

BioGeoBEARS: An R package for inference and model testing in historical biogeography

Nicholas J. Matzke

Postdoctoral Fellow in Mathematical Biology

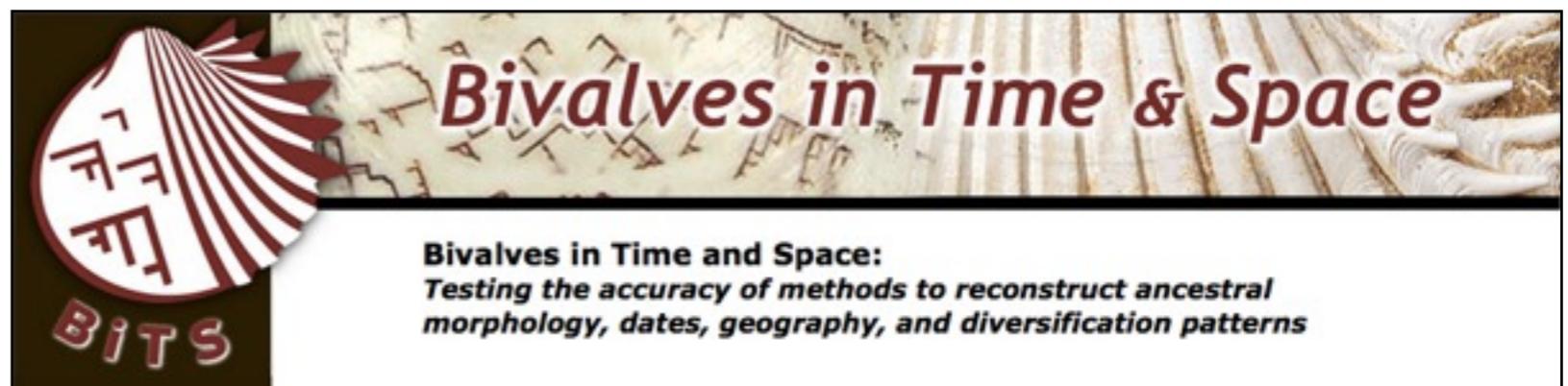
National Institute of Mathematical and Biological Synthesis

(NIMBioS, www.nimbios.org)

Ph.D. Research at:

Dept. Integrative Biology,
U.C. Berkeley

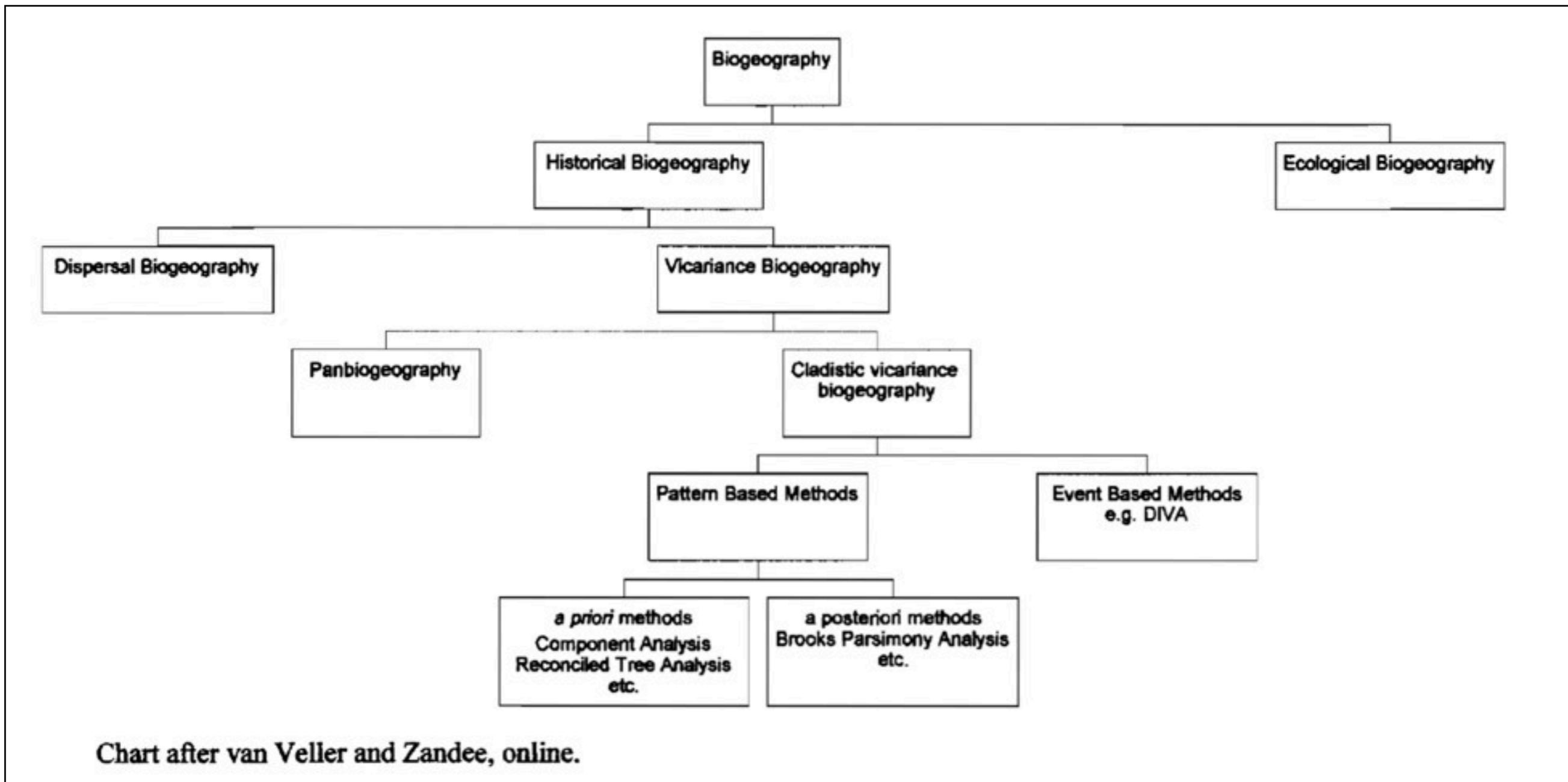
Nov. 1, 2013



Part I: Testing old and new models in biogeography

History of historical biogeography:

- heavy on assumptions about models
- light on model testing



For example: dispersal versus vicariance

For example: dispersal versus vicariance

Let's do a short review...

For example: dispersal versus vicariance

Let's do a short review...



D versus V

Opinion
TRENDS in Ecology and Evolution Vol.20 No.2 February 2005
Full text provided by www.electronicjournals.com
REVIEW

The resurrection of oceanic dispersal in historical biogeography

Alan de Queiroz
2695 Mineral Drive, Ely, NV 89301, USA

Geographical distributions of terrestrial or freshwater taxa that are broken up by oceans can be explained by either oceanic dispersal or vicariance in the form of fragmentation of a previously contiguous landmass. The validation of plate-tectonics theory provided a global vicariance mechanism and, along with cladistic arguments for the primacy of vicariance, helped create a view of oceanic dispersal as a rare phenomenon and an explanation of last resort. Here, I describe recent work that suggests that the importance of oceanic dispersal has been strongly underestimated. In particular, molecular dating of lineage divergences favors oceanic dispersal over tectonic vicariance as an explanation for disjunct distributions in a wide variety of taxa, from frogs to beetles to baobab trees. Other evidence, such as substantial gene flow among island populations of *Anolis* lizards, also indicates unexpectedly high frequencies of oceanic dispersal. The resurrection of oceanic dispersal is the most striking aspect of a major shift in historical biogeography toward a more even balance between vicariance and dispersal explanations. This new view implies that biotas are more dynamic and have more recent origins than had been thought previously. A high frequency of dispersal also suggests that a fundamental methodological assumption of many biogeographical studies – that vicariance is *a priori* a more probable explanation than dispersal – needs to be re-evaluated and perhaps discarded.

Glossary

Area cladogram: a cladogram in which the taxa have been replaced by the areas in which they occur. From an area cladogram, various algorithms can be used to derive a resolved area cladogram in which a single area is associated with each terminal node and each area is represented only once. A resolved area cladogram is meant to reflect the history of biotic connections among areas for that group [7]. For example, if areas A and B are grouped together in a resolved area cladogram to the exclusion of area C, this implies that A and B had a more recent biotic connection to each other than either did to C. Area

“A high frequency of dispersal also suggests that a fundamental methodological assumption of many biogeographical studies -- that vicariance is *a priori* a more probable explanation than dispersal -- needs to be re-evaluated and perhaps discarded.”

de Queiroz, A. (2005). "The resurrection of oceanic dispersal in historical biogeography." TREE, 20(2).

Journal of Biogeography U. Biogeogr.) (2006) 33, 193–198

GUEST EDITORIAL



Dispersal is fundamental to biogeography and the evolution of biodiversity on oceanic islands

Robert H. Cowie* and Brenden S. Holland

ABSTRACT

Vicariance biogeography emerged several decades ago from the fusion of cladistics and plate tectonics, and quickly came to dominate historical biogeography. The field has since been largely constrained by the notion that only processes of vicariance and not dispersal offer testable patterns and refutable hypotheses, dispersal being a random process essentially adding only noise to a vicariant system. A consequence of this thinking seems to have been a focus on the biogeography of continents and continental islands, considering the biogeography of oceanic islands less worthy of scientific attention because, being dependent on stochastic dispersal, it was uninteresting. However, the importance of dispersal is increasingly being recognized, and here we stress its fundamental role in the generation of biodiversity on oceanic islands that have been created *in situ*, never connected to larger land masses. Historical dispersal patterns resulting in modern distributions, once considered unknowable, are now being revealed in many plant and animal taxa, in large part through the analysis of polymorphic molecular markers. We emphasize the profound evolutionary insights that oceanic island biodiversity has provided, and the fact that, although small in area, oceanic islands harbour disproportionately high biodiversity and numbers of endemic taxa. We further stress the importance of continuing research on mechanisms generating oceanic island biodiversity, especially detection of general, non-random patterns of dispersal, and hence the need to acknowledge oceanic dispersal as significant and worthy of research.

Keywords
Biodiversity, dispersal, endemism, historical biogeography, hot spot islands, land snails, oceanic islands, Pacific Ocean.

INTRODUCTION

A consequence of this thinking seems to have been a focus on the biogeography of continents and continental islands.

“We hope, therefore, that the trend identified by de Queiroz (2005) – the resurrection of oceanic dispersal as important in historical biogeography – is real and that the straightjacket of strict vicariance biogeography is being loosened to include once again the plurality of mechanisms and processes...”

p. 197 of: Cowie, R.H.; Holland, B.S. (2006). "Dispersal is fundamental to biogeography and the evolution of biodiversity on oceanic islands." Journal of Biogeography, 33, 193–198. Emphasis added.

D versus V

Journal of Biogeography (J. Biogeogr.) (2007) 34, 1470–1471

CORRESPONDENCE



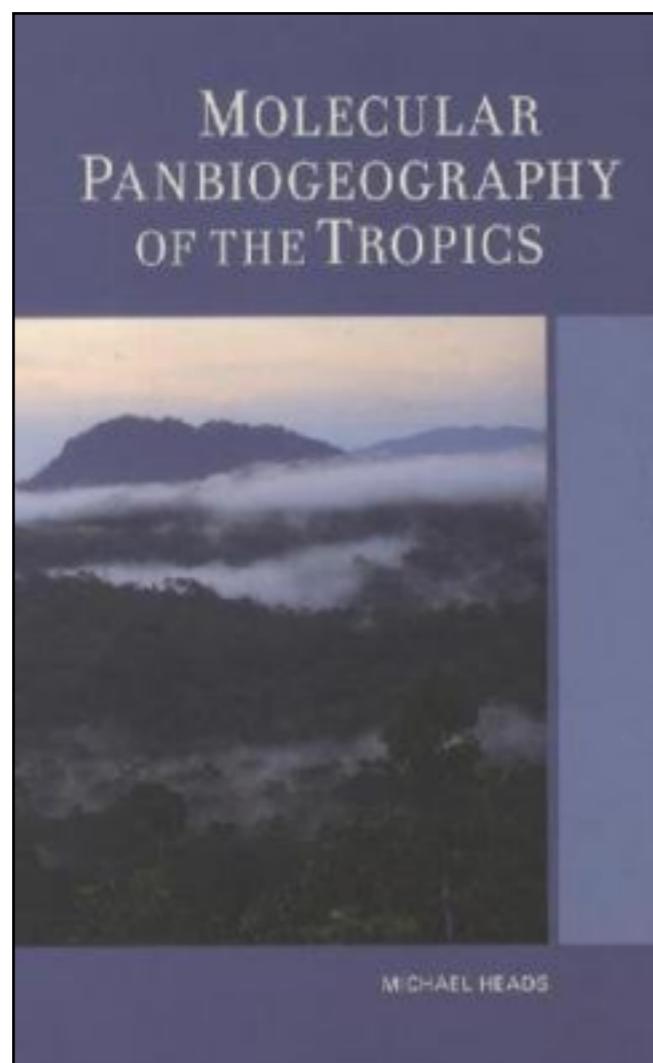
On basal clades and ancestral areas

During the last 40 years, historical biogeography has been the arena for an apparently endless struggle: the so-called dispersal versus vicariance debate. It is a widespread view that after the 1970s and early 1980s, vicariance dominated historical biogeographical explanations, with dispersal reduced to biogeographical noise wherein no common pattern is able to elucidate the relationships among the areas under inquiry. Most authors see

It is not the aim of the present correspondence to repeat Ebach's (1999) or Morrone's (2002) argumentation against ancestral area analysis, but to reinforce the criticisms already made, adding the points raised by Krell & Cranston (2004) and Crisp & Cook (2005) concerning the relative position of 'more basal' branches in cladograms, which is critical to Bremer's (1992) analytical procedure.

According to Ebach (1999), central to Bremer's ancestral area analysis is Hennig's (1966) 'progression rule', which

analysis. An equivalent tree is presented in Fig. 1. The calculation of gains and losses shows that area A has the highest G : L ratio compared to the other areas (A has a G : L value of 1, while B has a G : L value of 0.5, and both C and D have G : L values of 0.33). This way, area A is the most probable ancestral area and, according to Bremer's method, it has to be topologically basal, as stated by the progression rule. Nevertheless, as pointed out by Krell & Cranston (2004), 'a clade branching off near the base is a basal



“During the last 40 years, historical biogeography has been the arena for an apparently endless struggle: the so-called dispersal versus vicariance debate.”

“The resurrection of dispersalism, as de Queiroz (2005) has envisioned, is reactionary and with little (or even no) empirical foundation.”

Santos, C.M.D. (2007). “On basal clades and ancestral areas.” *Journal of Biogeography*, 34, 1470–1469.

“Most modern biogeographers follow Mayr...in accepting that allopatry can be found by vicariance (dichopatry) or by founder dispersal (peripatry), but only vicariance is accepted here.”

“founder dispersal...is controversial and may not exist.”

p. 15 in: Heads, M.J. (2012) *Molecular panbiogeography of the tropics*. University of California Press, Berkeley, CA.

**It would be nice if we could do
statistical model choice here.**

Let the data decide.

State-of-the-art method: LAGRANGE DEC: Dispersal-extinction cladogenesis

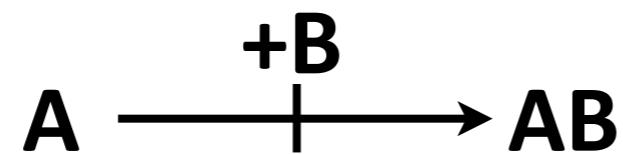
- d = “dispersal” = range extension
- e = “extinction” = range loss
- cladogenesis process

Implemented in the program LAGRANGE

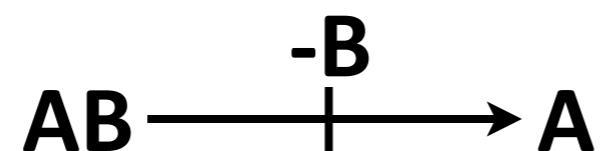
(Likelihood Analysis of Geographic RANGe Evolution)

by Rick Ree & Stephen Smith (2008)

DEC: Dispersal-extinction cladogenesis



**“dispersal” event
(really, range expansion)**



**“extinction” event
(really, extirpation or
range contraction)**

DEC: Dispersal-extinction cladogenesis

- d = “dispersal” = range extension
- e = “extinction” = range loss

Instantaneous rate matrix:

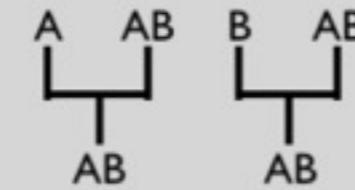
		ending range							
		null	A	B	C	AB	BC	AC	ABC
starting range	null	-	0	0	0	0	0	0	0
	A	e	-	0	0	d	0	d	0
	B	e	0	-	0	d	d	0	0
	C	e	0	-	-	0	d	d	0
	AB	0	e	e	-	-	0	0	$d+d$
	BC	0	0	e	e	-	-	0	$d+d$
	AC	0	e	0	e	0	-	-	$d+d$
		0	0	0	0	e	e	e	-

DEC: Dispersal-extinction cladogenesis

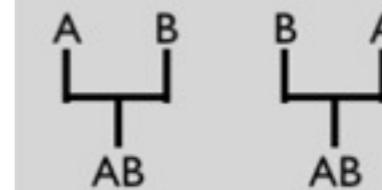
**Cladogenesis -- range evolution at
speciation events**



**Sympatric
speciation
(range
duplication)**



**Sympatric
speciation
(subset)**



**Vicariant
speciation
(range
division)**

LAGRANGE has a *fixed* speciation model (grey areas below)

Types of speciation, and example descendant ranges:

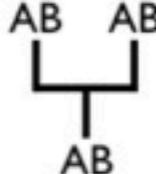
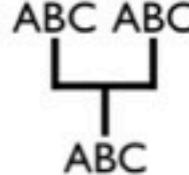
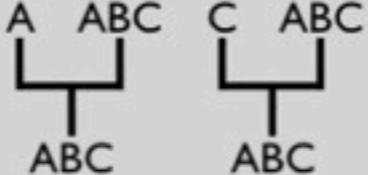
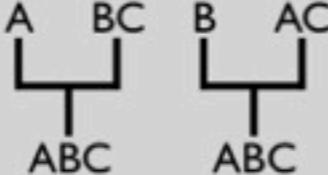
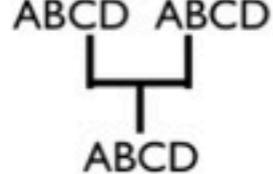
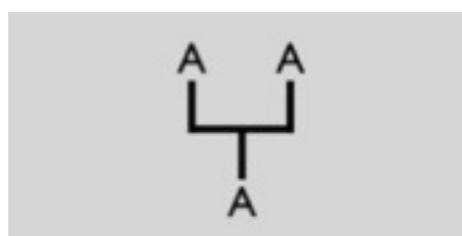
Ancestral ranges:	Sympatric (range copying)	Sympatric (subset)	Vicariance
A		--	--
AB			
ABC			
ABCD			

Figure 1. Various models for the evolution of geographic range at cladogenesis events. The events allowed by LAGRANGE are highlighted in gray. Each allowed event is fixed to have equal probability in the LAGRANGE algorithm.

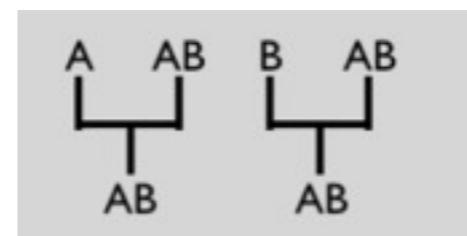
DEC: Dispersal-extinction cladogenesis

Cladogenesis -- range evolution at speciation events (LAGRANGE** model)**

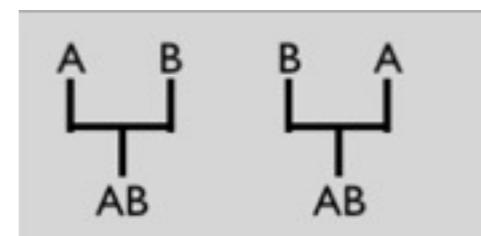
Probabilités per speciation event:



sympatric



subset



vicariance

		ending range																																			
starting range	left:	A	A	A	A	A	A	B	B	B	B	B	C	C	C	C	AB	AB	AB	AB	AB	BC	BC	BC	BC	AC	AC	AC	AC	AC	ABC	ABC	ABC	ABC	ABC	ABC	count 1 1 1 6 6 6 12
	right:	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	
	A	sym																																			
	B	sym																																			
	C	sym																																			
	AB	vic	sub	vic		sub	vic		sub	vic		sub	sub		sub																						
BC		vic		sub		vic		sub		vic		sub		sub		sub		sub		sub		sub		sub		sub											
AC		vic		sub		vic		sub		vic		sub		sub		sub		sub		sub		sub		sub		sub											
ABC		vic		sub		vic		sub		vic		sub		sub		sub		sub		sub		sub		sub		sub											

LAGRANGE: all events equiprobable

Problems with the state-of-the-art analysis

- Complete range data is assumed
- “true” vicariance is not allowed
 - new species *always* start out at rangesize=1
- Species always start inside their ancestral range
- In LAGRANGE, this cladogenesis model is *fixed and unchangeable*

Unresolved fundamental issues

What is the right model?

- “Vicariance vs. dispersal”
- Range extension / contraction
- Cladogenic vs. anagenic change
- Founder-event speciation

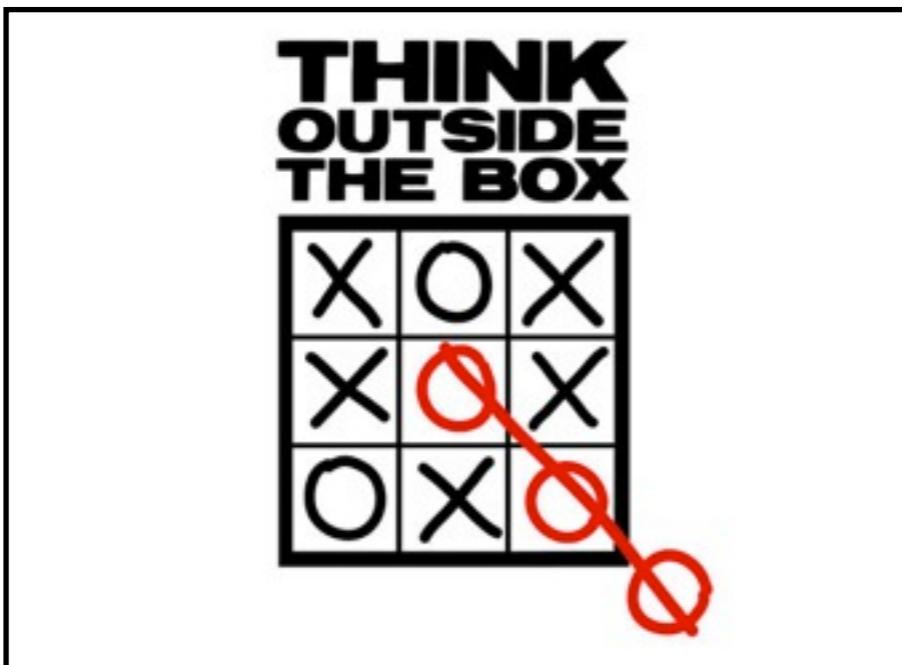
Unresolved fundamental issues

What is the right model?

- “Vicariance vs. dispersal”
- Range extension / contraction
- Cladogenic vs. anagenic change
- **Founder-event speciation**

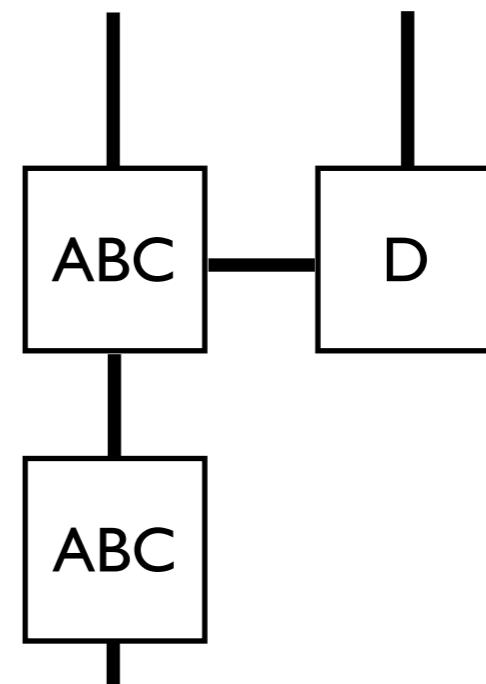
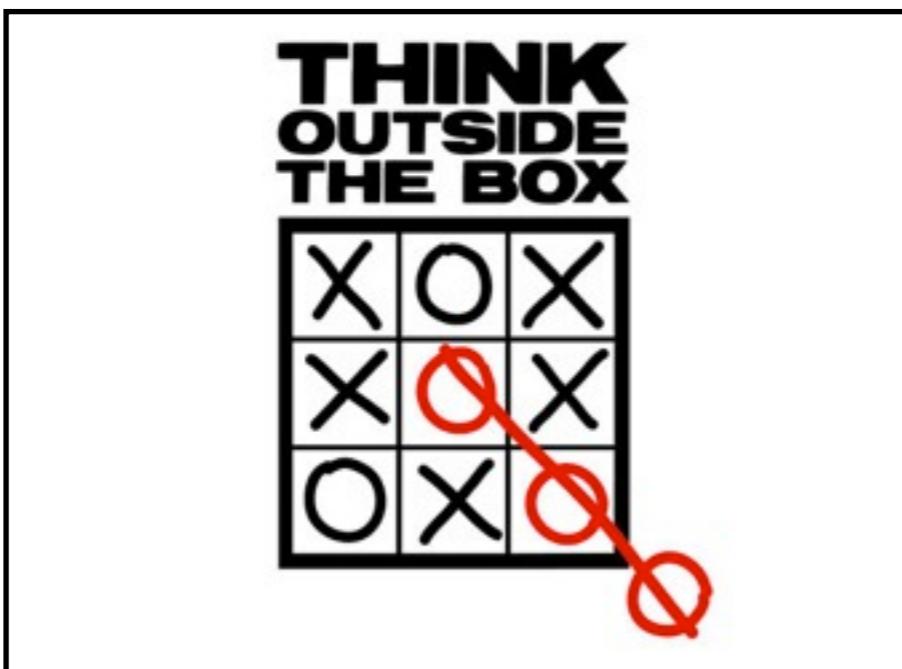
Founder-event speciation

- In founder-event speciation, a rare dispersal event “instantaneously” establishes a geographically isolated new lineage with one or a few individuals
- Widely discussed in population genetics, speciation literature, island biogeography
- However, so far ignored in historical biogeography models



Founder-event speciation

- In founder-event speciation, a rare dispersal event “instantaneously” establishes a geographically isolated new lineage with one or a few individuals
- Widely discussed in population genetics, speciation literature, island biogeography
- However, so far ignored in historical biogeography models



LAGRANGE has a *fixed* speciation model (grey areas below)

Types of speciation, and example descendant ranges:

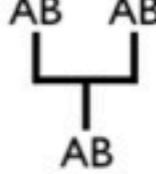
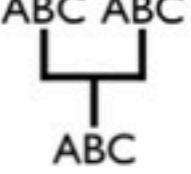
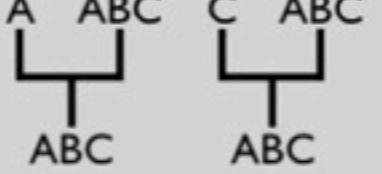
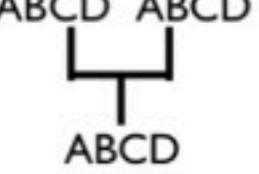
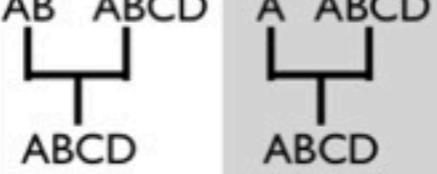
Ancestral ranges:	Sympatric (range copying)	Sympatric (subset)	Vicariance
A		--	--
AB			
ABC			
ABCD			

Figure 1. Various models for the evolution of geographic range at cladogenesis events. The events allowed by LAGRANGE are highlighted in gray. Each allowed event is fixed to have equal probability in the LAGRANGE algorithm.

Let's think outside the box!

Types of speciation, and example descendant ranges:

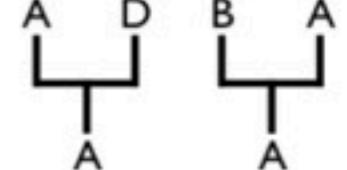
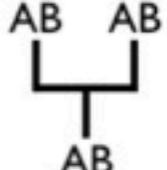
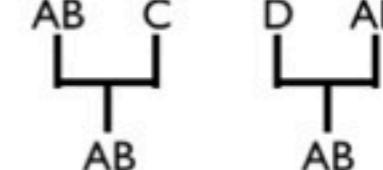
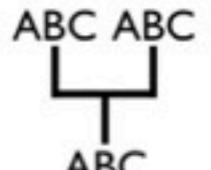
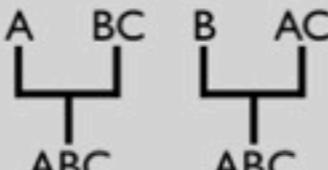
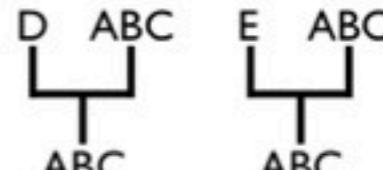
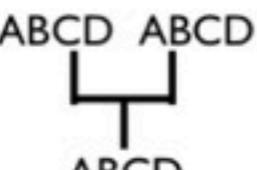
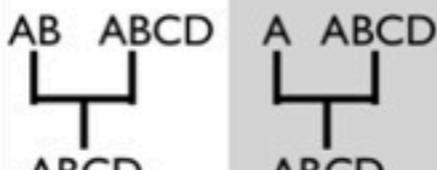
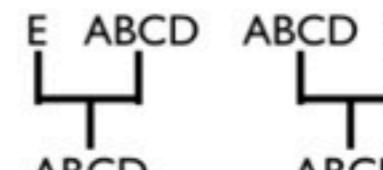
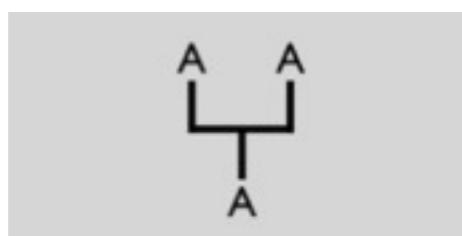
Ancestral ranges:	Sympatric (range copying)	Sympatric (subset)	Vicariance	Founder Event
A		--	--	
AB				
ABC				
ABCD				

Figure 1. Various models for the evolution of geographic range at cladogenesis events. The events allowed by LAGRANGE are highlighted in gray. Each allowed event is fixed to have equal probability in the LAGRANGE algorithm.

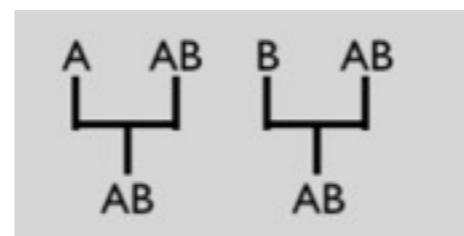
DEC: Dispersal-extinction cladogenesis

Cladogenesis -- range evolution at speciation events (LAGRANGE** model)**

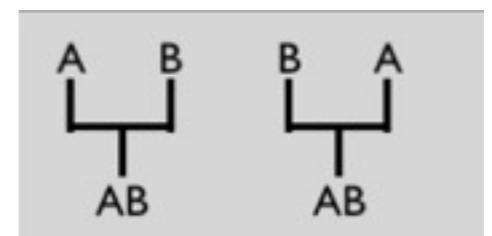
Probabilités per speciation event:



sympatric



subset

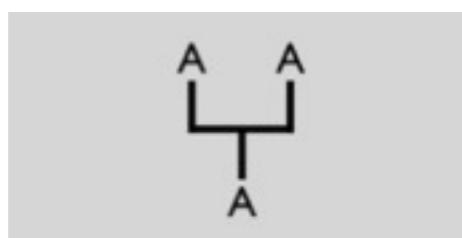


vicariance

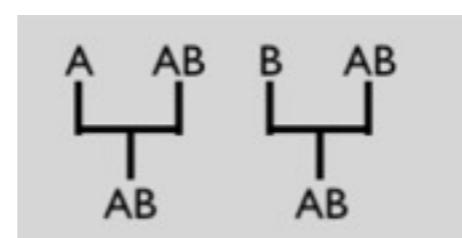
LAGRANGE: all events equiprobable

DECj: Dispersal-extinction cladogenesis PLUS founder-event speciation

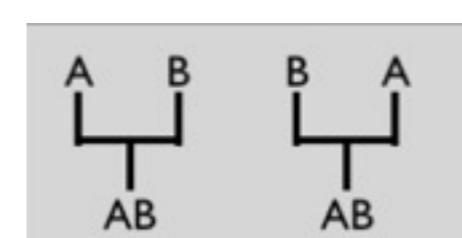
*Probabilités
per speciation
event:*



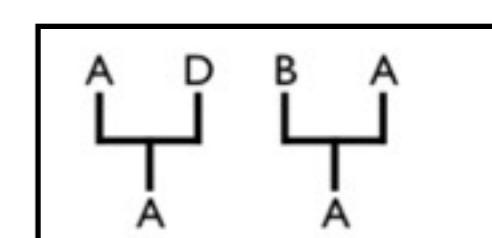
sympatric



subset



vicariance

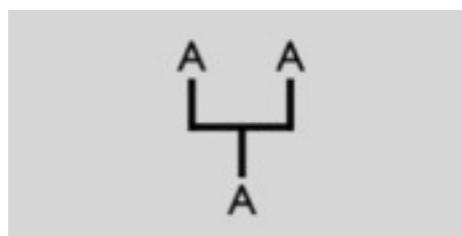


Founder Event

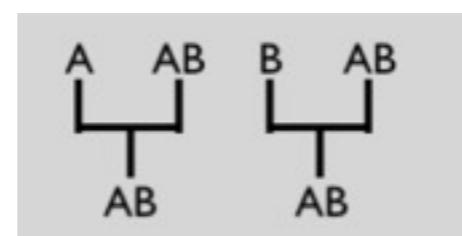
LAGRANGE: all events equiprobable

DECj: Dispersal-extinction cladogenesis PLUS founder-event speciation

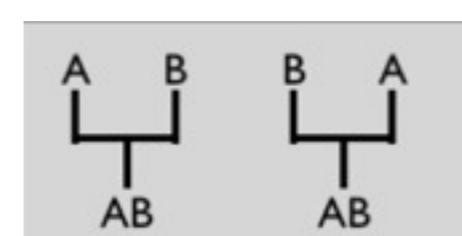
Probabilités
per speciation
event:



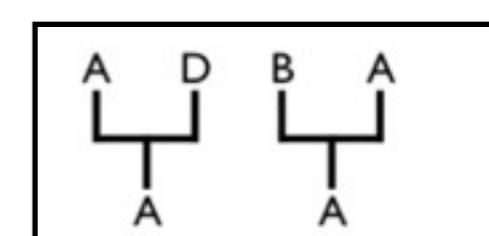
sympatric



subset



vicariance



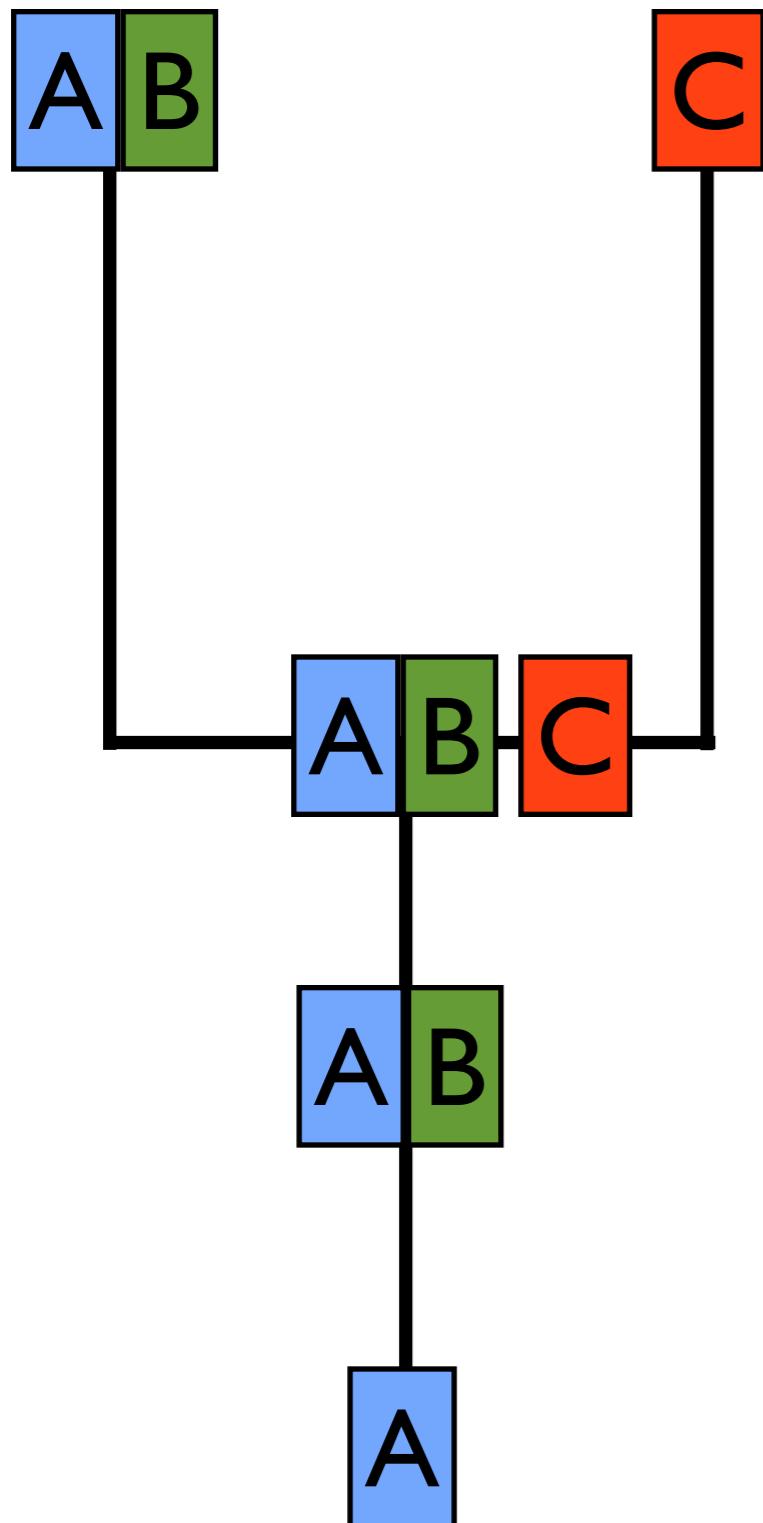
Founder Event

			ending range																																				
left:	A	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C	C	AB	AB	AB	AB	AB	BC	BC	BC	BC	BC	AC	AC	AC	AC	AC	ABC	ABC	ABC	ABC	ABC	ABC	count
right:	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC				
A	sym	j	j									j																											
B	j											j	sym	j																									
C		j											j	j	sym																								
AB				vic	sub																																		
BC						j																																	
AC							vic		sub				j	vic																									
ABC																																							

The probability of different events:
controlled by value of parameter j

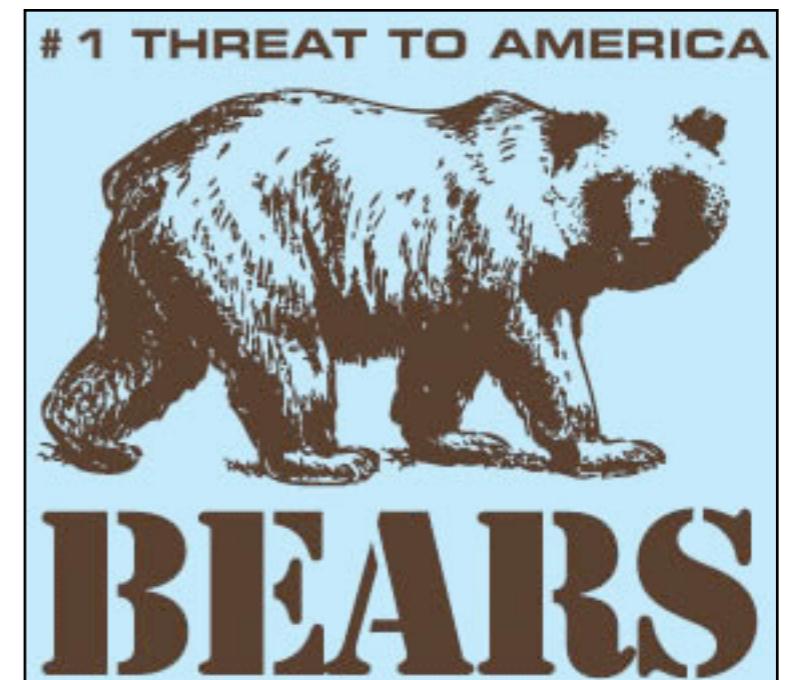
Implementing founder-event speciation

I implemented this method in *three* R packages:



Package #3, **BioGeoBEARS**, does the ML/Bayesian searches, model testing, etc.

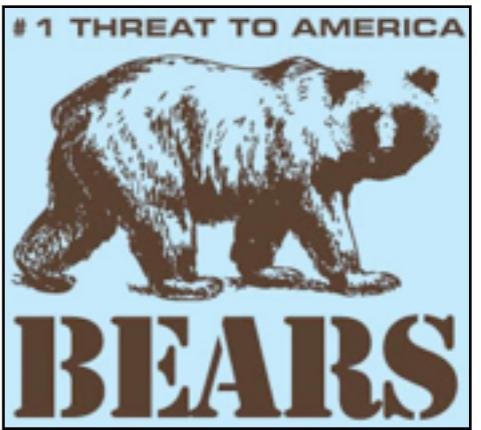
BioGeographic Bayesian
Evolutionary Analysis with
R Scripts



BioGeoBEARS features

As in LAGRANGE, user can load:

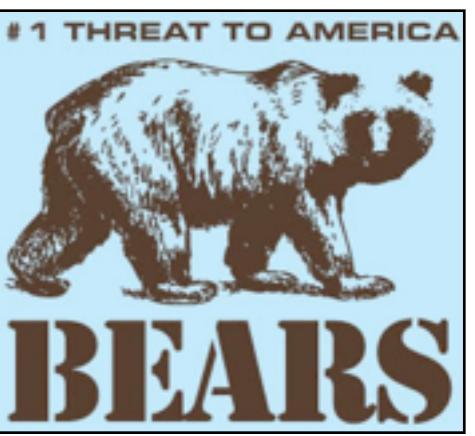
- Time strata
- Dispersal multipliers
- Connectivity matrices
- New: Distance matrix (as in BayArea), if you think
 $\text{Prob}(\text{dispersal/founder event}) \sim \text{distance}$
- New: Area size list (as in SHIBA, Webb & Ree 2011), if
you think $\text{Prob}(\text{extinction event}) \sim \text{area size}$



BioGeoBEARS features

Ancestral state (ancestral range) estimation:

LAGRANGE does marginal ancestral state estimates
(fix each state, then re-optimize)



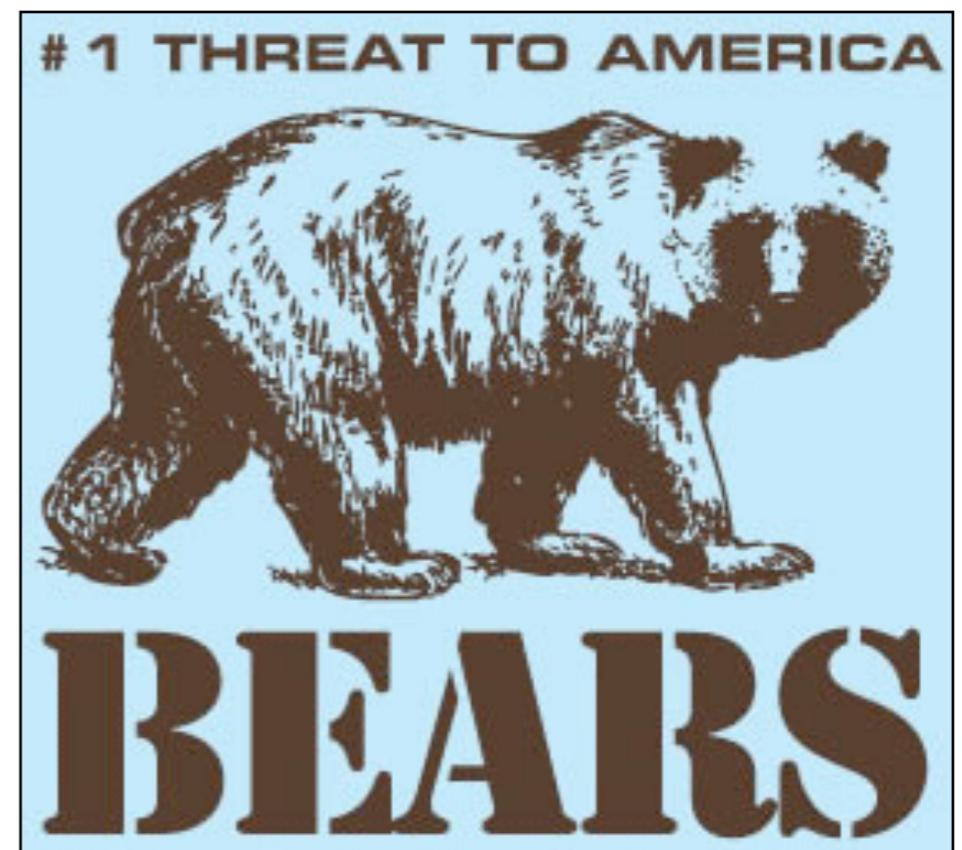
BioGeoBEARS features

Forward simulations under your model

Display & summary functions

Try it here:

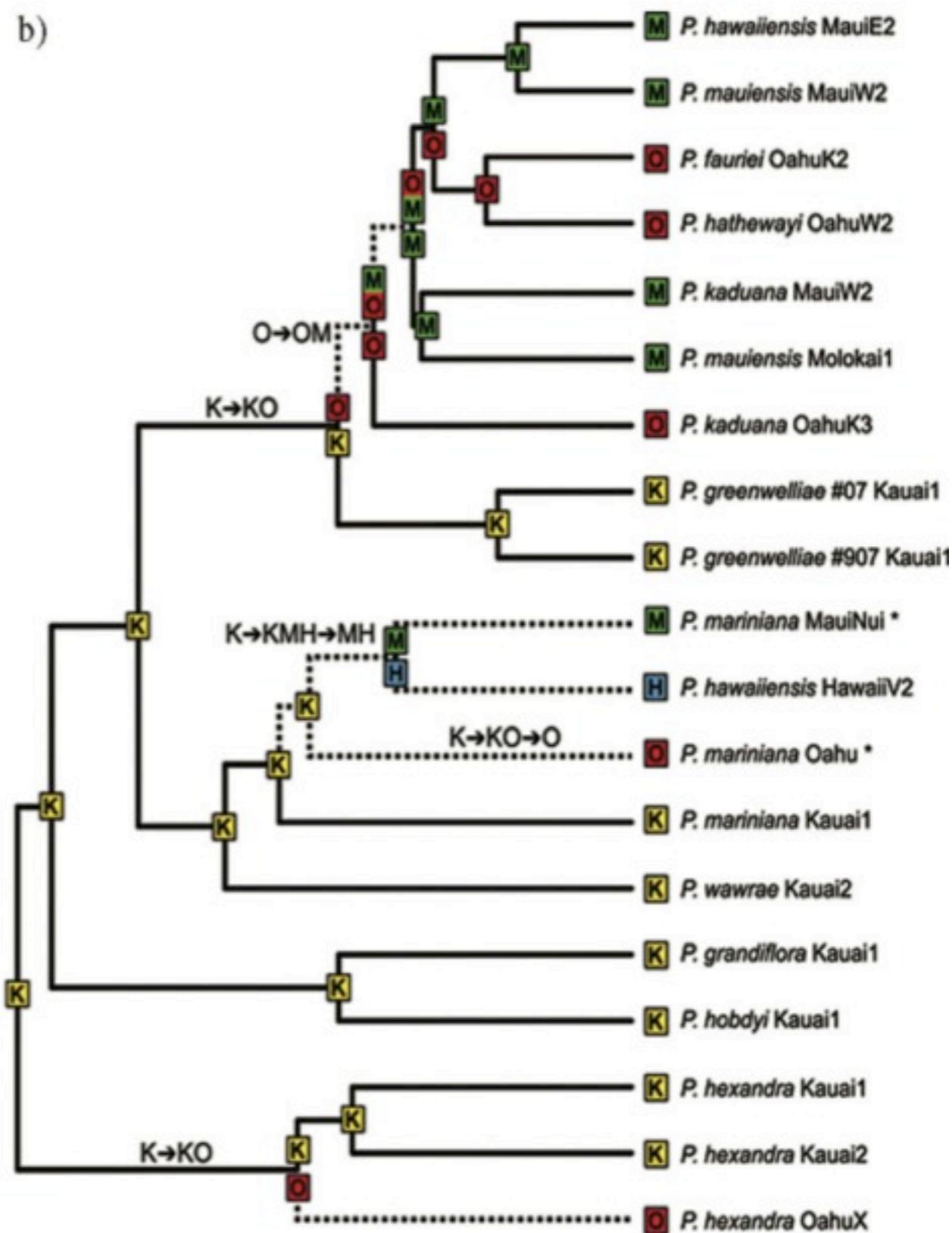
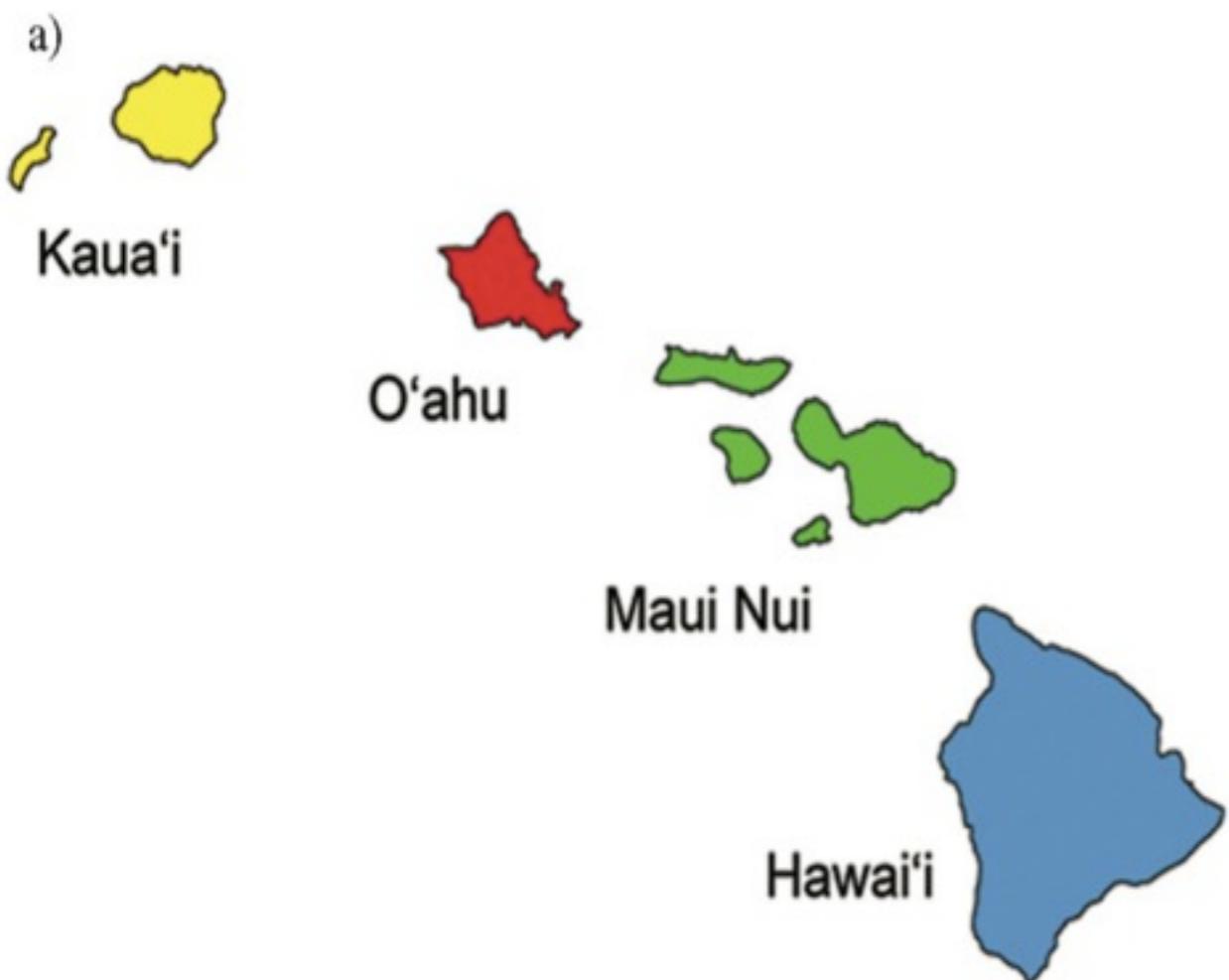
<http://phylo.wikidot.com/biogeobears>



BioGeoBEARS testing

Run BioGeoBEARS on the same example datasets used for LAGRANGE

e.g., Ree & Smith (2008) Hawaiian *Psychotria*



BioGeoBEARS testing

LAGRANGE has 2 free parameters:

d , dispersal rate per region

e , extinction rate per region

BioGeoBEARS with 2 parameters gets exactly the same ML inference and log-likelihoods:

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542

BioGeoBEARS testing

LAGRANGE has 2 free parameters:

d , dispersal rate per region

e , extinction rate per region

BioGeoBEARS with 2 parameters gets exactly the same ML inference and log-likelihoods:

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542
				Lagrange (Python)	0.035053361	0.028310169	-34.54

BioGeoBEARS testing

LAGRANGE has 2 free parameters:

d , dispersal rate per region

e , extinction rate per region

BioGeoBEARS with 2 parameters gets exactly the same ML inference and log-likelihoods:

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542
				Lagrange (Python)	0.035053361	0.028310169	-34.54
				BioGeoBEARS	0.035054699	0.028298009	-34.54196

BioGeoBEARS testing

LAGRANGE has 2 free parameters:

d , dispersal rate per region

e , extinction rate per region

BioGeoBEARS with 2 parameters gets exactly the same ML inference and log-likelihoods:

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542
				Lagrange (Python)	0.035053361	0.028310169	-34.54
				BioGeoBEARS	0.035054699	0.028298009	-34.54196

BioGeoBEARS testing

		Region/ constraint	Source	# areas	Method	d	e	lnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)		4	Lagrange (C++)	0.035051338	0.0282923	-34.542
					Lagrange (Python)	0.035053361	0.028310169	-34.54
					BioGeoBEARS	0.035054699	0.028298009	-34.54196
<i>Psychotria</i>	Hawaii - M1	Smith & Ree (2008)		4	Lagrange (C++)	0.042734756	0.042884793	-35.2111
					Lagrange (Python)	0.042729362	0.042886703	-35.21
					BioGeoBEARS	0.042736465	0.042891221	-35.21093
<i>Psychotria</i>	Hawaii - M2	Smith & Ree (2008)		4	Lagrange (C++)	0.122739586	1.94522E-09	-31.92
					Lagrange (Python)	0.1227261	5.68002E-09	-31.92
					BioGeoBEARS	0.122739586	1.94522E-09	-31.92015
<i>Psychotria</i>	Hawaii - stratified	Smith & Ree (2008)		4	Lagrange (C++)	0.0342543	1.15E-07	-36.6318
					BioGeoBEARS (no bug)	"	"	-36.24472
					Lagrange (Python)	0.03687	7.22E-08	-39.8
					BioGeoBEARS (with bug)	"	"	-39.79887
<i>Lonicera</i>	Northern Hemisphere	Smith & Donoghue (2010)		4	Lagrange (C++)	0.00581186	0.00111119	-124.977
					Lagrange (Python)	0.005812	0.001049	124.976821
					BioGeoBEARS	0.005878539	0.001159741	-124.979
<i>Cyrtandra</i>	Pacific Islands	Clark et al. 2008		7	Lagrange (C++)	0.00110037	2.90E-05	-46.4624
					Lagrange (Python)	0.001251	4.29E-09	-46.425059
					BioGeoBEARS	0.00110037	2.90E-05	-46.46235

BioGeoBEARS replicates LAGRANGE

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542
				Lagrange (Python)	0.035053361	0.028310169	-34.54
				BioGeoBEARS	0.035054699	0.028298009	-34.54196
<i>Psychotria</i>	Hawaii - M1	Smith & Ree (2008)	4	Lagrange (C++)	0.042734756	0.042884793	-35.2111
				Lagrange (Python)	0.042729362	0.042886703	-35.21
				BioGeoBEARS	0.042736465	0.042891221	-35.21093
<i>Psychotria</i>	Hawaii - M2	Smith & Ree (2008)	4	Lagrange (C++)	0.122739586	1.94522E-09	-31.92
				Lagrange (Python)	0.1227261	5.68002E-09	-31.92
				BioGeoBEARS	0.122739586	1.94522E-09	-31.92015
<i>Psychotria</i>	Hawaii - stratified	Smith & Ree (2008)	4	Lagrange (C++)	0.0342543	1.15E-07	-36.6318
				BioGeoBEARS (no bug)	"	"	-36.24472
				Lagrange (Python)	0.03687	7.22E-08	-39.8
				BioGeoBEARS (with bug)	"	"	-39.79887
<i>Lonicera</i>	Northern Hemisphere	Smith & Donoghue (2010)	4	Lagrange (C++)	0.00581186	0.00111119	-124.977
				Lagrange (Python)	0.005812	0.001049	124.976821
				BioGeoBEARS	0.005878539	0.001159741	-124.979
<i>Cyrtandra</i>	Pacific Islands	Clark et al. 2008	7	Lagrange (C++)	0.00110037	2.90E-05	-46.4624
				Lagrange (Python)	0.001251	4.29E-09	-46.425059
				BioGeoBEARS	0.00110037	2.90E-05	-46.46235

BioGeoBEARS replicates LAGRANGE

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542
				Lagrange (Python)	0.035053361	0.028310169	-34.54
				BioGeoBEARS	0.035054699	0.028298009	-34.54196
<i>Psychotria</i>	Hawaii - M1	Smith & Ree (2008)	4	Lagrange (C++)	0.042734756	0.042884793	-35.2111
				Lagrange (Python)	0.042729362	0.042886703	-35.21
				BioGeoBEARS	0.042736465	0.042891221	-35.21093
<i>Psychotria</i>	Hawaii - M2	Smith & Ree (2008)	4	Lagrange (C++)	0.122739586	1.94522E-09	-31.92
				Lagrange (Python)	0.1227261	5.68002E-09	-31.92
				BioGeoBEARS	0.122739586	1.94522E-09	-31.92015
<i>Psychotria</i>	Hawaii - stratified	Smith & Ree (2008)	4	Lagrange (C++)	0.0342543	1.15E-07	-36.6318
				BioGeoBEARS (no bug)	"	"	-36.24472
				Lagrange (Python)	0.03687	7.22E-08	-39.8
				BioGeoBEARS (with bug)	"	"	-39.79887
<i>Lonicera</i>	Northern Hemisphere	Smith & Donoghue (2010)	4	Lagrange (C++)	0.00581186	0.00111119	-124.977
				Lagrange (Python)	0.005812	0.001049	124.976821
				BioGeoBEARS	0.005878539	0.001159741	-124.979
<i>Cyrtandra</i>	Pacific Islands	Clark et al. 2008	7	Lagrange (C++)	0.00110037	2.90E-05	-46.4624
				Lagrange (Python)	0.001251	4.29E-09	-46.425059
				BioGeoBEARS	0.00110037	2.90E-05	-46.46235

BioGeoBEARS replicates LAGRANGE

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542
				Lagrange (Python)	0.035053361	0.028310169	-34.54
				BioGeoBEARS	0.035054699	0.028298009	-34.54196
<i>Psychotria</i>	Hawaii - M1	Smith & Ree (2008)	4	Lagrange (C++)	0.042734756	0.042884793	-35.2111
				Lagrange (Python)	0.042729362	0.042886703	-35.21
				BioGeoBEARS	0.042736465	0.042891221	-35.21093
<i>Psychotria</i>	Hawaii - M2	Smith & Ree (2008)	4	Lagrange (C++)	0.122739586	1.94522E-09	-31.92
				Lagrange (Python)	0.1227261	5.68002E-09	-31.92
				BioGeoBEARS	0.122739586	1.94522E-09	-31.92015
<i>Psychotria</i>	Hawaii - stratified	Smith & Ree (2008)	4	Lagrange (C++)	0.0342543	1.15E-07	-36.6318
				BioGeoBEARS (no bug)	"	"	-36.24472
				Lagrange (Python)	0.03687	7.22E-08	-39.8
				BioGeoBEARS (with bug)	"	"	-39.79887
<i>Lonicera</i>	Northern Hemisphere	Smith & Donoghue (2010)	4	Lagrange (C++)	0.00581186	0.00111119	-124.977
				Lagrange (Python)	0.005812	0.001049	124.976821
				BioGeoBEARS	0.005878539	0.001159741	-124.979
<i>Cyrtandra</i>	Pacific Islands	Clark et al. 2008	7	Lagrange (C++)	0.00110037	2.90E-05	-46.4624
				Lagrange (Python)	0.001251	4.29E-09	-46.425059
				BioGeoBEARS	0.00110037	2.90E-05	-46.46235

BioGeoBEARS replicates LAGRANGE

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542
				Lagrange (Python)	0.035053361	0.028310169	-34.54
				BioGeoBEARS	0.035054699	0.028298009	-34.54196
<i>Psychotria</i>	Hawaii - M1	Smith & Ree (2008)	4	Lagrange (C++)	0.042734756	0.042884793	-35.2111
				Lagrange (Python)	0.042729362	0.042886703	-35.21
				BioGeoBEARS	0.042736465	0.042891221	-35.21093
<i>Psychotria</i>	Hawaii - M2	Smith & Ree (2008)	4	Lagrange (C++)	0.122739586	1.94522E-09	-31.92
				Lagrange (Python)	0.1227261	5.68002E-09	-31.92
				BioGeoBEARS	0.122739586	1.94522E-09	-31.92015
<i>Psychotria</i>	Hawaii - stratified	Smith & Ree (2008)	4	Lagrange (C++)	0.0342543	1.15E-07	-36.6318
				BioGeoBEARS (no bug)	"	"	-36.24472
				Lagrange (Python)	0.03687	7.22E-08	-39.8
				BioGeoBEARS (with bug)	"	"	-39.79887
<i>Lonicera</i>	Northern Hemisphere	Smith & Donoghue (2010)	4	Lagrange (C++)	0.00581186	0.00111119	-124.977
				Lagrange (Python)	0.005812	0.001049	124.976821
				BioGeoBEARS	0.005878539	0.001159741	-124.979
<i>Cyrtandra</i>	Pacific Islands	Clark et al. 2008	7	Lagrange (C++)	0.00110037	2.90E-05	-46.4624
				Lagrange (Python)	0.001251	4.29E-09	-46.425059
				BioGeoBEARS	0.00110037	2.90E-05	-46.46235

BioGeoBEARS replicates LAGRANGE

Clade	Region/ constraint	Source	# areas	Method	d	e	LnL
<i>Psychotria</i>	Hawaii - M0	Smith & Ree (2008)	4	Lagrange (C++)	0.035051338	0.0282923	-34.542
				Lagrange (Python)	0.035053361	0.028310169	-34.54
				BioGeoBEARS	0.035054699	0.028298009	-34.54196
<i>Psychotria</i>	Hawaii - M1	Smith & Ree (2008)	4	Lagrange (C++)	0.042734756	0.042884793	-35.2111
				Lagrange (Python)	0.042729362	0.042886703	-35.21
				BioGeoBEARS	0.042736465	0.042891221	-35.21093
<i>Psychotria</i>	Hawaii - M2	Smith & Ree (2008)	4	Lagrange (C++)	0.122739586	1.94522E-09	-31.92
				Lagrange (Python)	0.1227261	5.68002E-09	-31.92
				BioGeoBEARS	0.122739586	1.94522E-09	-31.92015
<i>Psychotria</i>	Hawaii - stratified	Smith & Ree (2008)	4	Lagrange (C++)	0.0342543	1.15E-07	-36.6318
				BioGeoBEARS (no bug)	"	"	-36.24472
				Lagrange (Python)	0.03687	7.22E-08	-39.8
				BioGeoBEARS (with bug)	"	"	-39.79887
<i>Lonicera</i>	Northern Hemisphere	Smith & Donoghue (2010)	4	Lagrange (C++)	0.00581186	0.00111119	-124.977
				Lagrange (Python)	0.005812	0.001049	124.976821
				BioGeoBEARS	0.005878539	0.001159741	-124.979
<i>Cyrtandra</i>	Pacific Islands	Clark et al. 2008	7	Lagrange (C++)	0.00110037	2.90E-05	-46.4624
				Lagrange (Python)	0.001251	4.29E-09	-46.425059
				BioGeoBEARS	0.00110037	2.90E-05	-46.46235

Model testing on a cladogenesis model including founder-event speciation

LAGRANGE has 2 free parameters:

- d , dispersal rate per region
- e , extinction rate per region

BioGeoBEARS with 3 free parameters:

- d , dispersal rate per region
- e , extinction rate per region
- j , weighting factor for jump dispersal (“founder-event speciation”) vs. standard LAGRANGE cladogenesis

(as j increases, the probability of the standard LAGRANGE cladogenesis events goes down; the weight of each event depends on how many “allowed” events there are)

Comparing the two models

BioGeoBEARS with 2 parameters (replicates LAGRANGE's results)

$d = 0.03505$ $e = 0.02829$ (j fixed at 0) $\text{LnL} = -34.542$

Comparing the two models

BioGeoBEARS with 2 parameters (replicates LAGRANGE's results)

$d = 0.03505$ $e = 0.02829$ (j fixed at 0) $\text{LnL} = -34.542$

BioGeoBEARS with 3 parameters:

$d = 0.000\dots$ $e = 1.00e-30$ $j = 0.10566$ $\text{LnL} = -20.92879$

Comparing the two models

BioGeoBEARS with 2 parameters (replicates LAGRANGE's results)

$d = 0.03505$ $e = 0.02829$ (j fixed at 0) $\text{LnL} = -34.542$

BioGeoBEARS with 3 parameters:

$d = 0.000\dots$ $e = 1.00e-30$ $j = 0.10566$ $\text{LnL} = -20.92879$

$2 * \Delta \text{LnL} = 13.61 * 2 = 27.22$

Likelihood ratio test p-value = $9.4e-08$

Comparing the two models

BioGeoBEARS with 2 parameters (replicates LAGRANGE's results)

$d = 0.03505$ $e = 0.02829$ (j fixed at 0)

$\text{LnL} = -34.542$

BioGeoBEARS with 3 parameters:

$d = 0.000...$ $e = 1.00e-30$ $j = 0.10566$

$\text{LnL} = -20.92879$

$2 * \Delta \text{LnL} = 13.61 * 2 = 27.22$

Likelihood ratio test p-value = $9.4e-08$

AIC, 2 parameter model = 73.084

AIC, 3 parameter model = 47.858

Relative probability, 3 param = 0.99999667

Results

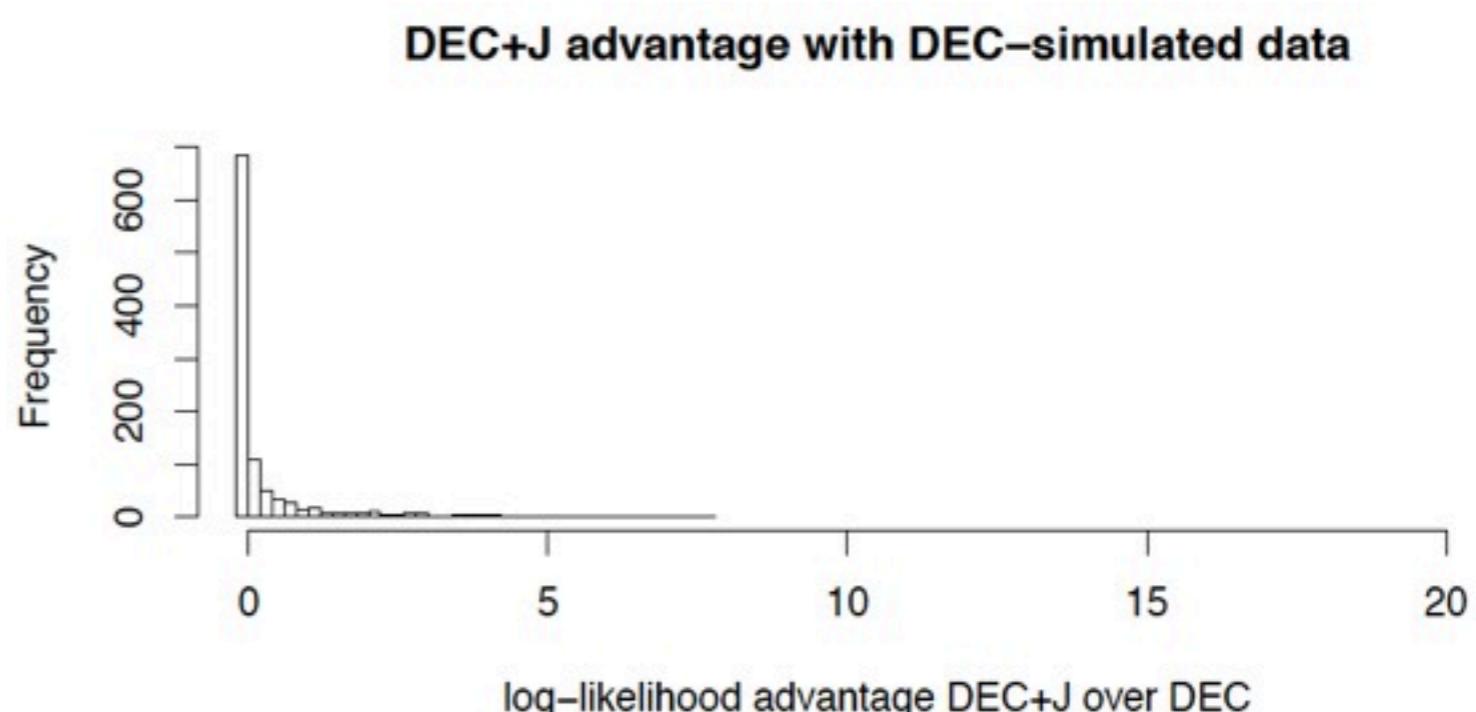
This is a fairly impressive improvement in likelihood for a model with a single additional parameter.

How about other datasets / analyses?

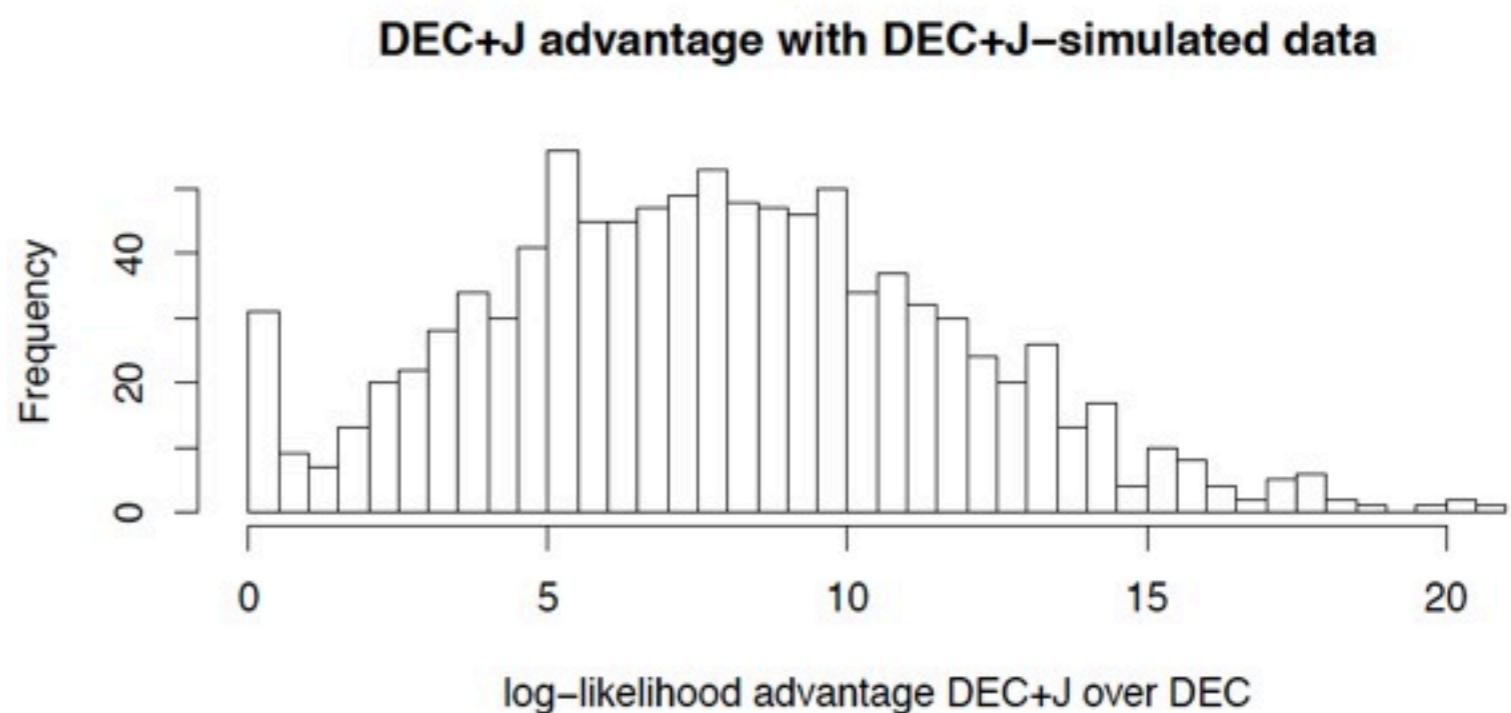
Reverse situation: the 3-parameter model is the truth.

$$d=0, e=0, j=0.105$$

Truth = DEC:



Truth = DEC+J



False positives and false negatives

Table 5. False positive rates (A) and false negative rates (B) for the Likelihood Ratio Test applied to the DEC and DEC+J models, on 2000 biogeographical histories simulated on the *Psychotria* phylogeny using the parameters inferred under DEC and DEC+J ML estimation.

A.

True model	Inference model	# sims	mean LnL	mean LnL difference	# false positives in LRT	False Positive Rate
DEC	DEC	1000	-22.055			
DEC	DEC+J		-21.805	0.247	43	0.043

B.

True model	Inference model	# sims	mean LnL	mean LnL difference	# false negatives in LRT	False Negative Rate
DEC+J	DEC	1000	-24.401			
DEC+J	DEC+J		-16.542	7.86	59	0.059

False positives and false negatives

Table 5. False positive rates (A) and false negative rates (B) for the Likelihood Ratio Test applied to the DEC and DEC+J models, on 2000 biogeographical histories simulated on the *Psychotria* phylogeny using the parameters inferred under DEC and DEC+J ML estimation.

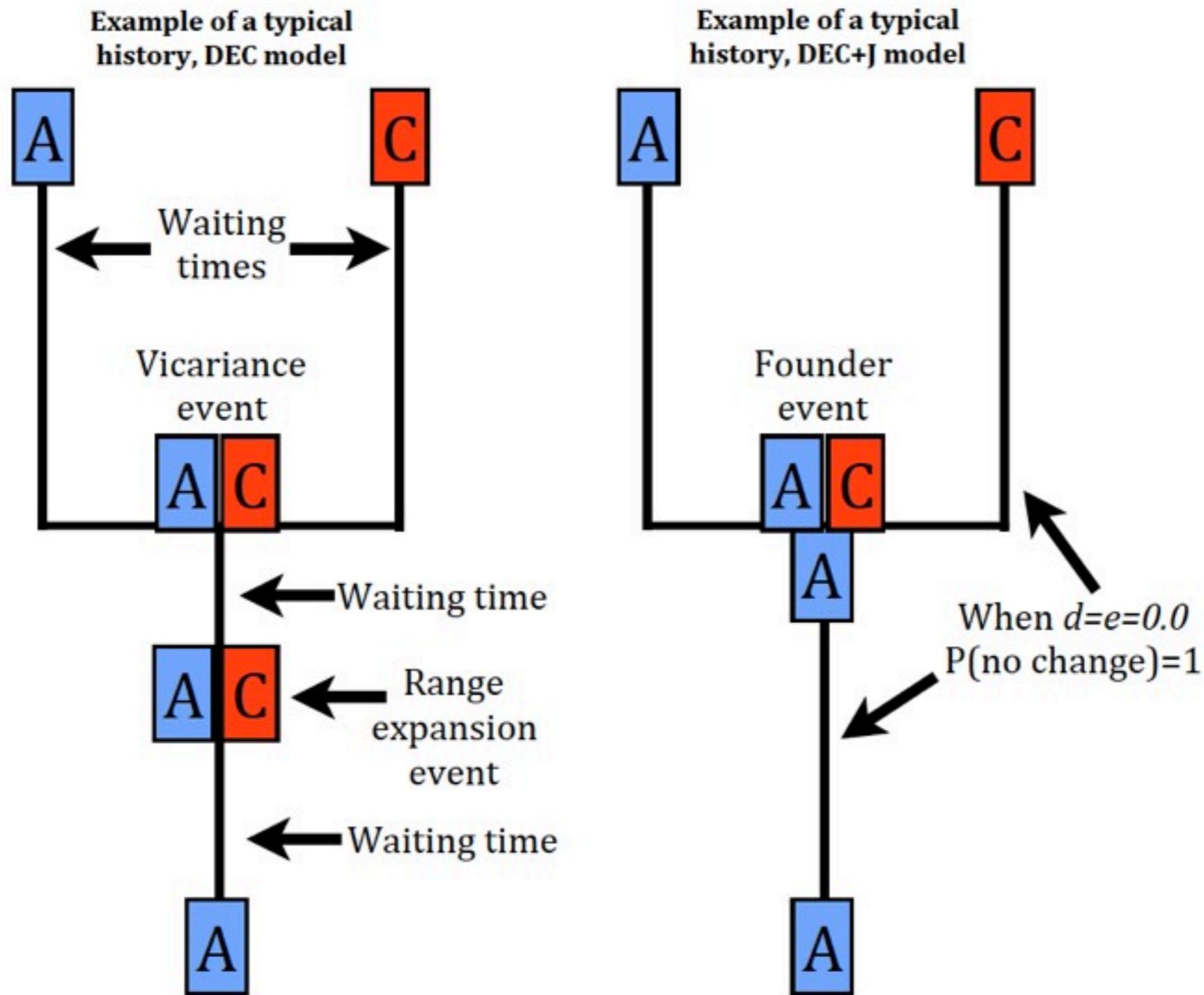
A.

True model	Inference model	# sims	mean LnL	mean LnL difference	# false positives in LRT	False Positive Rate
DEC	DEC	1000	-22.055			
DEC	DEC+J		-21.805	0.247	43	0.043

B.

True model	Inference model	# sims	mean LnL	mean LnL difference	# false negatives in LRT	False Negative Rate
DEC+J	DEC	1000	-24.401			
DEC+J	DEC+J		-16.542	7.86	59	0.059

Why founder-event speciation improves data likelihoods

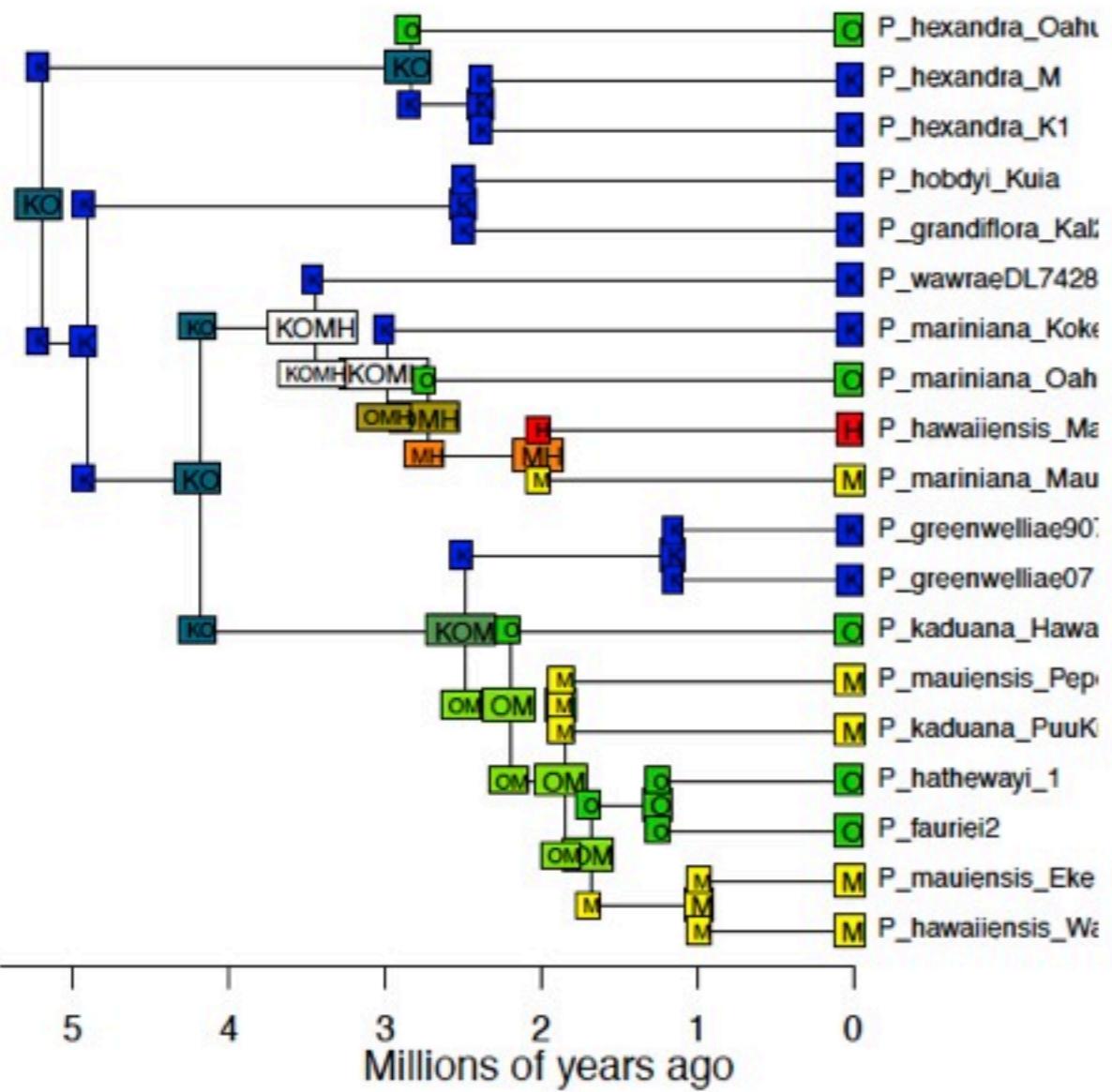


Does model choice affect
inference of ancestral geographic
ranges?

YES

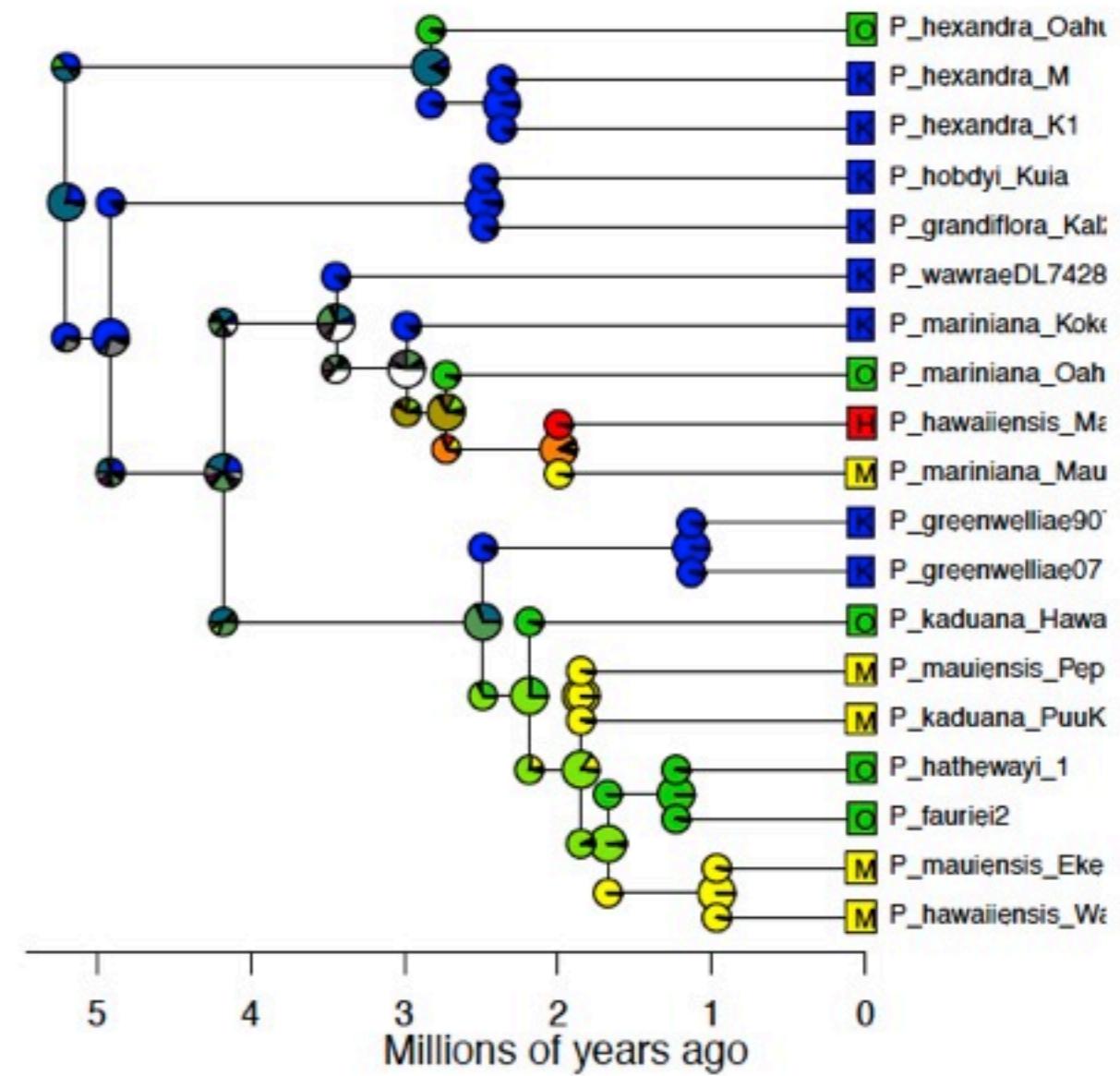
BioGeoBEARS DEC on Psychotria M0

anstates: global optim, 4 areas max. d=0.035; e=0.028; j=0; LnL=-34.5

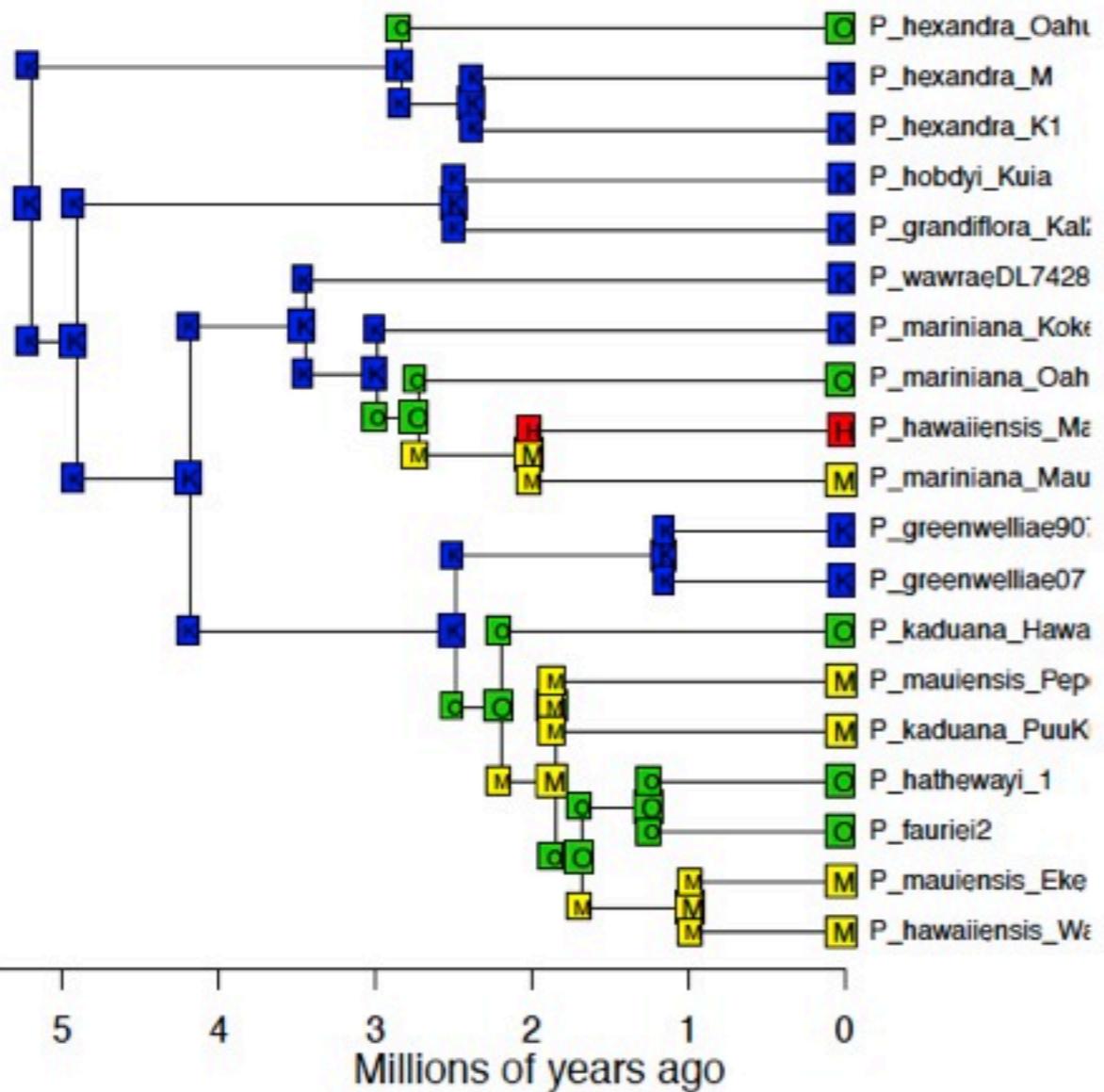


BioGeoBEARS DEC on Psychotria M0

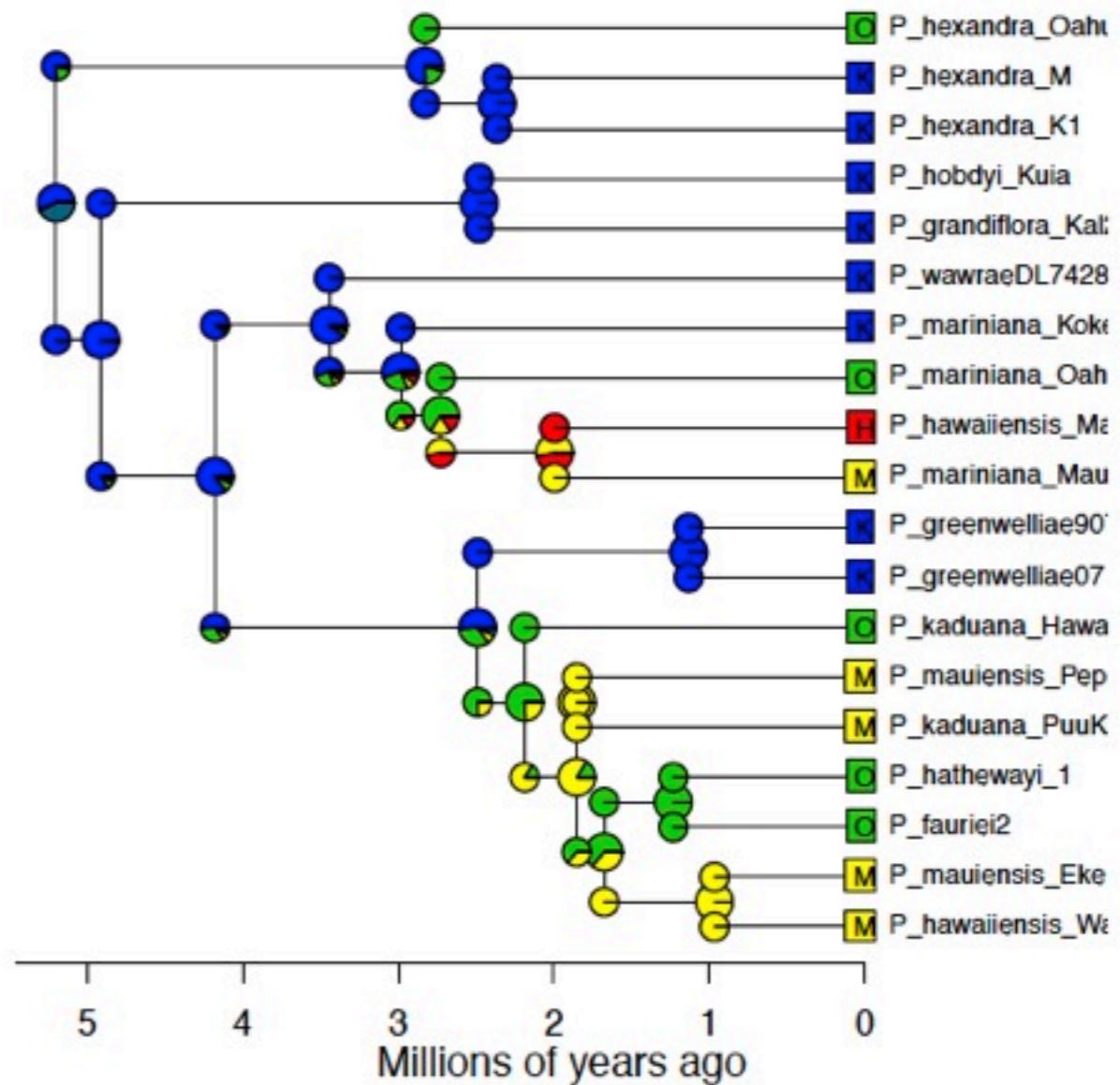
anstates: global optim, 4 areas max. d=0.035; e=0.028; j=0; LnL=-34.5



BioGeoBEARS DEC+J on Psychotria M0
anstates: global optim, 4 areas max. d=0; e=0; j=0.114; LnL=-20.9



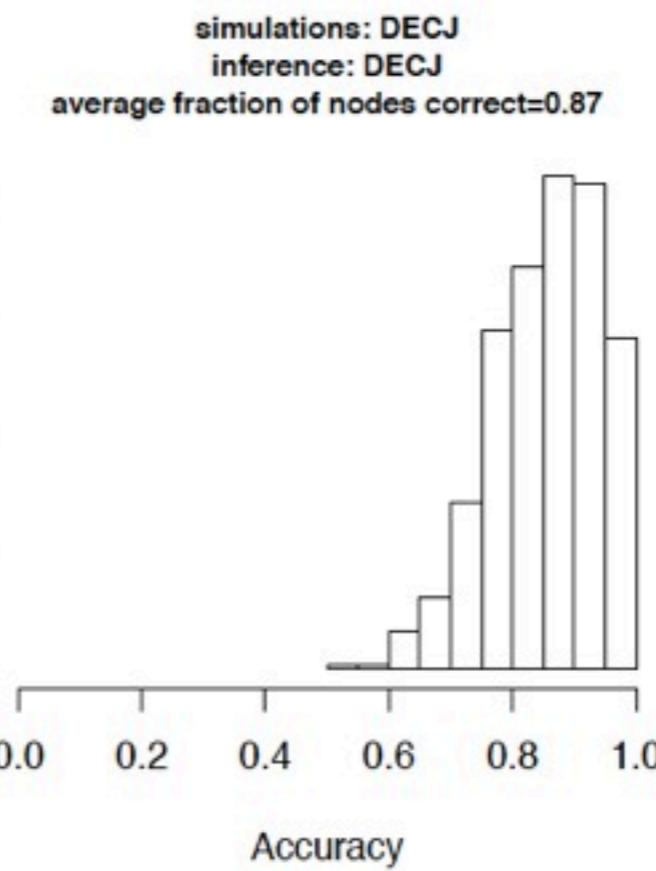
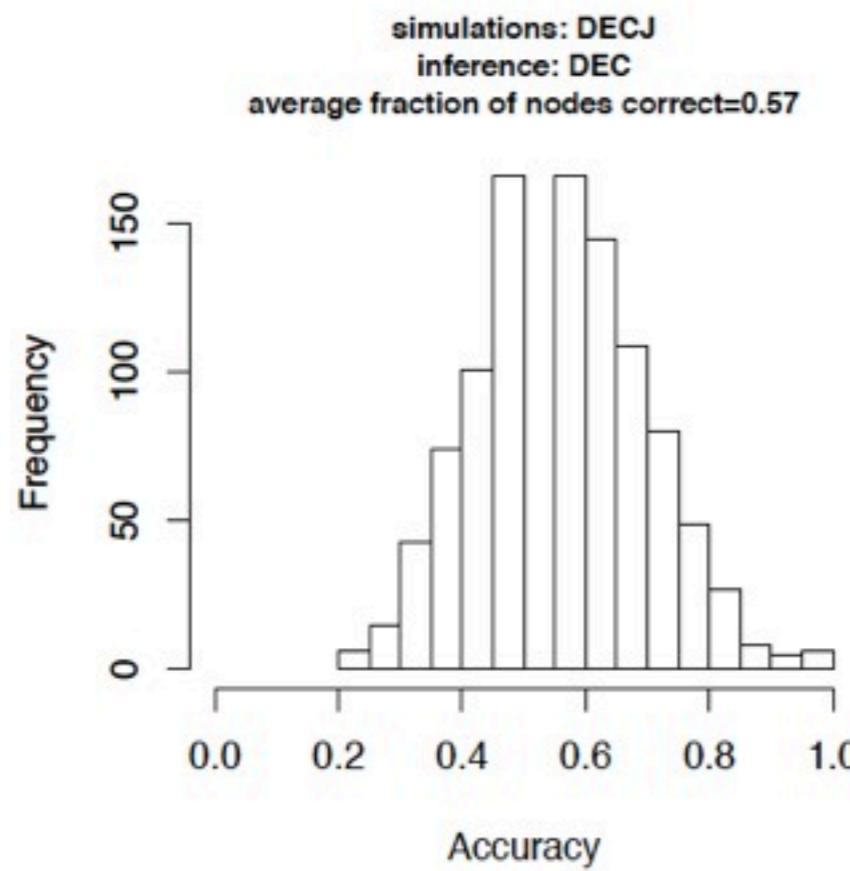
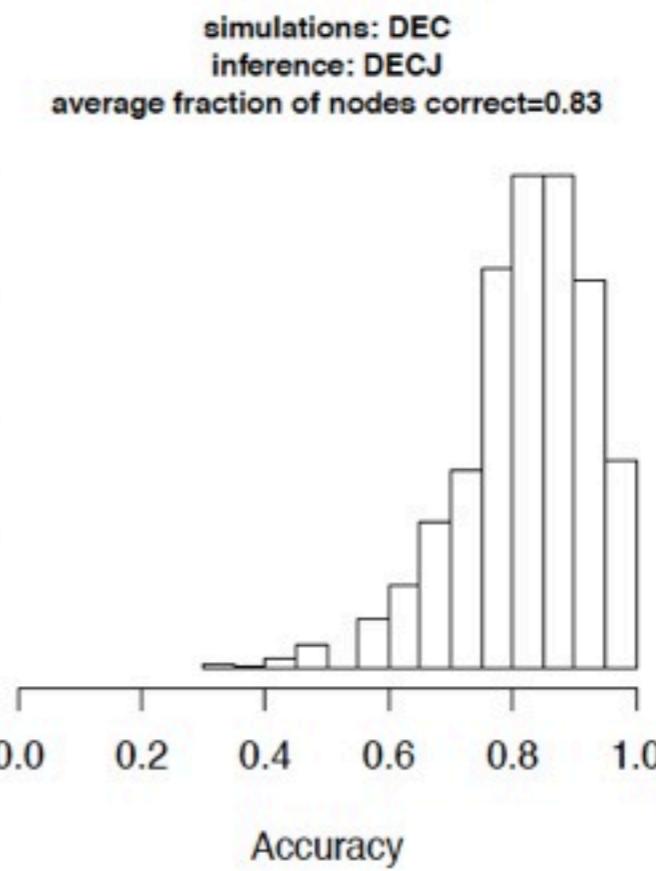
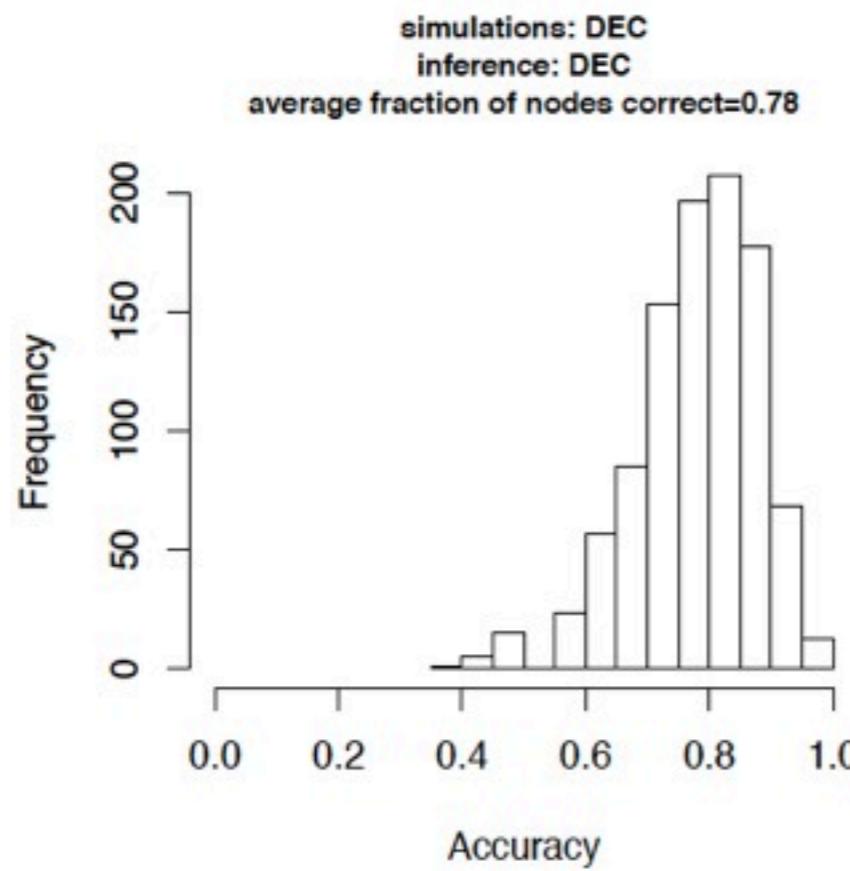
BioGeoBEARS DEC+J on Psychotria M0
anstates: global optim, 4 areas max. d=0; e=0; j=0.114; LnL=-20.9



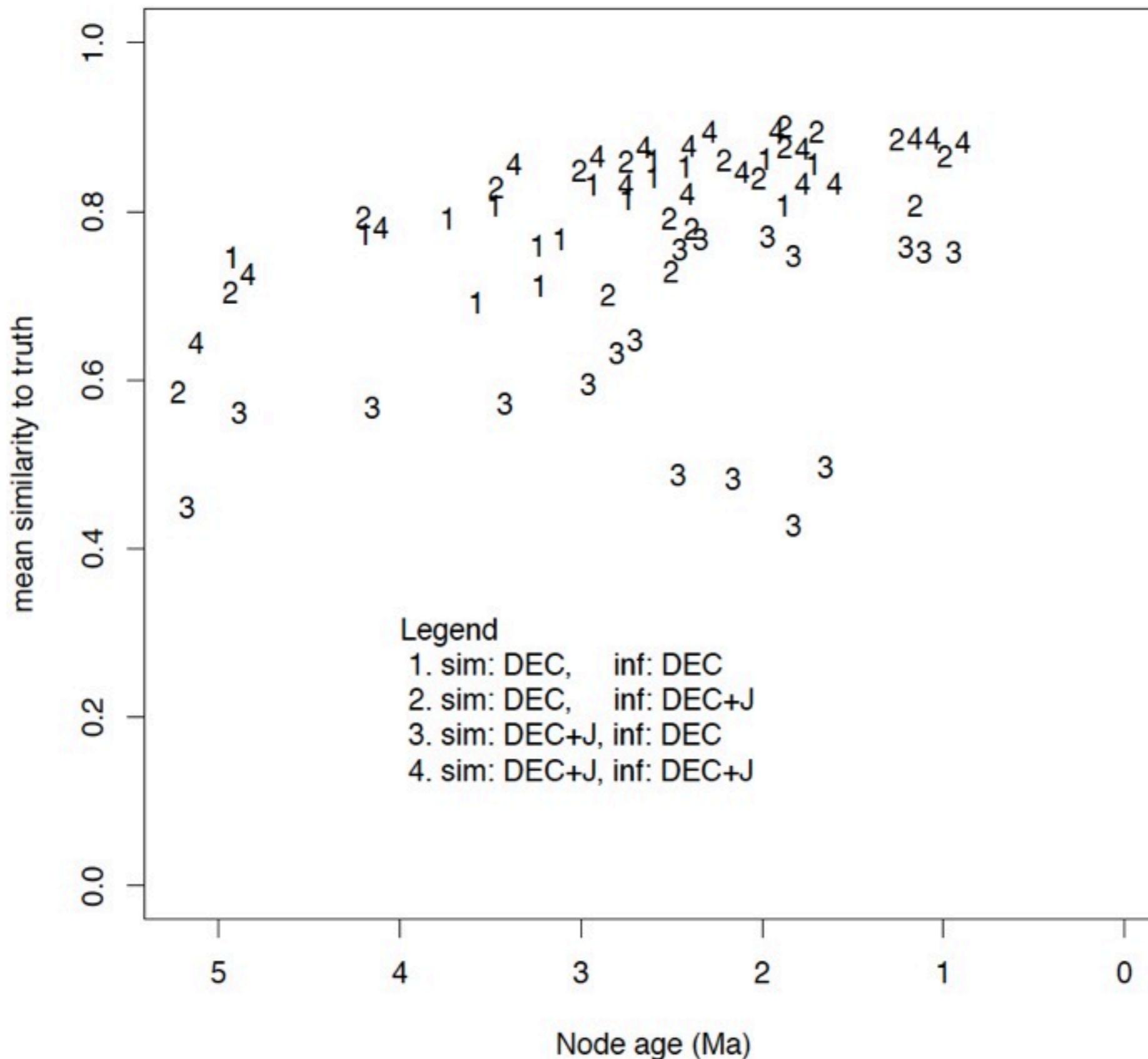
Does model choice affect
accuracy of ancestral geographic
ranges?

YES

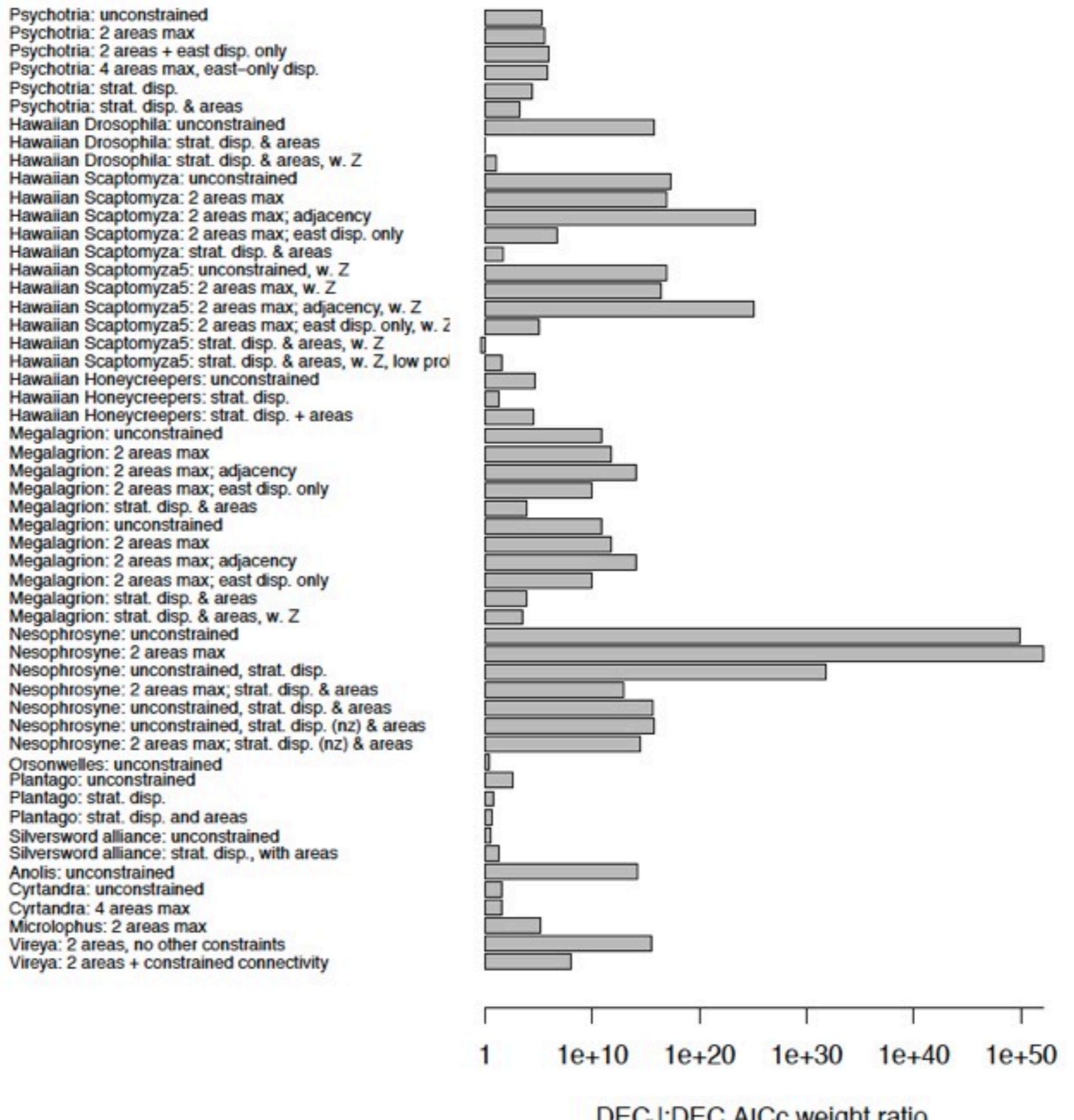
Accuracy of ancestral range inference

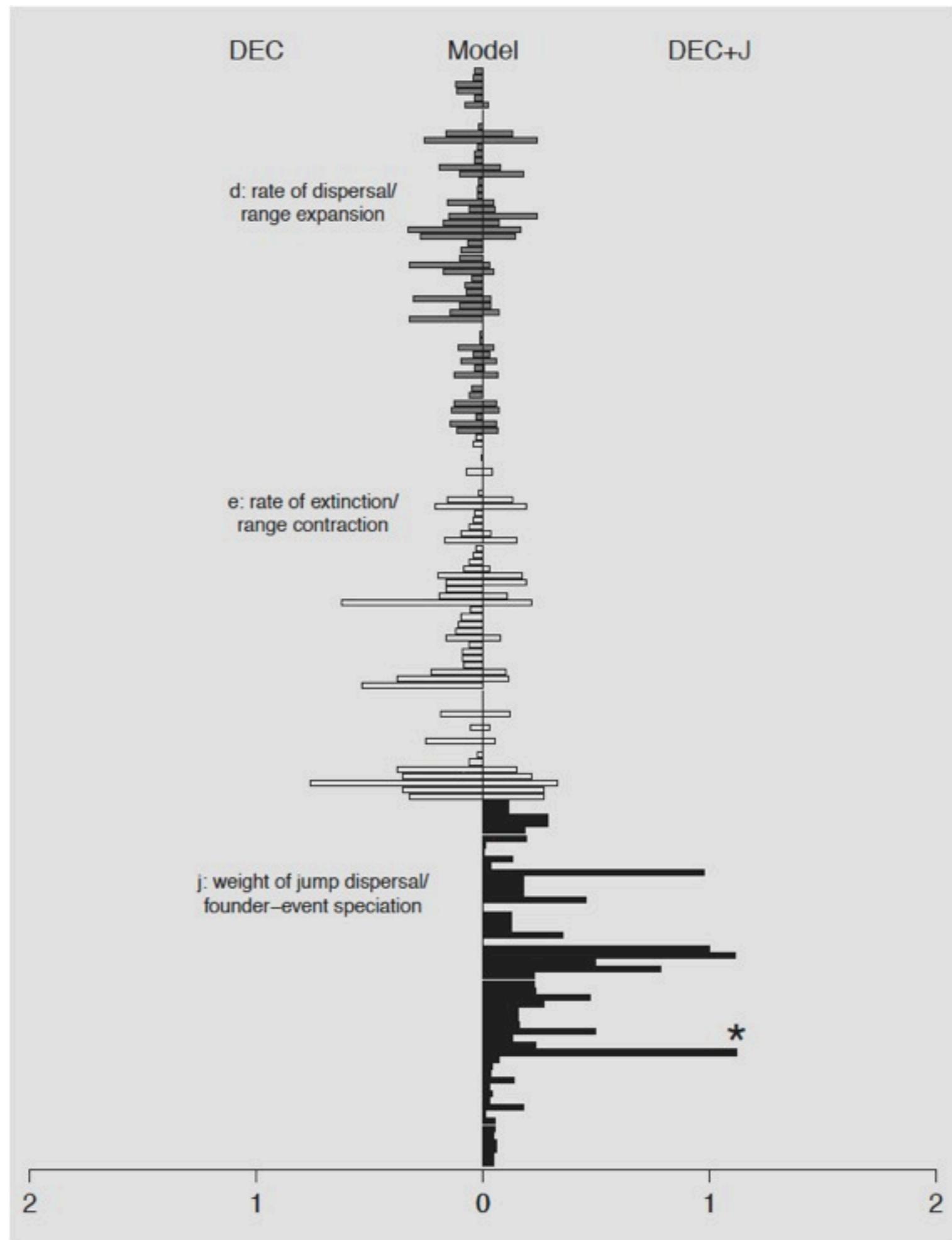


Accuracy of ancestral range inference

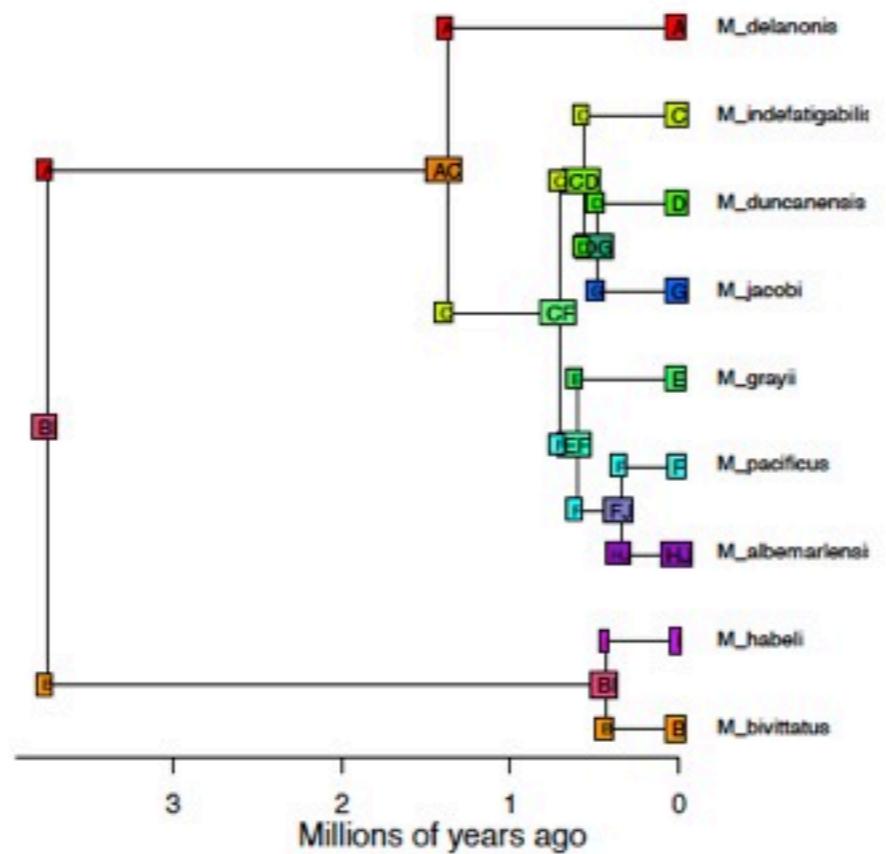


Test on lots of island clades

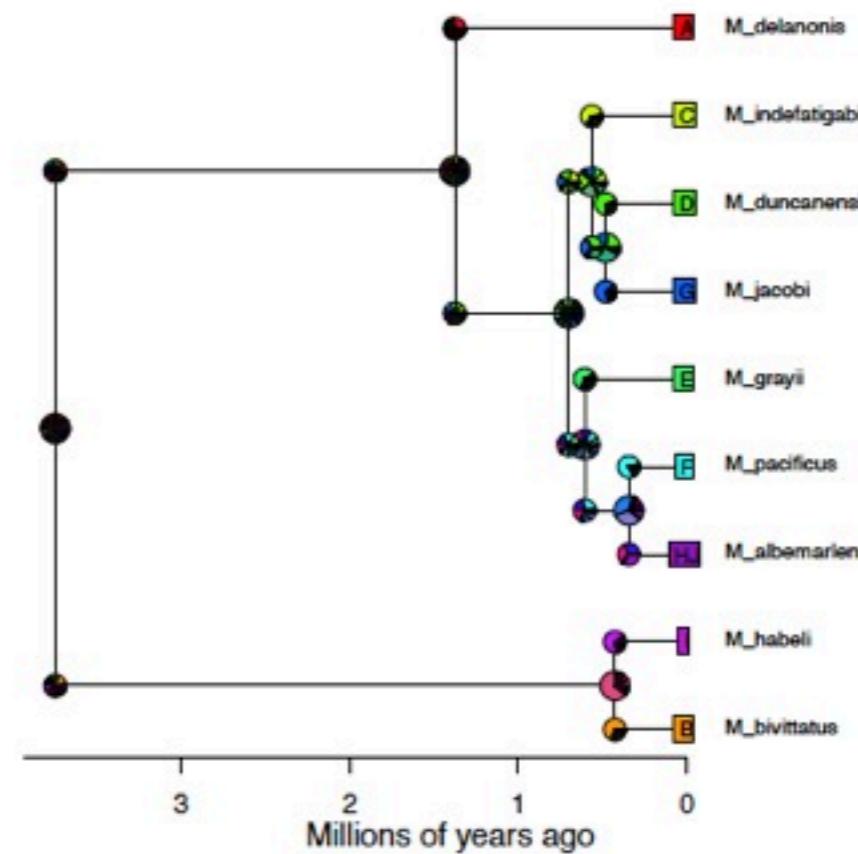




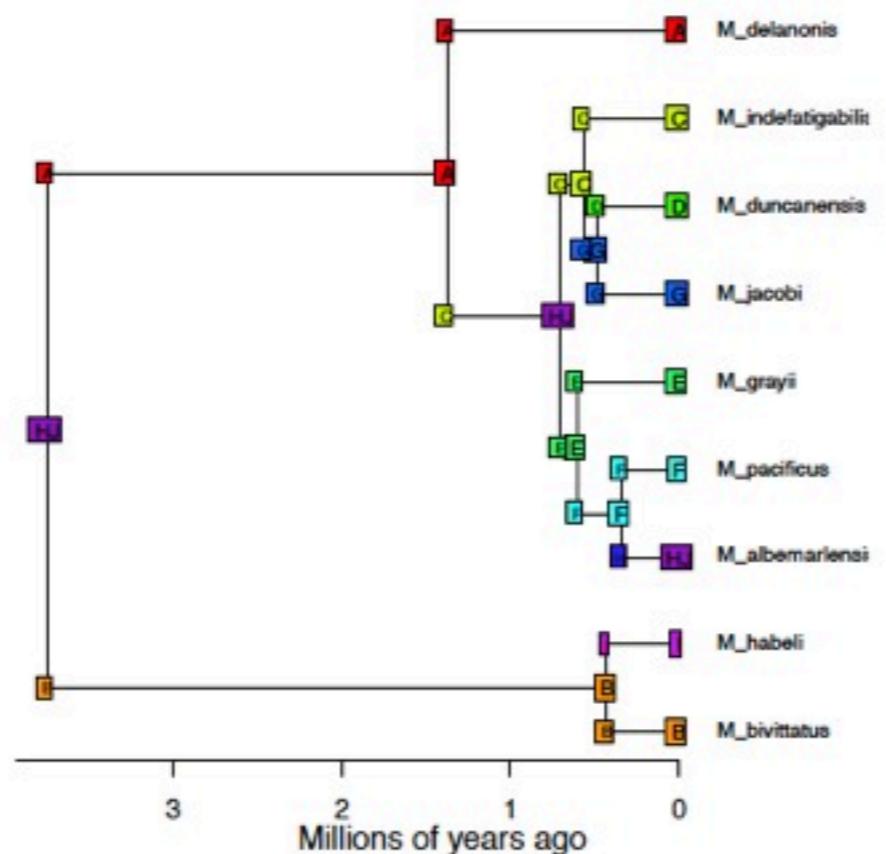
BioGeoBEARS DEC on Microlophus M1_2areas
anstates: global optim, 2 areas max. d=0.324; e=0.533; j=0; LnL=-34.8



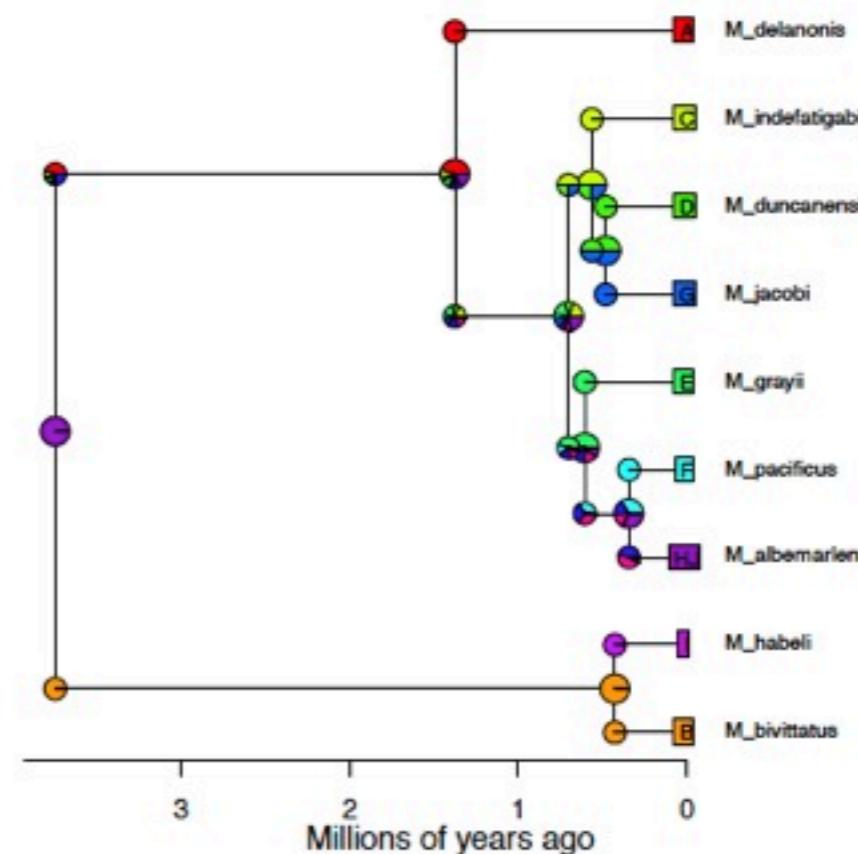
BioGeoBEARS DEC on Microlophus M1_2areas
anstates: global optim, 2 areas max. d=0.324; e=0.533; j=0; LnL=-34.8

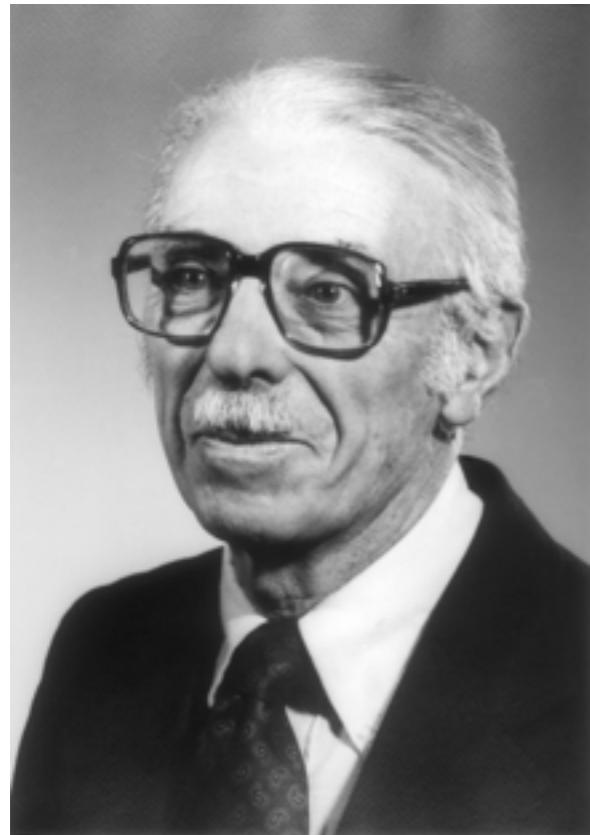


BioGeoBEARS DEC+J on Microlophus M1_2areas
anstates: global optim, 2 areas max. d=0; e=0; j=3; LnL=-20.5



BioGeoBEARS DEC+J on Microlophus M1_2areas
anstates: global optim, 2 areas max. d=0; e=0; j=3; LnL=-20.5





Thomas H. Jukes (1906-1999)

Co-inventor of Jukes-Cantor (JC) model of DNA substitution

	A	C	G	T
A	-	a	a	a
C	a	-	a	a
G	a	a	-	a
T	a	a	a	-

The Jukes-Cantor (JC) was simple, but it led to:

D.Posada and K.A.Crandall

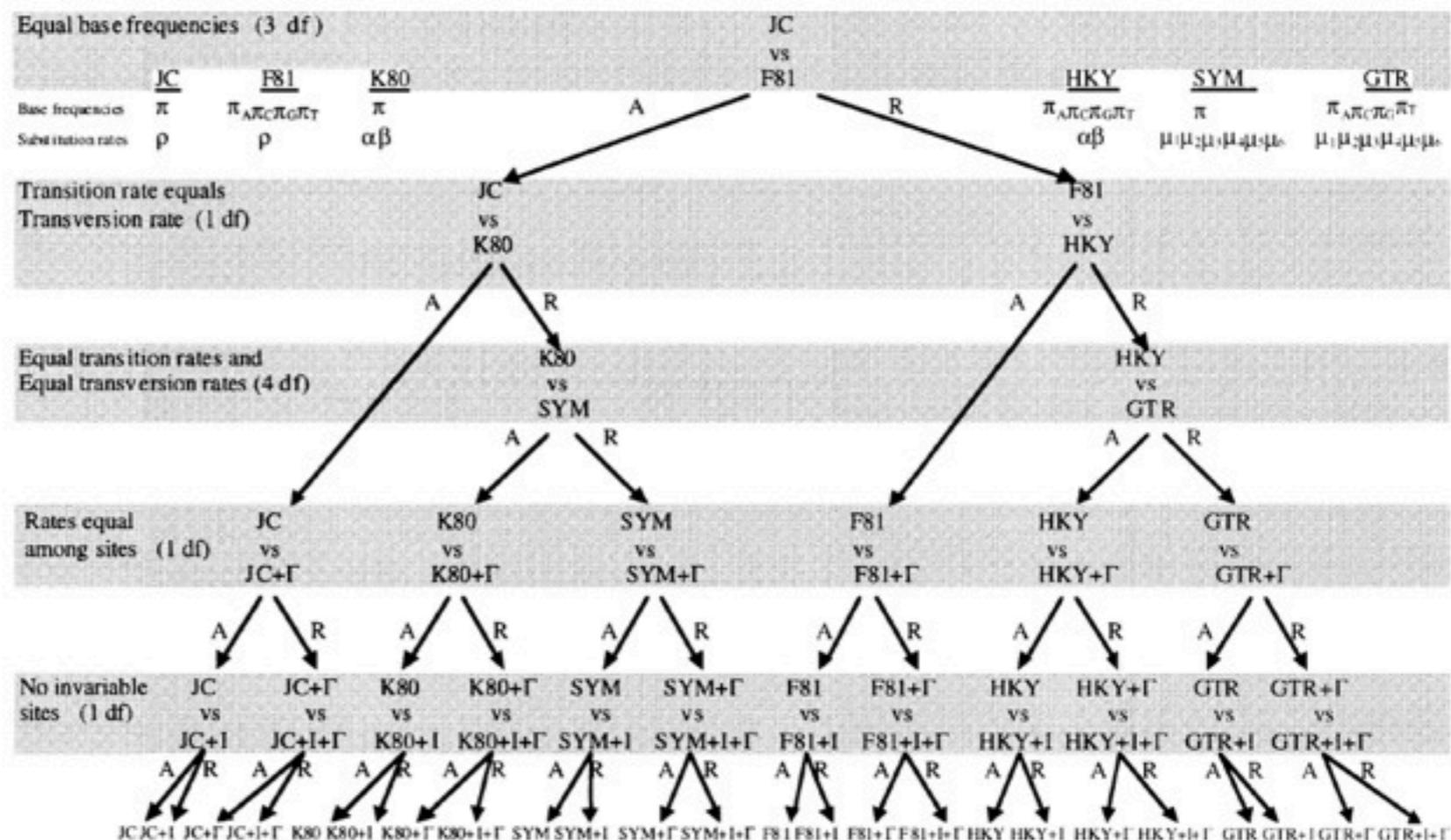
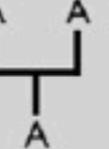
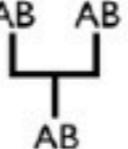
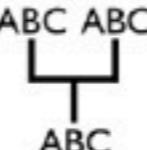
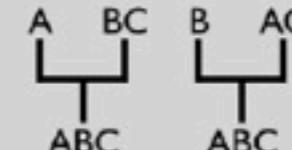
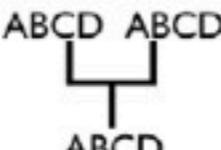
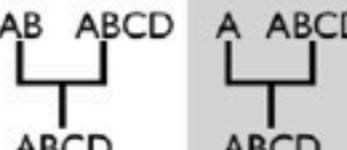
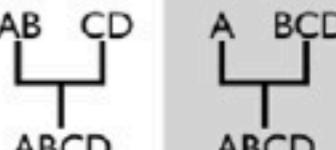


Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al.*, 1985), and GTR (Rodríguez *et al.*, 1990). Γ : shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. 1: equal base frequencies (0.25), π_A : frequency of adenine, π_C : frequency of cytosine, π_G : frequency of guanine, π_T : frequency of thymine. ρ : equal substitution rate, α : transition rate, β : transversion rate; μ_1 : $A \Rightarrow C$ rate, μ_2 : $A \Rightarrow G$ rate, μ_3 : $A \Rightarrow T$ rate, μ_4 : $C \Rightarrow G$ rate, μ_5 : $C \Rightarrow T$ rate, μ_6 : $G \Rightarrow T$ rate.

JC is just a special case of the GTR+I+gamma model

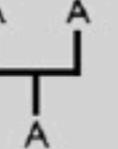
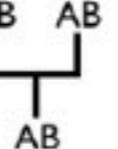
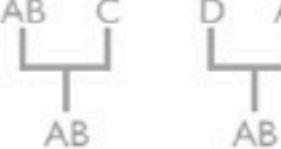
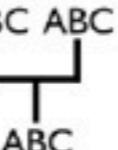
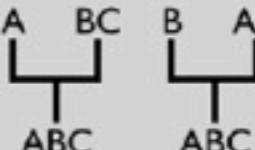
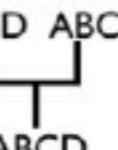
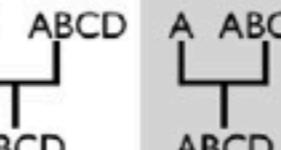
More cladogenesis models!

DEC cladogenesis model

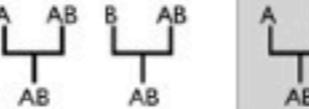
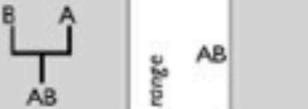
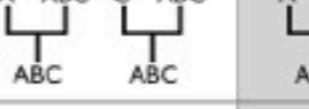
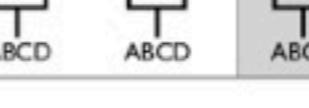
	Sympatric (range copying)	Sympatric (subset)	Vicariance	Founder Event
A		--	--	
AB				
ABC				
ABCD				

More cladogenesis models!

DEC cladogenesis model

	Sympatric (range copying)	Sympatric (subset)	Vicariance	Founder Event
Ancestral range	A A 	--	--	A D 
	AB AB 	A AB B AB 	A B B A 	AB C D AB 
	ABC ABC 	A ABC C ABC 	A BC B AC 	D ABC E ABC 
	ABCD ABCD 	AB ABCD A ABCD 	AB CD A BCD 	E ABCD ABCD E 

DIVA cladogenesis model

	Sympatric (range copying)	Sympatric (subset)	Vicariance
Ancestral range	A A 	--	--
	AB AB 	A AB B AB 	A B B A 
	ABC ABC 	A ABC C ABC 	A BC B AC 
	ABCD ABCD 	AB ABCD A ABCD 	AB CD A BCD 

BAYAREA cladogenesis model

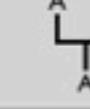
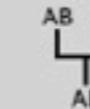
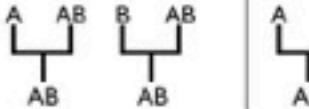
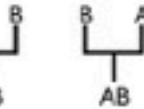
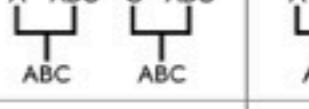
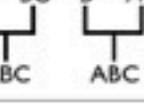
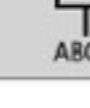
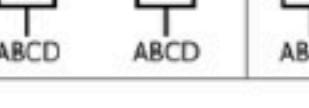
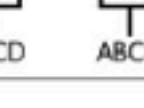
	Sympatric (range copying)	Sympatric (subset)	Vicariance
Ancestral range	A A 	--	--
	AB AB 	A AB B AB 	A B B A 
	ABC ABC 	A ABC C ABC 	A BC B AC 
	ABCD ABCD 	AB ABCD A ABCD 	AB CD A BCD 

Figure 1. The types of cladogenesis events allowed (gray) and disallowed (white) under the DEC, DIVA, and BAYAREA cladogenesis models. The “+J” version of each model adds founder-event speciation, as illustrated for DEC+J.

DEC cladogenesis model

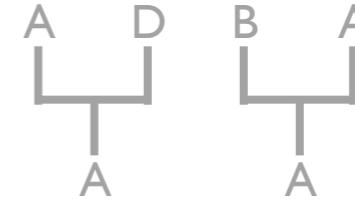
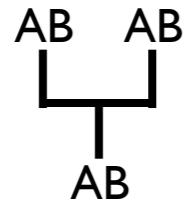
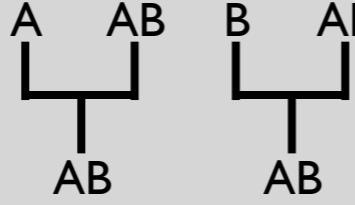
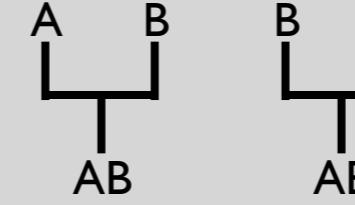
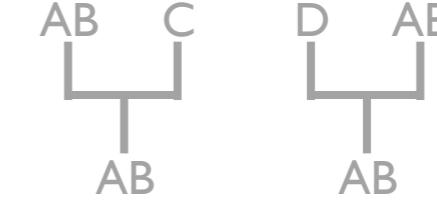
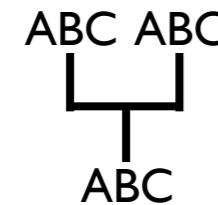
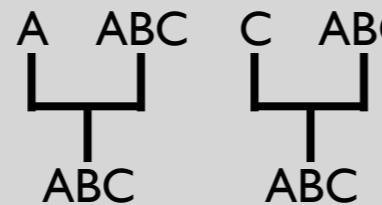
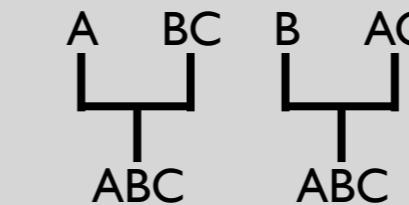
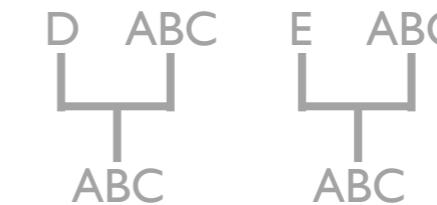
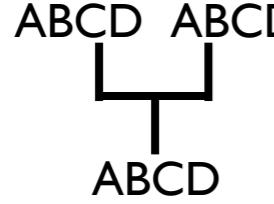
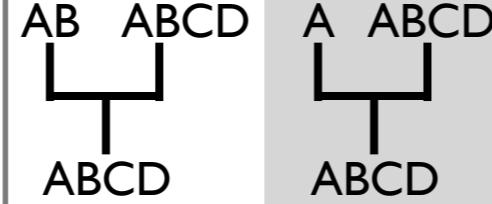
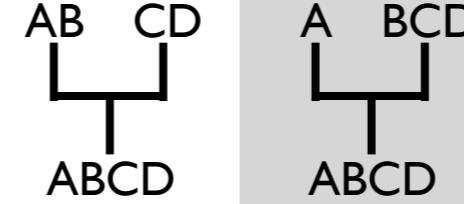
	Sympatric (range copying)	Sympatric (subset)	Vicariance	Founder Event
A		--	--	
AB				
ABC				
ABCD				

Figure 1A. The types of cladogenesis events allowed (gray) and disallowed (white) under the cladogenesis model of LAGRANGE DEC. The DEC+J model adds founder-event speciation.

DIVA cladogenesis model

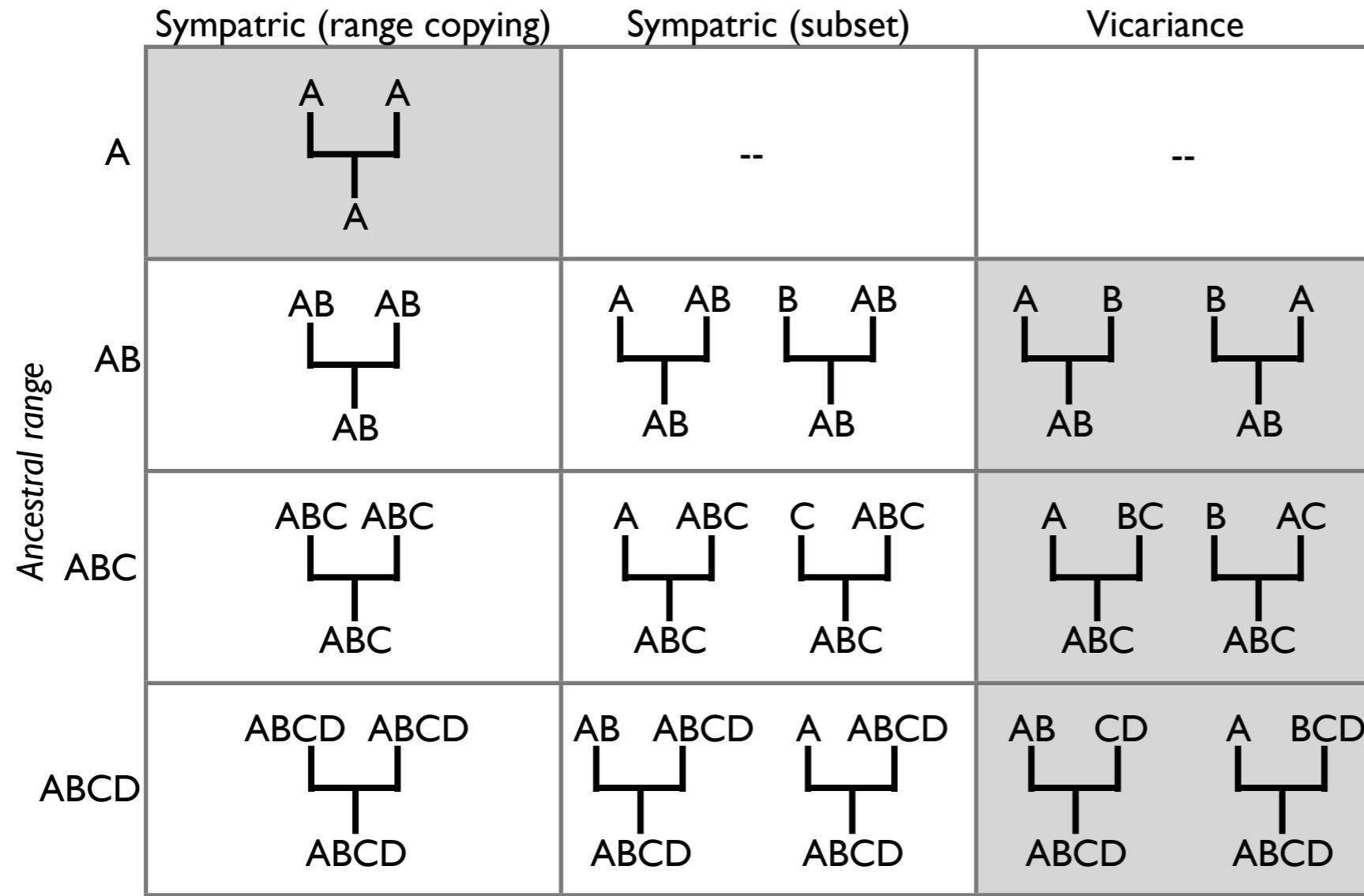


Figure 1B. DIVA cladogenesis model. DIVA+J would add founder-event speciation, as with DEC+J.

BAYAREA cladogenesis model

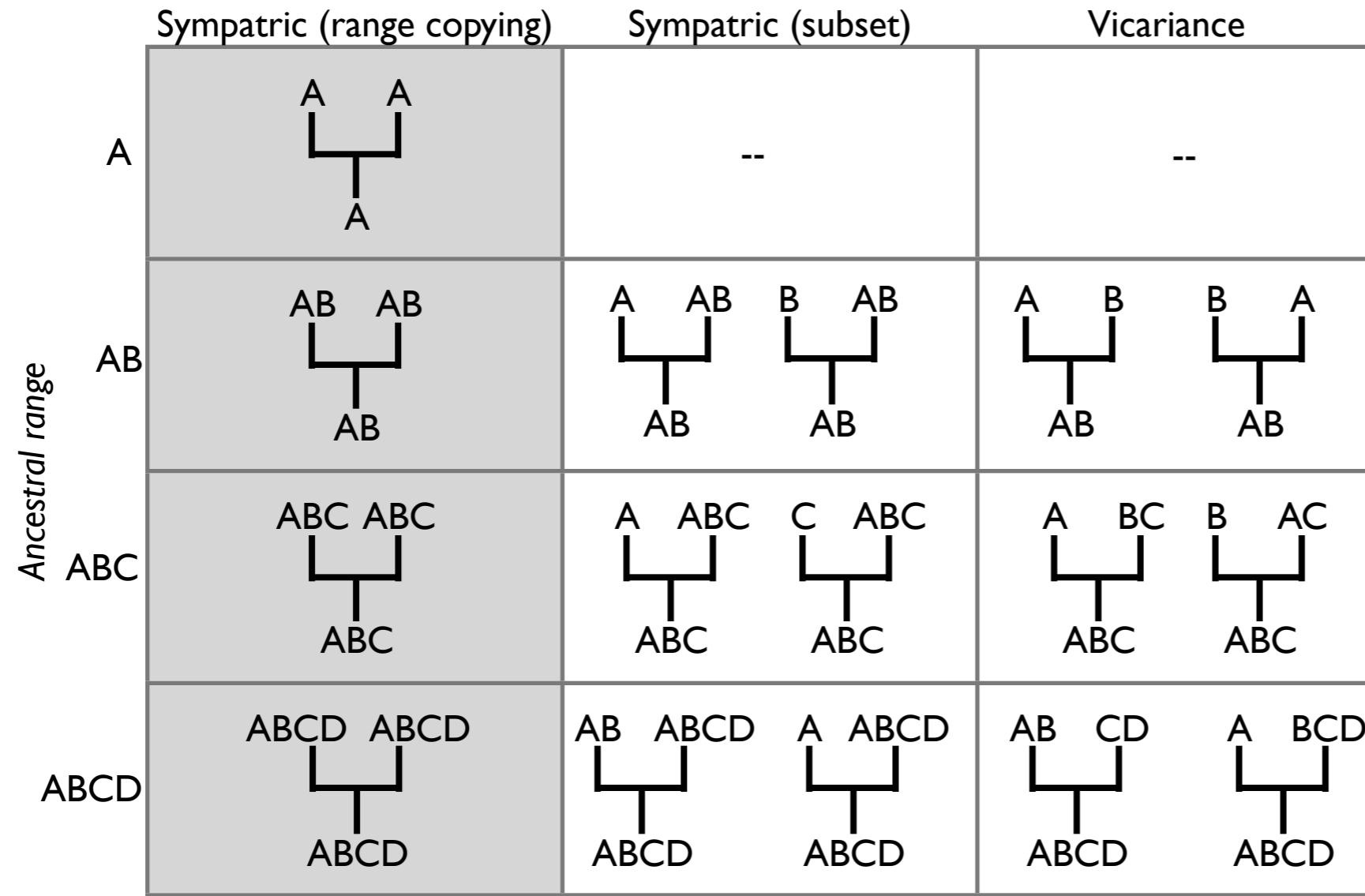


Figure 1C. BAYAREA cladogenesis model. BAYAREA+J would add founder-event speciation, as with BAYAREA+J.

DEC cladogenesis model

	Sympatric (range copying)	Sympatric (subset)	Vicariance	Founder Event
A		--	--	
AB				
ABC				
ABCD		 		

DIVA cladogenesis model

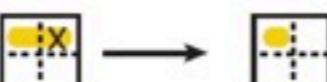
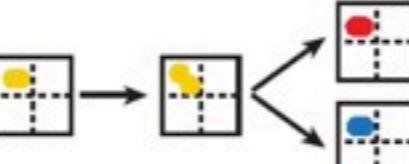
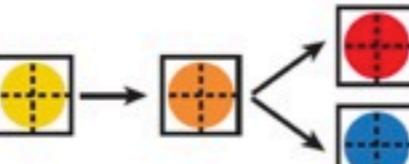
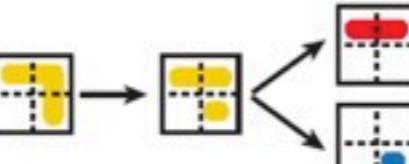
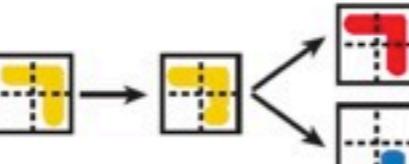
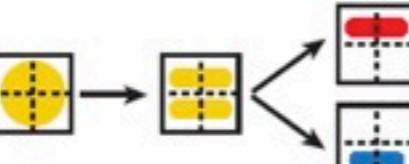
	Sympatric (range copying)	Sympatric (subset)	Vicariance
A		--	--
AB			
ABC			
ABCD		 	

BAYAREA cladogenesis model

	Sympatric (range copying)	Sympatric (subset)	Vicariance
A		--	--
AB			
ABC			
ABCD		 	

Figure 1. The types of cladogenesis events allowed (gray) and disallowed (white) under the DEC, DIVA, and BAYAREA cladogenesis models. The “+J” version of each model adds founder-event speciation, as illustrated for DEC+J.

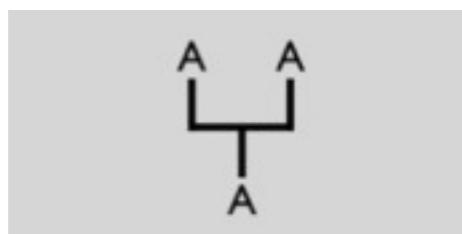
Range evolution at cladogenesis

	a	Before	After	Event	DIVA	DEC
Anagenetic events				Range expansion (geodispersal)	Yes	Yes
				Range contraction (local extinction)	Yes	Yes
Cladogenetic events				Duplication (within-area diversification)	Yes	Yes
				Sympatric speciation (across multiple areas)	No	No
				Alloperipatric speciation (one daughter with one area)	Yes	Yes
				Peripatric speciation (speciation within one area)	No	Yes
				Classical vicariance (both daughters in >1 area)	Yes	No

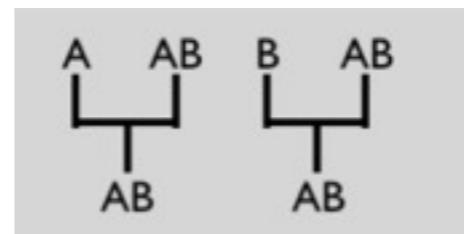
DEC: Dispersal-extinction cladogenesis

Cladogenesis -- range evolution at speciation events (LAGRANGE** model)**

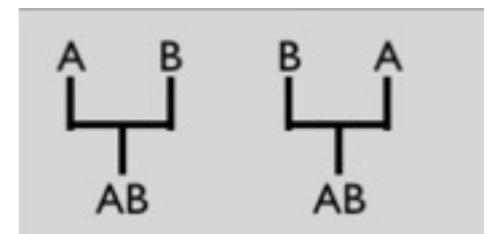
Probabilités per speciation event:



sympatric



subset

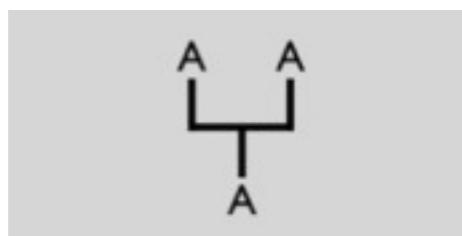


vicariance

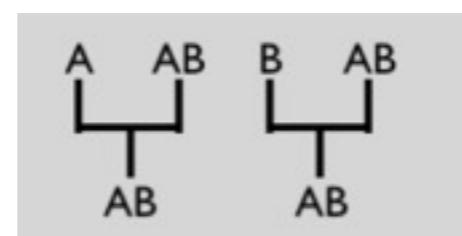
		ending range																																			
starting range	left:	A	A	A	A	A	A	B	B	B	B	B	C	C	C	C	AB	AB	AB	AB	AB	BC	BC	BC	BC	AC	AC	AC	AC	AC	ABC	ABC	ABC	ABC	ABC	ABC	count
	right:	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	
	A	sym																																			
	B	sym																																			
	C	sym																																			
	AB	vic	sub	vic		sub	vic		sub	vic		sub	vic		sub	sub		sub	sub		sub		sub		sub		sub		sub								
BC	vic		sub		vic		sub		vic		sub		vic		sub		sub		sub		sub		sub		sub		sub										
AC	vic		sub		vic		sub		vic		sub		sub		sub		sub		sub		sub		sub		sub		sub										
ABC	vic		sub		vic		sub		vic		sub		vic		sub		sub		sub		sub		sub		sub		sub										

DECj: Dispersal-extinction cladogenesis PLUS founder-event speciation

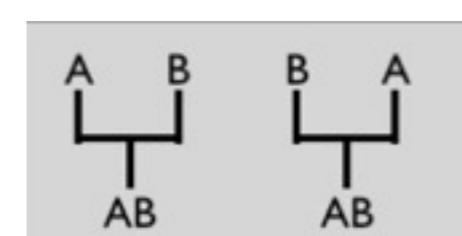
*Probabilités
per speciation
event:*



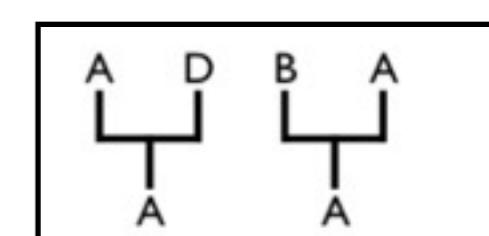
sympatric



subset



vicariance



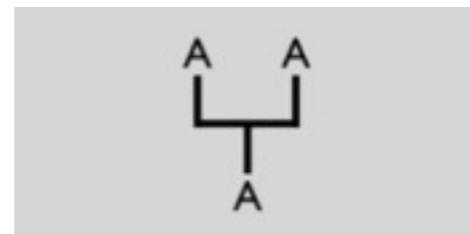
Founder Event

		ending range																																	
left:	A	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C	AB	AB	AB	AB	AB	BC	BC	BC	BC	AC	AC	AC	AC	ABC	ABC	ABC	ABC	ABC	count
right:	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC
A	sym	j	j					j					j																						
B	j							j	sym	j								j	j																
C		j							j				j	j	sym																				
AB	vic	sub	j				vic			sub						j				sub	sub	j										6			
BC			j					vic	sub				vic			sub							j	sub	sub							6			
AC		vic		sub					j		vic					sub																	6		
ABC			vic	sub						vic	sub				vic		sub		vic			vic				vic					sub sub sub	12			

BayArea cladogenesis model

Cladogenesis -- “no event” during speciation (i.e., pure continuous time)

*Probabilités
per speciation
event:*



Sympatric speciation (range duplication)

		ending range																																				
starting range	left:	A	A	A	A	A	A	B	B	B	B	B	B	C	C	C	C	C	AB	AB	AB	AB	AB	BC	BC	BC	BC	AC	AC	AC	AC	AC	ABC	ABC	ABC	ABC	ABC	count
	right:	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC	A	B	C	AB	BC	AC	ABC		
	A	sym																																				
	B	sym																																				
	C	sym																																				
	AB	sym																																				
		BC	sym																																			
		AC	sym																																			
		ABC	sym																																			

Island clades: DEC vs. DIVA vs. BAYAREA

Legend:

DEC DEC+J
 DIVA DIVA+J
 BAYAREA BAYAREA+J

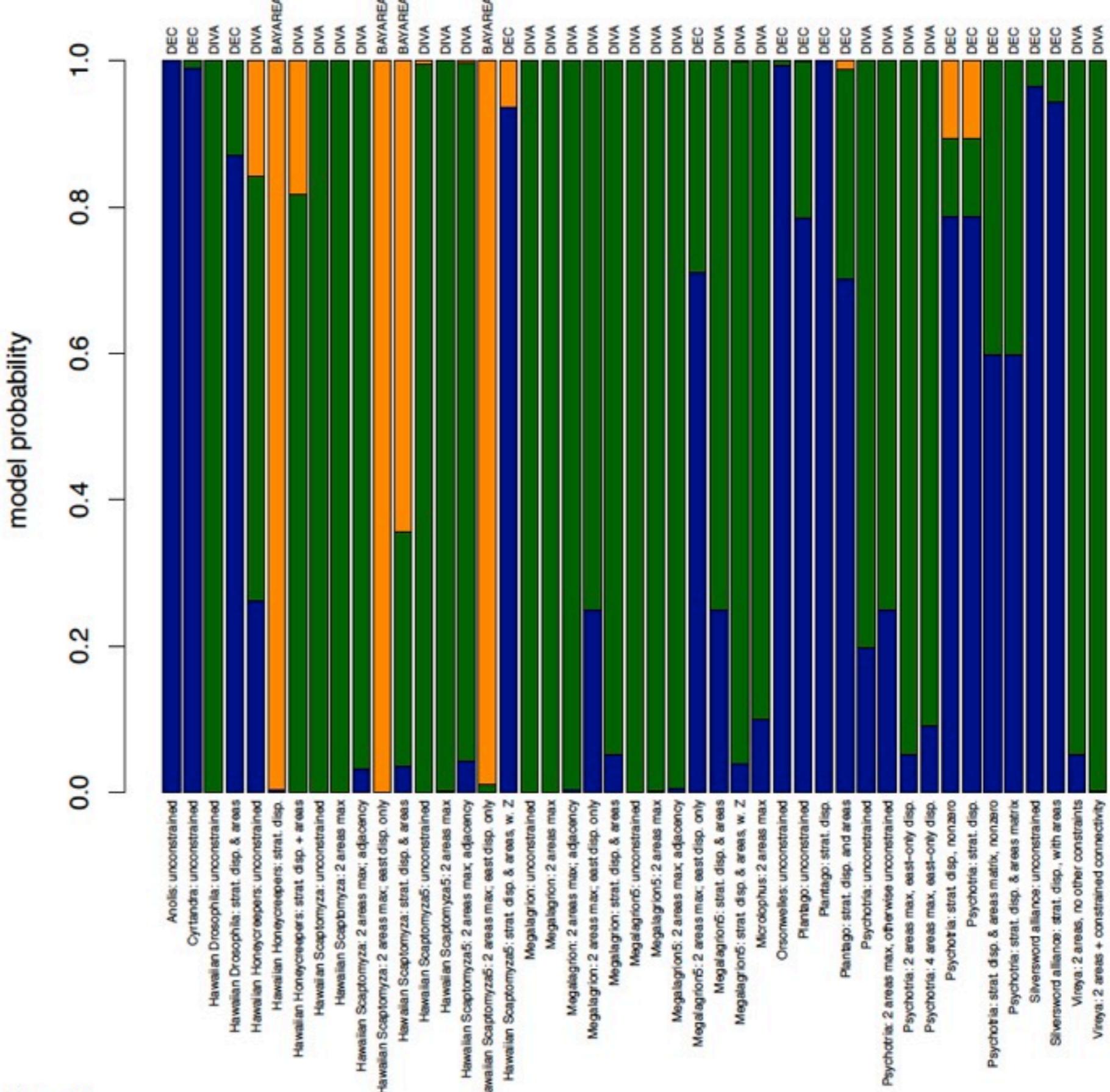


Figure 2

Non-island clades: DEC vs. DIVA vs. BAYAREA

Legend:

DEC DEC+J
 DIVA DIVA+J
 BAYAREA BAYAREA+J

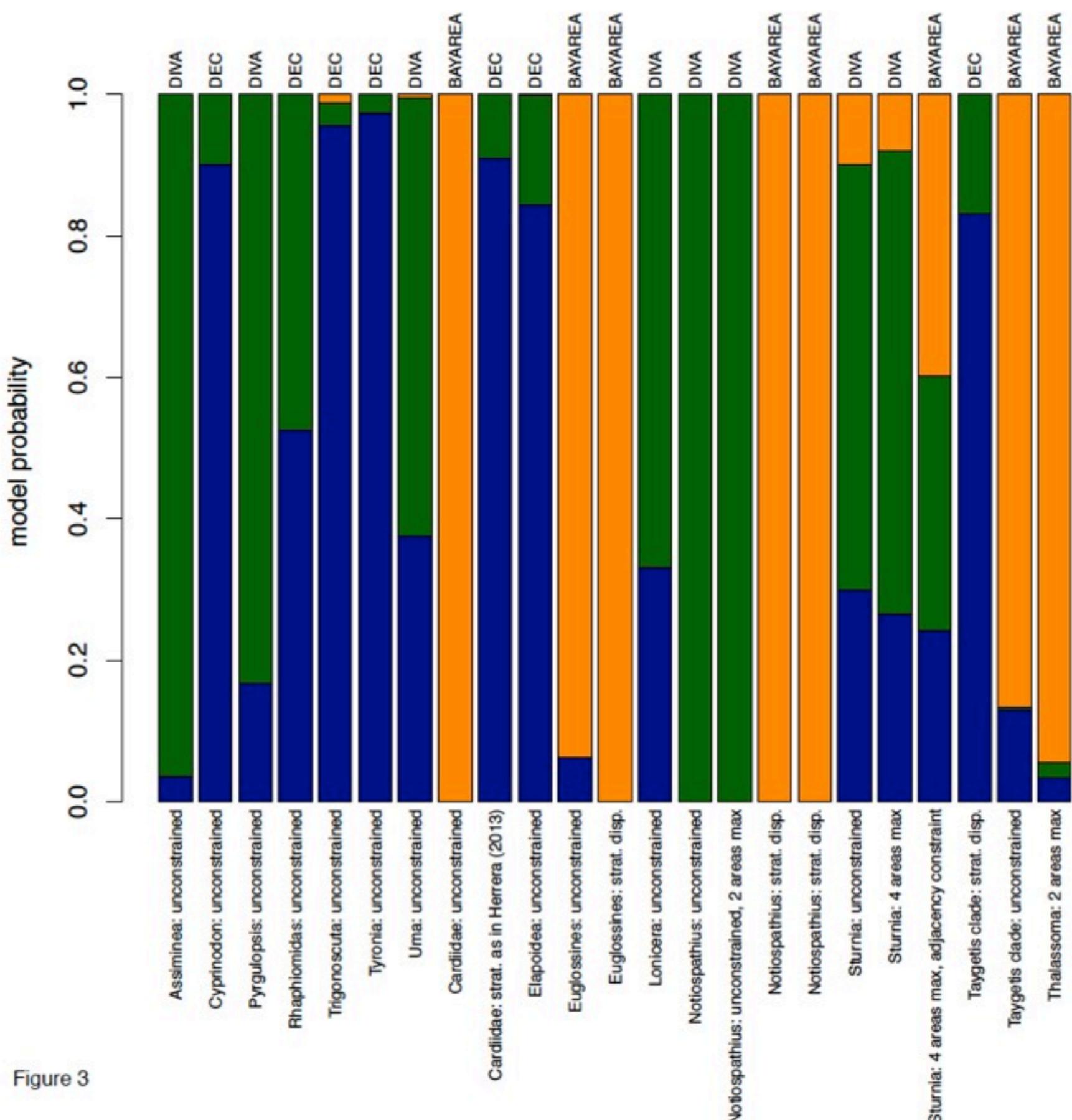


Figure 3

Island clades: DEC vs. DIVA vs. BAYAREA vs. DEC+J vs. DIVA+J vs. BAYAREA+J

Legend:

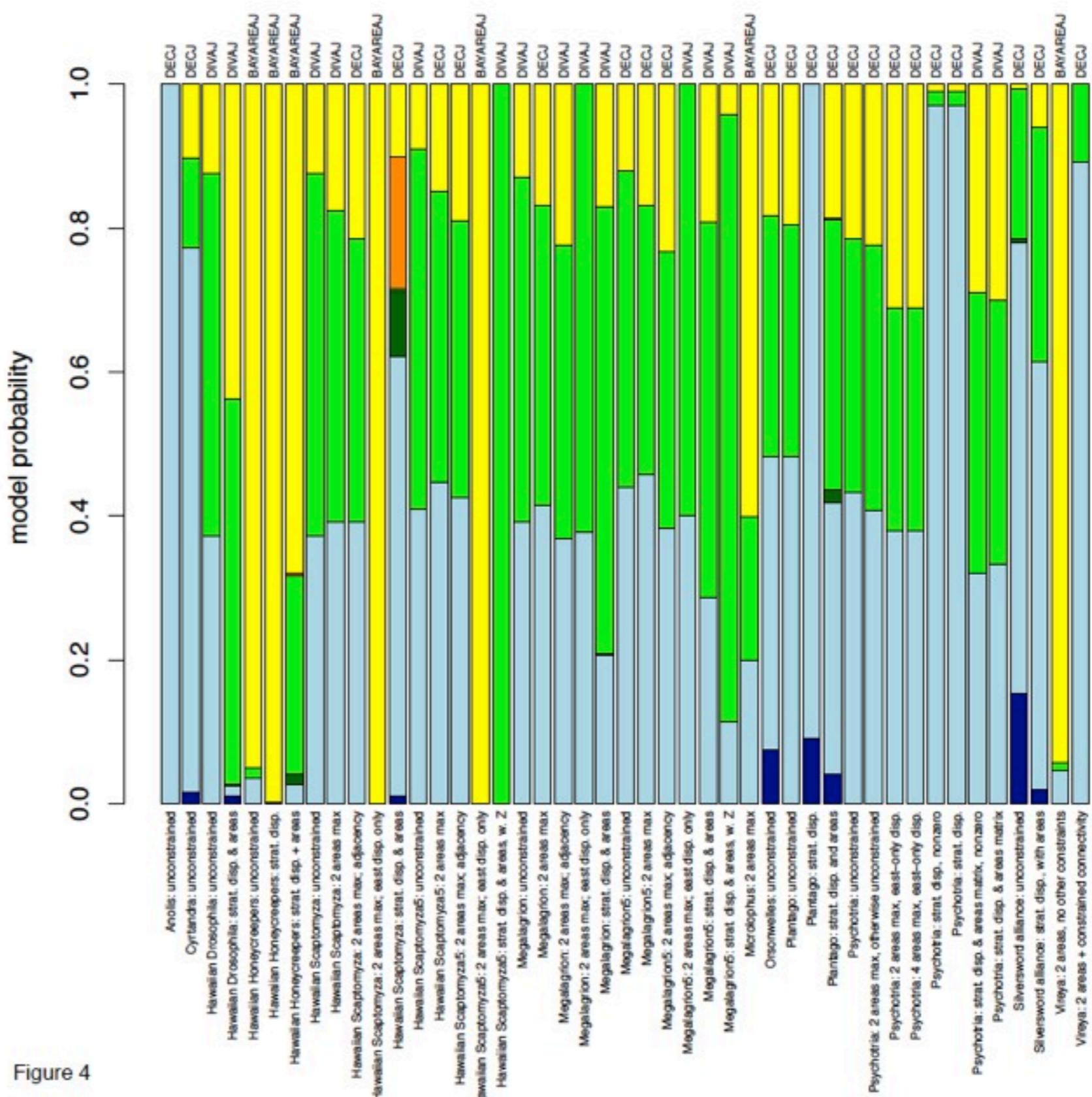


Figure 4

The Nick Matzke challenge:



Find me a dataset where the LAGRANGE
2-parameter model cannot be rejected in
favor of the 3-parameter model with
Founder-Event Speciation

The Nick Matzke challenge:



Find me a dataset where the LAGRANGE
2-parameter model cannot be rejected in
favor of the 3-parameter model with
Founder-Event Speciation

If you do, beer is on me.

Island clades: DEC vs. DIVA vs. BAYAREA vs. DEC+J vs. DIVA+J vs. BAYAREA+J

Legend:

DEC		DEC+J
DIVA		DIVA+J
BAYAREA		BAYAREA+J

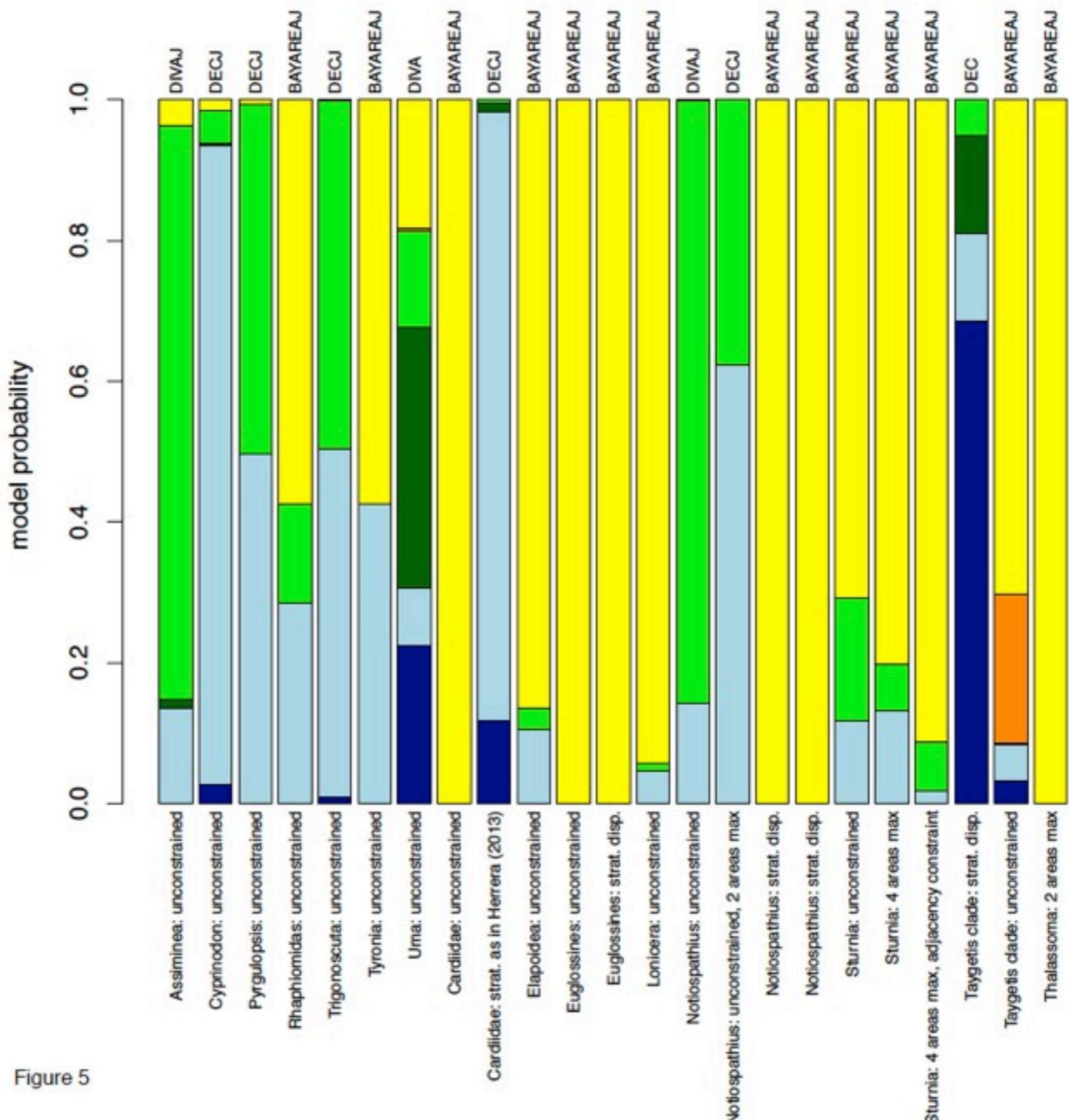


Figure 5

Inferences of j , all analyses

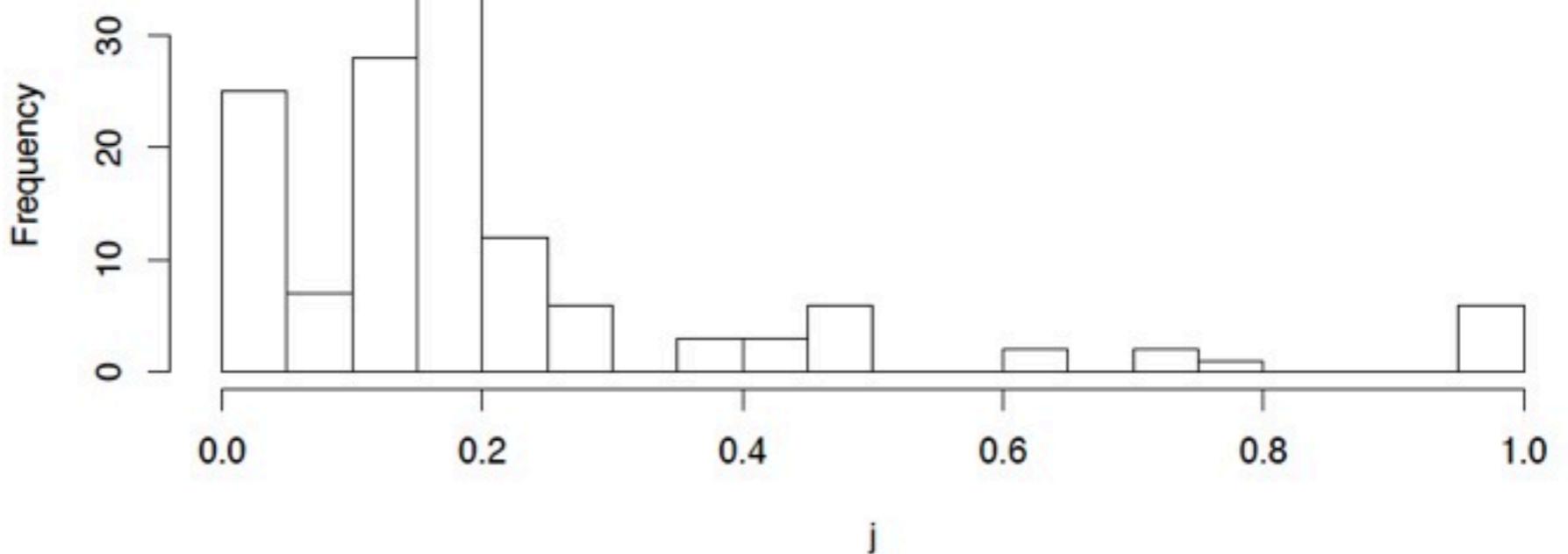
Kolmogorov-Smirnov
tests:

all analyses
 $p = 5.2e-6$

all unconstrained
analyses
 $p = 1.9e-15$

just unconstrained
DEC+J analyses
 $p = 3.4e-7$

island clades



non-island clades

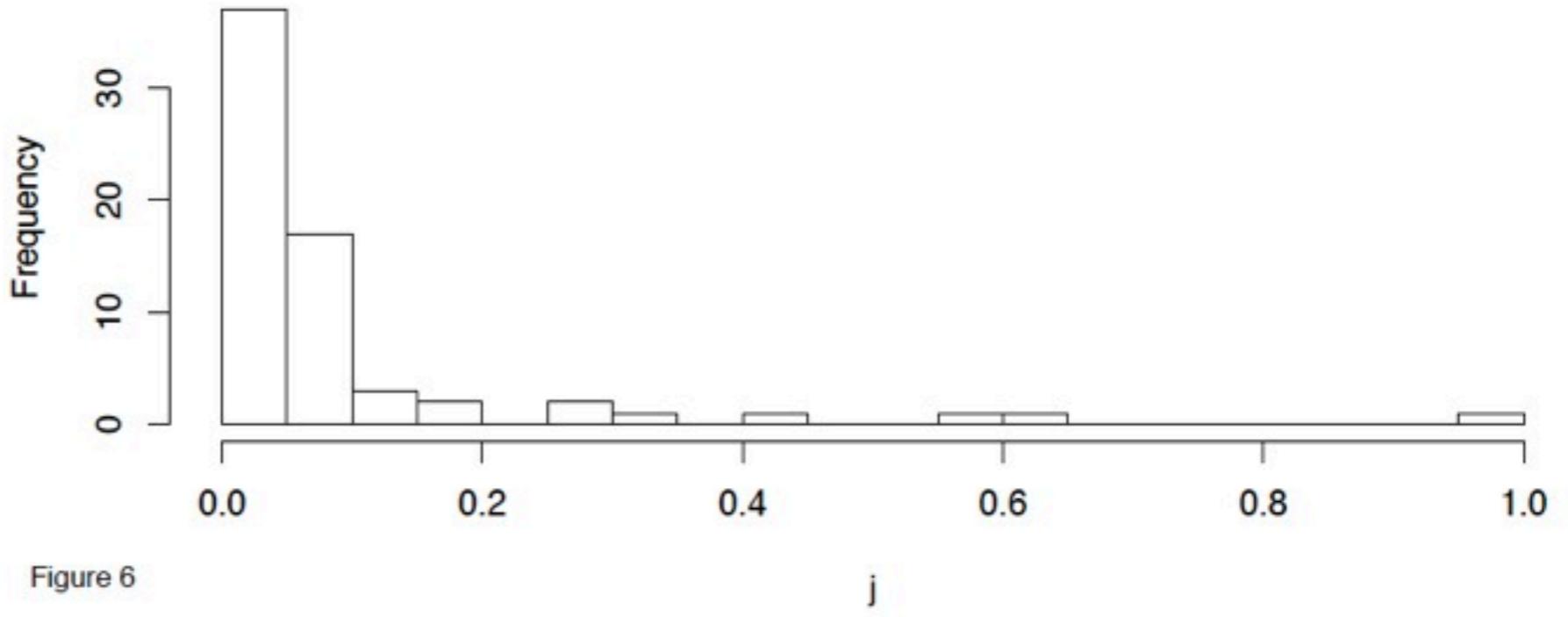


Figure 6

Inferences of j , all analyses

Kolmogorov-Smirnov tests:

all analyses
 $p = 5.2\text{e-}6$

all unconstrained
analyses
 $p = 1.9\text{e-}15$

just unconstrained
DEC+J analyses
 $p = 3.4\text{e-}7$

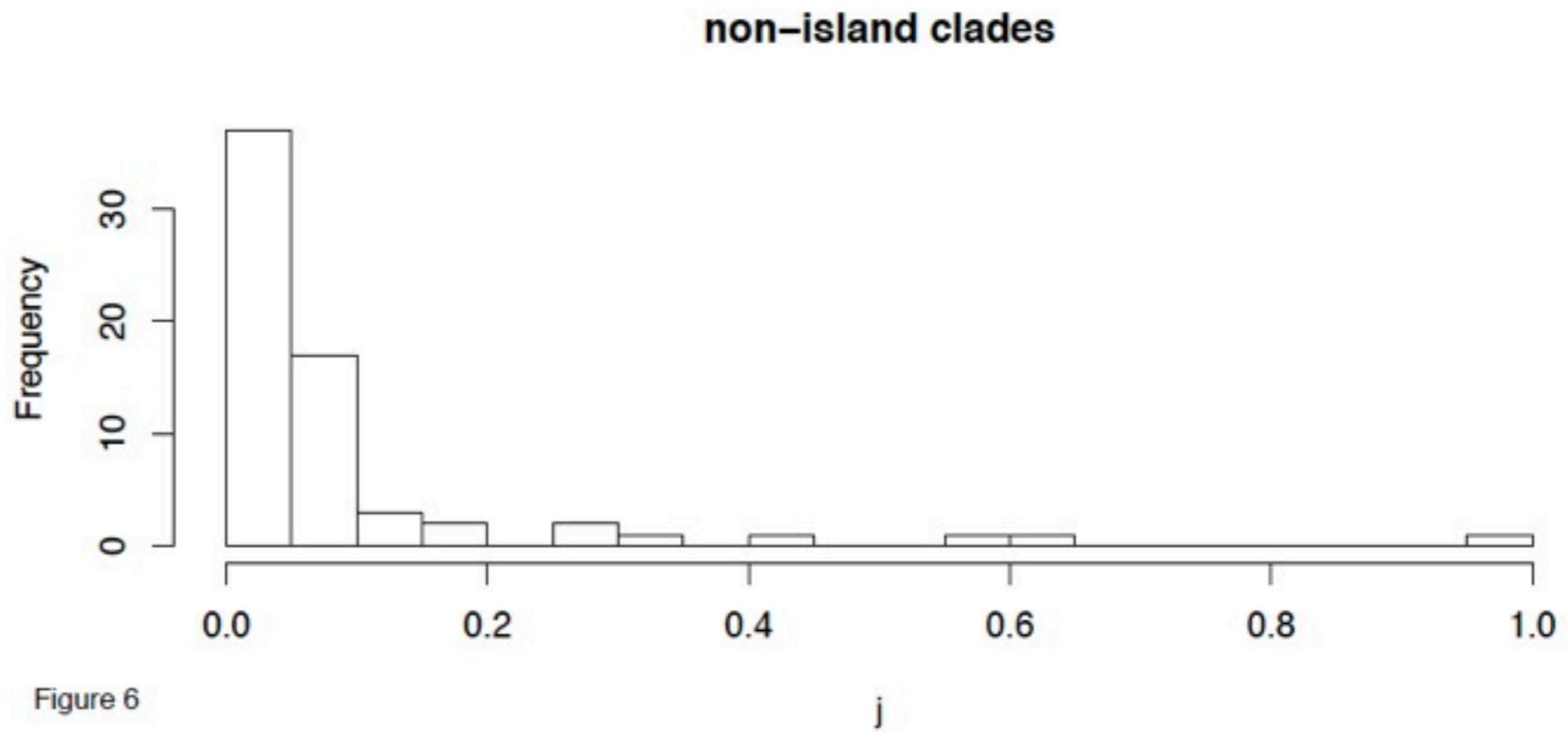
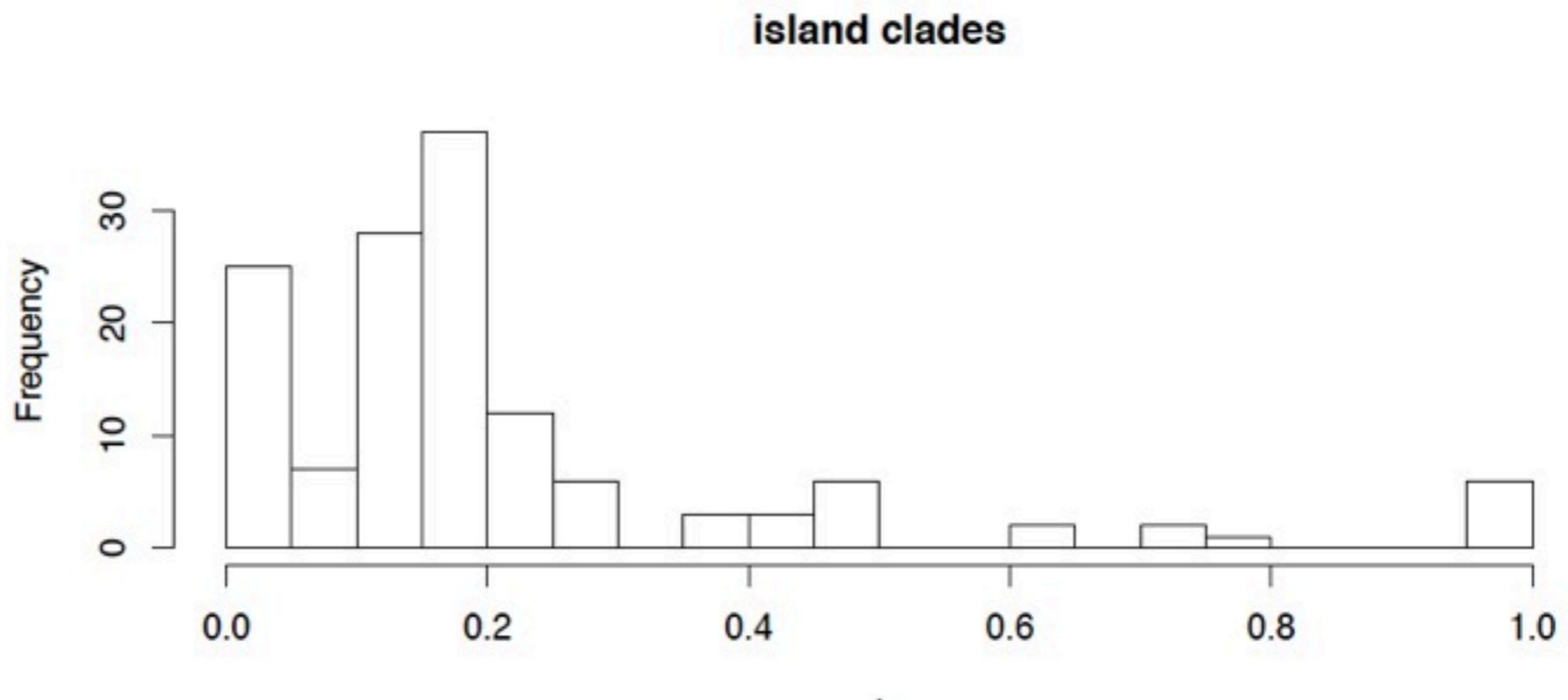


Figure 6

Founder-event speciation is 2-4 times more frequent on islands than non-islands, on a per-cladogenesis-event basis

Implications #2

“All models are wrong, but some models are useful.”



George E. P. Box
(1919-2013)

Implications #2

“All models are wrong, but some models are useful.”

-- George Box



George E. P. Box
(1919-2013)

I say: we'll never know the usefulness / misleading qualities unless we do statistical model comparison.

Conclusions

1. Founder-event speciation is important almost everywhere
2. Almost all datasets falsify the models currently in highest use -- DEC, DIVA, BayArea, in favor of +J models
3. With BioGeoBEARS, we can *test* all of these models and let the data tell us what processes are most important

Conclusions

Our models for historical biogeography appear to be under-parameterized.

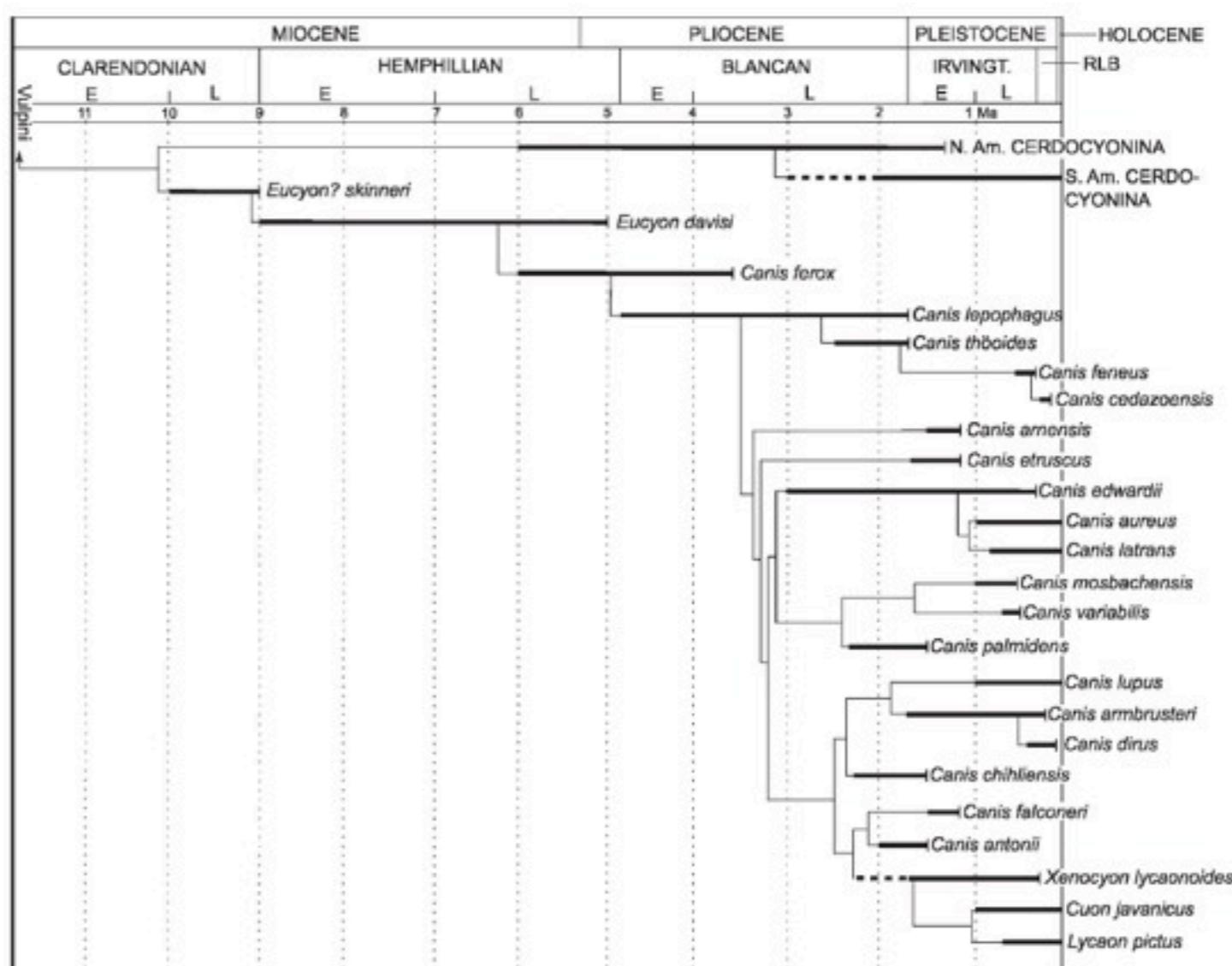
With BioGeoBEARS, we can attempt to estimate the relative probability of different cladogenesis processes (sympatric, subset, vicariance, founder-event speciation), or at least determine if models are distinguishable based on the data

Putting fossils into phylogenetic biogeography

Control of the cladogenesis model

allows fossil species to be included, and their ranges at multiple time points

This then means we can include detection probability, climate change, etc. in models



E.g., North American fossil Canidae

Putting fossils into phylogenetic biogeography

Control of the cladogenesis model
allows fossil species to be
included, and their ranges at
multiple time points

This then means we can include
detection probability, climate
change, etc. in models

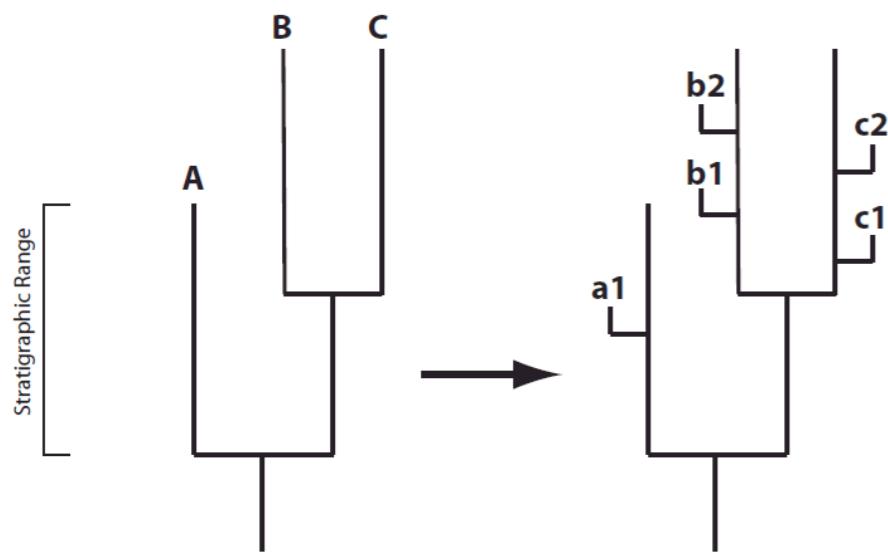


E.g., North American fossil Canidae

Putting fossils into phylogenetic biogeography

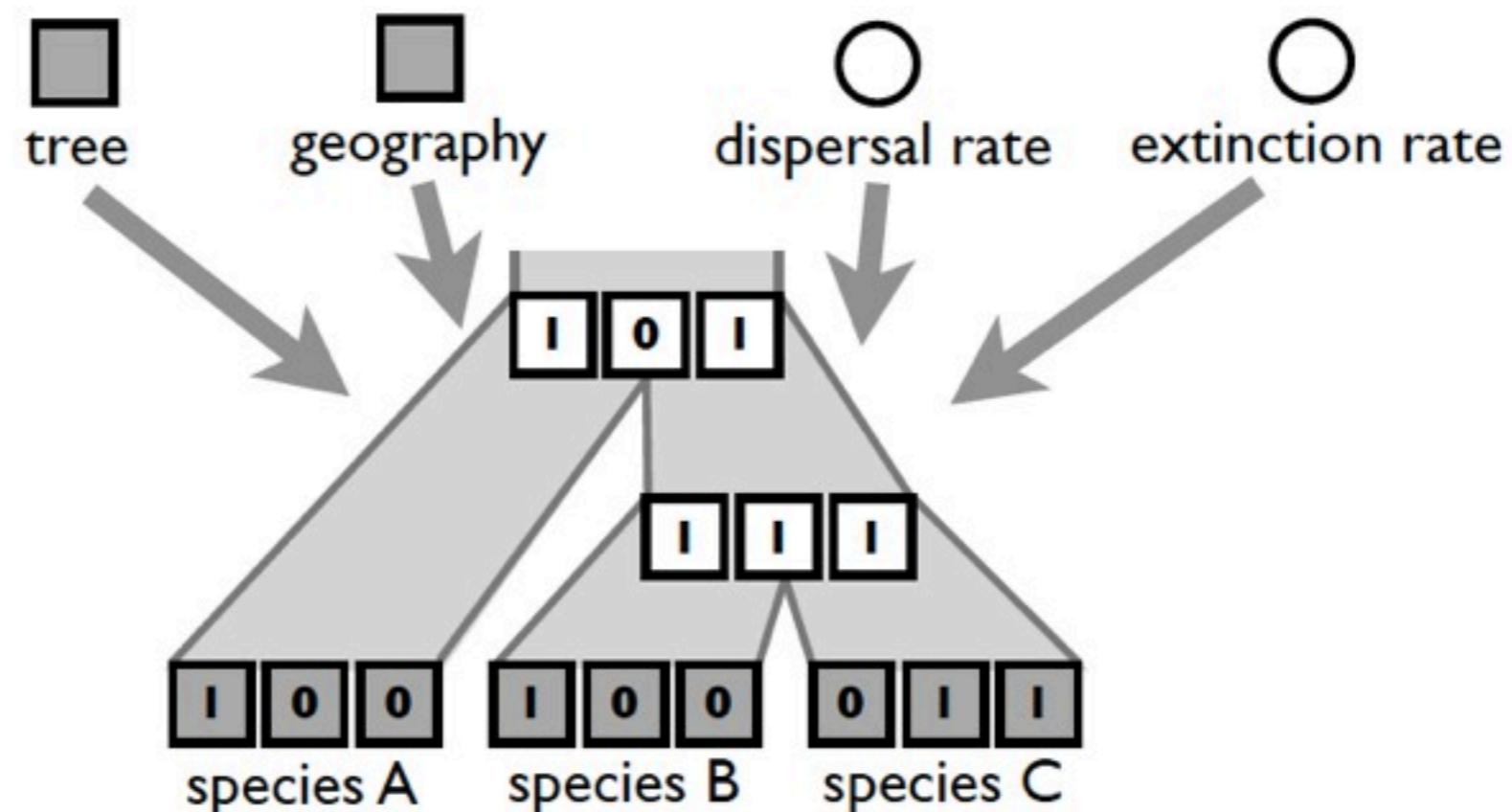
Control of the cladogenesis model
allows fossil species to be
included, and their ranges at
multiple time points

This then means we can include
detection probability, climate
change, etc. in models



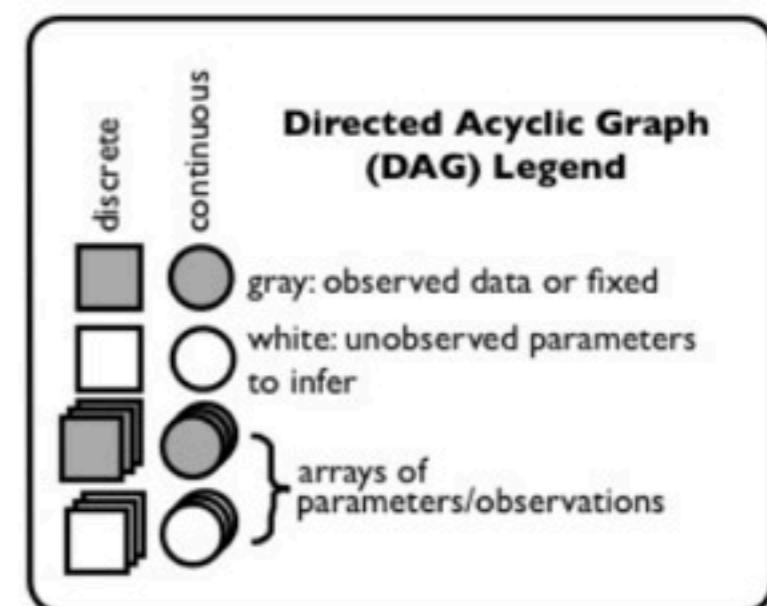
E.g., North American fossil Canidae

model of historical
biogeography (e.g. DEC)



present distributions

Figure 1A. Traditional inference procedure in likelihood analyses of historical biogeography on a phylogeny.



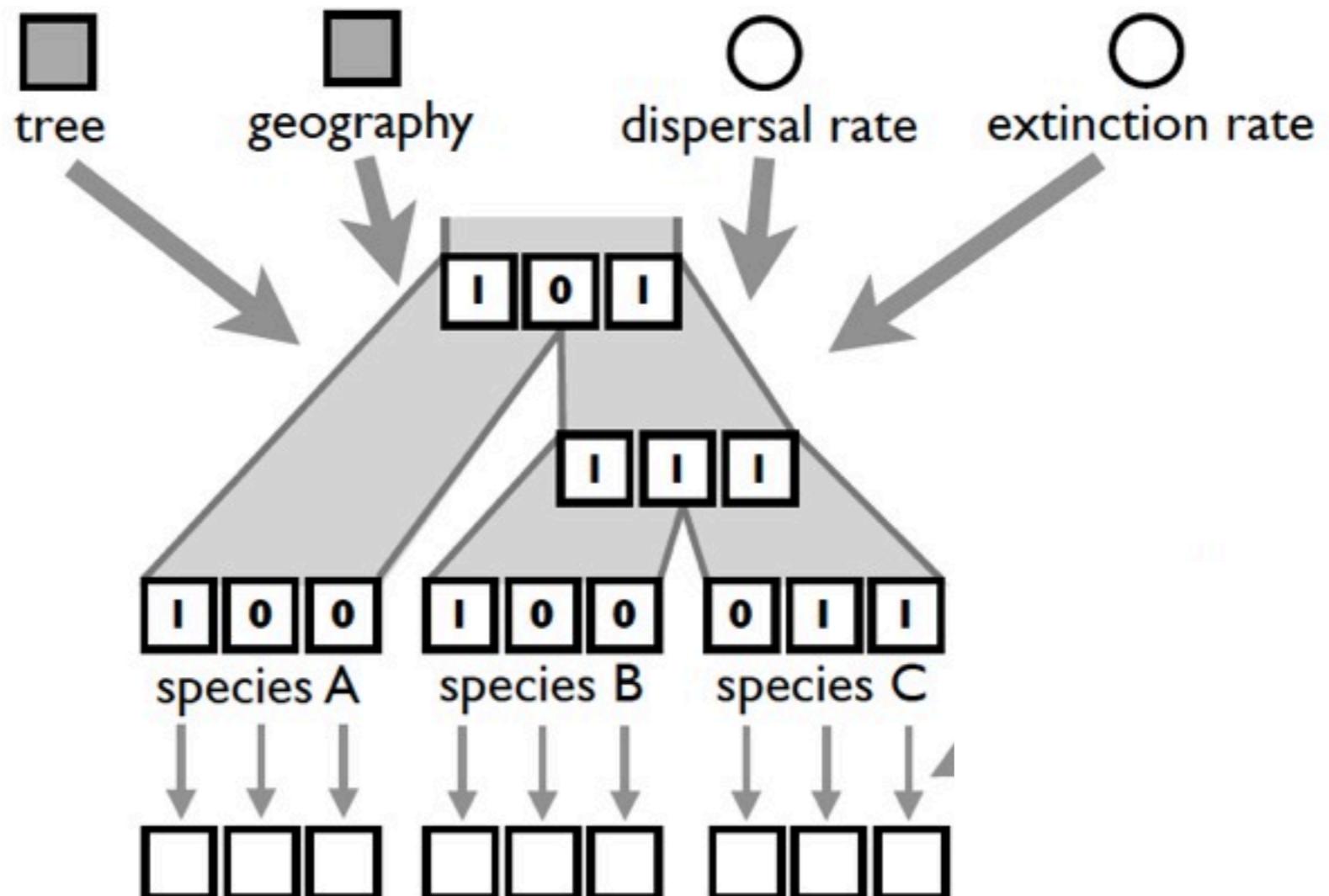
model of historical
biogeography (e.g. DEC)

inferred history

present distributions

sampling

actual taxa discovered



model of historical
biogeography (e.g. DEC)

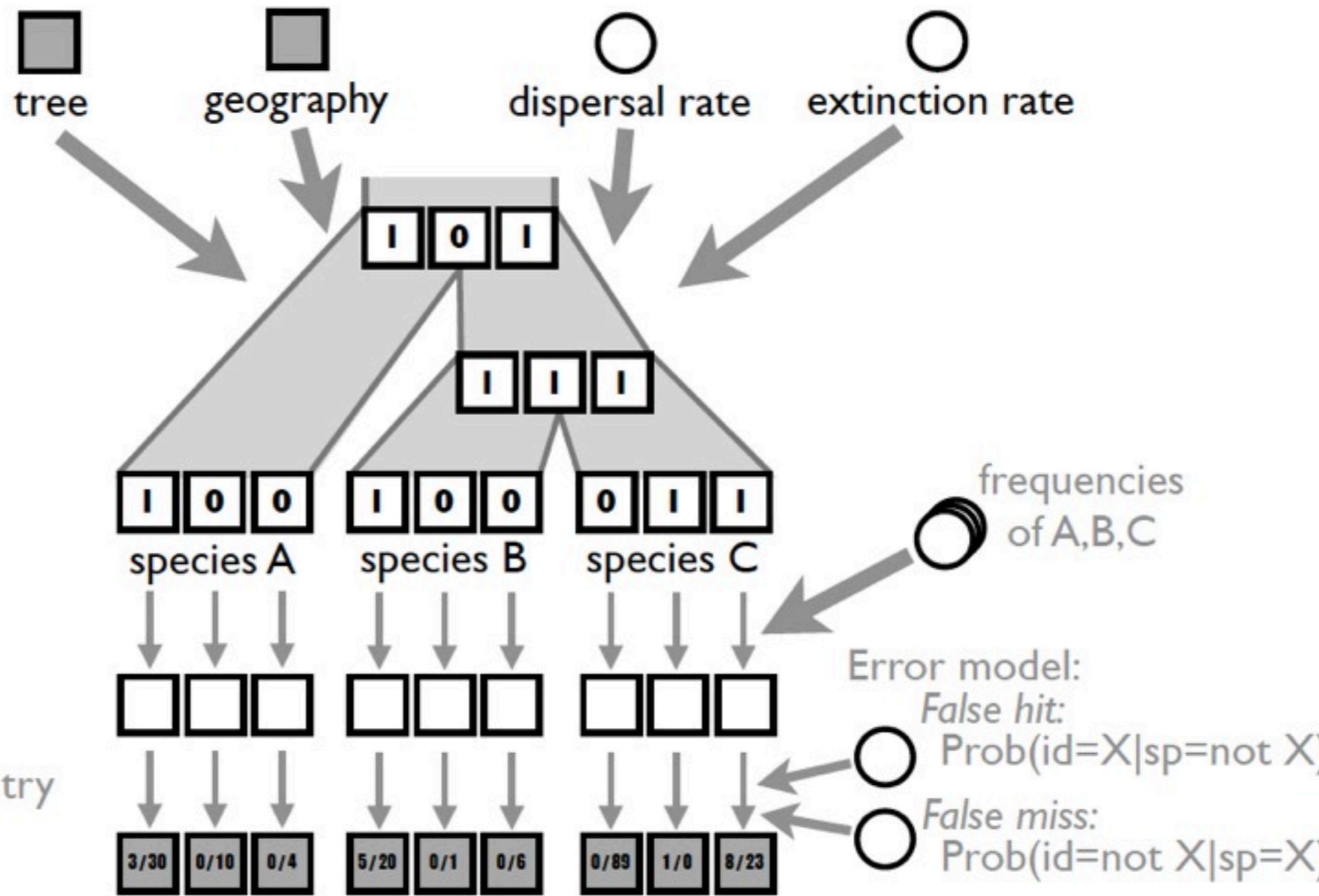


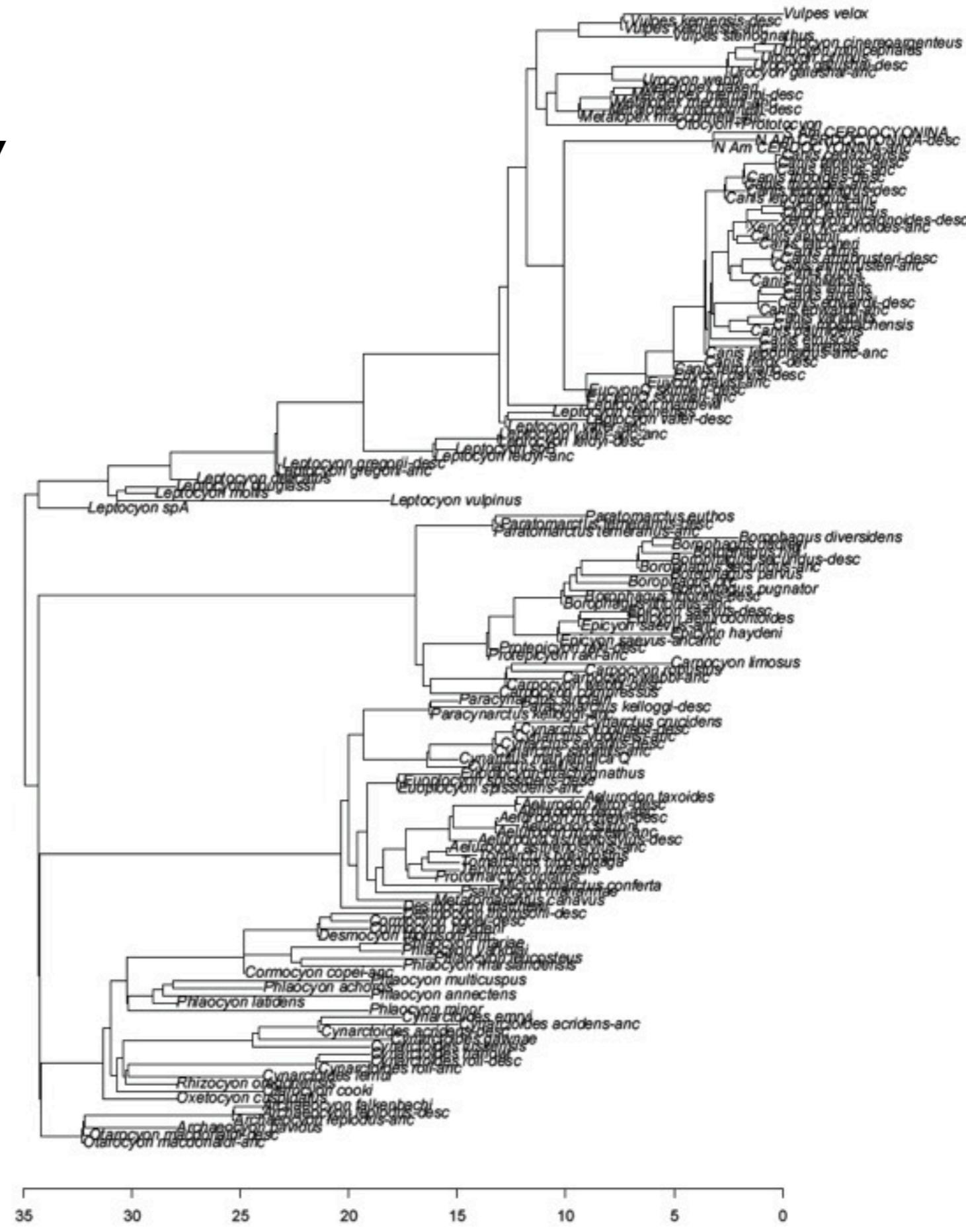
Figure 1B. DAG (Directed Acyclic Graph) for the model of imperfect detection implemented in BioGeoBEARS.

Putting fossils into phylogenetic biogeography

I. Take digitized phylogenies of North American fossil Equinae and Canidae

Canidae phylogeny:

I digitized and combined the Tedford & Wang time-scaled fossil phylogenies for Caninae, Borophaginae, and Hesperocyoninae



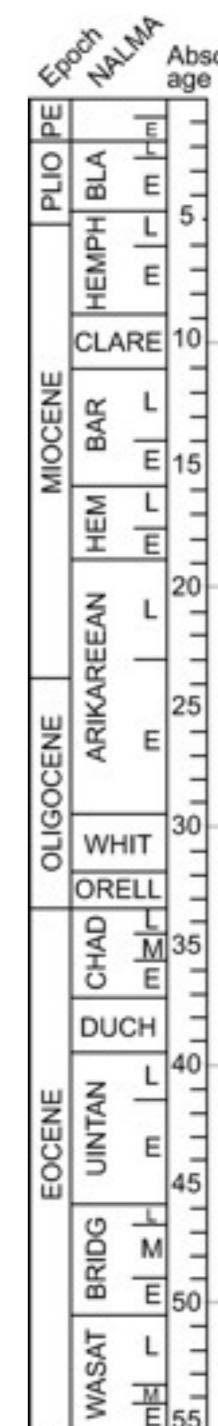
Putting fossils into phylogenetic biogeography

2. Dicreteize

2a. Discretize biogeography into regions

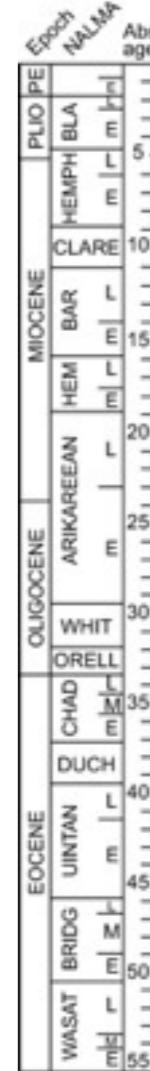


2b. Discretize time by NALMAs (North American Land Mammal Ages)



Putting fossils into phylogenetic biogeography

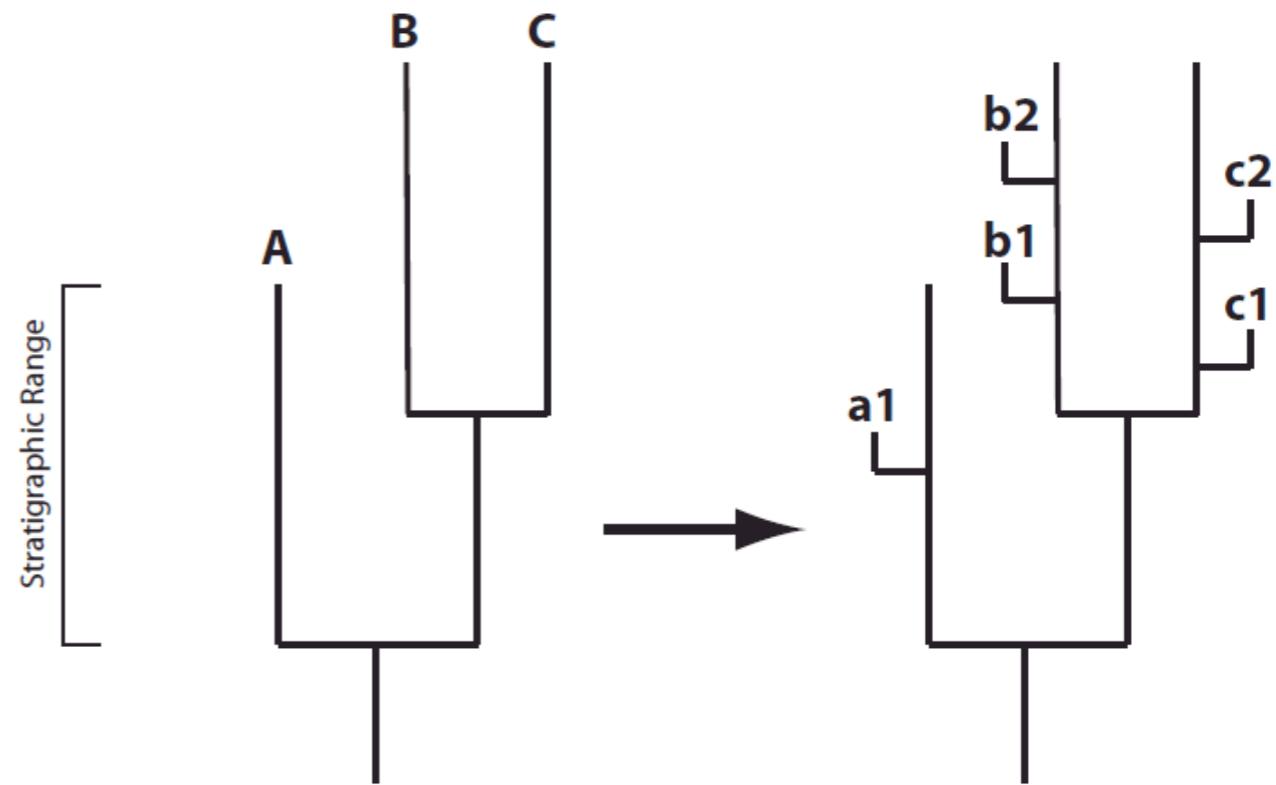
3. Take database (here, MIOMAP+FAUNMAP), and count occurrences of each fossil OTU in each area+time-bin



Putting fossils into phylogenetic biogeography

4. Here, each branch has a taxon name. For the mid-point of every time-bin, add a “hook” to each extant branch

BioGeoBEARS will treat these as “direct ancestor OTUs”, i.e. they will not apply the cladogenesis model to connect these to the tree



Putting fossils into phylogenetic biogeography

5. Taphonomic control groups (Bottjer & Jablonski 1988).

In each area+time-bin,
count the number of
fossil occurrences of
your taphonomic
controls

Agriochoeridae,
Amphicyonidae,
Antilocapridae,
Bovidae,
Camelidae,
Canidae,
Cervidae,
Chalicotheriidae,
Dichobunidae,
Entelodontidae,
Equidae,
Felidae,
Feliformia,
Gelocidae,
Gomphotheriidae,
Hyaenodontidae,
Hypertragulidae,
Hyracodontidae,
Mammutidae,
Moschidae,
Mustelidae,
Nimravidae,
Oreodontidae,
Oreodontoidea,
Palaeomerycidae,
Procyonidae,
Protoceratidae,
Rhinocerotidae,
Ruminantia,
Tapiridae,
Tayassuidae,
and
Ursidae

Putting fossils into phylogenetic biogeography

6. Apply detection model

We are interested in the likelihood of the observation data in R_t under the hypotheses that the species is truly present (T) or truly absent (F) from R_t . The likelihood under hypothesis T is:

$$P(\text{data}|T) = f^{ns}(1-f)^{n_s} \quad (1)$$

where f is the fraction abundance of S individuals in $R_{m,t}$, when S is present. Under hypothesis F the likelihood of the observations is:

$$P(\text{data}|F) = 0^{ns}(1)^{n_s} \quad (2)$$

In other words, under the hypothesis F that a species S is absent from R_t , the likelihood of the data is 0 if S has in fact been observed, and 1 otherwise. This is

Putting fossils into phylogenetic biogeography

7. Convert these likelihoods into geographic range likelihoods

With the ability to calculate the likelihood of the observation data for a particular species S in a particular region $R_{m,t}$ under hypotheses T and F, all that is required to calculate the likelihood of the data for a particular hypothesized *geographic range*, G_t , is to atomize the range into a series of presences and absences, where $R_{m,t} = T$ if S is hypothesized to be present in $R_{m,t}$, and F if absent:

$$G_t = \{R_{m=1,t}, \dots, R_{M,t}\} \quad (3)$$

where M is the number of discrete areas being used in the analysis, and m is the index identifying a particular area. The likelihood of the count data for a particular G_t is then:

$$P(\text{data}_{G_t} | G_t) = \prod_{m=1}^M P(\text{data}_{R_{m,t}} | R_{m=1,t}) \quad (4)$$

In a standard likelihood analysis of geographic range evolution on a phylogeny, likelihood 1 would be assigned for the known correct range at each tip on the tree, and likelihood 0 would be assigned for all other possible ranges. However, with the

model of historical
biogeography (e.g. DEC)

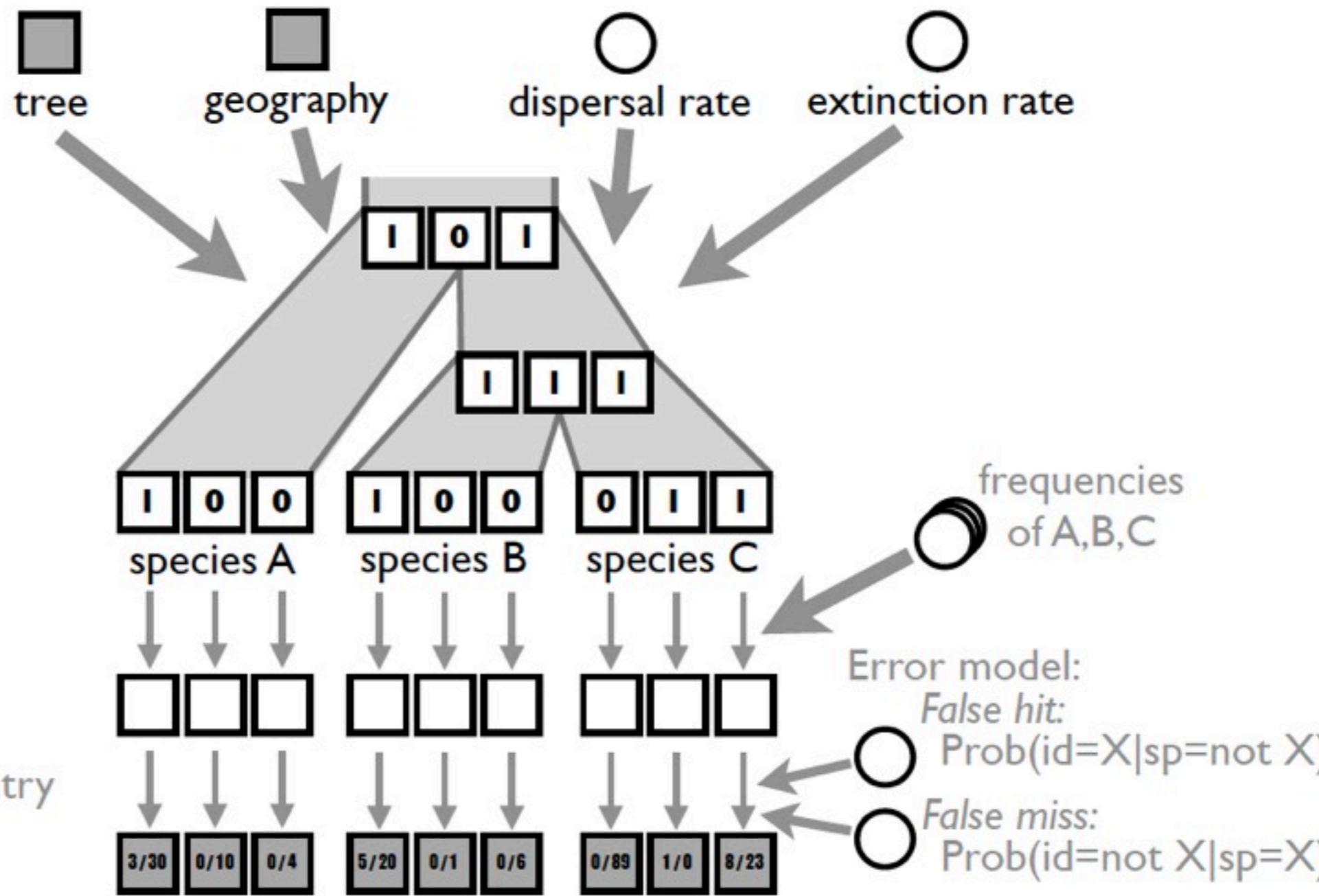
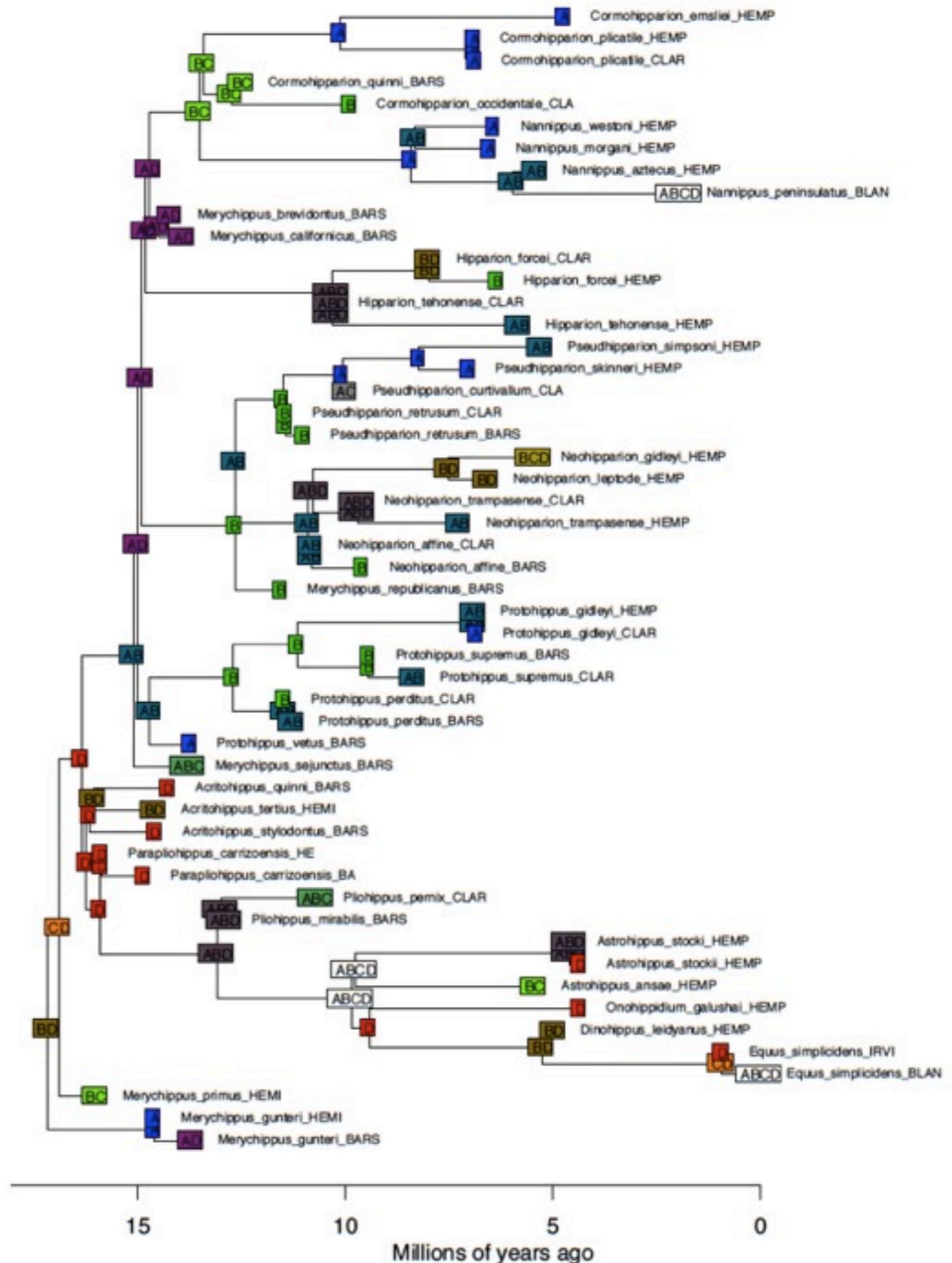


Figure 1B. DAG (Directed Acyclic Graph) for the model of imperfect detection implemented in BioGeoBEARS.

Putting fossils into phylogenetic biogeography

8. Run BioGeoBEARS models in the usual way

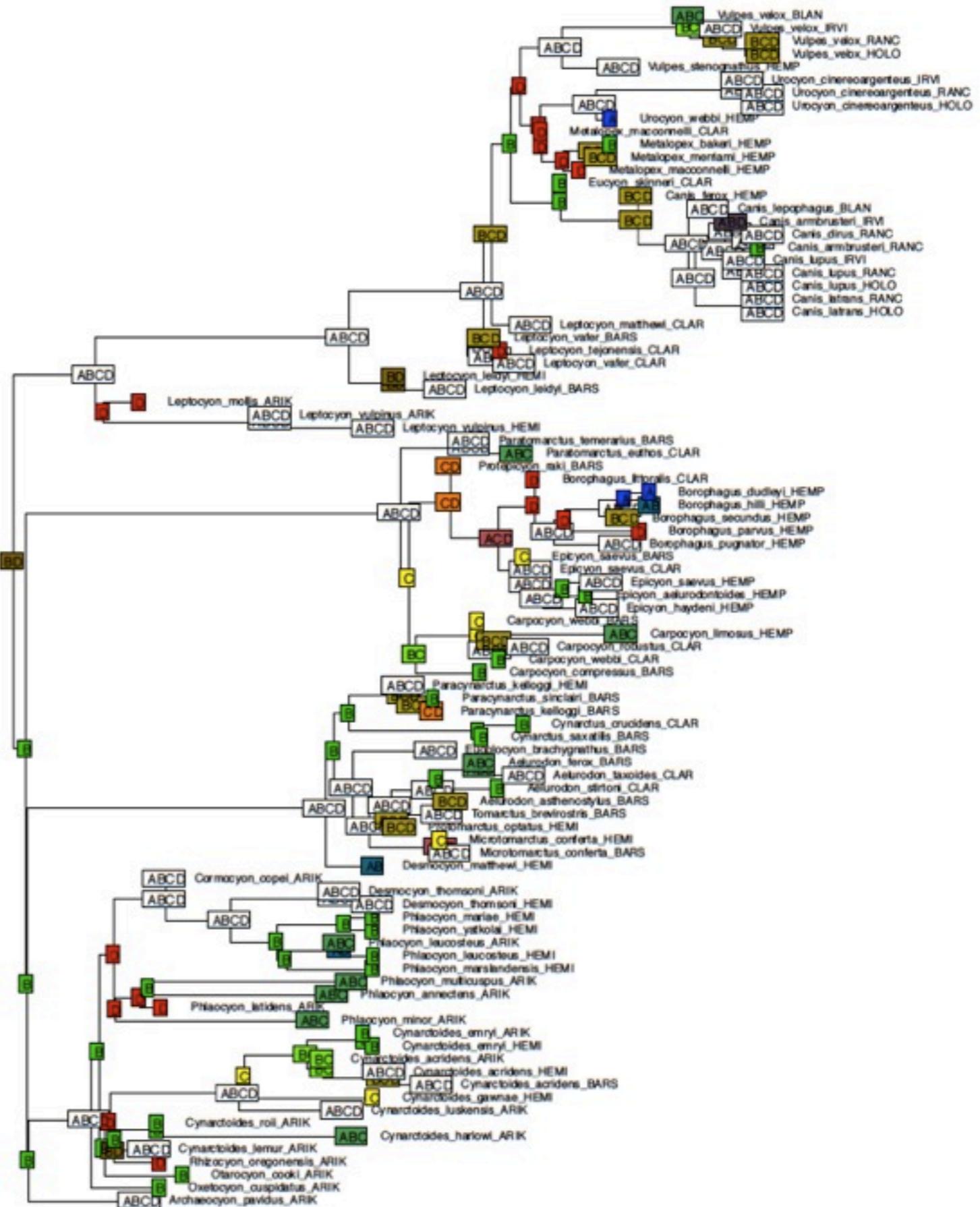
Equinae



Putting fossils into phylogenetic biogeography

Canidae

8. Run **BioGeoBEARS** models in the usual way



Parameter inference: pretty stable

(D1: with
detection
model)

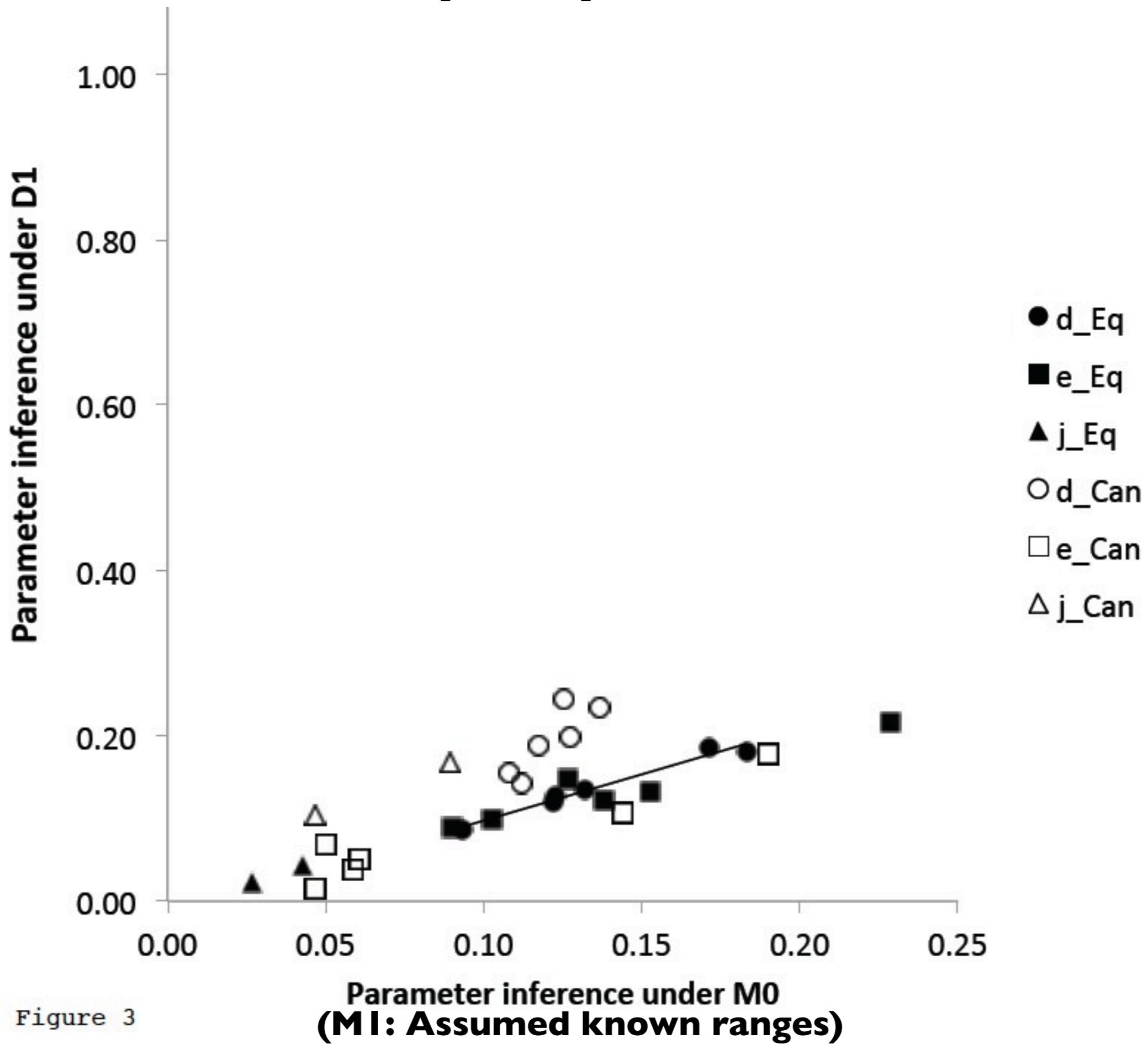


Figure 3

(M1: Assumed known ranges)

Table 3. Relative model probabilities, based on AICc.

Clade:	Equinae	Canidae
Dispersal:	M0	M0
DEC	0%	54%
DEC+J	0%	46%
DIVA	0%	0%
DIVA+J	0%	0%
BAYAREA	12%	0%
BAYAREA+J	88%	0%

Table 3. Relative model probabilities, based on AICc.

Clade:	Equinae	Equinae	Canidae	Canidae
Dispersal:	M0	D1	M0	D1
DEC	0%	0%	54%	43%
DEC+J	0%	0%	46%	50%
DIVA	0%	0%	0%	0%
DIVA+J	0%	0%	0%	0%
BAYAREA	12%	9%	0%	0%
BAYAREA+J	88%	91%	0%	7%

Conclusions: fossils in phylogenetic biogeography

1. Basically, this works great (*IF* you have a phylogeny with fossil tips, and optionally, direct ancestors)
2. Need to run on weaker fossil datasets (or, subsample this one)
3. More complex models possible

(see Wagner 2013, *Methods in Ecology & Evolution*. More complex sampling model, but simpler biogeography model.)

Conclusions: fossils in phylogenetic biogeography

4. Actually, this talk was a trick!

It was actually about: **Including Biogeography in Strato-Likelihood!**

Strato-likelihood models that do not jointly infer biogeographic range may be missing important information that will inform sampling probabilities

Big Conclusion:

5. Diversification models should include biogeographical processes.

A LOT of the nodes in trees literally represent geographic founder-event speciation

Having an abstract “speciation rate” seems a bit oversimple, if this is a complex of multiple underlying processes which actually can be pulled apart

(DEC, DIVA, BAYAREA, +J -- just the start of the models available)

Thanks! I welcome comments/collaborations at: matzke@nimbios.edu

Especially: Grad students & postdocs:

Ph.D. committee
John Huelsenbeck
Tony Barnosky

David Jablonski
Roger Byrne

Quals:
Paul Fine

Tony Barnosky
Roger Byrne
Rasmus Nielsen

IB200A/B:
Brent Mishler
David Lindberg
Kip Will
David Ackerly

Kevin Padian, for telling me to come to Berkeley.

Kaitlin Maguire
Chris Nasrallah
Michael Landis
Sarah Werning
Emily Lindsey
Jenny McGuire
Tracy Heath

Other folks:
Joe Felsenstein
Stephen Smith
Rick Ree
Scott Steppan
Eugenie C. Scott

Running head: FOUNDER-EVENT SPECIATION MODEL OF RANGE EVOLUTION

Title: Inclusion of Founder-event Speciation in Dispersal-Extinction-Cladogenesis (DEC) Analyses of Biogeography Drastically Alter Parameter Inference and Dramatically Improve Data Likelihoods

Author: Nicholas J. Matzke¹

¹ Department of Integrative Biology, University of California, Berkeley

Correspondence author contact information:

Nicholas J. Matzke, Department of Integrative Biology, 1005 Valley Life Sciences, Building #3140, Berkeley, CA 94720-3140. Phone: 510-301-0179. Email: matzke@berkeley.edu

TRY IT YOURSELF AT:
<http://phylo.wikidot.com/biogeobears>

Funding:

NSF “Bivalves in Time and Space”
UC Berkeley Wang Fellowship
UC Berkeley Tien Fellowship
Google Summer of Code