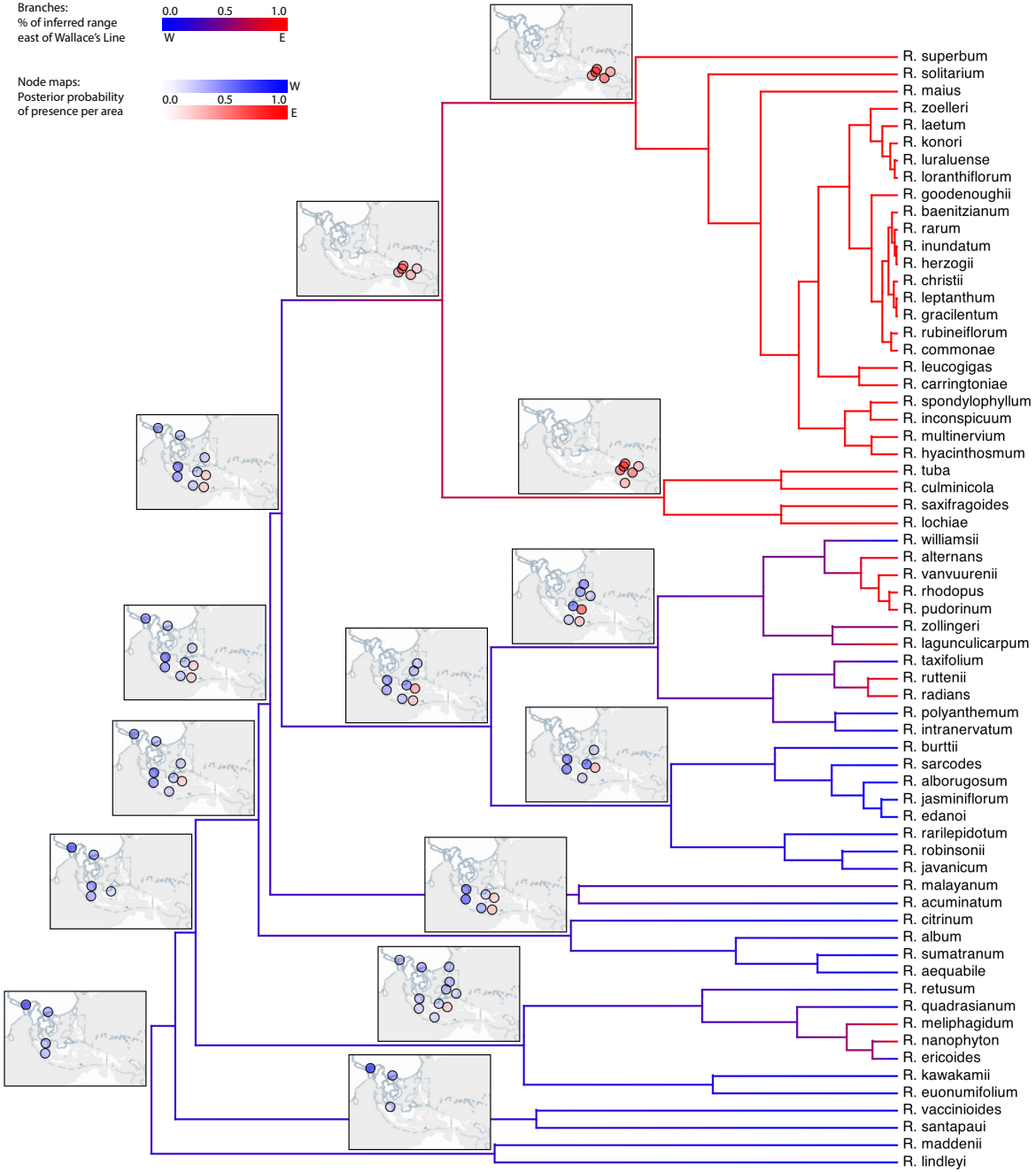
Bayesian inference vs truth



Posterior of ancestral ranges

East of Wallace's Line
West of Wallace's Line

Rhododendron goodenoughii



Continuous models

Brownian motion

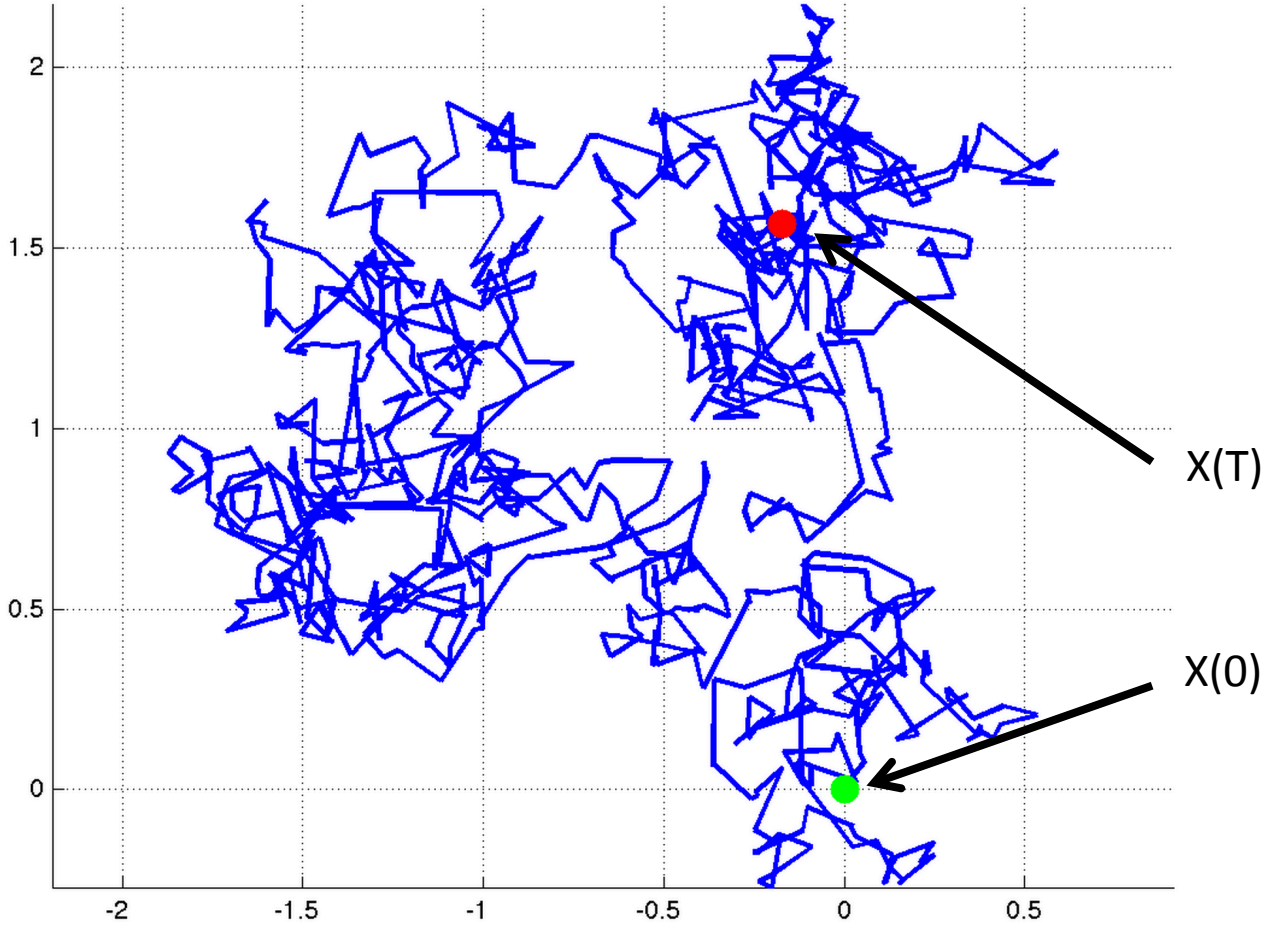
Each taxon is an individual sample

Epidemiology models

Work by:

- Lemmon & Lemmon, 2008 (Syst Biol)
- Lemey *et al.*, 2010 (Mol Biol Evol)

2D Brownian motion



Relaxed random walk

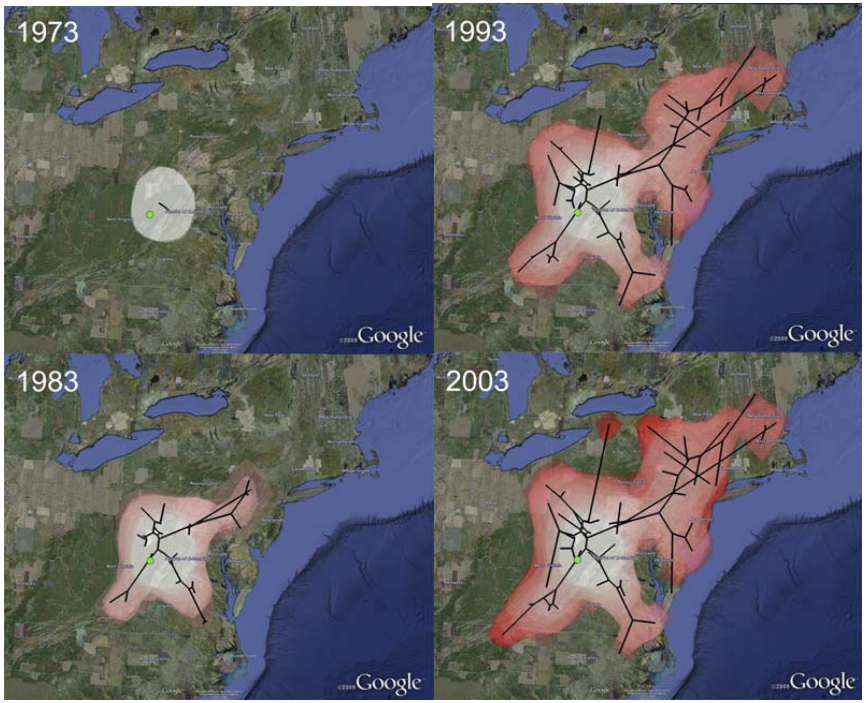
Joint inference of gene tree using relaxed molecular clock

Latitude, longitude diffuse by Brownian motion for each branch:

$$X_b \sim N\left(X_{\mathrm{pa}(b)}, t_b \phi_b \Sigma\right)$$

Branch rate rescaled ("relaxed"):

$$\phi_b \sim \mathrm{Gamma}(\nu/2, \nu/2)$$



Lemey *et al.,* 2009 (Mol Biol Evol)

Continuous models for ranges or multiple individuals

Diffusion of set of individual coordinates

???

Diffusion of range as polygon

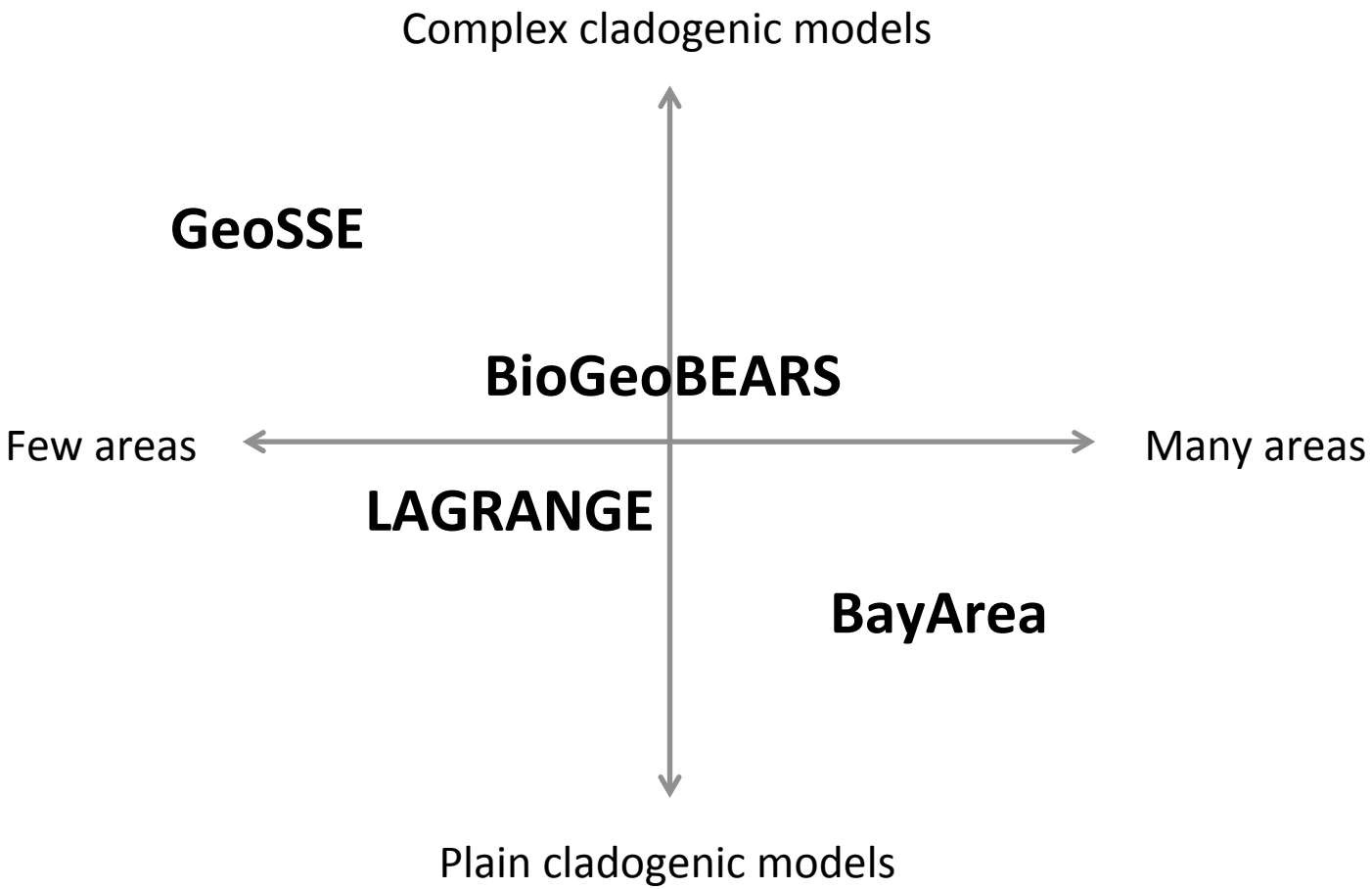
???

Hard, underexplored

Discrete vs. continuous models

	Discrete	**Continuous**
Data	Transformed	As is
Model	CTMC Asymmetry easy	Diffusion (BM) Asymmetry hard
Individual/Endemic	Yes	Yes
Range	Yes (scales poorly)	No (currently)
Dispersal/Extirpation	Yes	NA
Cladogenesis	Yes	NA
Speciation/Extinction	Yes (for 2-3 areas)	Yes (for individuals)
Spatial heterogeneity	Easy	Hard
Temporal heterogeneity	Easy	Easy

Tradeoffs


Complex cladogenic models
GeoSSE
BioGeoBEARS
Few areas
Many areas
LAGRANGE
BayArea
Plain cladogenic models

Lab

BayArea 1.0.2

- Input
- Analysis
- Output

Phylowood, BayArea-Fig

- Visualization

Papers

Bielejec, F., A. Rambaut, M. A. Suchard, and P. Lemey. 2011. Spread: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics 27:2910–2912.

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.

Lamm, K. S. and B. D. Redelings. 2009. Reconstructing ancestral ranges in historical biogeography: properties and prospects. Journal of Systematics and Evolution 47:369–382.

Landis, M. J. and T. Bedford. 2014. Phylowood: interactive web-based animations of biogeographic and phylogeographic histories. Bioinformatics 30:123–124.

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.

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 incontinuous space and time. Molecular 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.

Matzke, N. J. 2013. Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. Frontiers of Biogeography 5.

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. Sanmartin. 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.

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

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

Sanmartin, I. and F. Ronquist. 2004. Southern hemisphere biogeography inferred by event-based models: plant versus animal patterns. Systematic Biology 53:216–243.

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.

Software

BayArea	https://code.google.com/p/bayarea/
BEAST	http://beast.bio.ed.ac.uk/Main_Page
BioGeoBEARS	http://cran.r-project.org/web/packages/BioGeoBEARS/index.html
GeoSSE	http://www.zoology.ubc.ca/prog/diversitree/
LAGRANGE	https://code.google.com/p/lagrange/
SHIBA	http://phylodiversity.net/shiba/