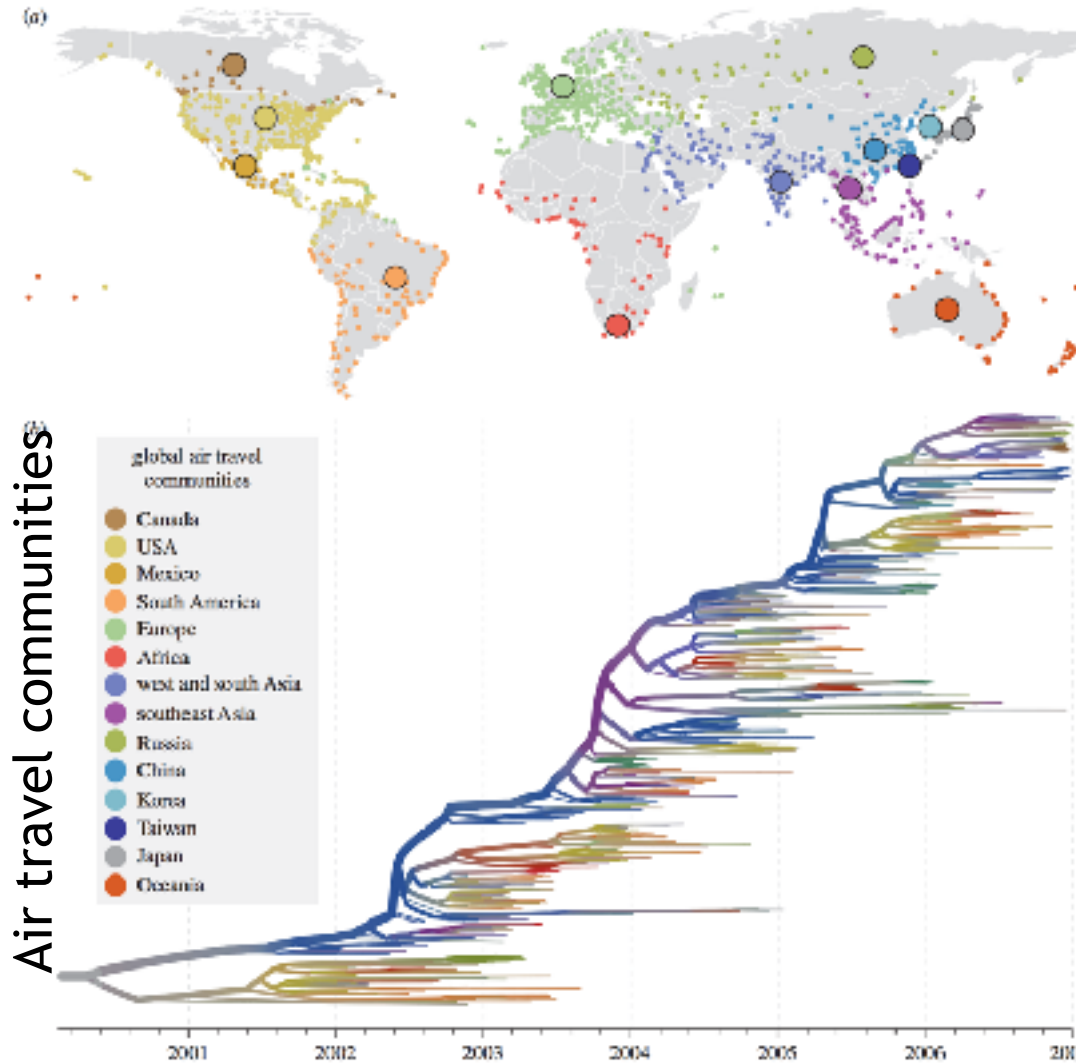


# A primer on phylogenetic biogeography and DEC models

March 13, 2017  
Michael Landis  
Bodega Bay Workshop  
Sunny California

# Epidemiology



H3N2 Influenza  
virus samples

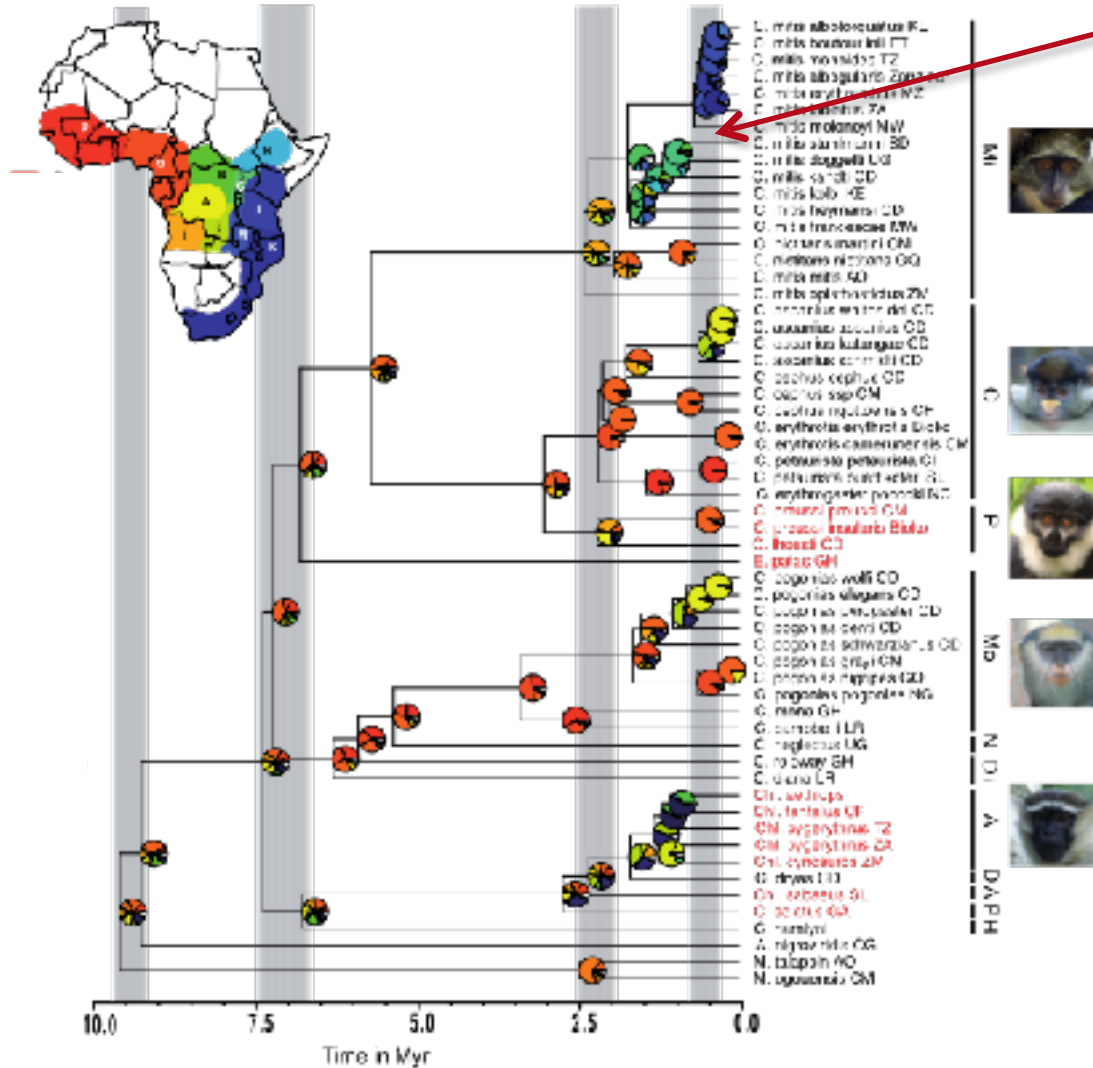
2000 - 2007 CE

Flu, H3N2

Pybus et al, 2015

# Conservation

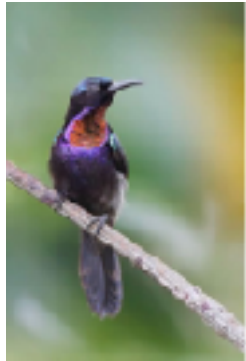
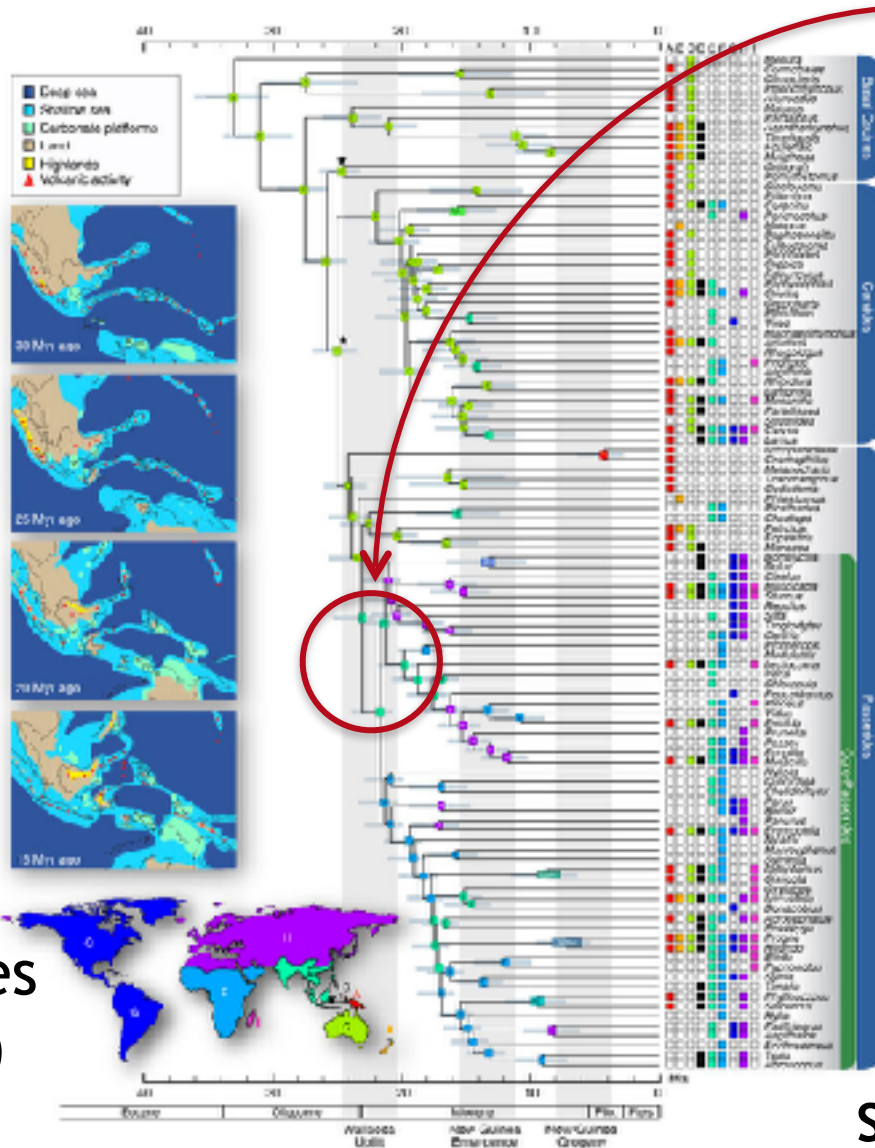
Eastern forests relatively new habitat



*Cercopithecidae*  
(Old World Monkeys)

# Origin stories

Out-of-Australia Radiation



Wallacea rises (20 - 25 Ma)

Songbirds

Moyle et al, 2011

# Biogeographic events

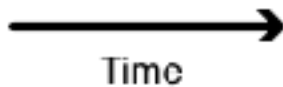
## Anagenesis



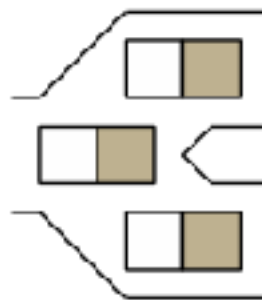
a. Dispersal



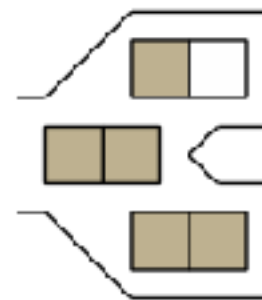
b. Extirpation



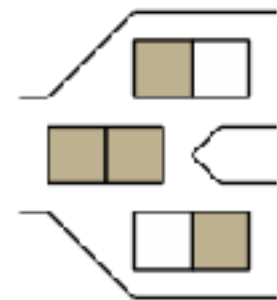
## Cladogenesis



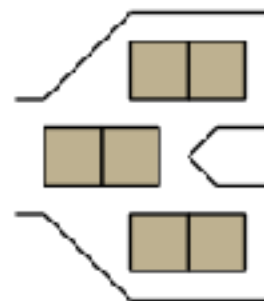
c. Narrow sympatry



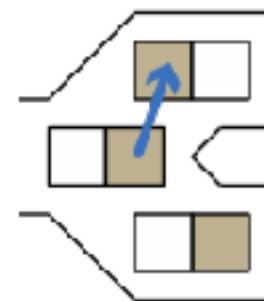
d. Subset sympatry



e. Allopatry



f. Full sympatry



g. Jump dispersal

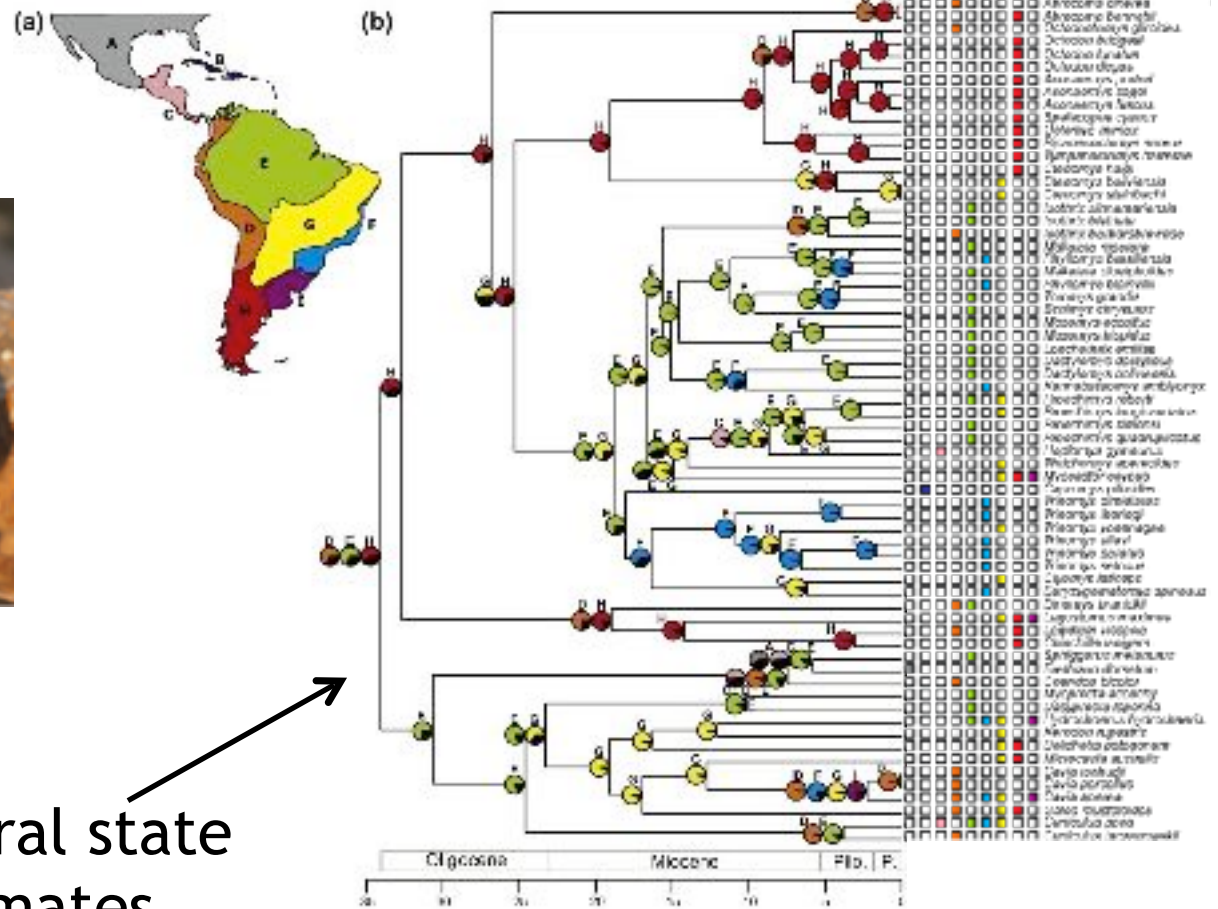
Discrete areas  
(character states)

Time-calibrated  
phylogeny

Data matrix



*Octodon degus*



Ancestral state  
estimates



# Data matrix

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

## Continuous

e.g. latitude-longitude

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

## Discrete

e.g. single area presence-absence  
(range)  $X_i = \text{Africa}$   $X_i = (0, 0, 1, 0, 0, 1, 1, 1)$



# Models

## Continuous

e.g. Brownian motion (Gaussian)

$$X \sim N(\mu, \Sigma)$$

## Discrete

e.g. continuous-time Markov chain

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

# Discrete Island Model

Dispersal-only model

One area per taxon (endemic/individual)

Learn favored dispersal routes

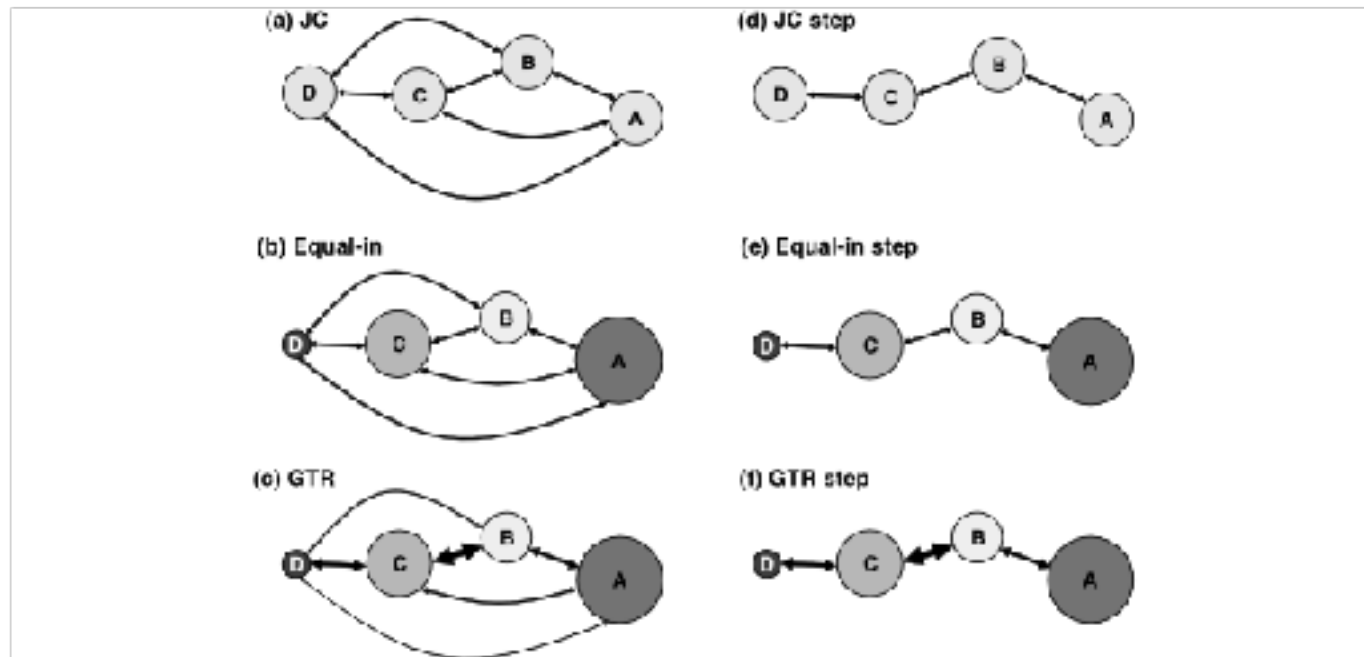
Work by:

Sanmartín *et al.*, 2008 (J Biogeog)

Lemey *et al.*, 2009 (PLoS Comp Biol)

# Dispersal on a graph

I. Sanmartin, 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–Gantor (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)

$$Q = \begin{matrix} & A & B & C & D \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \left( \begin{array}{cccc} - & \pi_{B^r_{AB}} & \pi_{C^r_{AC}} & \pi_{D^r_{AD}} \\ \pi_{A^r_{AB}} & - & \pi_{C^r_{BC}} & \pi_{D^r_{BD}} \\ \pi_{A^r_{AC}} & \pi_{B^r_{BC}} & - & \pi_{D^r_{CD}} \\ \pi_{A^r_{AD}} & \pi_{B^r_{BD}} & \pi_{C^r_{CD}} & - \end{array} \right) \end{matrix}$$

Stepping  
Stone  
(b - f)

$$Q = \begin{matrix} & A & B & C & D \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \left( \begin{array}{cccc} - & \pi_{B^r_{AB}} & 0 & 0 \\ \pi_{A^r_{AB}} & - & \pi_{C^r_{BC}} & 0 \\ 0 & \pi_{B^r_{BC}} & - & \pi_{D^r_{CD}} \\ 0 & 0 & \pi_{C^r_{CD}} & - \end{array} \right) \end{matrix}$$

# Continuous models

Brownian motion

Each taxon is an individual sample

Population and epidemiology models

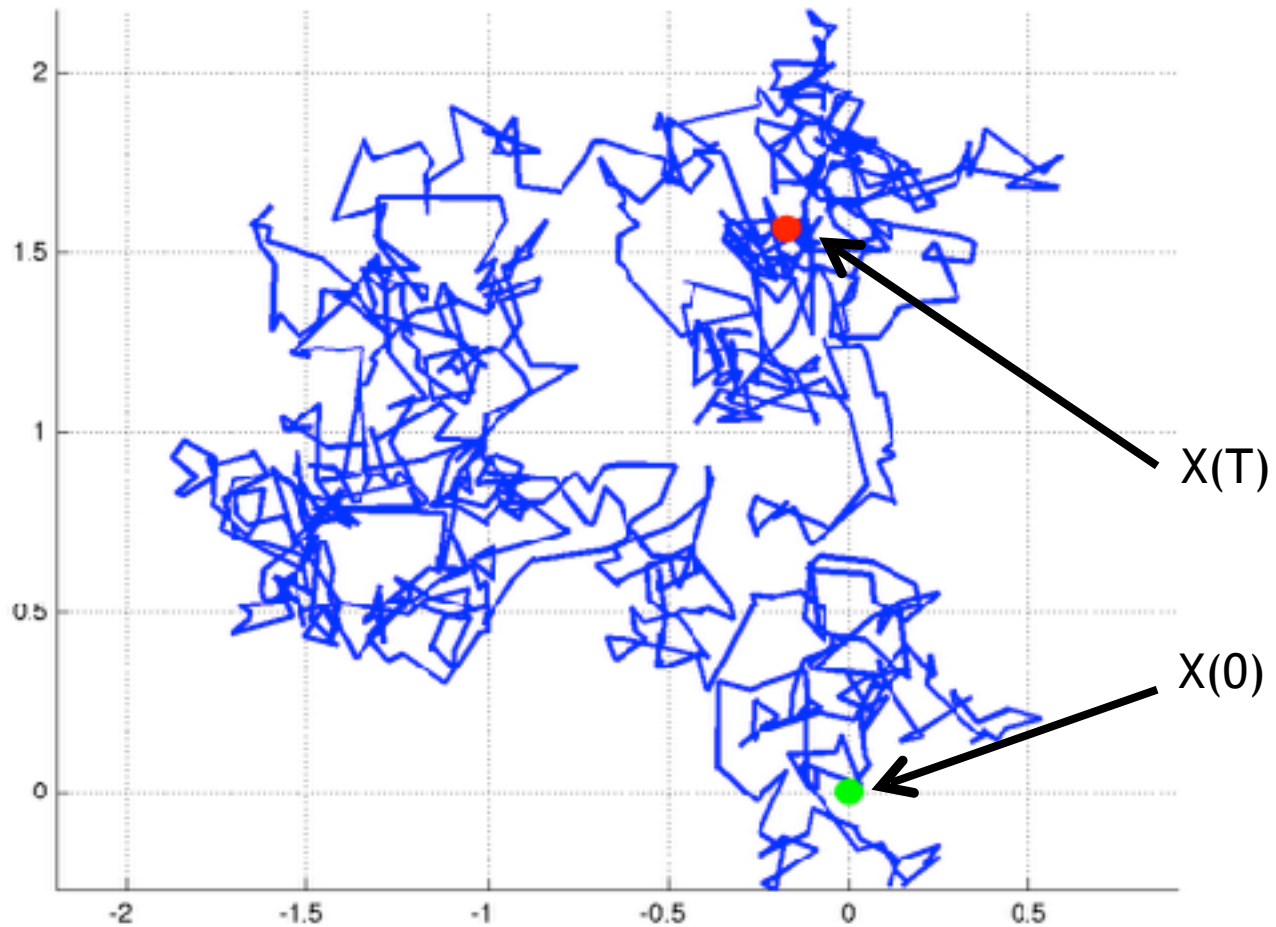
Work by:

Lemmon & Lemmon, 2008 (Syst Biol)

Lemey *et al.*, 2010 (Mol Biol Evol)

Bouckaert *et al.* 2012 (Science)

# 2D Brownian motion



# Relaxed random walk

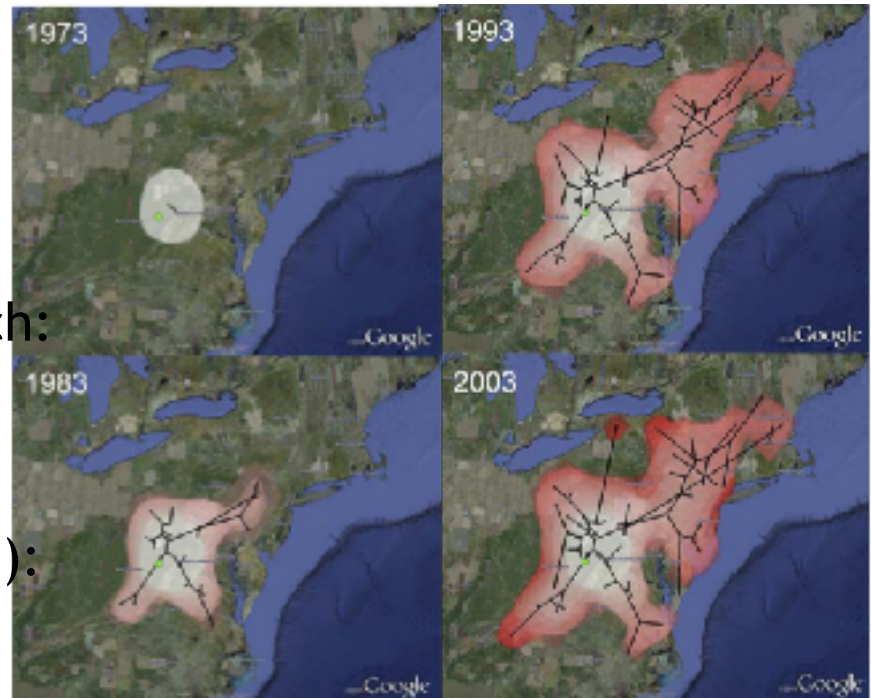
Joint inference of gene tree using relaxed molecular clock

Latitude, longitude diffuse by Brownian motion for each branch:

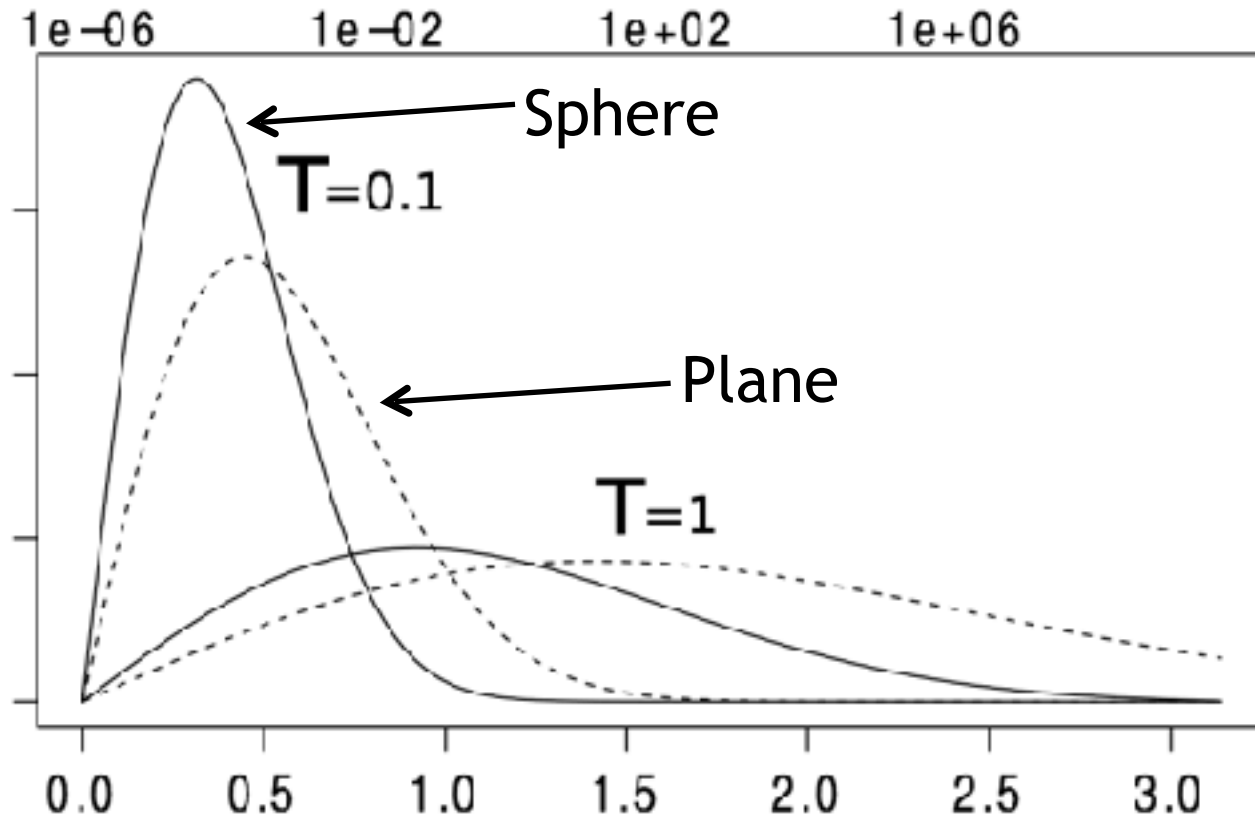
$$X_b \sim N(X_{pa(b)}, t_b \phi_b \Sigma)$$

Branch rate rescaled (“relaxed”):

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



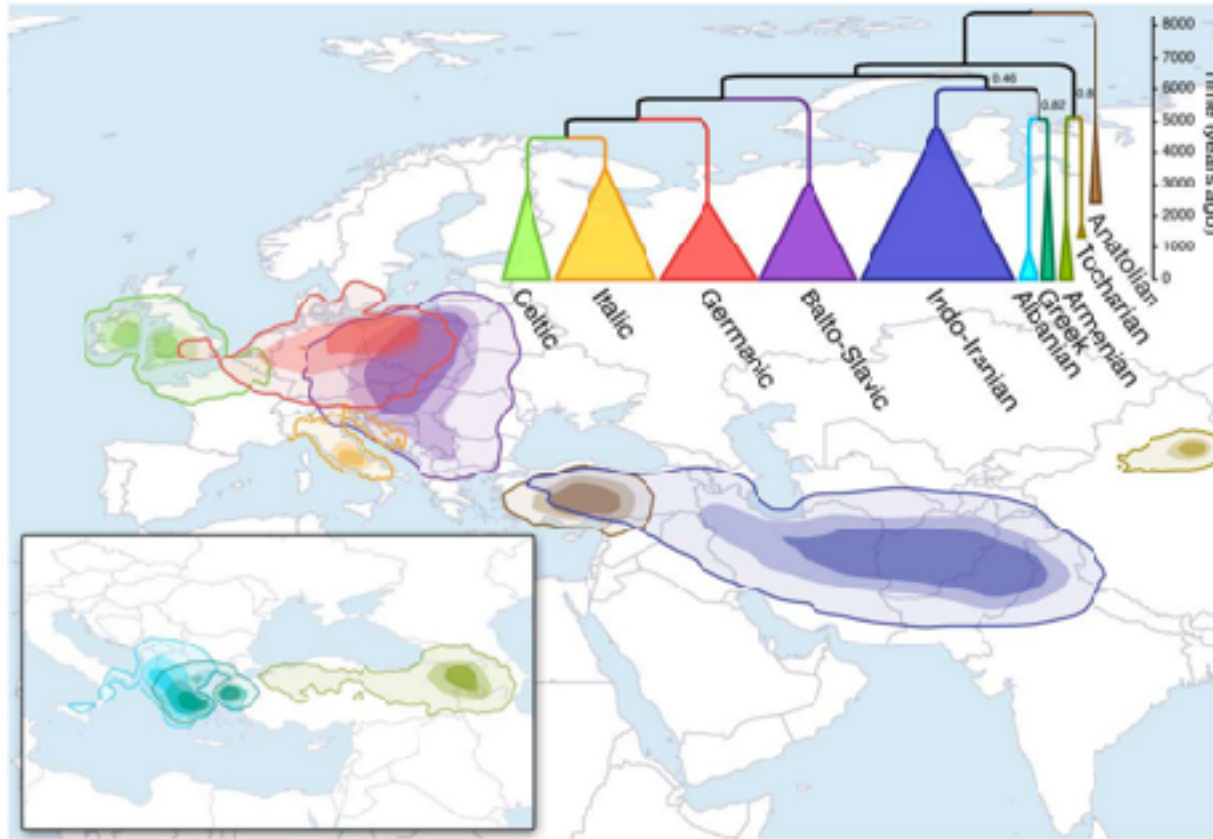
# Earth is round



Diffusion distance for  $T=0.1$ ,  $T=1$



# Earth is wet



Numerical integration + rasters to forbid certain “areas”

# Continuous range evolution models

Diffusion of set of individual coordinates

???      °8 → 8°

Diffusion of range as polygon

???      ▭ → ⬡

Harder problem, underexplored

# DEC

Dispersal-(Local) Extinction-Cladogenesis  
(DEC)

Many areas per taxon (range)

D,E as parameterized event types

Work by:

Ree *et al.*, 2005 (Evolution)

Ree & Smith, 2008 (Syst Biol)

# DEC event types

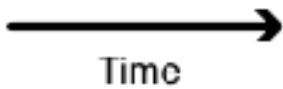
## Anagenesis



a. Dispersal



b. Extirpation



## Cladogenesis



c. Narrow sympatry



d. Subset sympatry



e. Allopatry



f. Full sympatry



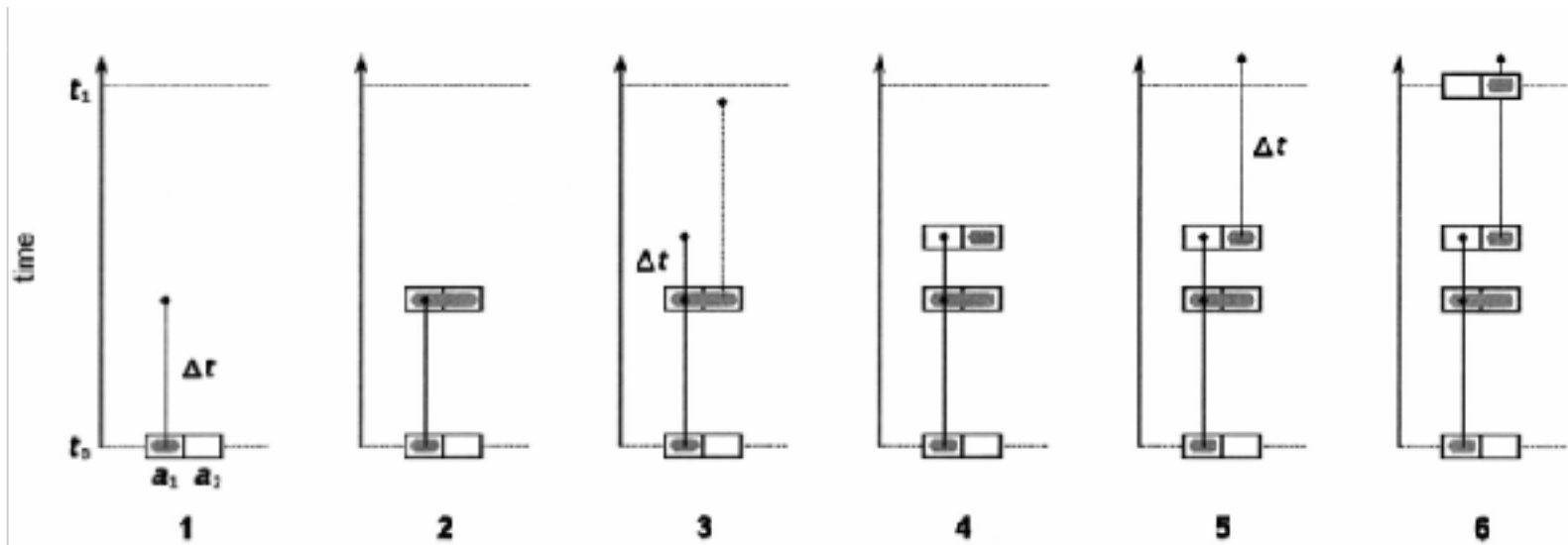
g. Jump dispersal

# Rate matrix for anagenesis

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

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

# Dispersal & Extirpation



Exponentially-distributed times between events

# DEC event types

## Anagenesis



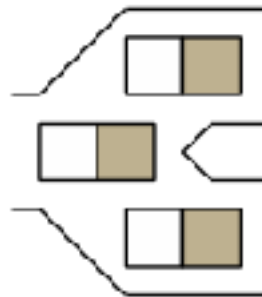
a. Dispersal



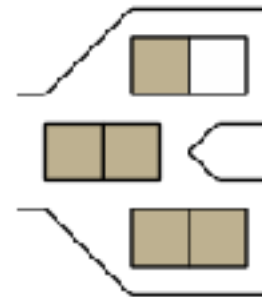
b. Extirpation



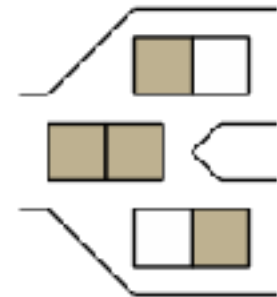
## Cladogenesis



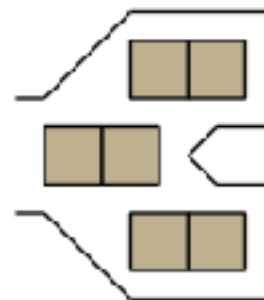
c. Narrow sympatry



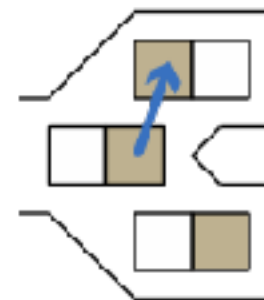
d. Subset sympatry



e. Allopatry



f. Full sympatry

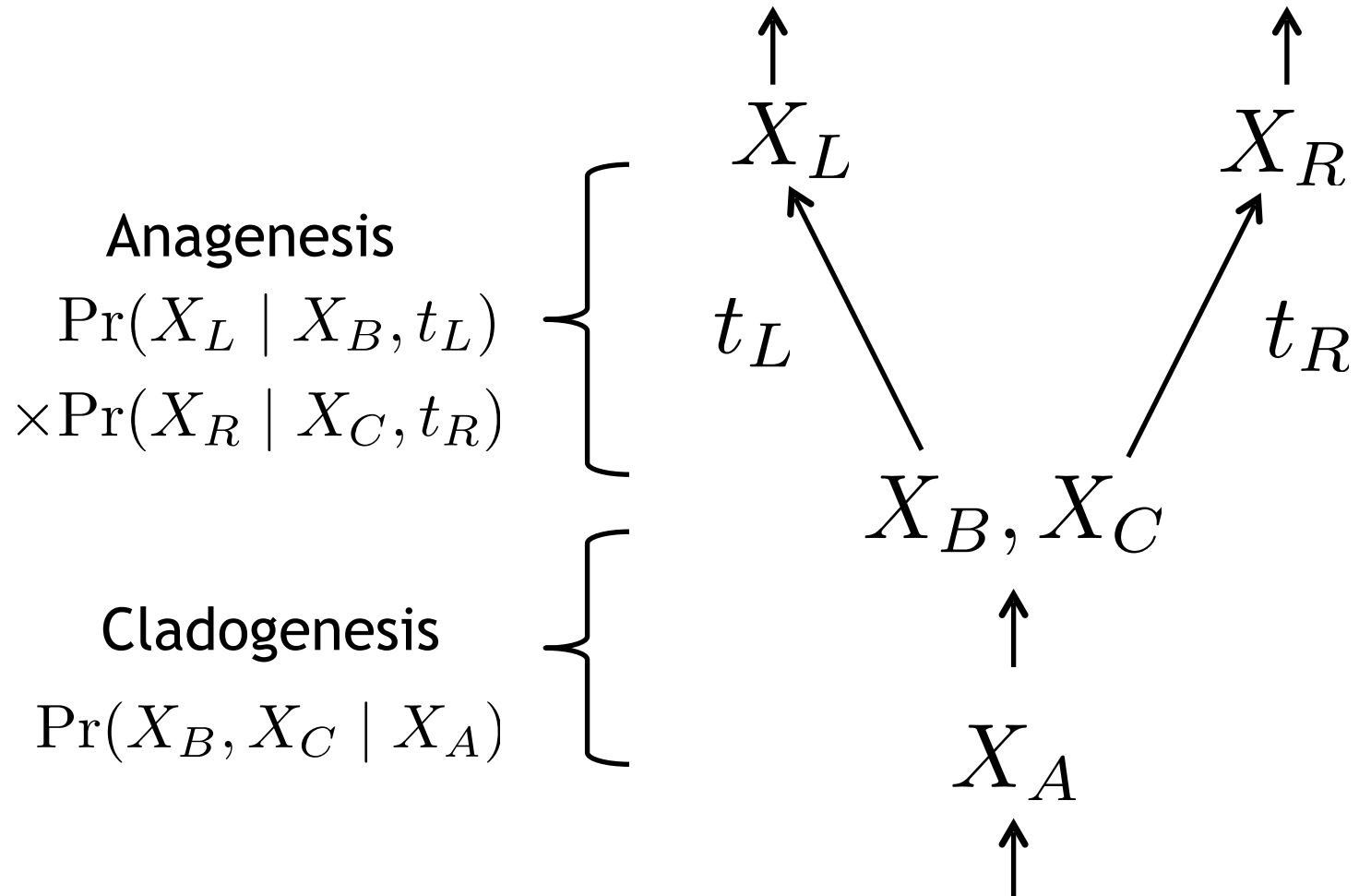


g. Jump dispersal

	Ancestral range	Subdivided range after speciation	Inheritance scenarios for subdivided range	Prob
Sympatry (narrow)		Scenario 1 		1/1
				1/1
Allopatry		Scenario 2 		1/6, 1/6
		Scenario 3 		
Sympatry (subset)				1/6, 1/6
				1/6, 1/6



# DEC likelihood

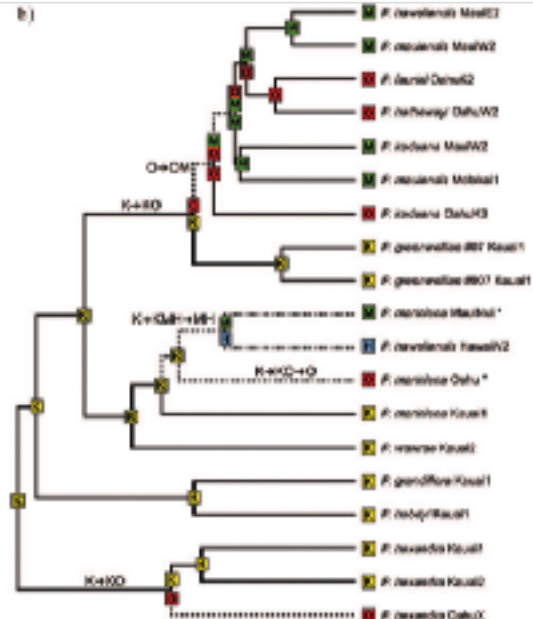
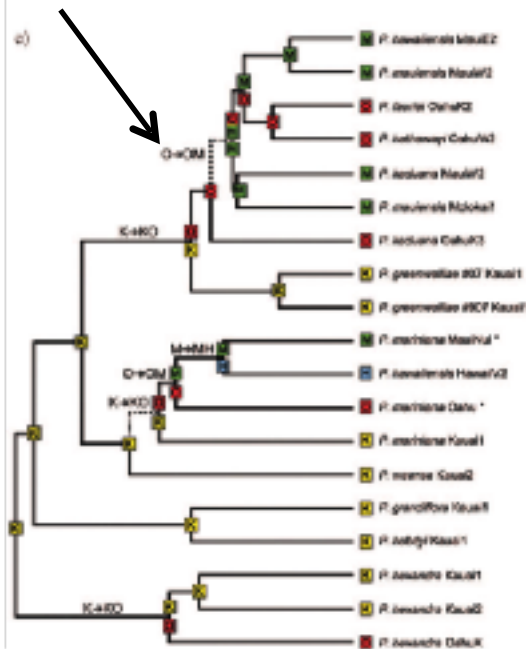


# Generalized DEC model

	Process	Ranges		Character mapping	DIVA	DEC (GILBERT, LAGRANGE)	BayArea, BBM (NASP)	Parameter of BioGeoBEARS Supermodel
		Before	After					
Anagenetic	Dispersal				✓	✓	✓	$d$ ( $\lambda, c$ )
	Extinction				✓	✓	✓	$e$ ( $\lambda, \omega$ )
	Range-switching			✓				$a$ ( $\lambda, s$ )
Cladogenetic	Sympatry (narrow)			✓	✓	✓	✓	$y$ ( $\lambda, \tau, \omega$ )
	Sympatry (widespread)						✓	$y$ ( $\lambda, \tau, \omega$ )
	Sympatry (subset)					✓		$s$ ( $\lambda, \tau, \omega$ )
	Vicariance (narrow)				✓	✓		$v$ ( $\lambda, \tau, \omega$ )
	Vicariance (widespread)				✓			$v$ ( $\lambda, \tau, \omega$ )
	Founder							$j$ ( $\lambda, c, \tau, \omega$ )



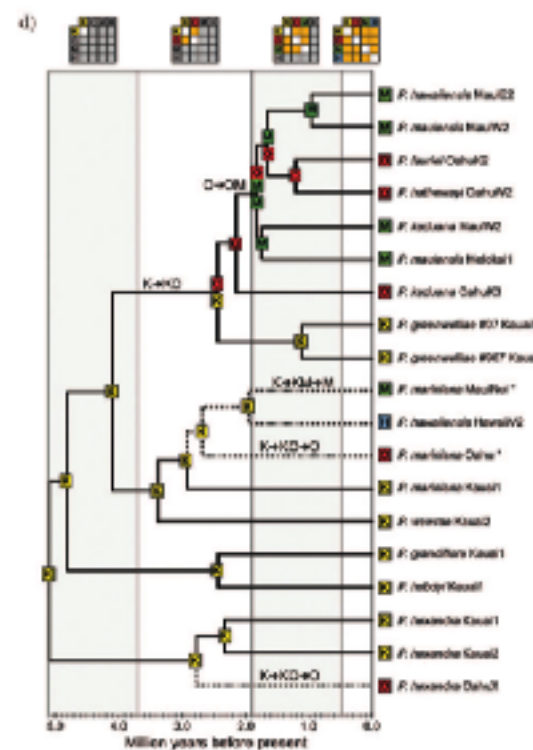
Stepping stone  
(Small adjacent ranges:  
K, O, M, H, KO, OM, MH)



Unconstrained  
(constant,  
equal rates)

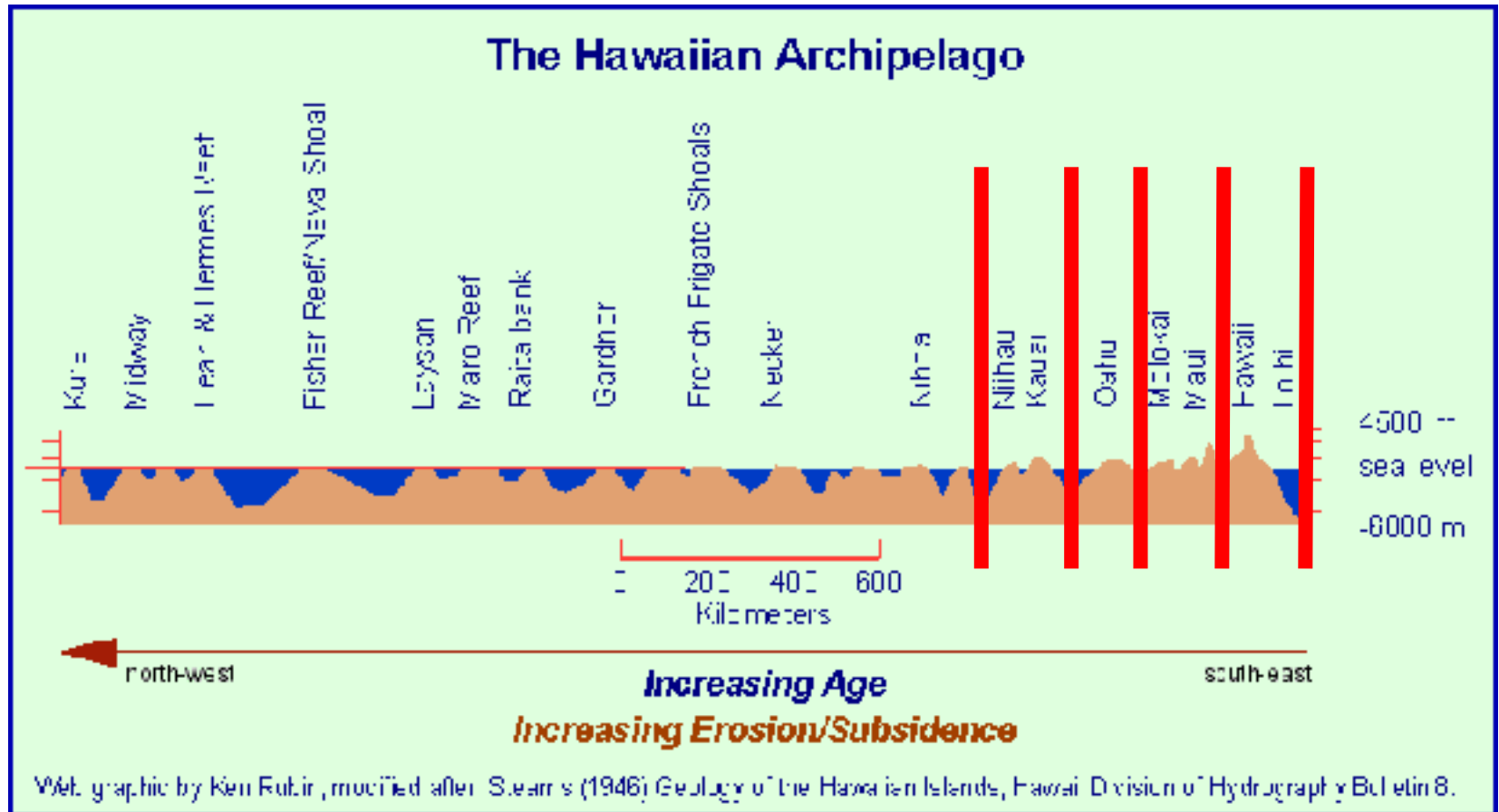


*Psychotria mariniana*

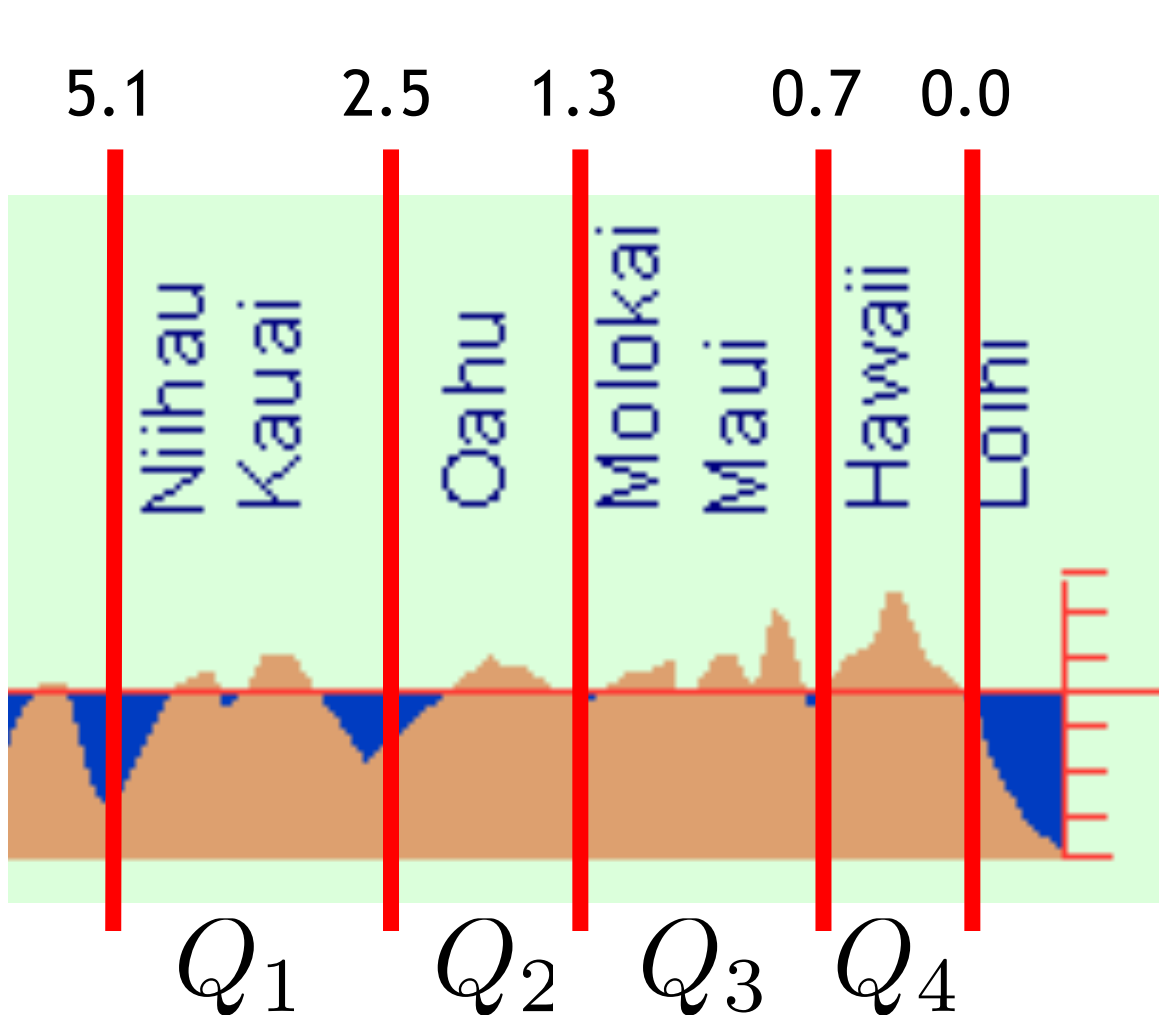


Stratified  
(time-dependent probs)

# Stratified model



Island origin times (Ma)

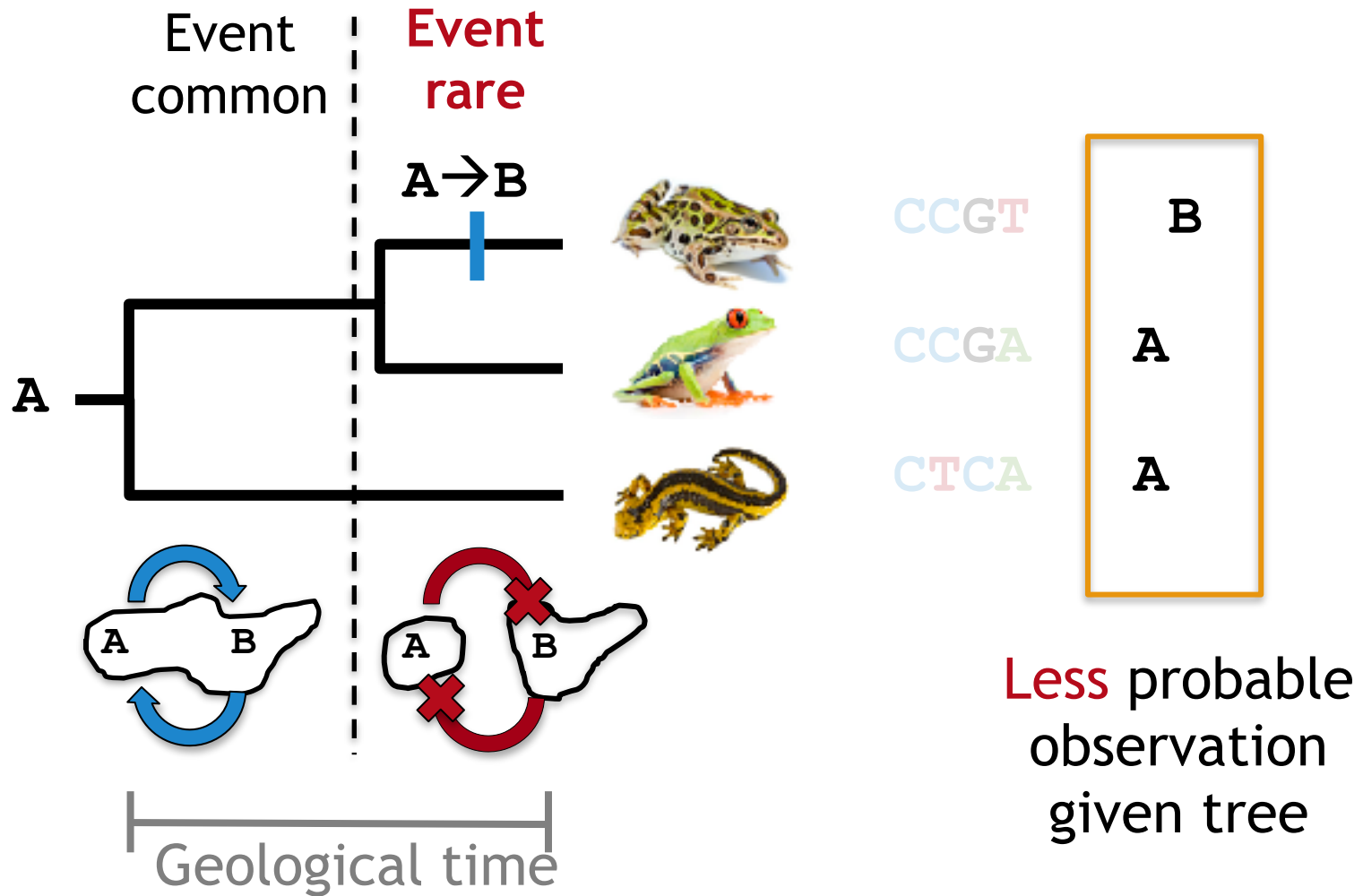


$Q_i$  varies over time!

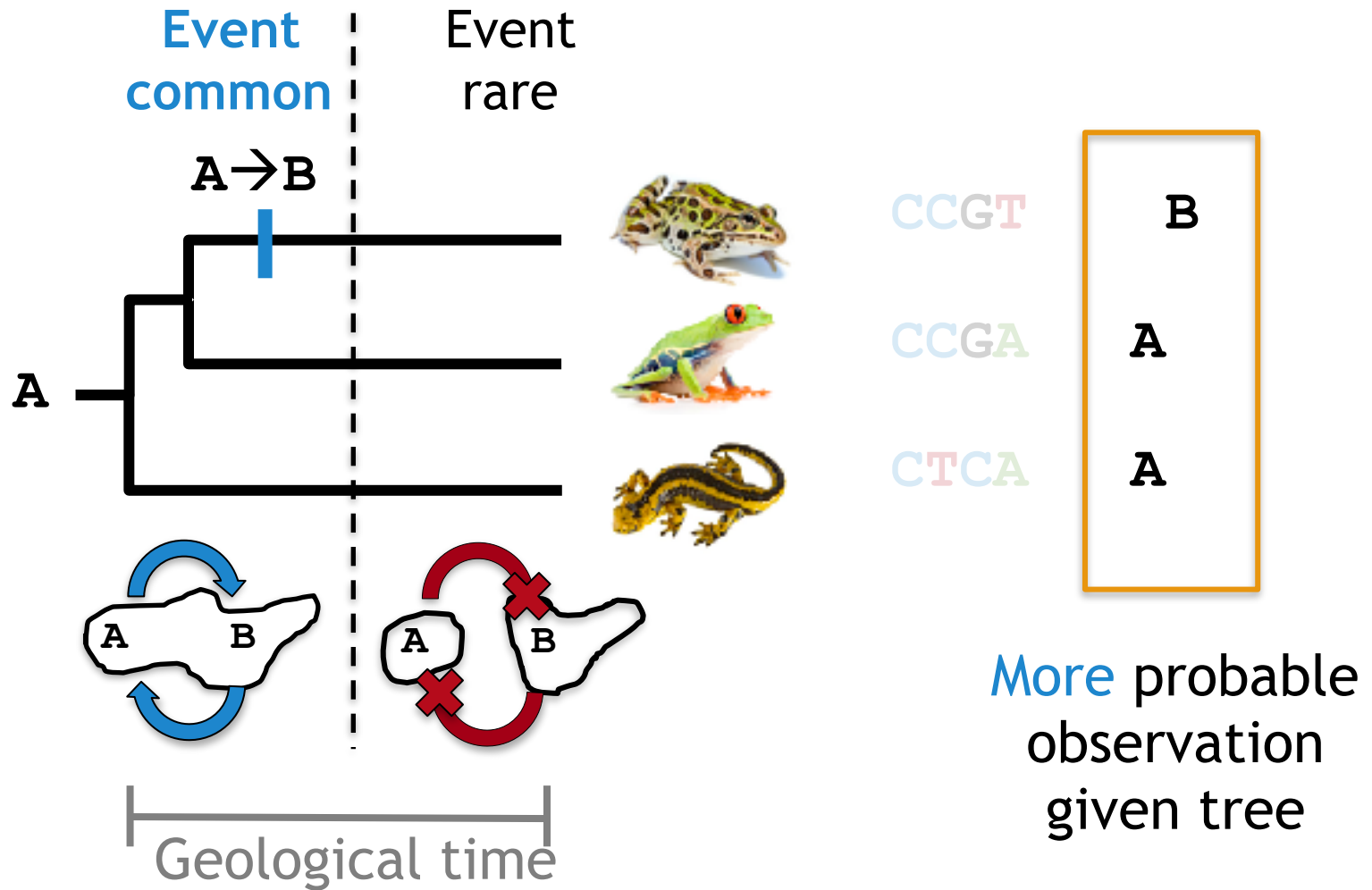
Transition probability over epochs

$$P(x \rightarrow y; t) = \left[ \prod e^{Q_i \delta t_i} \right]_{x,y}$$

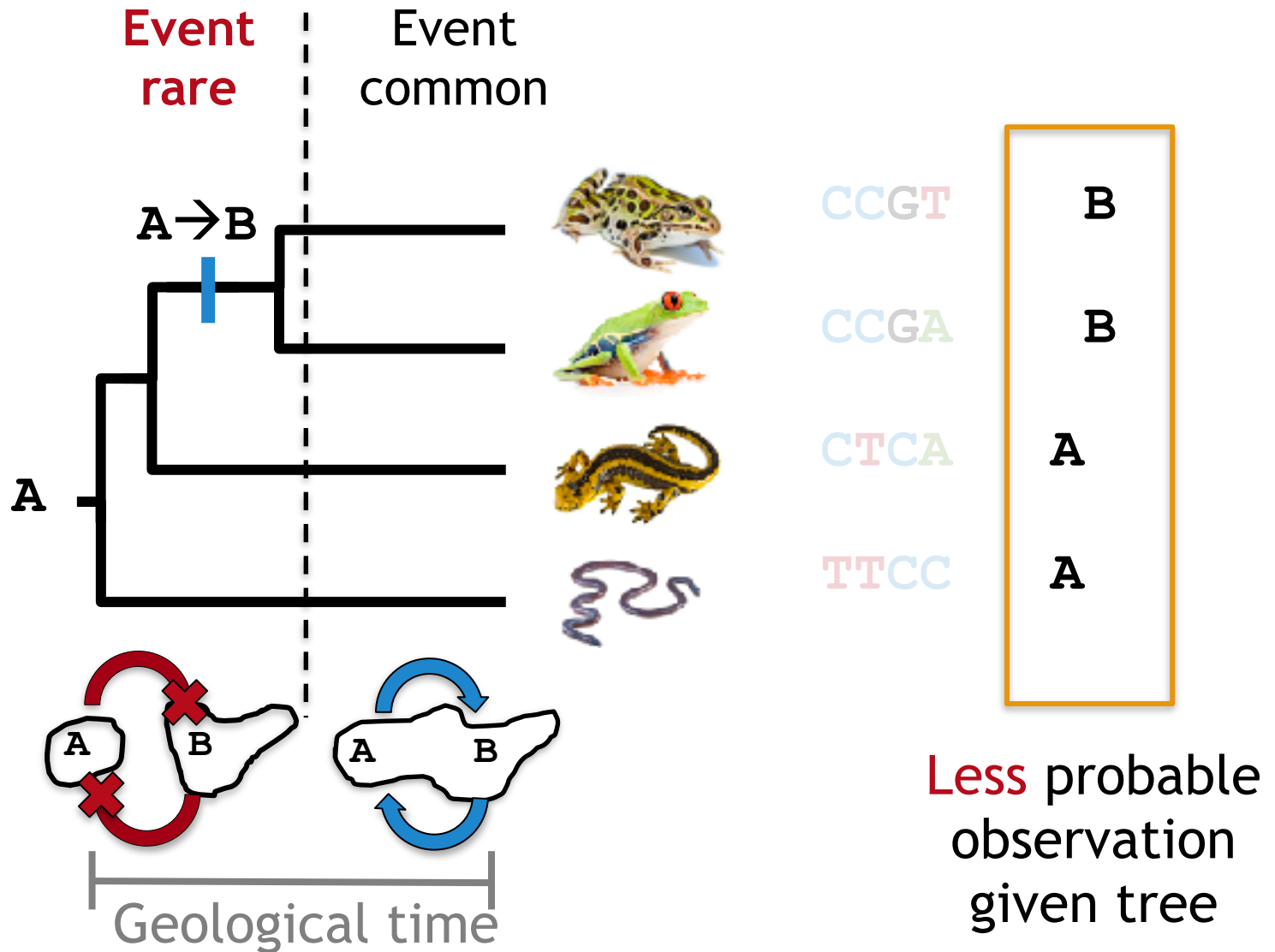
# Dispersal should occur *before* areas split



# Dispersal should occur *before areas split*

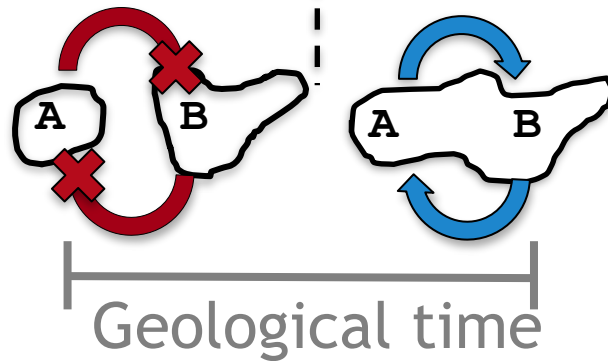
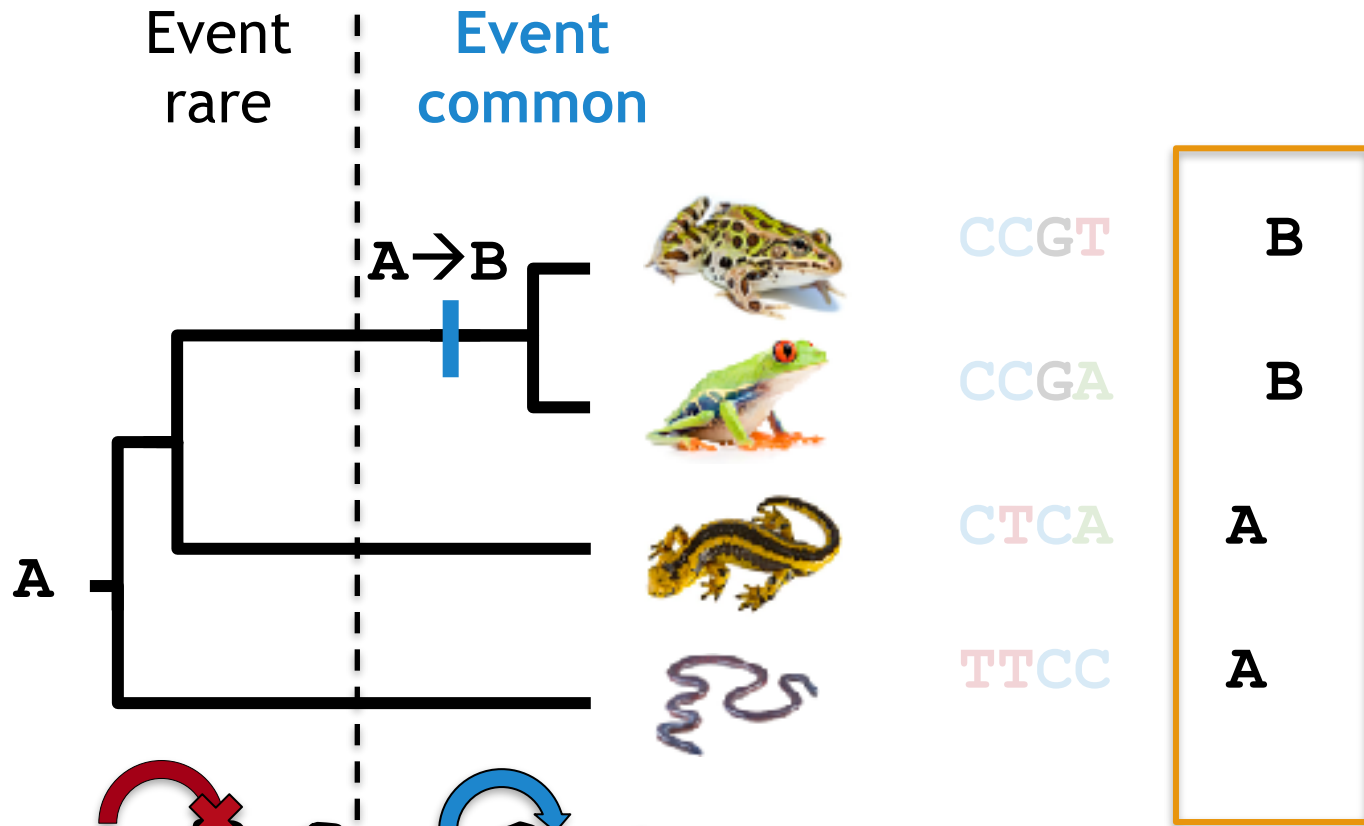


# Dispersal should occur *after* areas merge

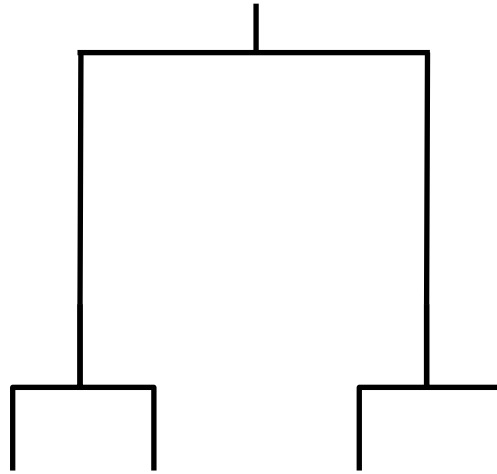




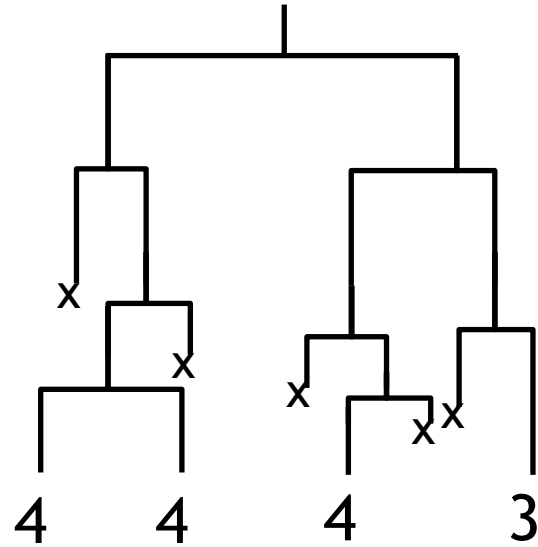
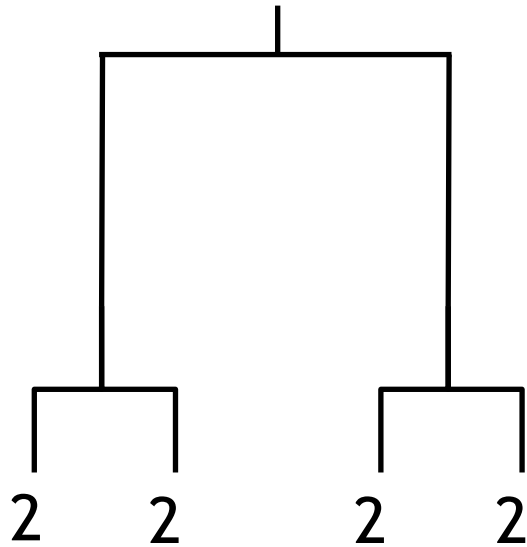
# Dispersal should occur *after* areas merge



More probable observation given tree



# Speciation hidden by extinction



# Geographic State Speciation Extinction (GeoSSE)

DEC model

Joint birth-death process & range evolution

Accounts for “hidden” speciation

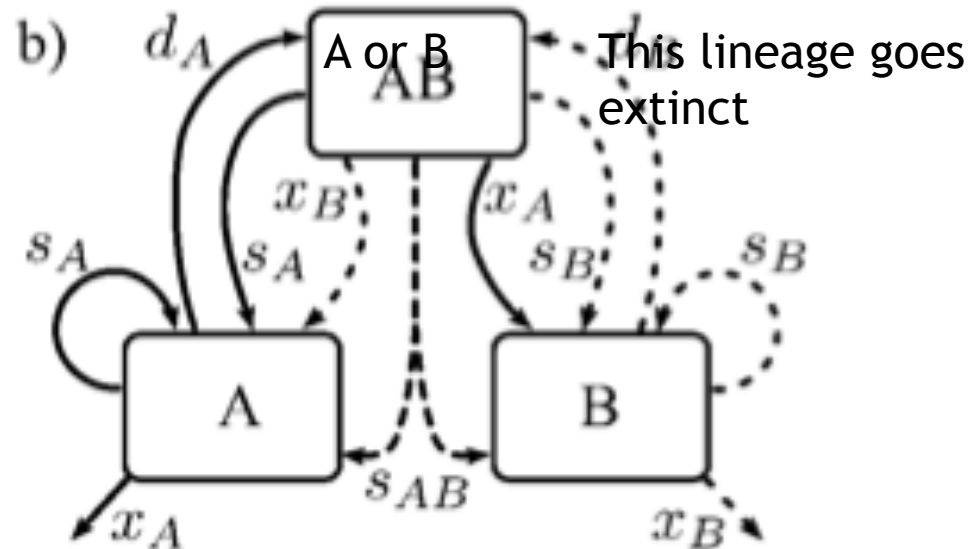
Range evolution, speciation, extinction  
intertwined

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 and/or B	This lineage gains an area
	$x_A, x_B$	AB	This lineage loses an area



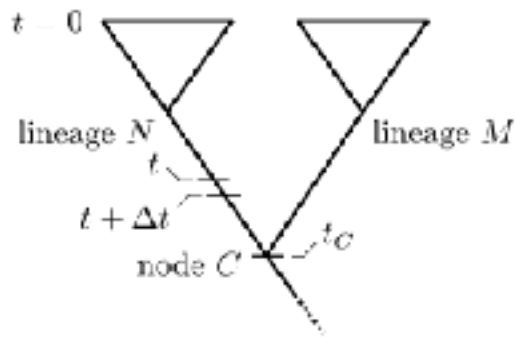
# GeoSSE likelihood

## Likelihood of anagenetic change

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

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

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



## Likelihood of extinction

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

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

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

## Likelihood at “observed” speciation

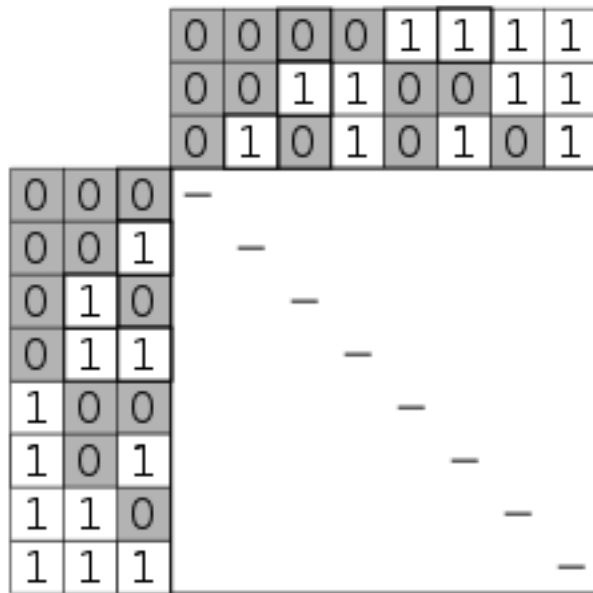
$$D_{CA}(t_C) = D_{NA}(t_C)D_{MA}(t_C)s_A, \quad (2a)$$

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

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

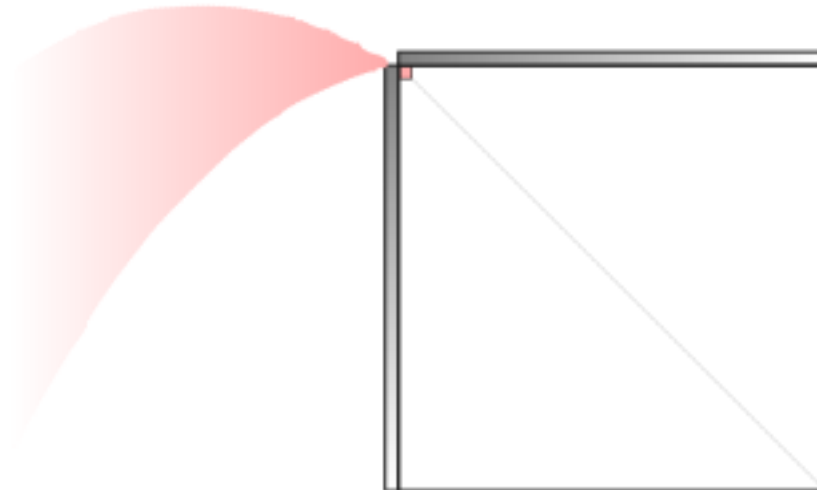
# Q explodes with many areas

3 areas



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

10 areas



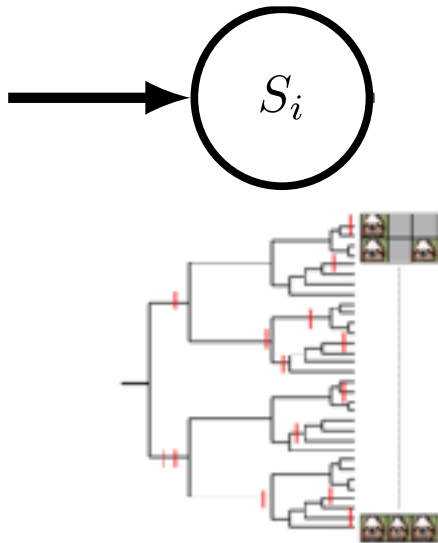
$$2^{10} \times 2^{10} = 1024 \times 1024$$

Each additional area doubles the number of ranges

# MCMC proposal to sample histories

Efficient sampling under iid model

Efficient likelihood under non-iid model

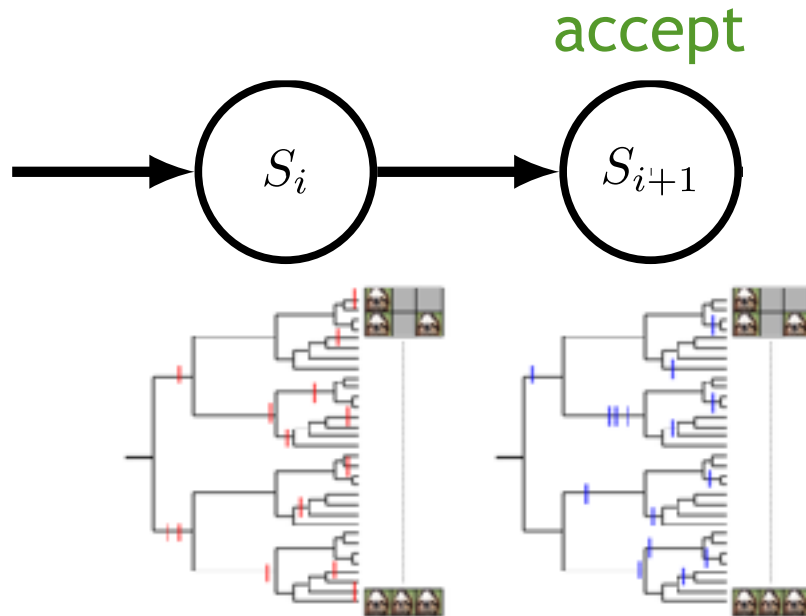




# MCMC proposal to sample histories

Efficient sampling under iid model

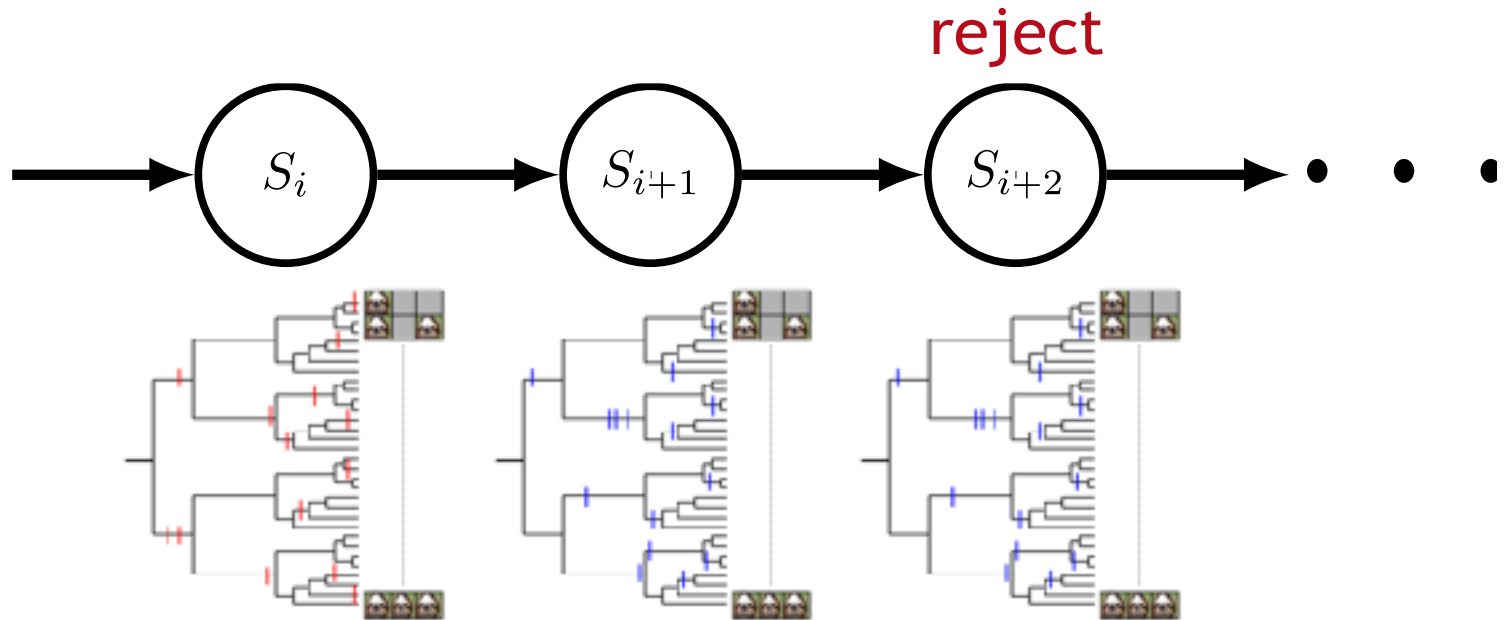
Efficient likelihood under non-iid model



# MCMC proposal to sample histories

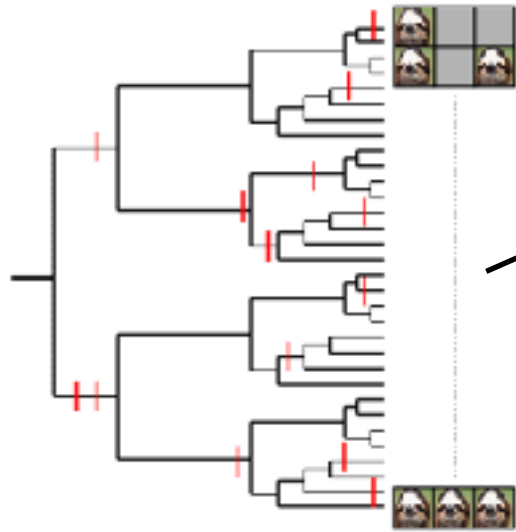
Efficient sampling under iid model

Efficient likelihood under non-iid model

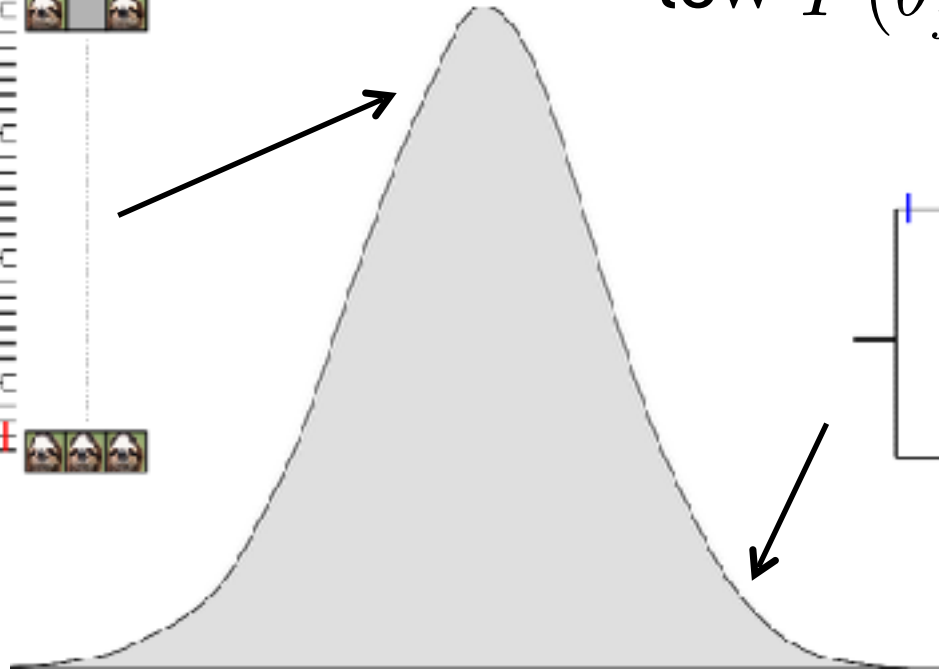


# Estimate $P(\theta, H \mid D)$ using MCMC

high  $P(\theta_i, H_i \mid D)$



low  $P(\theta_j, H_j \mid D)$



$\theta, H$

BayArea

Landis et al. (2013)

# Discrete biogeography in RevBayes

	Range size	# Areas	Anagenesis	Cladogenesis
<b>Island</b>	One	Many	Area-switching	No
<b>DEC</b>	Many	< 10	Dispersal-extirpation	Yes
<b>GeoSSE</b>	Many	< 4	Dispersal-extirpation Linked to diversification	Yes Linked to diversification
<b>BayArea</b>	Many	100+	Dispersal-extirpation	Some (experimental)

Choose the right model for your question!