Lecture 25: Estimating Biogeographic Histories

Nicholas J. Matzke

I. Background: Very short history

Historical biogeography has a fascinating history stretching back to Darwin and before. We can't go into it in a huge amount of detail here, but just so you are aware of its existence:

1. Historical biogeography can be traced back to speculations about how critters got around the world after Noah's Ark landed on Mount Ararat after Noah's Flood. As European explorers began to sail around the world and catalog different floras and faunas in the 1600s and 1700s, it became increasingly difficult to fit all of the animals on the Ark, or to explain how they could have gotten to their present positions from Ararat.

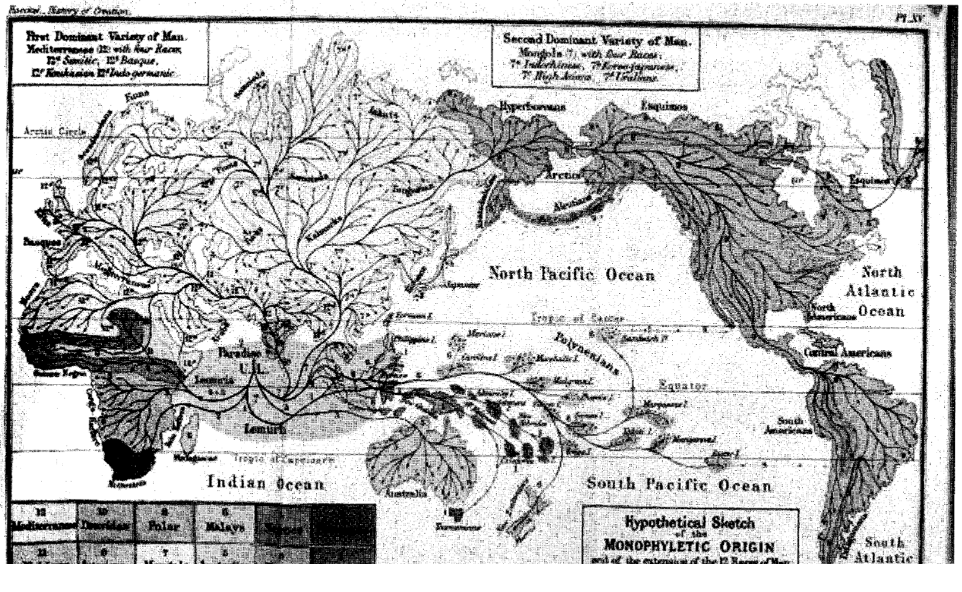
2. Linnaeus tried to place all diversity on one very tall mountain during the flood, with the altitudinal zones containing the plants and animals from different climate zones. See: Browne, Janet (1983). *The Secular Ark: Studies in the History of Biogeography*. New Haven & London: Yale University Press.

3. Various attempts to explain biogeography (and taxonomic structure in biogeography) via special creation became more and more attenuated during the 1800s, until Darwin and Wallace used biogeography as one of the strongest arguments that species must share common ancestry.

4. Common ancestry explained why taxonomically similar organisms lived in similar regions, and why faunas on oceanic islands were so skewed. But there were still many puzzles, especially disjunct distributions. Darwin in particular pushed for "dispersalist" explanations, invoking long-distance dispersal. He knew of e.g. strong similarities between southern floras (via Hooker), but disliked explanations invoking land bridges:

   E.g., a letter that Darwin wrote to Lyell (in 1856) complained of the "geological strides, which many of your disciples are taking" by creating land bridges "as easy as a cook does pancakes."

5. Haeckel was also a major developer of historical biogeography, putting trees on maps for the first time:



…here, Haeckel traces the origin of humans to the mythical sunken continent of "Lemuria"…I am not making this up…

6. The dispersalist tradition was dominant for the next 100 years. Major concepts included centers of origin, and Simpson's concepts of sweepstakes dispersal, corridors, land bridges, and filters. However, there was really no method here, beyond making maps of species, genera, and family distributions, and telling stories to explain them.

7. This all was challenged in the 1950s-1970s with (1) the acceptance of plate tectonics, which suggested that disjunctions might be explained by vicariance, and (2) Leon Croizat's polemics, which boiled down to the assertion that many genera/families had congruent distributions, and therefore Darwin and modern dispersalists were wrong and stupid.

8. Croizat was self-publishing in South America and kind of kooky, but was introduced to the mainstream by early cladists like Gareth Nelson, who argued that dispersalism was unscientific because it could explain anything, and that only vicariance offered a falsifiable hypothesis. This hypothesis could be tested by seeing if cladograms from different groups had congruent geographic structure.

9. Most of the assumptions above on all sides are pretty dubious:
   - dispersal isn't completely random
   - geographic congruence between groups isn't only explainable by dispersal
   - the distribution of Linnaean taxa doesn't necessarily indicate much about geographic history (e.g. if they are not monophyletic; this trips up several of Croizat's favorite tracks)
   - centers of diversity are not necessarily centers of origin
   - cladograms are not necessarily great evidence for/against vicariance (e.g. pseudocongruence, timing)
10. Nevertheless, the debate was useful in provoking the development of explicit methods. Of which there

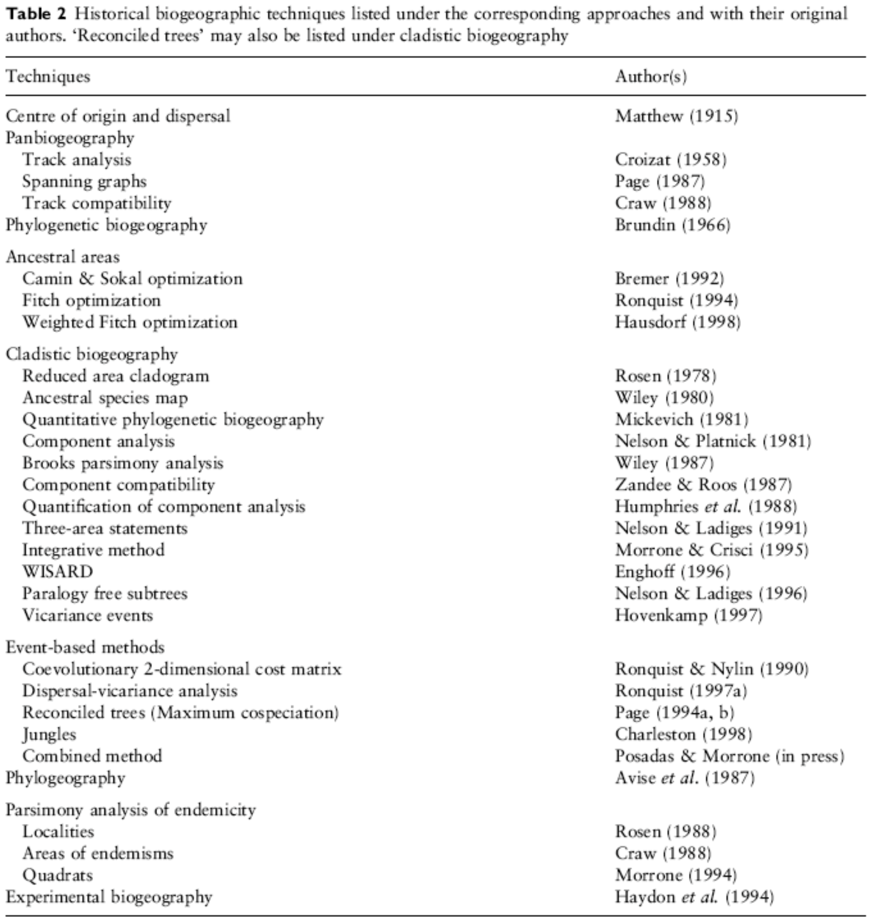


Table 2 Historical biogeographic techniques listed under the corresponding approaches and with their original authors. 'Reconciled trees' may also be listed under cladistic biogeography

Techniques	Author(s)
Centre of origin and dispersal	Matthew (1915)
Panbiogeography	
Track analysis	Croizat (1958)
Spanning graphs	Page (1987)
Track compatibility	Craw (1988)
Phylogenetic biogeography	Brundin (1966)
Ancestral areas	
Camin & Sokal optimization	Bremer (1992)
Fitch optimization	Ronquist (1994)
Weighted Fitch optimization	Hausdorf (1998)
Cladistic biogeography	
Reduced area cladogram	Rosen (1978)
Ancestral species map	Wiley (1980)
Quantitative phylogenetic biogeography	Mickevich (1981)
Component analysis	Nelson & Platnick (1981)
Brooks parsimony analysis	Wiley (1987)
Component compatibility	Zandee & Roos (1987)
Quantification of component analysis	Humphries *et al.* (1988)
Three-area statements	Nelson & Ladiges (1991)
Integrative method	Morrone & Crisci (1995)
WISARD	Enghoff (1996)
Paralogy free subtrees	Nelson & Ladiges (1996)
Vicariance events	Hovenkamp (1997)
Event-based methods	
Coevolutionary 2-dimensional cost matrix	Ronquist & Nylin (1990)
Dispersal-vicariance analysis	Ronquist (1997a)
Reconciled trees (Maximum cospeciation)	Page (1994a, b)
Jungles	Charleston (1998)
Combined method	Posadas & Morrone (in press)
Phylogeography	Avise *et al.* (1987)
Parsimony analysis of endemicity	
Localities	Rosen (1988)
Areas of endemisms	Craw (1988)
Quadrats	Morrone (1994)
Experimental biogeography	Haydon *et al.* (1994)

are many…

Source: Crisci, Jorge V. (2001). "The voice of historical biogeography." *Journal of Biogeography*. **28**(2): 157-168. URL: *http://dx.doi.org/10.1046/j.1365-2699.2001.00523.x*.

11. These led to event-based approaches like DIVA and Lagrange, and Bayesian elaborations on these methods, which will be discussed below.

12. As a final comment: it seems to me that historical biogeography is still not a completely mature discipline.
    - There are many famous patterns, but their explanation is still hotly disputed.

    - Philosophies and schools of thought still seem to have a strong influence on the conclusions that researchers reach. E.g. hardcore vicariance advocates, tied to pattern cladistics, suspicion of molecular phylogenetics and divergence time estimation, etc. (see some stuff from 2009 in *Journal of Biogeography* on the last pages of notes)

II. Event-based methods

An "event-based method" in historical biogeography (as opposed to a "pattern-based" method; Ronquist 1996) basically consists of explicitly considering a history of events (dispersal events, extirpation events, speciation events, etc.) and trying to find a history that invokes the minimum number of events (parsimony optimality criterion) or e.g. has the maximum likelihood. Obviously these methods are closely related to phylogenetics methods in general.

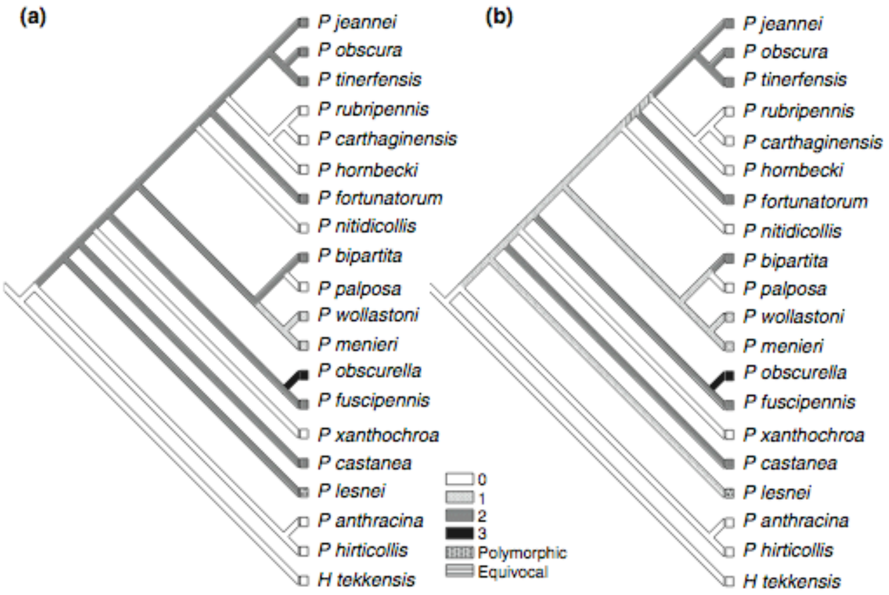


Figure 1 Parsimony-based optimization of the geographic distribution of the Canary Island species of *Pachydema* (Coleoptera, Scarabaeoidea) and related African species, onto a morphology-based phylogeny (one of three most parsimonious trees, I. Sanmartín, unpublished data); Outgroup: *Hemictenius tekkensis*. (a) Fitch (unordered) optimization; (b) Wagner (ordered) optimization. The Wagner optimization is four steps longer than the Fitch optimization. Area codes: (0) Mainland: Africa/Asia Minor; (1) eastern Canary Islands (Lanzarote and Fuerteventura); (2) central Canary Islands (Gran Canaria, Tenerife, La Gomera); (3) western Canary Islands (La Palma, El Hierro); (polymorphic) widespread in two or more island groups.

Some methods of estimating biogeographic history are just standard ancestral character reconstruction algorithms (e.g. Fitch parsimony, maximum likelihood, or stochastic mapping) with the character of interest being location. However, these methods require/assume that species live in only one region, and their ancestors lived in only one region as well. For some taxa and problems (e.g. tree species on different continents) this may well be a reasonable approximation. In such situations, just use these methods. However, these methods have no chance of inferring vicariance events, and do not deal with extinction or range expansion events.

DIVA

The first method that explicitly tried to take these factors into account was Fredrik Ronquist's DIVA (Dispersal-Vicariance Analysis) program (Ronquist was also a coauthor on MrBayes). It was partially inspired by Ronquist's earlier work on host-parasite coevolution. Here are some of the analogies between the different situations where we have lineages nesting within each other (or not!):

Host	Organism	Area
Parasite	Gene	Organism
Host switch	Horizontal transfer	Dispersal
Cospeciation	Orthology	Vicariance
Parasite speciation on one host	Gene duplication or allelic divergence	Sympatric speciation (kind of)
Parasite extinction	Gene loss or fixation	Extinction

DIVA basically tries to find a history that invokes the minimum number of extinction and dispersal events. Range inheritance events due to vicariance have no cost. The cost matrix is defined by:

1. Speciation is assumed to be by vicariance separating a wide distribution into two mutually exclusive sets of areas. This event costs nothing.

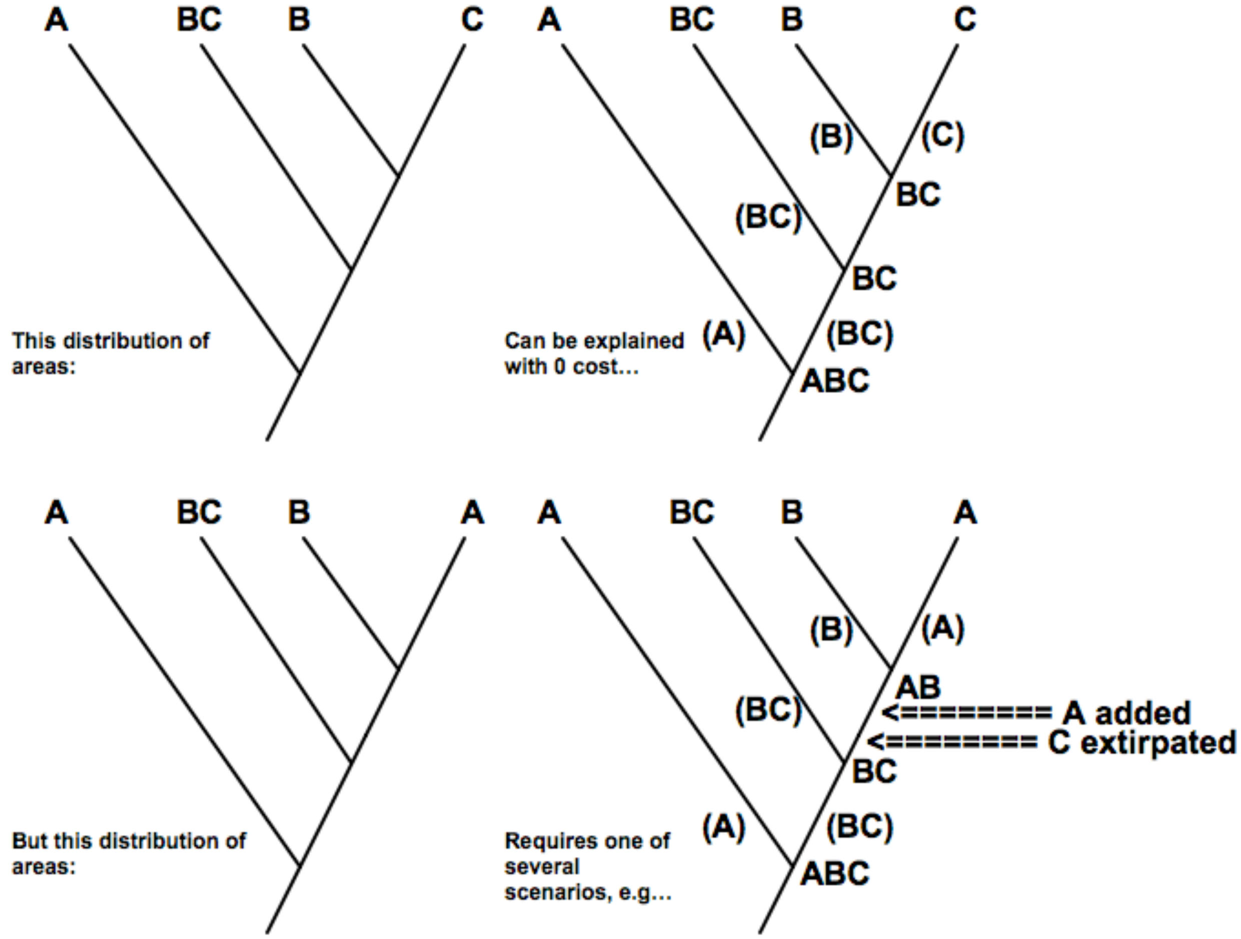
2. A species occurring in a single area may speciate within the area by allopatric (or possibly sympatric) speciation giving rise to two descendants occurring in the same area. The cost is zero.

3. Dispersal costs one per unit area added to a distribution.

4. Extinction costs one per unit area deleted from a distribution.

(DIVA manual, http://www.ebc.uu.se/systzoo/research/diva/manual/dmanual.html)

Here is a simple example of how DIVA would score two scenarios:


A
BC
B
C
This distribution of areas:
A
BC
B
C
(B)
(C)
BC
(BC)
BC
Can be explained with 0 cost...
(A)
(BC)
ABC
A
BC
B
A
But this distribution of areas:
A
BC
B
A
(B)
(A)
AB
(BC)
<========= A added
<========= C extirpated
BC
Requires one of several scenarios, e.g...
(A)
(BC)
ABC

Can you think of a lower cost scenario for the bottom phylogeny?

Here is one: **use all vicariance**, no extinction required. This is a key “feature” of DIVA: running on default costs, extinction is *never* inferred, because vicariance of a more widespread ancestor is always a cheaper (0 cost) explanation.

One way to think of what DIVA is doing is that it uses a 3-dimensional cost matrix to score different transitions, rather than the standard 2-dimensional matrix (Ronquist 1997):

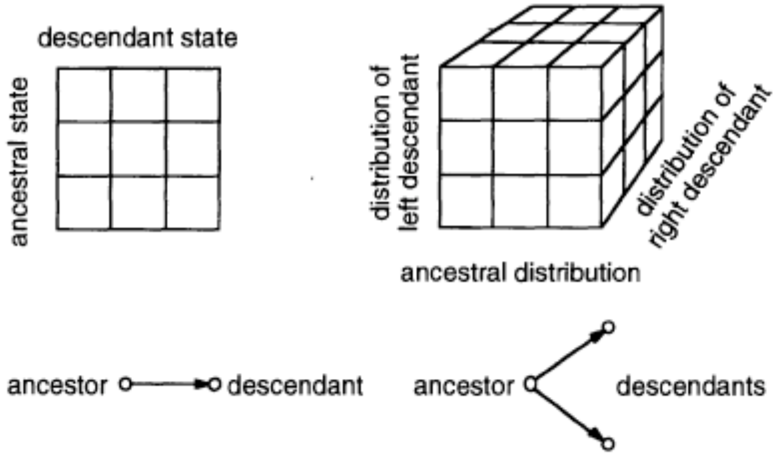
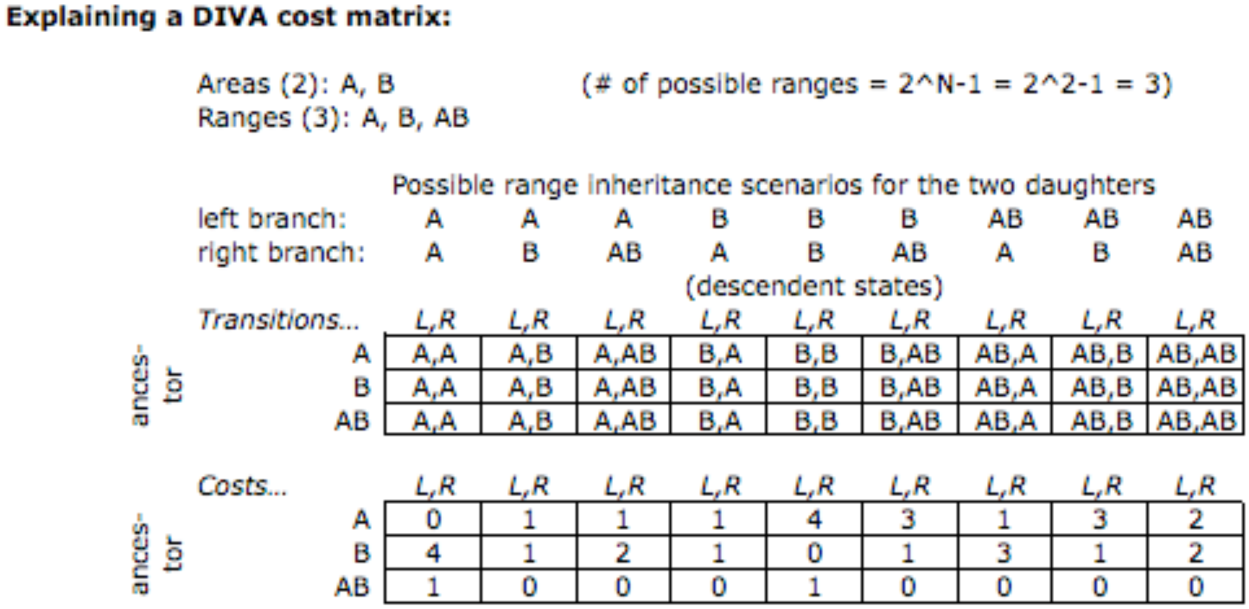


FIGURE 2. The difference between an ordinary step matrix and the cost matrix needed for the reconstruction of ancestral areas. An ordinary step matrix is two dimensional and specifies the cost of moving between states along an internode. The cost matrix used in dispersal–vicariance analysis is three dimensional and specifies the cost of combinations of ancestral, left descendant, and right descendant distributions.

…however, this may be more confusing than helpful. Any 3-D matrix can just be represented as a large 2-D matrix:



Explaining a DIVA cost matrix:

Areas (2): A, B (# of possible ranges = 2^N-1 = 2^2-1 = 3)
Ranges (3): A, B, AB

Possible range inheritance scenarios for the two daughters

left branch:		A	A	A	B	B	B	AB	AB	AB
right branch:		A	B	AB	A	B	AB	A	B	AB
						(descendent states)				
Transitions...		L,R	L,R	L,R	L,R	L,R	L,R	L,R	L,R	L,R
ances-tor	A	A,A	A,B	A,AB	B,A	B,B	B,AB	AB,A	AB,B	AB,AB
	B	A,A	A,B	A,AB	B,A	B,B	B,AB	AB,A	AB,B	AB,AB
	AB	A,A	A,B	A,AB	B,A	B,B	B,AB	AB,A	AB,B	AB,AB

Costs...		L,R	L,R	L,R	L,R	L,R	L,R	L,R	L,R	L,R
ances-tor	A	0	1	1	1	4	3	1	3	2
	B	4	1	2	1	0	1	3	1	2
	AB	1	0	0	0	1	0	0	0	0

(this is my initial guess at costs…should be close – Nick)

DIVA has had some significant uses, e.g. Donoghue & Smith (2004):

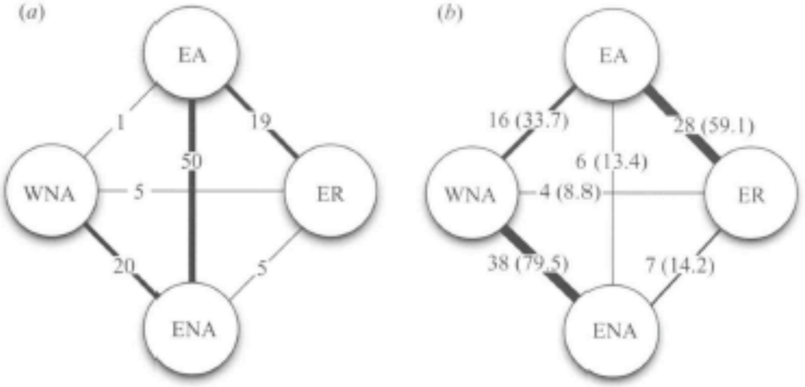


Figure 2. Comparison of disjunction patterns among the four major Holarctic areas of endemism (figure 1) for (*a*) plants (this study) and (*b*) animals (Sanmartín *et al.* 2001); line thickness is proportional to the percentage in each category. (*a*) Plants, from table 1; for 100 disjunctions the absolute number and percentage are equivalent. (*b*) Animals, redrawn from Sanmartín *et al.* (2001); the first number is the percentage of the total in that category; the number in parentheses is the absolute number from Sanmartín *et al.* (decimals result from the partitioning of reconstruction ambiguities in the original study; see § 2).

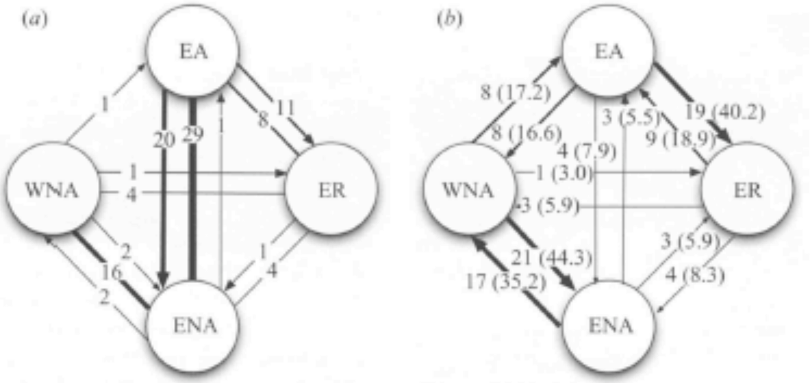


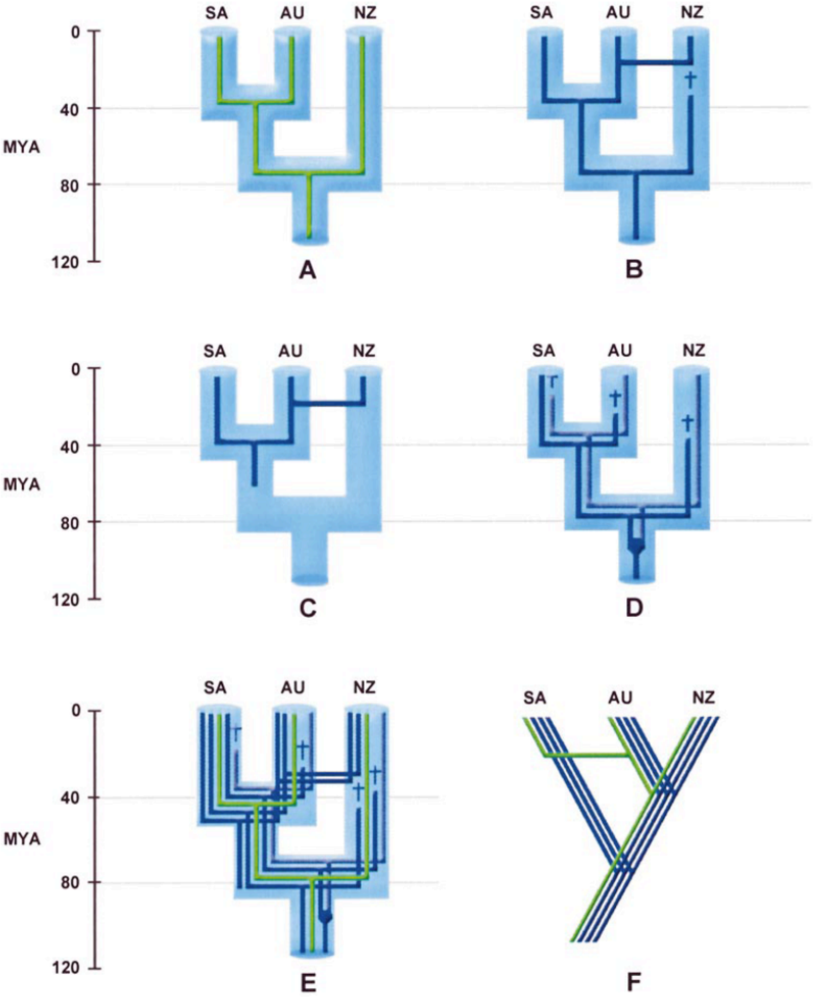
Figure 3. Inferred ancestral areas and directions of movement among the four major Holarctic areas of endemism (figure 1) for (*a*) plants (this study) and (*b*) animals (Sanmartín *et al.* 2001); line thickness is proportional to the percentage in each category, arrows point from the inferred ancestral area to the inferred derived area. (*a*) Plants, from table 1; for 100 disjunctions the absolute number and percentage are equivalent. Lines without arrowheads represent cases for which an unambiguous inference of ancestral area was not possible. (*b*) Animals, redrawn from Sanmartín *et al.* (2001); the first number is the percentage of the total in that category; the number in parentheses is the absolute number from Sanmartín *et al.* (decimals result from the partitioning of reconstruction ambiguities in the original study; see § 2).

LAGRANGE

On the other hand, the parsimony-based approach of DIVA was criticized heavily by Donoghue & Brian Moore:

Donoghue, M. J. and Moore, B. R., 2003. Toward an Integrative Historical Biogeography. *Integrative and Comparative Biology*. 43 (2), 261-270.

…which argued that biogeographical histories and patterns were not very useful without an explicit time component. E.g., the same pattern could be produced by different events and different times, and the available methods would not point this out. Congruence, typically taken as strong evidence of common history, could in biogeography very easily be due to "pseudocongruence." In addition, time estimates for biogeographic events were often either much too early or too late for the geological/climatic events that had been hypothesized to be behind inferred vicariance events (de Queiroz, 2005; Bush *et al.*, 2006).



From 2005-2008, Rick Ree, Stephen Smith, Brian Moore, and others have developed a maximum-likelihood method for inference in historical biogeography:

Ree, R. H., Moore, B. R., Webb, C. O. and Donoghue, M. J., 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*. 59 (11), 2299-2311.

Ree, R. H. and Smith, S. A., 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst Biol*. 57 (1), 4-14.

Moore, B. R., Smith, S. A., Ree, R. H. and Donoghue, M. J., 2009. Incorporating Fossil Data in Biogeographic Inference: A Likelihood Approach. *Evolution*. In press.

The currently available program “Lagrange” (**L**ikelihood **A**nalysis of **G**eographic **R**ange **E**volution) is from Ree & Smith (2008) (the 2005 version was very complex and much slower). The figures below are from this paper.

The Lagrange program takes as input:

1. an ultrametric phylogeny (nodes are dated)
2. locations of the tips
3. a list of possible ranges (area 1, area 2, area 1+2, etc.)
4. area adjacency matrix (which areas are connected such that they could share the same species)
5. dispersal matrix (relative probability of dispersal between regions; note that adjacent areas will not have a higher rate of dispersal unless you specify this explicitly here)

Unlike DIVA, which calculates the number of dispersal and extinction events and tries to minimize them, Lagrange works down the tree to calculate the relative likelihood of each possible ancestral range at each node, given a particular probability of dispersal and extinction. Here is the rate matrix:

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

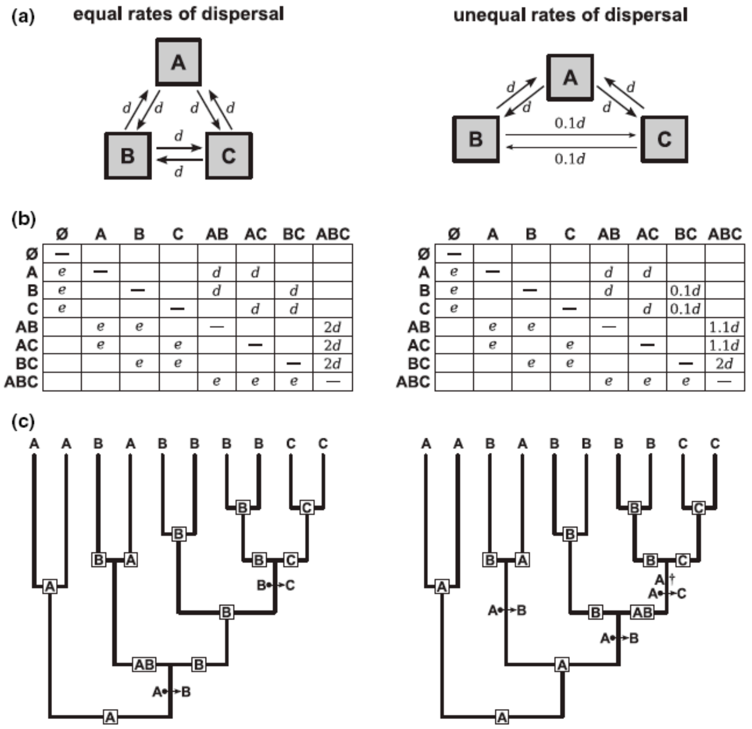
E1-E3 are instantaneous extinction rates (all the same in our example), the Ds are the instantaneous dispersal rates. This rate matrix is exponentiated to give the probability of change as a function of time (branch length):

$\mathbf{P}(t) = e^{-Qt}$

Thus, for an ancestral node, the likelihood of it being in Area 1 can be calculated given the ranges of the two daughter nodes, and their branch lengths (distance in time) to the ancestral node.

Using the above, the algorithm can calculate the likelihood for a whole history on a phylogeny, and then vary the extinction and dispersal parameters, calculate again, etc., optimizing for the ML estimates of dispersal and extinction rates. The output consists of the history resulting in the maximum likelihood, its log likelihood, and the estimated rates.

Via the input files, the user can prohibit certain histories (i.e., if an island doesn't exist at a certain point in time) or events (i.e., disallow certain dispersals), or put different relative probabilities on different dispersal events and then compare the likelihoods with a less constrained model.



(b) equal rates of dispersal

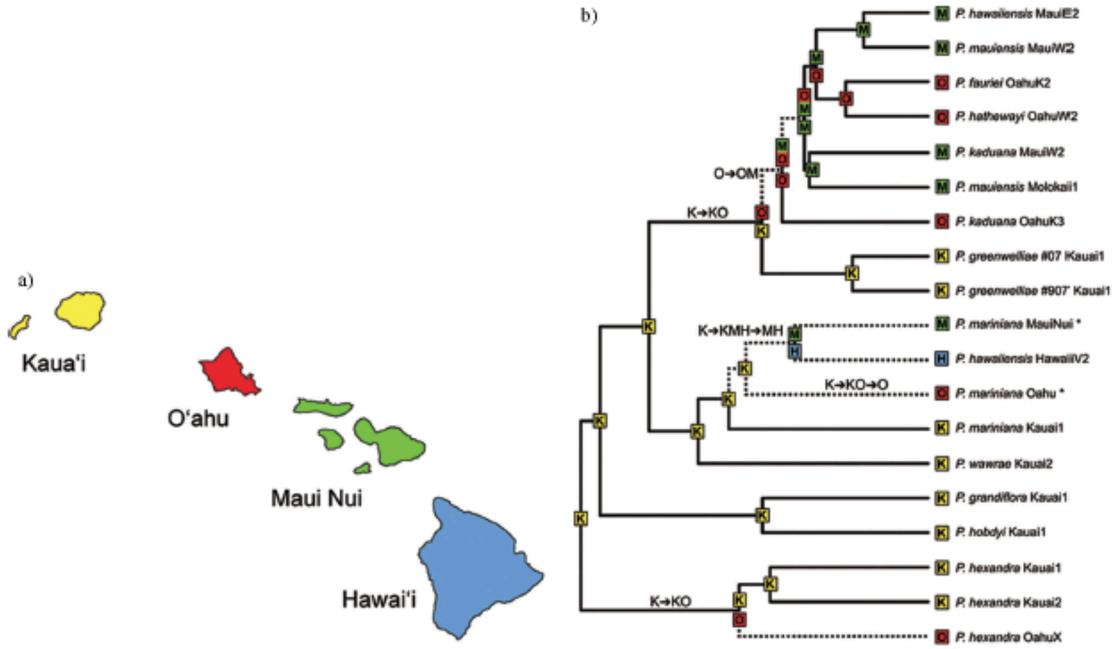
	Ø	A	B	C	AB	AC	BC	ABC
Ø	—							
A	*e*	—			*d*	*d*		
B	*e*		—		*d*		*d*	
C	*e*			—		*d*	*d*	
AB		*e*	*e*		—			2*d*
AC		*e*		*e*		—		2*d*
BC			*e*	*e*			—	2*d*
ABC					*e*	*e*	*e*	—

(b) unequal rates of dispersal

	Ø	A	B	C	AB	AC	BC	ABC
Ø	—							
A	*e*	—			*d*	*d*		
B	*e*		—		*d*		0.1*d*	
C	*e*			—		*d*	0.1*d*	
AB		*e*	*e*		—			1.1*d*
AC		*e*		*e*		—		1.1*d*
BC			*e*	*e*			—	2*d*
ABC					*e*	*e*	*e*	—

Figure 1 The effect of assuming equal and unequal (constrained) rates of dispersal on inferences of ancestral ranges and biogeographical events using DEC (dispersal, extinction and cladogenesis) models. The three component areas are labelled A, B and C; parameters for dispersal and local extinction rates are denoted by *d* and *e*, respectively. For each model, from top to bottom is shown (a) a schematic diagram of dispersal rates between areas, (b) the corresponding matrix of instantaneous transition rates between geographic ranges, and (c) a hypothetical phylogeny with observed species ranges, on which maximum likelihood ancestral ranges and implied dispersal and extinction events are mapped. The constrained model scales dispersal between B and C to one-tenth of the overall rate, as might be assumed based on relative distances. The transition matrices show rates between ranges separated by a single dispersal or extinction event. All other transitions have an instantaneous rate of zero, and elements along the diagonal are defined such that the sum of rates across a row is zero. For transitions involving dispersal, the rate is the sum of rates from areas in the starting range to the target area. At internal phylogenetic nodes, identical range inheritance is shown as a single area, whereas non-identical inheritance is shown by each daughter's range at the base of descendant branches, with the ancestral range being the union of these ranges. Dispersal events implied by ancestral ranges are shown by arrows between the source and destination areas; extinction is denoted by †. Assuming equal dispersal favours one dispersal event from A to B and one from B to C. Assuming a lower rate between B and C favours two dispersal events from A to B, one from A to C, and one local extinction event in A.

Here is Ree & Smith's inference for their example dataset, *Psychotria*, with an unconstrained model (no blockage of certain dispersals, range can be any combination of islands):



Here is the log likelihood of each possible ancestral range for the root of *Psychotria*, for the unconstrained (M0) and more constrained models:

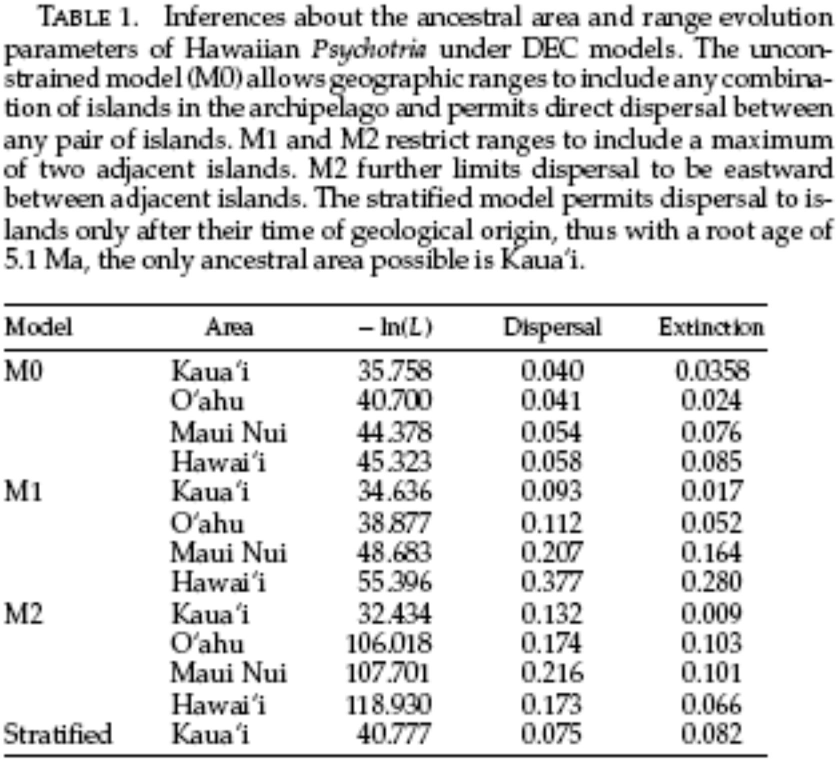


TABLE 1. Inferences about the ancestral area and range evolution parameters of Hawaiian *Psychotria* under DEC models. The unconstrained model (M0) allows geographic ranges to include any combination of islands in the archipelago and permits direct dispersal between any pair of islands. M1 and M2 restrict ranges to include a maximum of two adjacent islands. M2 further limits dispersal to be eastward between adjacent islands. The stratified model permits dispersal to islands only after their time of geological origin, thus with a root age of 5.1 Ma, the only ancestral area possible is Kaua'i.

Model	Area	− ln(*L*)	Dispersal	Extinction
M0	Kaua'i	35.758	0.040	0.0358
	O'ahu	40.700	0.041	0.024
	Maui Nui	44.378	0.054	0.076
	Hawai'i	45.323	0.058	0.085
M1	Kaua'i	34.636	0.093	0.017
	O'ahu	38.877	0.112	0.052
	Maui Nui	48.683	0.207	0.164
	Hawai'i	55.396	0.377	0.280
M2	Kaua'i	32.434	0.132	0.009
	O'ahu	106.018	0.174	0.103
	Maui Nui	107.701	0.216	0.101
	Hawai'i	118.930	0.173	0.066
Stratified	Kaua'i	40.777	0.075	0.082

Bayesian methods

In the last year or two, some Bayesian approaches have been tried. Basically, they consist of:

(1) Running DIVA (Nylander et al. 2008) or Lagrange (Smith, unpublished, Evolution2009 talk) on a collection of MrBayes trees.

(2) Requiring that lineages occupy a single area, and treating lineage location as a character with several character states, as with e.g. DNA, then optimizing it in MrBayes (Sanmartin et al. 2008):

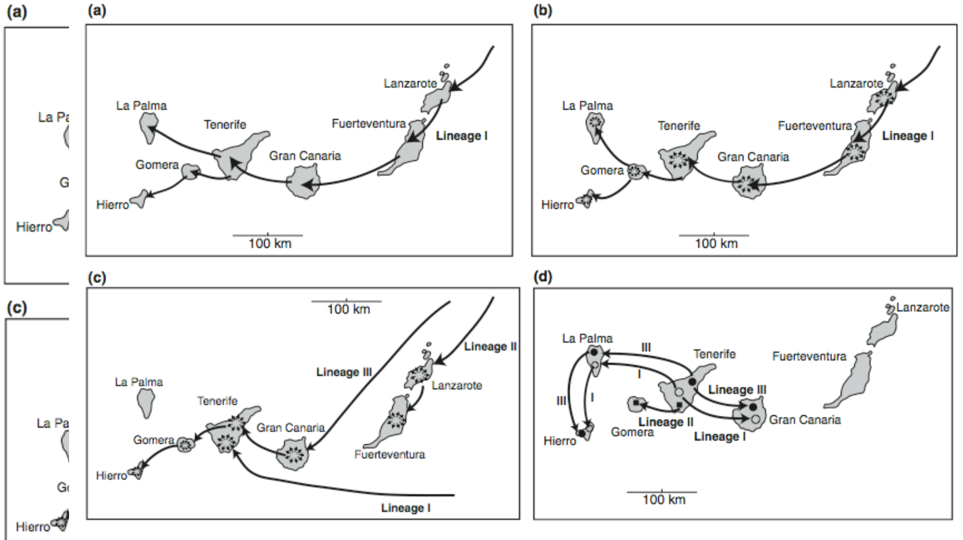


Figure 5 Some common modes of species diversification in Canarian animals and plants. (a) Model I: Stepwise colonization with concomitant speciation resulting in a single species on each island. (b) Model II: Stepwise colonization with speciation followed by within-island speciation; each species has its closest relative in the same island. (c) Model III: Multiple independent colonization events from the mainland (or even back-colonization events of continental areas), followed by within-island speciation. (d) Model IV: Inter-island colonization between similar ecological habitats; each species has its closest relative in a different island but occupying a similar habitat (geometric symbols). The last mode of speciation is common in plants.

Figure 5
concomita
within-isla
the mainland (or even back-colonization events of continental areas), followed by within-island speciation. (d) Model IV: Inter-island colonization between similar ecological habitats; each species has its closest relative in a different island but occupying a similar habitat (geometric symbols). The last mode of speciation is common in plants.

Treat area like a character state in DNA (Ronquist & Sanmartin 2008). Here, the parameters are:

"Carrying capacity" of each island (equivalent to DNA base frequencies)

Dispersal rate between each island (equivalent to transition probability)

Parameters can be set to all be equal (like a DNA Jukes-Cantor model)
Or all different (like a DNA GTR model)

Extinction not explicit in model

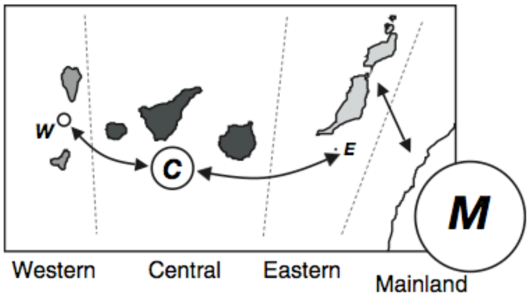


Figure 6 Estimated relative carrying capacities for the Canary Islands based on a data set of 13 Canarian plant and animal phylogenies and using the Equal-in step Bayesian island model. The Canary Islands were divided into three island-groups ('Eastern', Lanzarote and Fuerteventura; 'Central', Gran Canaria, Tenerife and La Gomera; and 'Western', La Palma and El Hierro); 'Mainland' represents non-Canarian distributions (continental areas and Macaronesia). The size of the circles is roughly proportional to the estimated relative carrying capacity for each island-group (see Table 4). The arrow width represents the relative dispersal rate, here 1/3 because the dispersal rate is the same for all island groups and dispersal is only allowed between adjacent island groups ('step model', see Fig. 2e).

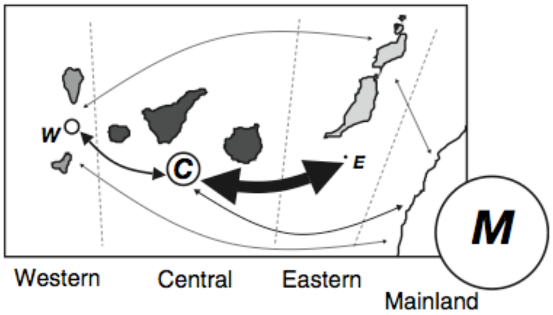


Figure 7 Relative carrying capacities and dispersal rates estimated for the GTR Bayesian island model based on a data set of 13 Canarian plant and animal phylogenies (see Fig. 6 for further explanation).

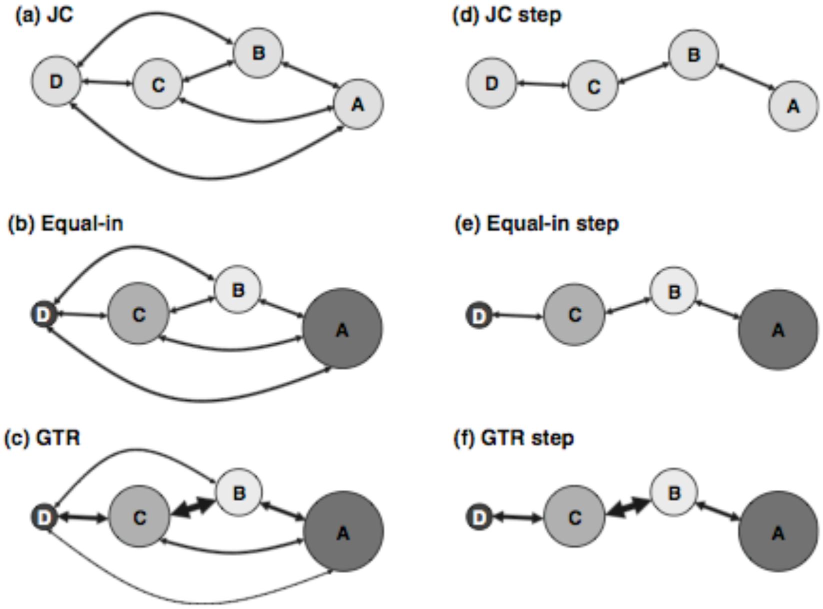


Figure 2 Bayesian Island Models: Each circle represents an island; circle size represents the relative carrying capacity of the island (expected number of lineages at equilibrium); arrow width represents the relative dispersal rate between two single islands. (a) Jukes–Cantor (JC) model: all carrying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

Comparison of methods (Clark et al. 2008)

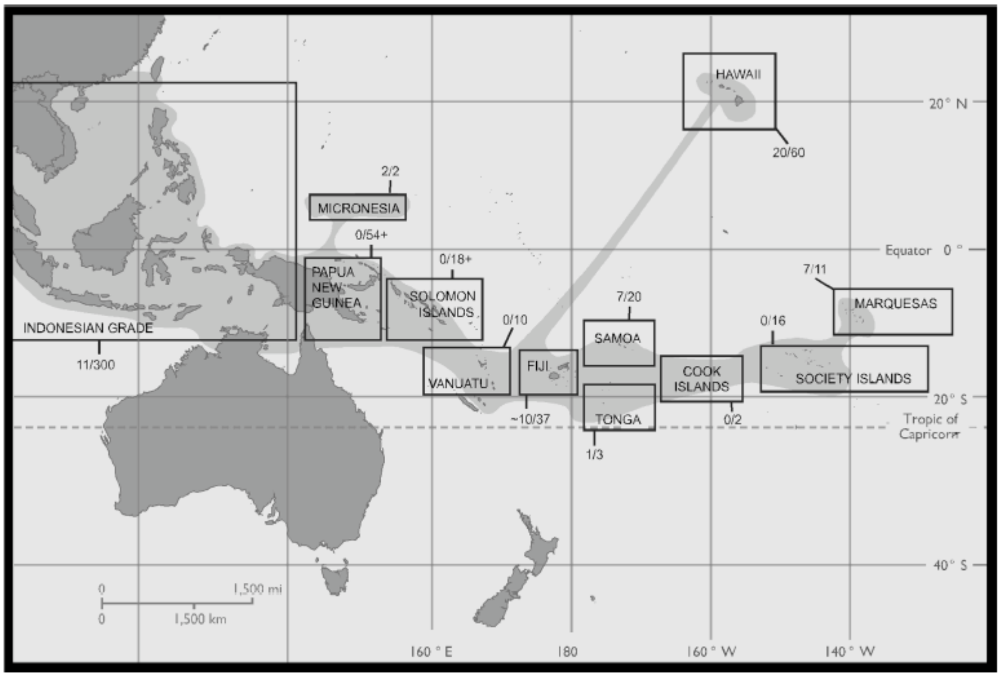


FIGURE 1. Southeast Asian and Pacific distribution of *Cyrtandra*. Numbers before the forward slash are approximate number of species sampled in this study; numbers after slash are conservative estimates for species numbers in the defined areas based on herbarium records (Skog and Boggan, 2007).

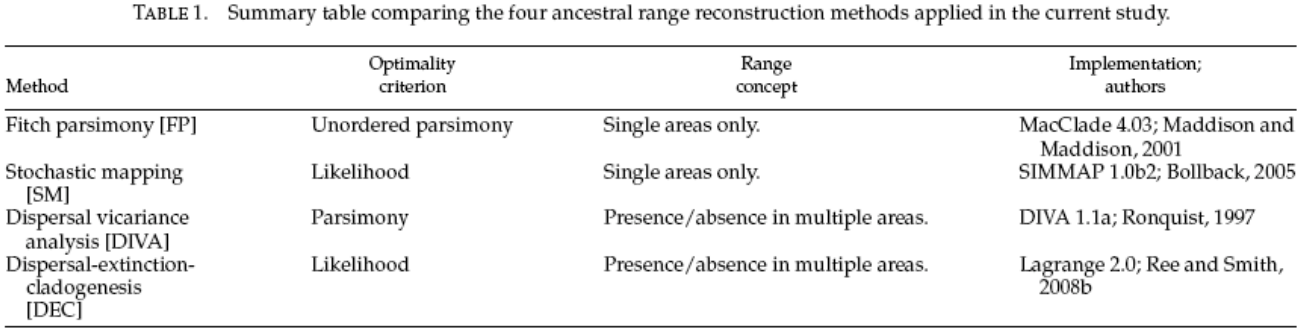


TABLE 1. Summary table comparing the four ancestral range reconstruction methods applied in the current study.

Method	Optimality criterion	Range concept	Implementation; authors
Fitch parsimony [FP]	Unordered parsimony	Single areas only.	MacClade 4.03; Maddison and Maddison, 2001
Stochastic mapping [SM]	Likelihood	Single areas only.	SIMMAP 1.0b2; Bollback, 2005
Dispersal vicariance analysis [DIVA]	Parsimony	Presence/absence in multiple areas.	DIVA 1.1a; Ronquist, 1997
Dispersal-extinction-cladogenesis [DEC]	Likelihood	Presence/absence in multiple areas.	Lagrange 2.0; Ree and Smith, 2008b

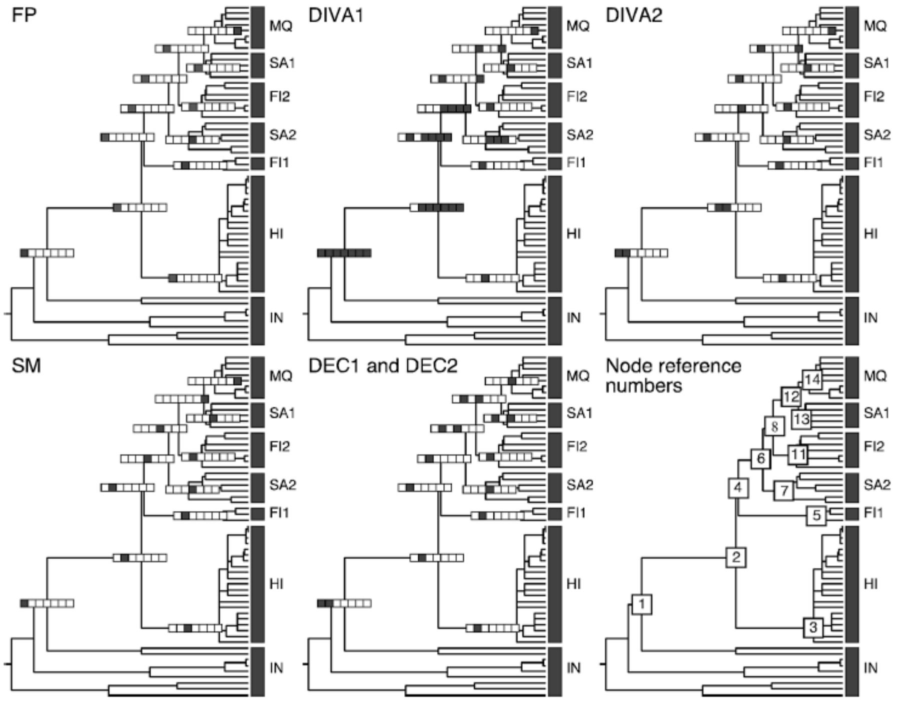


FIGURE 4. Summary and comparison of results from the four ancestral range reconstruction methods (details in Table 3). Area reconstructions are represented by open or shaded blocks (open=absent; shaded=present) in the order of Indonesian grade, Fiji, Hawaii, Samoa, Tonga, Micronesia, Marquesas. In instances of more than one reconstruction, only the first reconstruction is shown for simplicity. FP = Fitch parsimony, DIVA1 = dispersal vicariance analysis (unrestricted), DIVA2 = dispersal vicariance analysis (restricted to ≤ 2 areas per node), SM = stochastic mapping; DEC1 = dispersal-extinction-cladogenesis (unrestricted), DEC2 = dispersal-extinction-cladogenesis (restricted to ≤ 2 areas per node). Node reference numbers are those nodes referred to in Figures 2 and 3 and in the text.

Major issues in estimating biogeographic histories with event-based methods:

1. Both DIVA and Lagrange tend to overestimate ancestral ranges, the further back you go down the phylogeny, although this can be limited by manually setting a limit on the maximum ancestral range size.

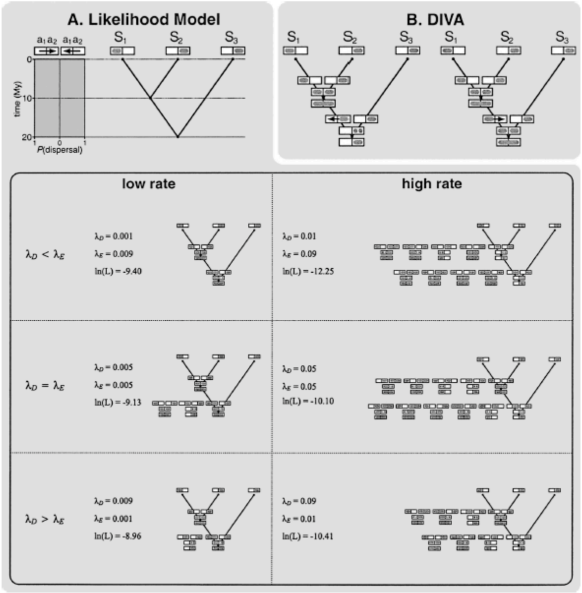


FIG. 4. (A) Likelihood inference of lineage geohistory for a hypothetical three-species phylogeny. The model includes two areas (a_1 and a_2) sharing a single connection, through which the probability of successful dispersal is constant over time and symmetrical with respect to direction. Likelihood estimates of ancestral ranges and subdivision-inheritance scenarios were obtained for different relative dispersal rates ($\lambda_D < \lambda_E$, $\lambda_D = \lambda_E$, and $\lambda_D > \lambda_E = 0.1$) and two overall rates, "low" ($\lambda_D + \lambda_E = 0.01$) and "high" ($\lambda_D + \lambda_E = 0.1$). Adjacent to every internal node are scenarios within two log-likelihood units of the maximum, ordered left to right from lowest to highest likelihood. The highest overall likelihood is obtained at a low rate with $\lambda_D > \lambda_E$. (B) Ancestral ranges and implied speciation scenarios inferred with dispersal-vicariance analysis (DIVA). The two equally most-parsimonious reconstructions correspond to the plausible scenarios inferred from the model parameters yielding the highest likelihood, although in the latter case a widespread ancestor at the root is favored.

2. It seems pretty crude to estimate ancestral ranges simply by a few huge areas. Could we do better, i.e. with an approach modeled on species distribution modeling?

3. Approaches that allow any combination of areas have the problem that the number of effective character states exponentially increases the memory and processing time of the algorithm. I.e., computation time is proportional to 2^N-1. For only 10 regions, there are 2^10-1 possible ranges for which costs or likelihoods have to be calculated. There are some tricks to improve things a bit, and one can manually set an upper limit on number of ranges but none can overcome the fundamental issue.

In practice, 6 or 7 ranges is about the absolute limit of what you can analyze in a reasonable amount of time in Lagrange. Above that, you are forced to merge ranges etc.

4. Neither method takes into account fossils. There is allegedly a paper in press which will attempt this, but it has yet to come out, and the online draft indicates there is room for more approaches. One might be to allow fossil lineages to go from a range of (AB) to (A) to (), i.e. have local extirpations add up to a global extinction.

5. This, by the way, would help fix another major issue with Lagrange, which is that it typically infers near-zero extinction rates.

6. Plate tectonics has not really been incorporated in a satisfactory way. Ideally, a "continuous" plate tectonic history estimate would be part of the input into the model, and then e.g. dispersal probability would be a function of distance, climate zones that need to be crossed, etc.

7. The fundamental difficulty with all of these methods is the limited amount of data upon which to estimate models with many parameters. With DNA, for a 100 species phylogeny and 1000-base alignment, we have, in a sense, 1000 repetitions of a DNA evolution experiment. From this we can estimate many parameters. With the biogeography character, though, we have only 1 repetition.

   This might be improved by sharing parameters across many clades, and using a Bayesian clustering method to reduce the number of parameters (work with Michael Landis, Ginger Jui).

New ideas

Geophylogenies (David Kidd)

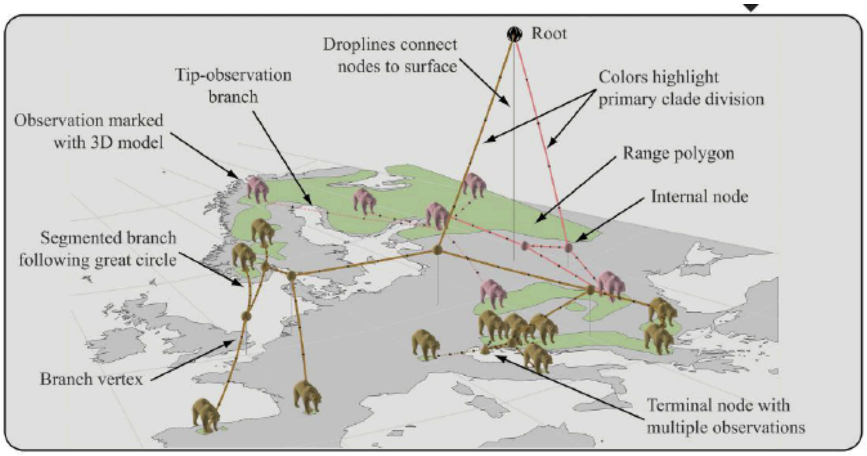
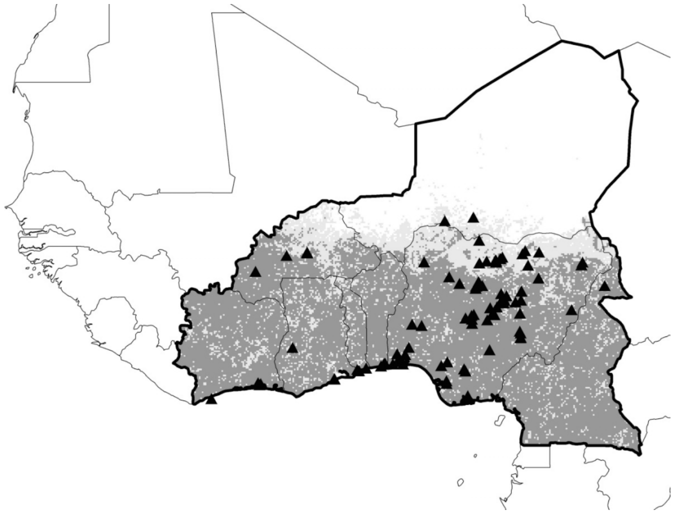

Root
Droplines connect nodes to surface
Tip-observation branch
Colors highlight primary clade division
Observation marked with 3D model
Range polygon
Internal node
Segmented branch following great circle
Branch vertex
Terminal node with multiple observations

FIGURE 1. Building the *Ursus arctos* geophylogeny. The tree, the location map, and $n : m$ link table between tree and locations are digitized. Terminal nodes are placed at the centroid of the spatial envelope of the locations where the tip was observed, and internal nodes are similarly placed at the centroid of the envelope of daughter node positions. Branches trace the shortest path between locations (great-circles). Color is used to highlight difference in the geographical spread of the two deepest clades. Some aspects of the geophylogeny and visualization mentioned in the text are labeled in the lower panel.

Strengths: sexy graphics
Weakness: not really an inference method

Inference in continuous space (Nick's new kick)

A totally different approach to all of the above would be to attempt to map ancestral ranges in continuous space, rather than in a small number of discrete regions – as is done currently with "species distribution modeling" and "ecological niche modeling".



Regional projection across West Africa of HPAI-H5N1 ecological niche model. Results based on OIE case occurrence points and environmental layers for the Middle East and northeastern Africa. Model predictions are shown as ramps of model agreement in predictions: light grey = 5–9 models predict potential presence, dark grey = all models agree in predicting potential presence. Black diamonds indicate independent test data (N = 101) from the region [10,11]. Study area is delineated by bold border.

Williams and Peterson (2009), *International Journal of Health Geographics* 8:47 doi: http://dx.doi.org/10.1186/1476-072X-8-47

Why don't we have "Lineage Distribution Modeling?"

What would it take?

1. Climate predictors for current species distributions can be modeled for currently extant species, using a variety of approaches (MaxEnt, etc.)

2. Inferring the climatic preferences of ancestral lineages is basically a matter of standard ancestral character estimation. This is being worked on, e.g. phyloclim package, and:

Evans, M. E. K., S. A. Smith, R. S. Flynn, and M. J. Donoghue. 2009. Climate, niche evolution, and diversification of the 'bird-cage evening primroses' (Oenothera, sections Anogra and Kleinia). *Am. Nat.* 173: 225-240.

3. Such models produce estimates of probability of presence at any given pixel.

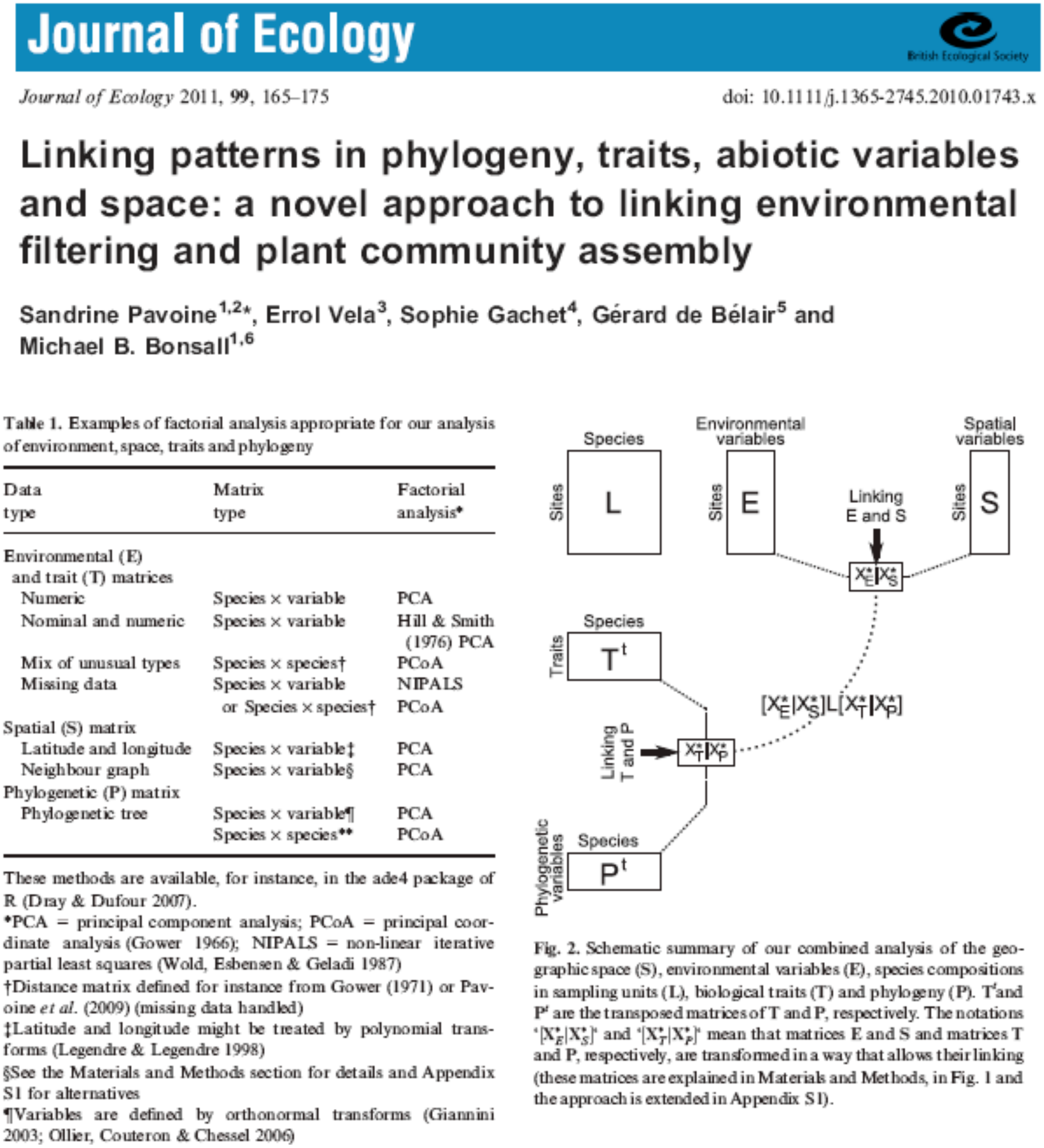
What has not been worked on much, if at all:

4. This approach can also incorporate fossils, and the problem of detection probability. There is a lot of work on this in the ecological literature, but little to none in historical biogeography.

5. Spatial and phylogenetic autocorrelation in location can be incorporated to improve ancestral range estimates (i.e., if you have an observation of a lineage nearby in space or time, the chance of it or a close relative existing nearby is higher than it would be otherwise). ("Phylogenetic kriging")

6. Spatial autocorrelation does not necessarily have to use euclidean distance as the measurement of space. Connectivity between continents might be modeled with e.g. path networks.

The community assembly literature should serve as inspiration:



Journal of Ecology

British Ecological Society

Journal of Ecology 2011, **99**, 165–175 doi: 10.1111/j.1365-2745.2010.01743.x

Linking patterns in phylogeny, traits, abiotic variables and space: a novel approach to linking environmental filtering and plant community assembly

Sandrine Pavoine[1,2*], Errol Vela[3], Sophie Gachet[4], Gérard de Bélair[5] and Michael B. Bonsall[1,6]

Table 1. Examples of factorial analysis appropriate for our analysis of environment, space, traits and phylogeny

Data type	Matrix type	Factorial analysis*
Environmental (E) and trait (T) matrices		
Numeric	Species × variable	PCA
Nominal and numeric	Species × variable	Hill & Smith (1976) PCA
Mix of unusual types	Species × species†	PCoA
Missing data	Species × variable or Species × species†	NIPALS PCoA
Spatial (S) matrix		
Latitude and longitude	Species × variable‡	PCA
Neighbour graph	Species × variable§	PCA
Phylogenetic (P) matrix		
Phylogenetic tree	Species × variable¶ Species × species**	PCA PCoA

These methods are available, for instance, in the ade4 package of R (Dray & Dufour 2007).
*PCA = principal component analysis; PCoA = principal coordinate analysis (Gower 1966); NIPALS = non-linear iterative partial least squares (Wold, Esbensen & Geladi 1987)
†Distance matrix defined for instance from Gower (1971) or Pavoine *et al.* (2009) (missing data handled)
‡Latitude and longitude might be treated by polynomial transforms (Legendre & Legendre 1998)
§See the Materials and Methods section for details and Appendix S1 for alternatives
¶Variables are defined by orthonormal transforms (Giannini 2003; Ollier, Couteron & Chessel 2006)

Fig. 2. Schematic summary of our combined analysis of the geographic space (S), environmental variables (E), species compositions in sampling units (L), biological traits (T) and phylogeny (P). T^t and P^t are the transposed matrices of T and P, respectively. The notations '$[X_E^*|X_S^*]$' and '$[X_T^*|X_P^*]$' mean that matrices E and S and matrices T and P, respectively, are transformed in a way that allows their linking (these matrices are explained in Materials and Methods, in Fig. 1 and the approach is extended in Appendix S1).

References

Bush, M. B., Gosling, W. D. and Colinvaux, P. A., 2006. Climate change in the lowlands of the Amazon Basin in: Flenley, J. R. and Bush, M. B. (Eds.), *Tropical Rainforest Responses to Climatic Change*, USA and UK: Springer, jointly published with Praxis Publishing, UK, pp. 55–79.

Clark, J. R., Ree, R. H., Alfaro, M. E., King, M. G., Wagner, W. L. and Roalson, E. H. (2008). "A comparative study in ancestral range reconstruction methods: retracing the uncertain histories of insular lineages." *Syst Biol*. **57**(5): 693-707.

de Queiroz, A., 2005. The resurrection of oceanic dispersal in historical biogeography. *Trends Ecol Evol*. 20 (2), 68-73.

Donoghue, M. J. and Moore, B. R., 2003. Toward an Integrative Historical Biogeography. *Integrative and Comparative Biology*. 43 (2), 261-270.

Donoghue, M. J. and Smith, S. A., 2004. Patterns in the assembly of temperate forests around the Northern Hemisphere. *Philosophical Transactions of the Royal Society of London B Biological Sciences*. 359 (1450), 1633-1644.

Moore, B. R., Smith, S. A., Ree, R. H. and Donoghue, M. J., 2009. Incorporating Fossil Data in Biogeographic Inference: A Likelihood Approach. *Evolution.* In press

Nylander, J. A., Olsson, U., Alstrom, P. and Sanmartin, I. (2008). "Accounting for phylogenetic uncertainty in biogeography: a Bayesian approach to dispersal-vicariance analysis of the thrushes (Aves: Turdus)." *Syst Biol*. **57**(2): 257-268.

Ree, R. H., Moore, B. R., Webb, C. O. and Donoghue, M. J., 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*. 59 (11), 2299-2311.

Ree, R. H. and Smith, S. A., 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst Biol*. 57 (1), 4-14.

Ronquist, F., 1996. DIVA version 1.1. Computer program and manual, accessed online. URL: http://www.ebc.uu.se/systzoo/research/diva/diva.html.

Ronquist, F., 1997. Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography. *Syst Biol*. 46 (1), 195-203.

Sanmartin, I. and Ronquist, F., 2004. Southern hemisphere biogeography inferred by event-based models: plant versus animal patterns. *Syst Biol*. 53 (2), 216-243.

Sanmartin, I., Van der Mark, P. and Ronquist, F. (2008). "Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands." *Journal of Biogeography*. **35**(3): 428-449.

Smith, Stephen A. (2009). "Taking into account phylogenetic and divergence-time uncertainty in a parametric biogeographical analysis of the Northern Hemisphere plant clade Caprifolieae." *Journal of Biogeography*. **36**(12): 2324-2337.