thesis abstract

Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing

Nicholas Joseph Matzke

PhD Thesis, Department of Integrative Biology, University of California, Berkeley. 3060 Valley Life Sciences Building, Berkeley, CA, 94720; matzke@berkeley.edu.

Current address: National Institute for Mathematical and Biological Synthesis (NIMBioS, www.nimbios.org). 1122 Volunteer Blvd., Suite 106, University of Tennessee, Knoxville, TN 37996-3410; matzke@nimbios.org.

Abstract. Historical biogeography has been characterized by a large diversity of methods and unresolved debates about which processes, such as dispersal or vicariance, are most important for explaining distributions. A new R package, BioGeoBEARS, implements many models in a common likelihood framework, so that standard statistical model selection procedures can be applied to let the data choose the best model. Available models include a likelihood version of DIVA (“DIVALIKE”), LAGRANGE’s DEC model, and BAYAREA, as well as “+J” versions of these models which include founder-event speciation, an important process left out of most inference methods. I use BioGeoBEARS on a large sample of island and non-island clades (including two fossil clades) to show that founder-event speciation is a crucial process in almost every clade, and that most published datasets reject the non-J models currently in widespread use. BioGeoBEARS is open-source and freely available for installation at the Comprehensive R Archive Network at http://CRAN.R-project.org/package=BioGeoBEARS. A step-by-step tutorial is available at http://phylo.wikidot.com/biogeobears.

Keywords. cladogenesis, DIVA, founder-event speciation, historical biogeography, jump dispersal, LAGRANGE, phylogenetics

The state of the field

The methods employed in historical biogeography are very diverse and include historical narrative, panbiogeography (Heads 2012, Waters et al. 2013), cladistic biogeography, multistate character methods, and ancestral state methods specialized for biogeography. The latter methods, which have become very popular, and been used and cited in hundreds of published analyses, include the parsimony-based Dispersal-Vicariance Analysis (DIVA; Ronquist 1997), and the likelihood-based Dispersal-Extinction Cladogenesis (DEC) model of the LAGRANGE program (Ree 2005, Ree and Smith 2008). A variety of new methods have also recently been proposed, including pseudo-Bayesian versions of DIVA and LAGRANGE (e.g., Wood et al. 2013). Another new method is BayArea, a Bayesian technique which samples geographical histories along phylogenetic branches jointly with the sampling of parameter values (Landis et al. 2013).

Each method described above relies on some model of geographic range evolution, explicitly or implicitly, and therefore makes some assumption about the processes that have produced the geographic ranges of observed taxa. These assumptions about process typically have a much larger impact on conclusions about biogeographical history than differences in statistical procedure (e.g., parsimony, likelihood, or Bayesian inference). However, thus far there has been no method to determine which processes are most important, nor to determine which available model best fits the geographical and phylogenetic data for any particular clade. This was the problem I addressed in my thesis (Matzke 2013a).

BioGeoBEARS: model testing and ancestral state estimation in historical biogeography

I have created an R package, "BioGeoBEARS" (Matzke 2013b) that implements in a likelihood framework several popular models, such as the LAGRANGE DEC model, a likelihood version of DIVA (here termed "DIVALIKE"), and a likelihood version of the range evolution model assumed by methods such as BayArea program and the Bayesian Binary Model (BBM) of RASP (Yu et al. 2013). BayArea and BBM have no special range evolution process occurring at cladogenesis, and this assumption is implemented in BioGeoBEARS in the "BAYAREA" model. Figure 1 shows the processes assumed by each major historical biogeography method, followed by the parameters of BioGeoBEARS which control the probability of each process. Setting a parameter to 0 allows the researcher to "turn off" the corresponding process. Alternatively, parameters may be fixed or set to be free parameters, to implement various models.

Figure 1 shows that DEC, DIVALIKE, and BAYAREA each have two free parameters (*d* and *e*) specifying the rate of "dispersal" (range expansion) and "extinction" (range contraction) along phylogeny branches. At cladogenesis events, DEC assumes that daughter lineages inherit the ancestral range if the ancestor lives in a single area (e.g., A->A,A), or if the ancestor is widespread, one daughter lineage will live in a subset of this area (ABCD->A, ABCD), or one area will split off by vicariance (ABCD->A,BCD). DEC assumes that one daughter lineage will always have a range of one area; thus ABCD->AB,CD is disallowed. DIVA, on the other hand, allows this form of vicariance, but disallows subset speciation (Ronquist and Sanmartín 2011). BayArea, for reasons of computational simplicity, assumes that no range evolution occurs at cladogenesis; the ancestral range is copied to both daughters (e.g., A->A,A; ABCD->ABCD,ABCD). These assumptions were hard-coded into each of the original source programs, and thus it can be difficult for researchers to realize the impact they are having. For example, all three of these programs leave out the process of founder-event speciation (Kodandaramaiah 2010, Buerki et al. 2011, Goldberg et al. 2011; also see below), where at cladogenesis one daughter lineage jumps to a new range outside the range of the ancestor (e.g., ABCD->ABCD, E). In BioGeoBEARS, founder-event speciation can be added to any of the previously-described models, and its relative probability can be hard-coded by the researcher, or can be left as a free parameter which is estimated from the data, creating "DEC+J", "DIVALIKE+J", and "BAYAREA+J" models.

Implementation in a common likelihood framework means that standard statistical techniques (Burnham and Anderson 2002) such as the Likelihood Ratio Test (LRT) and Akaike Information Criterion (AIC) may be used to directly compare how well the different biogeography models fit the data. LRT may be used to compare nested models—e.g., DEC nests within DEC+J, because the DEC+J model is identical to DEC when the *j* parameter is fixed to 0. AIC may be used to compare both nested and non-nested models.

BioGeoBEARS also implements the accessory features available in LAGRANGE, such as user-specified dispersal matrices controlling the relative probability of dispersal between regions, and time-stratification, where different dispersal matrices exist as geography changes over geological time. These may be used with all models; currently, dispersal probability modifiers are applied in the same way to traditional dispersal events and to founder-event speciation events. BioGeoBEARS also adds a number of features not previously available in most historical biogeography software, such as distance-based dispersal, a model of imperfect detection, the ability to simulate biogeographical evolution under each model, and the ability to include fossils either as ancestors or tips on a time-calibrated tree.

To improve computational speed, I wrote the accessory R packages "rexpokit" and "cladoRcpp" to conduct rapid probability calculations for the anagenetic and cladogenetic processes, respectively. These are particularly useful for analyses with large numbers of areas, where the state space and transition matrices rapidly become huge (see Figure 1 caption). Thus, for large analyses, BioGeoBEARS is much faster than

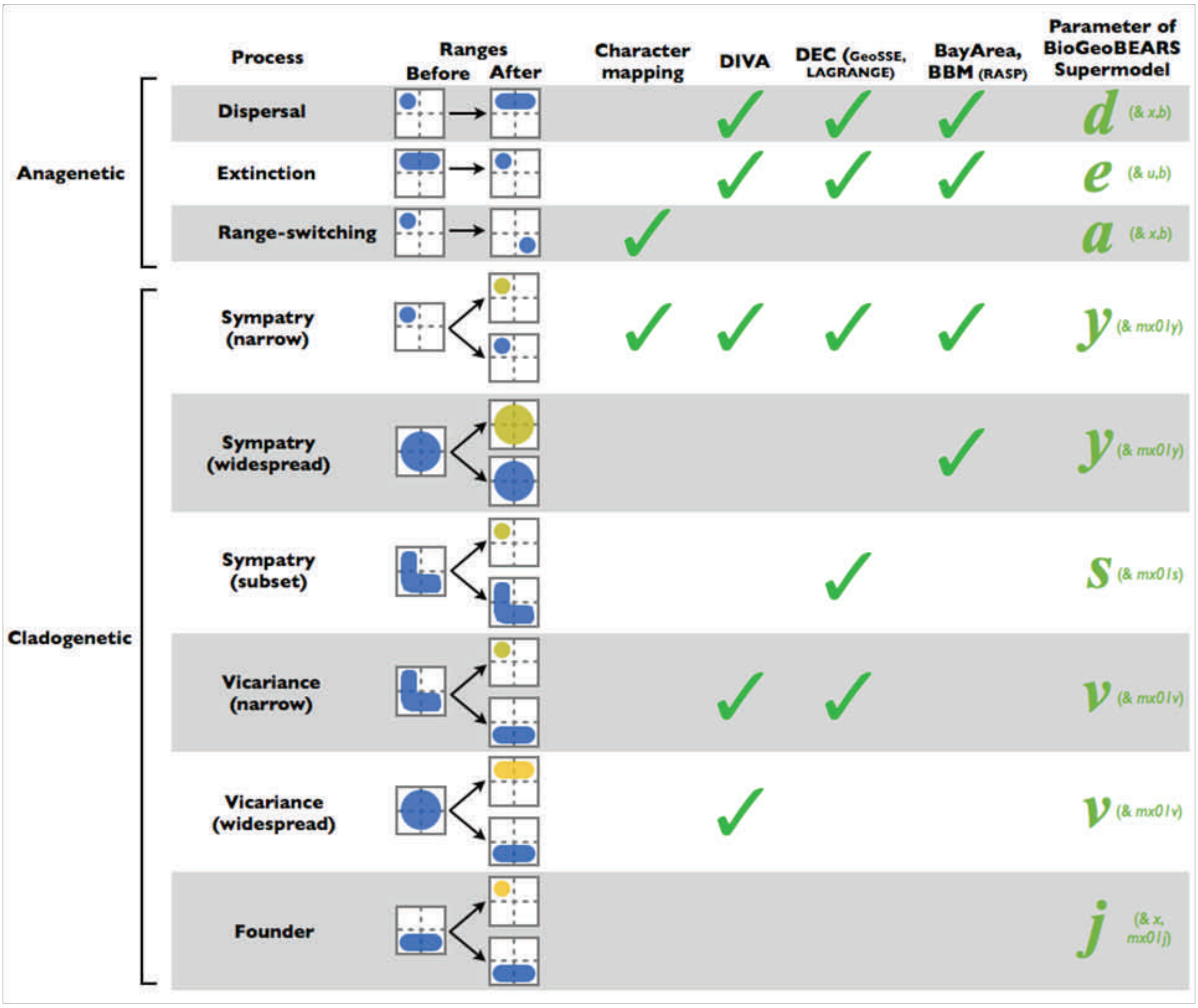

Figure 1. The processes assumed by different historical biogeography methods. Each of these processes is controlled by the specified parameter(s) in the BioGeoBEARS supermodel, allowing them to be turned on or off, or estimated from the data. Note that whether or not the data support a particular free parameter is an empirical question that should be tested with model choice procedures. Note also that this graphic deals only with the range-changing processes assumed by the different methods. BioGeoBEARS does not attempt to replicate the parsimony aspect of DIVA, just the processes allowed by DIVA, using "DIVALIKE." Similarly, BioGeoBEARS does not yet implement the "SSE" (state-based speciation and estimation rates) features of the GeoSSE model (Goldberg et al. 2011) of diversity. The ClaSSE model (Goldberg and Igić 2012) can in theory use a parameter to represent the probability of each possible combination of ancestor range, left descendant range, and right descendant range. In that sense ClaSSE is the ultimate supermodel, although users would have to develop their own parameterizations to produce a reasonable biogeographic model, and the number of parameters inflates dramatically with number of areas — on defaults, 9 areas means 2^9-1=511 possible ranges, and this means 511x511x511 = 133,432,831 possible combinations of ancestor/left descendant/right descendant. The cladoRcpp R package, a dependency of BioGeoBEARS, is designed to efficiently calculate probabilities for these combinations, under the implemented biogeography models.

Python LAGRANGE, and begins to approach C++ LAGRANGE in speed. However, R is well-known to be an inefficient language, so for small analyses, BioGeoBEARS will be slower, although small analyses will still complete in seconds or minutes. BioGeoBEARS uses the “snow” R package to allow multicore processing, so that a computer with four cores will run a BioGeoBEARS analysis approximately four times faster.

Formal model testing of the Dispersal-Extinction-Cladogenesis (DEC) model reveals importance of founder-event speciation

The process of founder-event speciation is widely thought to be crucial especially in island systems

(Cowie and Holland 2006, Templeton 2008, Gillespie et al. 2012), although the proposition receives vocal dissent from biogeographers that emphasize vicariance (Santos 2007, Heads 2012). However, probably for historical reasons (DEC was inspired by DIVA, and DIVA descends from maximum-vicariance methods), founder-event speciation has been left out of currently available historical biogeography software. To test the assumption that founder-event speciation does not occur, BioGeoBEARS was used to compare the DEC model (2 parameters, *d* and *e*) with the DEC+J model (3 parameters, adding a *j* parameter controlling the relative probability of founder-event speciation). I first validate BioGeoBEARS by showing that it exactly reproduces the log-likelihoods and parameter inferences made by the LAGRANGE DEC model on the LAGRANGE test dataset (Ree and Smith 2008) of the Hawaiian *Psychotria* clade. I further validate the method by taking the *Psychotria* phylogeny and simulating geographic range evolution under the DEC and DEC+J models, and then conducting inference under the two models. Model choice using LRT is highly accurate, with false positive and false negative rates of approximately 5%, indicating that the test has the desired frequentist properties, and also indicating that DEC and DEC+J are easy to distinguish from the data, even on a small phylogeny. The simulation results also indicate that when DEC+J is the true model, DEC+J has 87% accuracy in inferring ancestral states, while DEC has only 57% accuracy.

The DEC and DEC+J models are then applied to 13 island clades, most of them classic Hawaiian study systems (*Drosophila*, silverswords, etc.), under a variety of dispersal constraint scenarios. Under the Likelihood Ratio Test, DEC+J is vastly superior to standard DEC for all clades, for the first time verifying the importance of founder-event speciation in island clades via statistical model choice, and falsifying vicariance-dominated models of island biogeography. The case of *Psychotria* is typical: the DEC+J model is about 300,000 times more probable than the DEC model in an unconstrained analysis, according to AIC weights. Furthermore, the inferred maximum likelihood (ML) estimates of parameters often differ radically under the DEC+J model, with the "DE" part of the model sometimes playing no role (i.e., the parameters *d* and *e*, controlling anagenetic range expansion and range contraction, are inferred to be 0). Furthermore, under DEC+J, ancestral nodes are usually estimated to have ranges occupying only one island, avoiding the bias towards widespread ancestors often found with DEC. In conclusion, the many analyses where DEC has been used on island clades (e.g., Clark et al. 2008, Ree and Smith 2008, Bennett and O'Grady 2013, and many others) probably need to be reexamined.

Differences in cladogenesis processes in island versus continental clades

It is unsurprising that founder-event speciation is an important process in oceanic islands, although BioGeoBEARS is the first program to enable verification of this hypothesis through formal statistical model testing. It is interesting to ask whether the same pattern holds in non-island systems. Considering the variety of models currently in use, it would also be useful to perform direct head-to-head statistical model choice on the available models with a variety of datasets. Thus, I used BioGeoBEARS to compare the DEC, DIVALIKE, and BAYAREA cladogenesis models. These models, along with +J versions, were run on the island clades as well as a sample of 12 non-island (continental and oceanic) clades. Almost all analyses, including continental clades, strongly favored the "+J" models over the models without founder-event speciation, indicating that the process of founder-event speciation is important to consider in non-island clades as well as island clades. However, founder-event speciation appeared to be less frequent in non-island analyses, in that the estimates of the *j* parameter were on average 2–4 times lower in non-island clades, depending on which analyses were averaged. Making a rigorous statement about the relative frequency of founder -event speciation in island vs. non-island clades would require determination of what constitutes "representative sampling" from each, so this statement is intended as a heuristic one. The most important message of this work is that *only one* clade was found which favored the DEC model over all others: the "*Taygetis* clade" butterflies from the Neotropics (Matos-Maraví et al. 2013). The *Taygetis* species are often widespread and sympatric, a geographic pattern produced by simulations under the DEC model. However, such geographic patterns appear to be relatively rare in published analyses. This suggests that most published analyses have been using poorly-fitting models.

Incorporating fossils in historical biogeography using imperfect detection

BioGeoBEARS allows researchers to include both fossil tips and direct ancestors in their phylogeny and in their estimation of biogeographical history. The geographic ranges of fossil taxa may be input directly if they are known with confidence. Of course, detection of presence and absence will often be imperfect for fossil taxa. BioGeoBEARS users can turn on a probabilistic model of imperfect detection (Link and Barker 2009) in a hierarchical model (Suchard et al. 2003) along with the traditional likelihood calculations used for geographic range evolution. I present a proof-of-concept using two clades with good fossil records, namely, North American Canidae and Equinae. The NEOMAP database is used to provide counts of occurrences in each region and time bin, and is also used to provide counts of occurrences of taphonomic control groups (Bottjer and Jablonski 1988). The latter are used to measure relative sampling effort in each region and time bin. The two clades are found to prefer different models for cladogenesis: the equid dataset favors DEC and DEC+J about equally, indicating that jump dispersal is a weak process at best in this group, and that vicariance and subset speciation are fairly important. The canid dataset, on the other hand, strongly favors BAYAREA+J, indicating that vicariance and subset speciation were not important processes in canids. These results are found both with and without application of the imperfect detection model. Ironically, in test data chosen because of their high-quality fossil record, the record was so good that the model for imperfect detection had little impact. However, modeling imperfect detection is likely to be extremely useful in situations with poorer data, or with subsampled data.

Conclusions and looking forward

Several important conclusions may be drawn from my thesis research. First, formal model selection procedures can be applied in phylogenetic inferences of historical biogeography, and the relative importance of different processes can be measured. These techniques have great potential for strengthening quantitative inference in historical biogeography. No longer are biogeographers forced to simply assume from the start (whether consciously or through use of a model embedded in software) that some process, such as vicariance or dispersal, is important and others are not. Instead, this can be inferred from the data. Second, founder-event speciation appears to be a crucial explanatory process in most clades, the only exception being some intracontinental taxa showing a large degree of sympatry across widespread ranges. This is not the same thing as claiming that founder-event speciation is the *only* important process. Founder event speciation as the *only* important process was inferred in only one case (*Microlophus* lava lizards from the Galapagos). The importance of founder-event speciation will not be surprising to most island biogeographers. However, the results are important nonetheless, as there are still some vocal advocates of vicariance-dominated approaches to biogeography, such as Heads (2012) who allows vicariance and range-expansion to play a role in his historical inferences but explicitly excludes founder-event speciation *a priori*. The commonly used LAGRANGE DEC and DIVA programs actually make assumptions very similar to those of Heads (2012), even though many users of these programs likely consider themselves dispersalists or pluralists. Finally, the inclusion of fossils and imperfect detection within the same likelihood and model-choice framework clears the path for integrating paleobiogeography and neontological biogeography, strengthening inference in both.

Model choice is now standard practice in phylogenetic analysis of DNA sequences: a program such as ModelTest (Posada and Crandall

1998) is used to compare models such as Jukes-Cantor, HKY, and GTR+I+G, and to select the best model before inferring phylogenies or ancestral states. It is clear that the same should now happen in phylogenetic biogeography. BioGeoBEARS enables this procedure. Perhaps more importantly, however, is the potential for users to create and test new models. Probabilistic modeling of geographic range evolution on phylogenies is still in its infancy, and undoubtedly there are better models out there, waiting to be discovered. It is also likely that different clades and different regions will favor different processes, and that further improvements will be had by linking the evolution of organismal traits (e.g., loss of flight) with the evolution of geographic range. Another important step would be to include missing speciation and extinction events in the form of "SSE" models implemented in GeoSSE (Goldberg et al. 2011) and ClaSSE (Goldberg and Igić 2012).

In a world of rapid climate change and habitat loss, biogeographical methods must maximize both flexibility and statistical rigor if they are to be useful. This research takes several steps in that direction.

Acknowledgements

N.J.M. was supported in this work by a National Science Foundation (NSF) Grant DEB-0919451, a Wang Fellowship, and a Tien Fellowship. Full acknowledgements are given in the thesis. This paper was submitted to *Frontiers* as a thesis abstract; the thesis is being processed by ProQuest. Readers are invited to contact the author for preprints, as well as for BioGeoBEARS help and advice. An online tutorial, updates, hints, and notes, as well as the thesis, are available at http://phylo.wikidot.com/biogeobears.

References

Bennett, G.M. & O'Grady, P.M. (2013) Historical biogeography and ecological opportunity in the adaptive radiation of native Hawaiian leafhoppers (Cicadellidae: *Nesophrosyne*). Journal of Biogeography*,* 40, 1512–1523.

Bottjer, D.J. & Jablonski, D. (1988) Paleoenvironmental patterns in the evolution of post-Paleozoic benthic marine invertebrates. PALAIOS*,* 3, 540–560.

Buerki, S., Forest, F., Alvarez, N., Nylander, J.A.A., Arrigo, N. & Sanmartín, I. (2011) An evaluation of new parsimony-based versus parametric inference methods in biogeography: a case study using the globally distributed plant family Sapindaceae. Journal of Biogeography*,* 38, 531–550.

Burnham, K.P. & Anderson, D.R. (2002) Model selection and multimodel inference: a practical information-theoretic approach*,* 2nd edn. Springer, New York.

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

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.

Gillespie, R.G., Baldwin, B.G., Waters, J.M., Fraser, C.I., Nikula, R. & Roderick, G.K. (2012) Long-distance dispersal: a framework for hypothesis testing. Trends in Ecology & Evolution, 27, 47–56.

Goldberg, E.E. & Igić, B. (2012) Tempo and mode in plant breeding system evolution. Evolution, 66, 3701–3709.

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

Heads, M.J. (2012) Molecular panbiogeography of the tropics. University of California Press, Berkeley.

Kodandaramaiah, U. (2010) Use of dispersal-vicariance analysis in biogeography – a critique. Journal of Biogeography, 37, 3–11.

Landis, M., Matzke, N.J., Moore, B.R. & Huelsenbeck, J.P. (2013) Bayesian analysis of biogeography when the number of areas is large. Systematic Biology.

Link, W.A. & Barker, R.J. (2009) Bayesian inference: with ecological applications. Academic Press.

Matos-Maraví, P.F., Peña, C., Willmott, K.R., Freitas, A.V.L. & Wahlberg, N. (2013) Systematics and evolutionary history of butterflies in the "*Taygetis* clade" (Nymphalidae: Satyrinae: Euptychiina): Towards a better understanding of Neotropical biogeography. Molecular Phylogenetics and Evolution, 66, 54–68.

Matzke, N.J. (2013a) Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. Ph.D., University of California, Berkeley.

Matzke, N.J. (2013b) BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R scripts, CRAN: The Comprehensive R Archive Network, Berkeley, CA. http://CRAN.R-project.org/package=BioGeoBEARS

Posada, D., and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817–18.

Ree, R.H. (2005) Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. Evolution, 59, 257–265.

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

Ronquist, F. (1997) Dispersal-Vicariance Analysis: A new approach to the quantification of historical biogeography. Systematic Biology, 46, 195–203.

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

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

Suchard, M.A., Kitchen, C.M.R., Sinsheimer, J.S. & Weiss, R.E. (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. Systematic Biology, 52, 649–664.

Templeton, A.R. (2008) The reality and importance of founder speciation in evolution. BioEssays, 30, 470–479.

Waters, J.M., Trewick, S.A., Paterson, A.M., Spencer, H.G., Kennedy, M., Craw, D., Burridge, C.P. & Wallis, G.P. (2013) Biogeography off the tracks. Systematic Biology, 62, 494–498.

Wood, H.M., Matzke, N.J., Gillespie, R.G. & Griswold, C.E. (2013) Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the Palpimanoid spiders. Systematic Biology, 62, 264–284.

Yu, Y., Harris, A. J., and He, X.-J. (2013) RASP (Reconstruct Ancestral State in Phylogenies) 2.1 beta. http://mnh.scu.edu.cn/soft/blog/RASP/. Viewed August 4, 2013.

Edited by Ana Santos and Christophe Thébaud