

1

2

3

Running head: THE IMPACT OF THE NULL RANGE IN DEC

4

5

Title: Non-null Effects of the Null Range in Biogeographic Models: Exploring Parameter

6

Estimation in the DEC Model

7

8

Kathryn A. Massana¹, Jeremy M. Beaulieu², Nicholas J. Matzke³ and Brian C. O’Meara⁴

9

10

¹Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville,

11

TN, 37996-1610, USA, kmassana@vols.utk.edu

12

²National Institute for Biological and Mathematical Synthesis, University of Tennessee,

13

Knoxville, TN 37996, USA, jbeaulieu@nimbios.org

14

³National Institute for Biological and Mathematical Synthesis, University of Tennessee,

15

Knoxville, TN 37996, USA, matzke@nimbios.org

16

⁴Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville,

17

TN, 37996-1610, USA, bomeara@utk.edu

18

19

20

21

22

23

Email for correspondence: kmassana@vols.utk.edu

24 *Abstract.* – Historical biogeography seeks to understand the distribution of biodiversity in
25 space and time. The dispersal-extinction-cladogenesis (DEC) model, a likelihood-based
26 model of geographic range evolution, is widely used in assessing the biogeography of
27 clades. Robust inference of dispersal and local extinction parameters is crucial for
28 biogeographic inference, and yet a major caveat to its use is that the DEC model severely
29 underestimates local extinction. We suggest that this is mainly due to the way in which
30 the model is constructed to allow observed species to transition into being present in no
31 areas (i.e., null range). By prohibiting transitions into the null range in the transition rate
32 matrix, we were able to better infer local extinction and support this with simulations.
33 This modified model, DEC*, has higher model fit and model adequacy than DEC,
34 suggesting this modification should be considered for DEC and other models of
35 geographic range evolution.

36

37 **Keywords:** Lagrange, historical biogeography, local extinction, DEC, geographic range
38 evolution, BioGeoBEARS

39

40

41

42

43

44

45

46

47

INTRODUCTION

48

49

50

51

52

53

54

Historical biogeography has developed from simply observing the general patterns of species, to incorporating events that explain biogeographic processes (such as vicariance and dispersal), to developing explicit probabilistic approaches. With the advent of parametric methods based on maximum likelihood and Bayesian frameworks, researchers have been able to incorporate important information, such as branch lengths and fossils (e.g., allowing for better tree dating estimates) (Smith and Donoghue 2010; Wood et al. 2012; Beaulieu et al. 2013a; Pyron 2014).

55

56

57

58

59

60

61

62

63

64

65

66

A popular method (cited 499 times since its publication) to assess the historical biogeography of taxa is the dispersal-extinction-cladogenesis (DEC) model (Ree and Smith 2008), which estimates geographic range evolution for anagenetic (i.e., along branches) and cladogenetic (i.e., at nodes) change on a phylogeny. In the case of anagenetic change, range expansion and range contraction [modeled as parameters by way of the rate of dispersal from area i to area j (D_{ij}) and local extinction in area i (E_i)] are modeled as stochastic processes along branches. Published analyses with DEC assume the simplest model involving only a single dispersal rate and a single local extinction event, although it is possible to have more complex models (i.e., allowing different dispersal rates among each area pair by having n^2-n free parameters for the dispersal rates, and n free parameters for local extinction). A rate matrix can be assembled for a given number of geographic ranges and rate parameters (Fig. 1).

67

68

69

Ree and Smith (2008) carried out simulations to test the accuracy of the DEC model on parameter estimation. They found that although the model worked reasonably well, dispersal was underestimated and local extinction was severely underestimated,

70 often estimated as being effectively zero. Note that while there has been a robust
71 discussion of whether extinction rates can be estimated on molecular phylogenies (Nee et
72 al. 1994; Rabosky 2010; Beaulieu and O'Meara 2015), the extinction rate in that case
73 relates to speciation and extinction of entire species and its signal on a phylogeny's
74 topology and branch lengths. The extinction rate relevant for our purposes is the rate of
75 having a population no longer occurring in a particular area, using a fixed tree. In terms
76 of its role in a model, it is more like the rate of reduction in a meristic character than a
77 rate at which a species goes extinct, even though biologically it appears more similar to
78 the latter. Thus, its difficulty in being estimated is surprising.

79 One feature of the DEC model that has received little comment is that it includes
80 a null range (a geographic range of 0 areas) in the anagenesis transition matrix (Fig. 1). In
81 one sense, inclusion of the null range is a natural modeling decision, since the assumption
82 that local extirpation is a process directly implies that the same process can reduce a
83 single area geographic range to a range of size 0. However, the inclusion of a null range
84 in the state space has some peculiar properties. For instance, no sampled species will ever
85 occupy the null range state; even extinct species, if included in an analysis, are included
86 because they occurred in some area. We suspect that the only way to fit any data pattern
87 that does not observe null ranges is by driving down the rate of range contraction to the
88 point where the probability of such an event is effectively zero. Unfortunately, given that
89 all transitions to other extinction scenarios are linked through a global extinction
90 parameter it seems unavoidable that when null ranges are allowed in the model extinction
91 would generally be underestimated.

92 Of course, in some ways, the extinction rate is a nuisance parameter – that is, the
93 hundreds of studies using the DEC model primarily focus on ancestral state estimates
94 rather than on rates. However, given that this rate represents one of the two free
95 parameters that are then used for inferring ancestral states on the tree, we expect that
96 biased extinction estimates may result in errors in the ancestral state estimation. In other
97 words, given low extinction rate estimates, areas can only be lost at speciation events, so
98 we predict a greater number of areas as we move rootward on the tree and few area losses
99 along longer branches. This study attempts to modify the DEC model to improve
100 estimation of extinction rate and then tests using simulated and empirical data to see if
101 this results in a better model overall.

102

103

METHODS

104

105

106

107

108

109

110

111

112

113

114

We modified the DEC model (which we refer to DEC* hereafter) to omit transitions into the null state in the anagenesis transition rate matrix between ancestor and descendent pairs (Fig. 1). It has the same number of parameters as DEC (dispersal and local extinction, d and e respectively), with the only change being fixing the transition rate to 0 for transitions from ranges of size 1 to the null range. DEC* is distinct from the three-parameter DEC+J model which allows for founder-event speciation associated with lineage-splitting with the addition of the free j parameter (Matzke 2014b). DEC+J retains the DEC assumption that a null geographic range is a valid state. To implement the DEC* model, we modified the original lagrange DEC C++ code (<https://github.com/rhr/lagrange-cpp>) to omit the transition into the null range in the anagenesis transition rate matrix. The DEC* model is implemented as a modification to

115 the DEC C++ version, and is also allowed in the BioGeoBEARS R package (Matzke
116 2013) by changing the *include_null_range* setting from TRUE to FALSE in the
117 BioGeoBEARS model setup.

118 We implemented our own DEC simulator in R that follows the procedures
119 described by Ree and Smith (2008). The simulator produces birth-death phylogenetic
120 trees with concurrent range evolution, combining the DEC model and stochastic
121 cladogenesis. The simulator also does the same for the DEC* model. Trees were
122 produced with the same known dispersal and local extinction parameter, constrained to
123 vary between 0.01 and 0.2, while speciation was constrained to be 0.4 events per million
124 years. Geography consisted of three possible hard-coded geographic areas, meaning that
125 there were 8 possible geographic ranges in the state space of the DEC simulation (A, B,
126 C, AB, AC, BC, ABC, and null), and 7 possible ranges in the DEC* simulation (as the
127 null range state is excluded). At cladogenesis, when the lineage had a widespread range,
128 equal probabilities were assigned to each allowed range-inheritance scenario (vicariance
129 or subset sympatry). For both DEC and DEC*, we performed 2,000 simulation-inference
130 runs and compared dispersal and local extinction parameter estimates as well as the
131 number of correctly inferred number of areas at internal nodes for all simulations. The
132 simulations began by assigning the root node a range of a random single geographic area.
133 The phylogeny was allowed to grow according to the DEC or the DEC* model until it
134 reached 100 taxa (extant plus extinct). To match empirical datasets, the simulated
135 phylogenies were pruned of branches that went extinct.

136 Our main objective was to understand DEC* versus DEC analyses on empirical
137 datasets. Therefore, we searched the literature for published studies that used the DEC

138 model. Then we compiled the phylogenies and geography presence/absence data
139 available, which resulted in 15 empirical datasets. Most of these datasets were used in
140 Matzke (2014b), and followed any modifications made therein (Kambysellis et al. 1995;
141 Baldwin and Sanderson 1998; Hormiga et al. 2003; Jordan et al. 2003; Clark et al. 2008;
142 Dunbar-Co 2008; Ree and Smith 2008; Benavides et al. 2009; Clark et al. 2009; Gillespie
143 and Baldwin 2010; Smith and Donoghue 2010; Lerner et al. 2011; Nicholson et al. 2012;
144 Bennett et al. 2013; Lapoint et al. 2013; Matzke 2014a). However, we also assessed the
145 caecilian and salamander datasets from a recent published study using DEC+J (Pyron
146 2014), and a palpimanoid spider dataset (Wood et al. 2012). We performed unconstrained
147 analyses with C++ DEC and DEC* on each dataset. We compared analyses between
148 DEC, DEC*, and DEC+J for all 15 datasets. If the values were not available, we used the
149 package BioGeoBEARS (Matzke 2013) to run all DEC+J analyses. We also assessed
150 model adequacy for each dataset by comparing the number of areas estimated per node
151 between DEC and DEC* to the observed modern geographic range sizes.

152

153

RESULTS

154

Simulations

155

156

157

158

159

160

Figure 2 shows the observed parameter estimates of local extinction and dispersal for DEC and DEC* compared to the true estimates under a DEC simulation. Overall, the point estimate for local extinction was closer to the true value under DEC* than with DEC (Fig. 2), although with higher variance. With simulations under the DEC model we found that the median for local extinction under a DEC* inference ($e=0.0957$) was closer to the true local extinction median estimate ($e=0.0989$), while the median for local

161 extinction under DEC inference was close to zero ($e=1.287e-06$). Similarly, with
162 simulations under the DEC* model we found that the median local extinction estimated
163 under DEC* inference ($e=0.1028$) is almost identical to the true local extinction
164 parameter ($e=0.1030$), whereas, again, local extinction is grossly underestimated under
165 the DEC inference ($e=1.200e-6$).

166 Median estimates of dispersal under DEC simulations were closer to the median
167 of the true dispersal parameter ($d=0.0989$) under DEC* than under DEC inference
168 ($d=0.0766$) (Fig. 2). When simulating under DEC* (Supplemental Fig. 1), the median
169 dispersal under DEC* inference ($d=0.1053$) was again closer to the median dispersal rate
170 used in the simulation ($d=0.1030$) than dispersal inferred under DEC ($d=0.0789$).

171 We calculated the root mean square error (RMSE) of the estimated parameter
172 values for DEC and DEC*. The root mean square error gives the standard deviation
173 associated with the differences between the true parameter and the inferred parameter
174 estimates, and here a smaller value indicates less error in the parameter inference. Results
175 indicated that on the logarithmic scale the error for e was far better for DEC* than DEC
176 and nearly the same for d (RMSE of e was 10.9120 for DEC and 3.8363 for DEC*;
177 RMSE of d was 0.3784 for DEC and 0.3886 for DEC*). However, on a linear scale, error
178 is far better for both parameters for DEC than DEC*, due to some tremendously high
179 values of e . (RMSE of e was 0.1135 for DEC and 2.2345 for DEC*; RMSE of d was
180 0.0363 for DEC and 0.7816 for DEC*).

181 Finally, we assessed the accuracy of DEC against DEC* in estimating the
182 geographic area range at the root. Under DEC simulation, the root state was correctly
183 estimated 49.05% of the time, whereas under DEC*, 58.30% of root states were

184 accurately estimated.

185

186 *Empirical Datasets*

187 In comparisons of DEC and DEC* on empirical datasets, likelihood was always
188 better under DEC*. Thus, AIC always selected DEC* as the better model over DEC (as
189 the two models have the same number of parameters). In 10 out of 15 empirical datasets,
190 AIC selected DEC* over DEC+J (Supplementary Table 1). DEC+J has an extra
191 parameter relative to DEC*, so if likelihoods were equal between DEC+J and DEC*,
192 DEC* would be preferred by AIC. However, in all but one of the cases where AIC
193 preferred DEC* over DEC+J, the likelihood was itself better with DEC* (which is
194 possible, as the two models are not nested). The exception was *Psychotria*, where AIC
195 gives DEC* 50.2% of the model weight despite slightly higher likelihood for DEC+J
196 (model weight 20.4%).

197 Unlike the simulated data, for over half the empirical datasets the extinction rate
198 inferred by DEC was substantially higher than zero, ranging from 16% to 546% of the
199 estimated value of the dispersal rate. For DEC*, the extinction rates were even higher
200 relative to dispersal: only for one empirical dataset was the extinction rate
201 indistinguishable from zero, for the rest the extinction rate was between 3.2 and 1389-
202 fold higher than dispersal rate (median 104-fold higher). In some cases, the estimated
203 extinction rate was at the maximum allowed by the program; modifying it to increase the
204 bound by tenfold improved the likelihood by a median of 0.061 log likelihood units and
205 increased the extinction estimate up to the new maximum in most cases. The small
206 magnitude of improvement, with the large magnitude of change in the estimate, suggests

207 that the likelihood surface is very flat but that the unconstrained maximum likelihood
208 estimate would be even higher. More simply put, for these datasets, the best estimate of
209 extinction is extremely high, which would mean that after a species expands its range it
210 nearly instantly contracts it (into either the new region or back to the old region). In only
211 three of nine of these datasets was DEC+J chosen over DEC*, despite the apparent
212 evidence for a jump-like dispersal model.

213

214 *Model Adequacy*

215 In addition to model choice, a key question to examine with new models is model
216 adequacy: how well does the model fit overall? Even the best-fitting model may not do a
217 good job predicting the data, which would point to the need for new models to better
218 match reality. This has been increasingly emphasized in phylogenetics (Goldman 1993;
219 Bollback 2002; O'Meara 2012; Beaulieu et al. 2013b; Pennell et al. 2015). To see if the
220 DEC* model adequately describes the data, we counted the number of occupied areas
221 estimated for each node and compared this between DEC and DEC* for each empirical
222 dataset. We work under the assumption that the present should look like the past: a clade
223 of island endemics is more likely to have been island endemics for much of their history,
224 rather than being composed of very widespread species that only at the present suddenly
225 became endemic to single islands. Of course, there are processes that could make the
226 present not resemble the past (i.e., a sudden change in climate causing suitable habitat to
227 be divided into isolated patches), but this assumption should hold in most groups. For all
228 but two empirical datasets, the DEC* model was the more adequate model, with
229 estimated range sizes at ancestral nodes more closely matching the estimated the mean

230 range sizes observed at the tips of the phylogeny (Fig. 3). Inference under DEC usually
231 yields ancestral distributions that are very widespread, which is not the case under DEC*
232 (Fig. 3 and Fig. 4).

233

234 DISCUSSION

235 Given the results, we argue that DEC* should be considered for use in
236 biogeographic models. Testing models should be an intrinsic part of the research process,
237 so most users should try DEC, DEC*, DEC+J, and future models, but if one were limited
238 to just one model, in most cases DEC* would be preferred, based on the empirical results
239 presented here. DEC* does a more adequate job at estimating ancestral ranges than does
240 the canonical DEC model. However, while the median extinction and dispersal
241 parameters were better estimated under DEC* than with DEC, the RMSE of the estimates
242 on a linear scale was better under DEC, and DEC* often returns very high estimates for
243 extinction rate. For estimating ancestral areas, DEC* is probably the better model, but we
244 urge extreme caution when treating its rate parameters as parameters of interest rather
245 than nuisance parameters. Of course, ancestral states are known to be difficult to estimate
246 well (Cunningham 1999; Oakley and Cunningham 2000), so biologists should expect a
247 great deal of uncertainty with estimates. We also note that the estimates of uncertainty in
248 this model are always underestimates, due to other uncertainty (topology, branch lengths,
249 states) that is typically not accounted for.

250 Another caveat to the use of DEC* is its treatment of the phylogeny: it assumes
251 range evolves on a tree but that biogeography does not directly influence speciation or
252 extinction. Speciation often seems influenced by geographic context (Mayr 1963), such

253 as through the divergence of two isolated populations. While DEC, DEC*, and DEC+J
254 allow subdivision of ranges at speciation events, these models do not, for example, fit a
255 higher speciation rate to species with larger ranges. There are models that jointly fit the
256 diversification process and process of biogeographic evolution, such as GeoSSE
257 (Goldberg et al. 2011) and ClaSSE (Goldberg and Igić 2012). However, even though
258 these models are more realistic, they can require rather large data sets (Davis et al. 2013)
259 and are feasible only for very few areas. The empirical datasets used in biogeographic
260 models are typically small in comparison; the ones used in this study had an average of
261 75 taxa and 5.53 areas from all datasets, and a range of 4 to 10 areas and 9 to 469 taxa.
262 Recent work (Matzke 2014b) showed that DEC vs. DEC+J model choice appears robust
263 to some commonly-postulated SSE processes (speciation and extinction depending on
264 range size), and that ancestral range estimation is reasonably accurate if model choice is
265 performed and if the dispersal rate is low, suggesting that for datasets that limit the power
266 of SSE models, the DEC* model can still be used, with caution.

267 It is important to emphasize that the DEC* model is still relatively simple.
268 Though some complexity can be incorporated with different dispersal rates between areas
269 or at different time points, all species are treated as having the same rates of dispersal and
270 extinction at a given time. We know, however, that species in a clade may vary in traits
271 affecting successful dispersal (ability to inbreed, resting stages, wind versus animal
272 dispersal, tolerance of saltwater, and so forth) or extinction (body size, trophic level,
273 thermal tolerance, and so forth) and this variation is not yet incorporated in any of these
274 models. There are additional sources of heterogeneity that also may result in misleading
275 results if not incorporated.

276 The very high extinction estimates by DEC*, especially in empirical datasets, was
277 unexpected. One partial explanation may come from the empirical distribution of range
278 sizes in the empirical studies presented here, where the vast majority of species were
279 found in just one area. Under DEC* (or DEC) the only way for a species to change its
280 area is through expansion to a new area (i.e., dispersal), followed by other events. Given
281 that all the studies have species in different areas (there is little point to inferring
282 biogeographic history for a clade that only ever occupies one area), there must be a
283 nonzero dispersal rate. Lineages can reduce range size in two ways: at cladogenesis, or
284 through range contraction along branches. However, when e is high with respect to d and
285 the speciation rate, lineages will spend almost no time in widespread ranges. Therefore,
286 widespread ranges are essentially never available at cladogenesis, and all speciation will
287 be sympatric. Moreover, on terminal branches, which represent over half the branches on
288 the tree, any necessary dispersals cannot be “undone” by contraction at cladogenesis, and
289 so the only way to have some dispersals along terminal branches, but have observed
290 species in just one area each, is to have a substantial extinction rate. DEC* with high e is
291 a model with all range-change effectively occurring in anagenetic “jumps” along the
292 branches: any expansion is followed almost immediately by a contraction. The fact that
293 DEC* often outperforms DEC+J probably indicates that in these cases, the probability of
294 sister taxa living in different areas correlates better with the branch length between the
295 taxa than with the number of speciation events recorded in the observed tree. We expect a
296 DEC*+J model may incorporate the best of both of these models, and that adding “*” to
297 other models may also be beneficial.

298 The major use of parametric models in historical biogeography is for ancestral
299 state estimation. For model adequacy, we expect that the present tends to resemble the
300 past, so a model where past distributions are similar to present ones is probably a better
301 fit to the data. For empirical datasets used in biogeography, tip taxa are most often in one
302 area. But, DEC often estimates ancestors as being in many areas, because the DEC model
303 allows a transition into the null range, and since null ranges are not observed, the
304 inference is pushed towards a low extinction rate. In contrast, the DEC* model returns
305 estimates at internal nodes that usually resemble the number of areas present in the tips
306 (see Fig. 4). DEC* may return nearly equally likely single areas rather than a more
307 confident estimation of the ancestral state being a union of areas. In many cases,
308 especially given observed species that occupy individually few areas, this uncertainty
309 about which single area an ancestor occupied represents reality. However, even though
310 uncertainty exists in ancestral range estimates, statistical model choice is a fruitful way to
311 assess models against the data.

312

313 ACKNOWLEDGEMENTS

314 We thank the HOFF Lab Group, James Fordyce, Daniel Simberloff, and Sally
315 Horn for helpful discussions. Support for KAM has been provided by the National
316 Institutes of Health Program for Excellence and Equity in Research Grant [R25
317 5R25GM086761-06] at the University of Tennessee and from the University of
318 Tennessee, Knoxville Ecology and Evolutionary Biology Department Summer Funding
319 and Chancellor’s Funds. JMB and NJM were by the National Institute for Mathematical
320 and Biological Synthesis, an Institute sponsored by the National Science Foundation, the

321 U.S. Department of Homeland Security, and the U.S. Department of Agriculture through
322 NSF Award #EF-0832858, with additional support from The University of Tennessee,
323 Knoxville.

324

325

REFERENCES

- 326 Baldwin, B. G. and M. J. Sanderson. 1998. Age and rate of diversification of the
327 Hawaiian silversword alliance (Compositae). *Proceedings of the National*
328 *Academy of Sciences* 95:9402-9406.
- 329 Beaulieu, J., D. Tank, and M. Donoghue. 2013a. A Southern Hemisphere origin for
330 campanulid angiosperms, with traces of the break-up of Gondwana. *BMC*
331 *Evolutionary Biology* 13:80.
- 332 Beaulieu, J. M. and B. C. O'Meara. 2015. Extinction can be estimated from moderately
333 sized molecular phylogenies. *Evolution* 69:1036-1043.
- 334 Beaulieu, J. M., B. C. O'Meara, and M. J. Donoghue. 2013b. Identifying hidden rate
335 changes in the evolution of a binary morphological character: the evolution of
336 plant habit in campanulid angiosperms. *Systematic biology*.
- 337 Benavides, E., R. Baum, H. M. Snell, H. L. Snell, and J. W. Sites, Jr. 2009. Island
338 biogeography of Galapagos lava lizards (Tropiduridae: Microlophus): species
339 diversity and colonization of the archipelago. *Evolution* 63:1606-1626.
- 340 Bennett, G. M., P. M. O'Grady, and K. Triantis. 2013. Historical biogeography and
341 ecological opportunity in the adaptive radiation of native Hawaiian leafhoppers
342 (Cicadellidae:Nesophrosyne). *Journal of Biogeography* 40:1512-1523.

- 343 Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. *Molecular*
344 *Biology and Evolution* 19:1171-1180.
- 345 Clark, J. R., R. H. Ree, M. E. Alfaro, M. G. King, W. L. Wagner, and E. H. Roalson.
346 2008. A Comparative Study in Ancestral Range Reconstruction Methods:
347 Retracing the Uncertain Histories of Insular Lineages. *Systematic Biology*
348 57:693-707.
- 349 Clark, J. R., W. L. Wagner, and E. H. Roalson. 2009. Patterns of diversification and
350 ancestral range reconstruction in the southeast Asian–Pacific angiosperm lineage
351 *Cyrtandra* (Gesneriaceae). *Molecular Phylogenetics and Evolution* 53:982-994.
- 352 Cunningham, C. W. 1999. Some Limitations of Ancestral Character-State Reconstruction
353 When Testing Evolutionary Hypotheses. *Systematic Biology* 48:665-674.
- 354 Davis, M. P., P. E. Midford, and W. Maddison. 2013. Exploring power and parameter
355 estimation of the BiSSE method for analyzing species diversification. *Bmc*
356 *Evolutionary Biology* 13.
- 357 Dunbar-Co, S. 2008. Island evolution: phylogeny, adaptive radiation, and biogeography
358 of *Plantago* (Plantaginaceae) in the Hawaiian Islands.
- 359 Gillespie, R. G. and B. G. Baldwin. 2010. Island biogeography of remote archipelagoes.
360 The theory of island biogeography revisited:358-387.
- 361 Goldberg, E. E. and B. Igić. 2012. Tempo and Mode in Plant Breeding System Evolution.
362 *Evolution* 66:3701-3709.
- 363 Goldman, N. 1993. Simple diagnostic statistical tests of models for DNA substitution.
364 *Journal of molecular evolution* 37:650-661.

- 365 Hormiga, G., M. Arnedo, and R. G. Gillespie. 2003. Speciation on a Conveyor Belt:
366 Sequential Colonization of the Hawaiian Islands by Orsonwelles Spiders
367 (Araneae, Linyphiidae). *Systematic Biology* 52:70-88.
- 368 Jordan, S., C. Simon, and D. Polhemus. 2003. Molecular Systematics and Adaptive
369 Radiation of Hawaii's Endemic Damselfly Genus *Megalagrion* (Odonata:
370 *Coenagrionidae*). *Systematic Biology* 52:89-109.
- 371 Kambysellis, M. P., K.-F. Ho, E. M. Craddock, F. Piano, M. Parisi, and J. Cohen. 1995.
372 Pattern of ecological shifts in the diversification of Hawaiian *Drosophila* inferred
373 from a molecular phylogeny. *Current Biology* 5:1129-1139.
- 374 Lapoint, R. T., P. M. O'Grady, and N. K. Whiteman. 2013. Diversification and dispersal
375 of the Hawaiian *Drosophilidae*: the evolution of *Scaptomyza*. *Mol Phylogenet*
376 *Evol* 69:95-108.
- 377 Lerner, H. R., M. Meyer, H. F. James, M. Hofreiter, and R. C. Fleischer. 2011.
378 Multilocus resolution of phylogeny and timescale in the extant adaptive radiation
379 of Hawaiian honeycreepers. *Current biology : CB* 21:1838-1844.
- 380 Matzke, N. J. 2013. *BioGeoBEARS: BioGeography with Bayesian (and Likelihood)*
381 *Evolutionary Analysis in R Scripts*. University of California, Berkeley, Berkeley,
382 CA.
- 383 Matzke, N. J. 2014a. Data from: Model selection in historical biogeography reveals that
384 founder-event speciation is a crucial process in island clades. *Dryad Data*
385 *Repository*.
- 386 Matzke, N. J. 2014b. Model Selection in Historical Biogeography Reveals that Founder-
387 event Speciation is a Crucial Process in Island Clades. *Systematic Biology*.

- 388 Mayr, E. 1963. Animal species and evolution. *Animal species and their evolution*.
- 389 Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994. Extinction Rates Can Be
390 Estimated from Molecular Phylogenies. *Philosophical Transactions of the Royal
391 Society of London Series B-Biological Sciences* 344:77-82.
- 392 Nicholson, K. E., B. I. Crother, C. Guyer, and J. M. Savage. 2012. It is time for a new
393 classification of anoles (Squamata: Dactyloidae). *Zootaxa* 3477:1-108.
- 394 O'Meara, B. C. 2012. Evolutionary Inferences from Phylogenies: A Review of Methods.
395 Pp. 267-285. *Annual Review of Ecology, Evolution, and Systematics*, Vol 43.
- 396 Oakley, T. H. and C. W. Cunningham. 2000. Independent contrasts succeed where
397 ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution*
398 54:397-405.
- 399 Pennell, M. W., R. G. FitzJohn, W. K. Cornwell, and L. J. Harmon. 2015. Model
400 Adequacy and the Macroevolution of Angiosperm Functional Traits. *The
401 American Naturalist* 186:E33-E50.
- 402 Pyron, R. A. 2014. Biogeographic Analysis Reveals Ancient Continental Vicariance and
403 Recent Oceanic Dispersal in Amphibians. *Systematic Biology* 63:779-797.
- 404 Rabosky, D. L. 2010. Extinction Rates Should Not Be Estimated from Molecular
405 Phylogenies. *Evolution* 64:1816-1824.
- 406 Ree, R. H. and S. A. Smith. 2008. Maximum likelihood inference of geographic range
407 evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology*
408 57:4-14.

409 Smith, S. A. and M. J. Donoghue. 2010. Combining historical biogeography with niche
410 modeling in the Caprifolium clade of Lonicera (Caprifoliaceae, Dipsacales).
411 Systematic Biology 59:322-341.

412 Wood, H. M., N. J. Matzke, R. G. Gillespie, and C. E. Griswold. 2012. Treating fossils as
413 terminal taxa in divergence time estimation reveals ancient vicariance patterns in
414 the palpimanoid spiders. Systematic Biology 62:264-284.

415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

432 FIGURE LEGENDS

433

434 Figure 1. Diagram showing (a) the allowed cladogenetic events for DEC and DEC*, (b)
435 the anagenesis transition rate matrix for DEC, and (c) the DEC* anagenesis transition rate
436 matrix, assuming three geographic states (A, B, C). At a cladogenetic event, if a species
437 is in one area, the descendant species inherits that area. If the species is in multiple areas,
438 one species inherits one area, while the other species inherits all areas (peripatric
439 speciation; within area widespread) or it is allowed to inherit all areas but the area
440 occupied by the first species (vicariant allopatric speciation; between area widespread).
441 Note that extinction is not allowed in DEC* from one state to zero states. D = dispersal, E
442 = local extinction.

443

444 Figure 2. Plots showing parameter inference under the 2,000 DEC simulations. DEC
445 inference of dispersal (A) was not as effective as dispersal inferred under DEC* (B).
446 Local extinction under DEC inference (C) was highly underestimated. Local extinction
447 under DEC* (D) was better estimated in comparison to DEC inference, although with
448 more variance. Purple lines represent 95% confidence intervals; blue line shows the
449 median; orange line shows the 1:1 line.

450

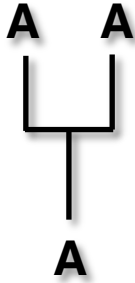
451 Figure 3. Model adequacy plots based on the number of areas occupied at nodes for the
452 empirical datasets of the Galapagan *Microlophus* (A), Hawaiian *Plantago* (B), Pacific
453 *Cyrtandra* (C) reconstructed with DEC or DEC* versus the tips which represent the
454 current number of areas for each group. In each empirical case (A, B, C), DEC* was able

455 to ancestrally infer the same number of areas as the tips in the current range. DEC
456 inferred ranges that were more widespread. The last plot depicts the mean number of
457 current areas at tips versus the mean number of areas estimated at nodes through DEC or
458 DEC* for each group (D). In each case except for two, DEC* was able to estimated the
459 ancestors to occupy the same number of areas as the tips.

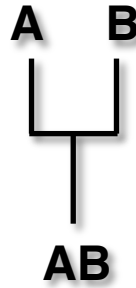
460

461 Figure 4. Most probable number of biogeographic areas estimated in the *Plantago* clade
462 under the DEC (A) versus the DEC* model (B). The number of areas at tips is also
463 shown. The estimated ancestral range probabilities under DEC versus DEC* for node 8
464 (C) and node 10 (D) are shown. Ancestral state estimates under the DEC model are more
465 widespread than under DEC*, therefore providing a different biogeographic history (C
466 and D). Nodes closer to the root provide more variance in the probable ranges estimated
467 (D).

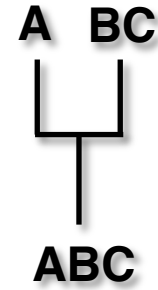
a) Fixed Cladogenesis



Single Area



Between area
widespread



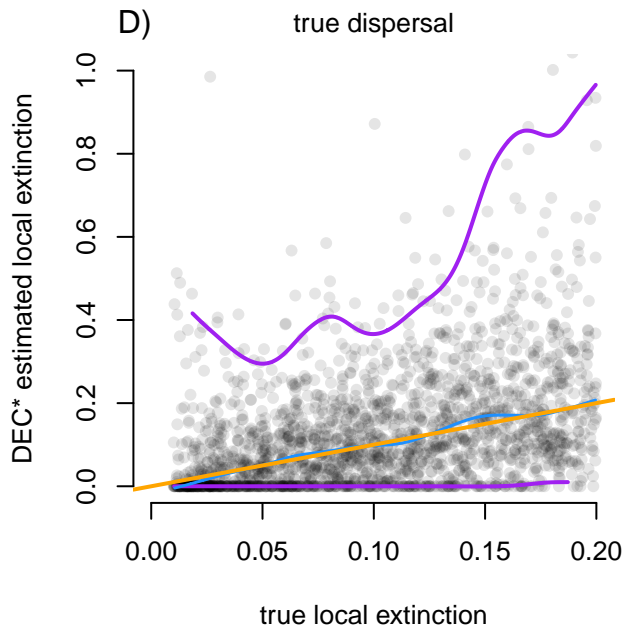
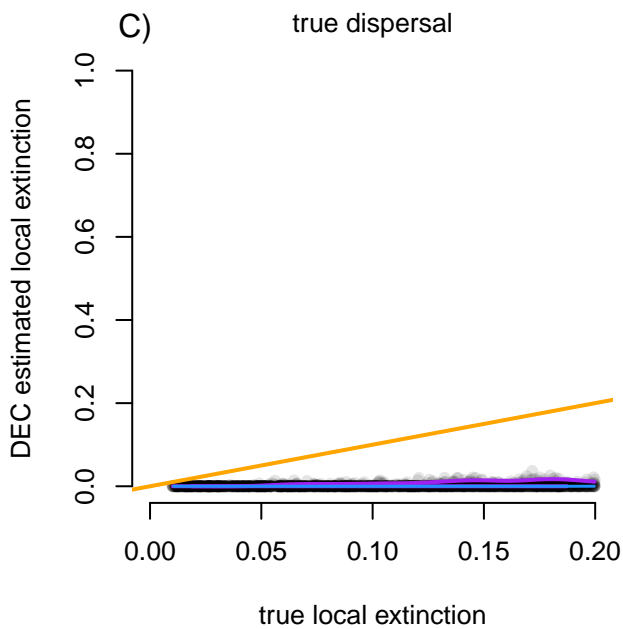
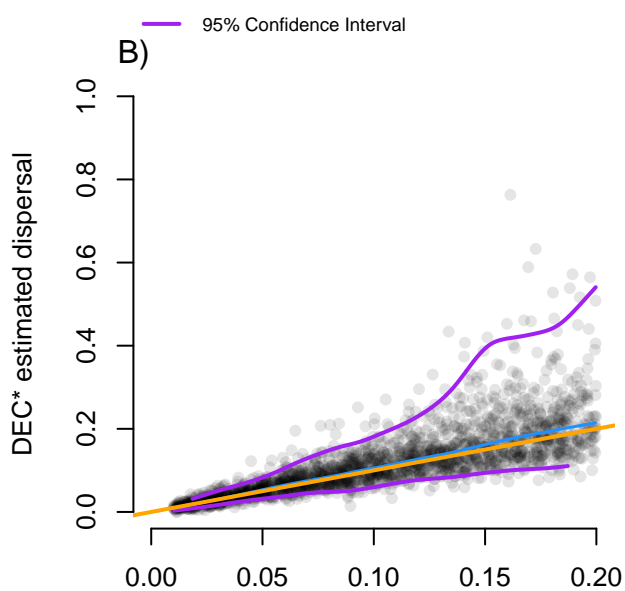
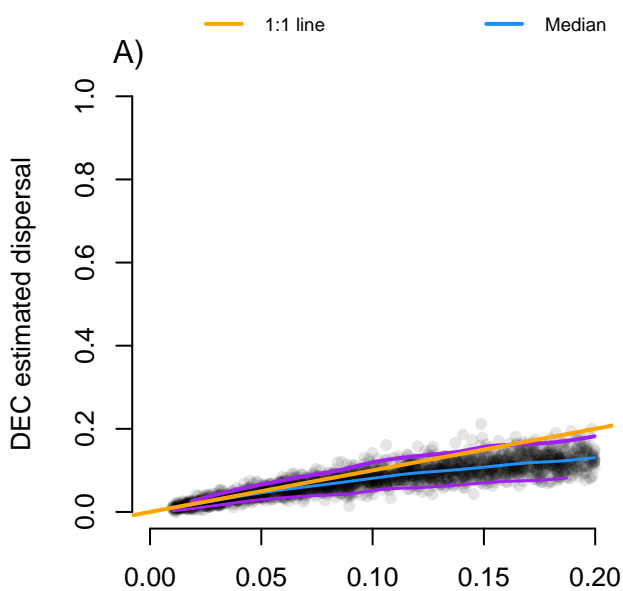
Within area
widespread

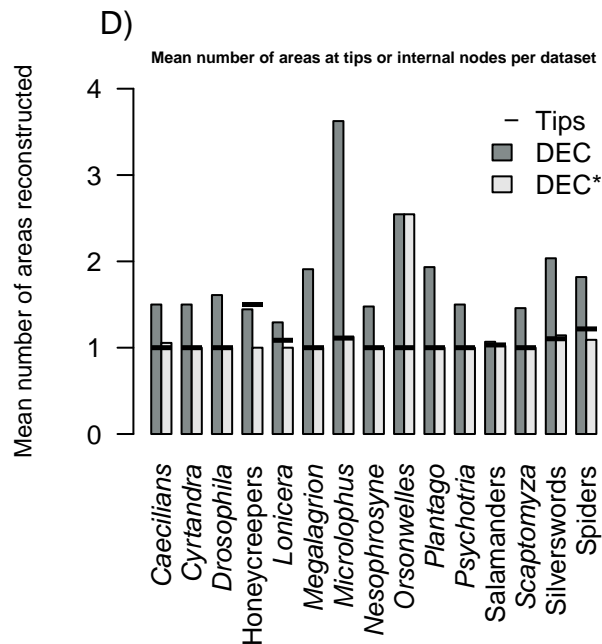
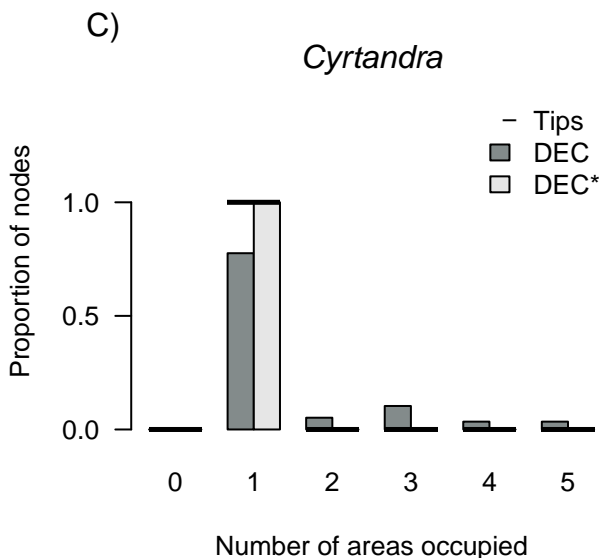
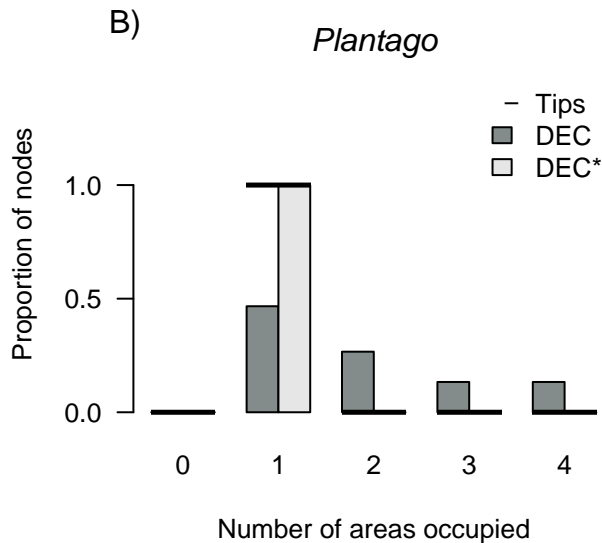
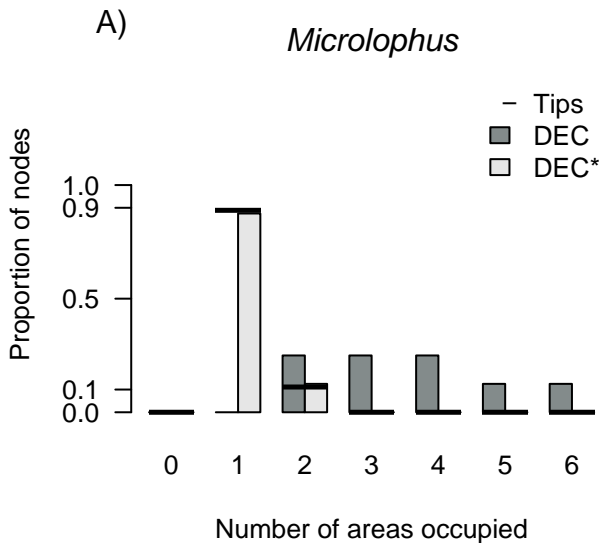
b) DEC Anagenesis

	\emptyset	A	B	C	AB	AC	BC	ABC
\emptyset	-	0	0	0	0	0	0	0
A	E_A	-	0	0	D_{AB}	D_{AC}	0	0
B	E_B	0	-	0	D_{BA}	0	D_{BC}	0
C	E_C	0	0	-	0	D_{CA}	D_{CB}	0
AB	0	E_B	E_A	0	-	0	0	$D_{AC} + D_{BC}$
AC	0	E_C	0	E_A	0	-	0	$D_{AB} + D_{CB}$
BC	0	0	E_C	E_B	0	0	-	$D_{BA} + D_{CA}$
ABC	0	0	0	0	E_C	E_B	E_A	-

c) DEC* Anagenesis

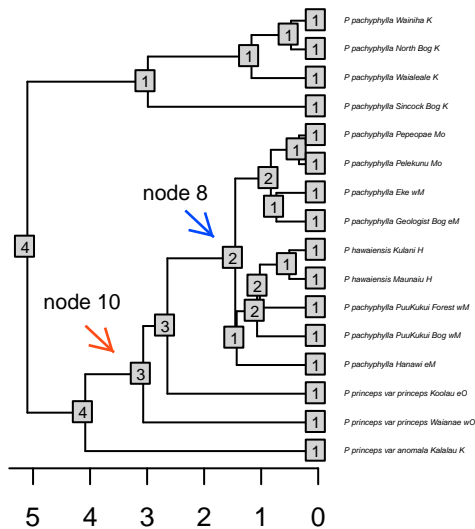
	\emptyset	A	B	C	AB	AC	BC	ABC
\emptyset	-	0	0	0	0	0	0	0
A	0	-	0	0	D_{AB}	D_{AC}	0	0
B	0	0	-	0	D_{BA}	0	D_{BC}	0
C	0	0	0	-	0	D_{CA}	D_{CB}	0
AB	0	E_B	E_A	0	-	0	0	$D_{AC} + D_{BC}$
AC	0	E_C	0	E_A	0	-	0	$D_{AB} + D_{CB}$
BC	0	0	E_C	E_B	0	0	-	$D_{BA} + D_{CA}$
ABC	0	0	0	0	E_C	E_B	E_A	-





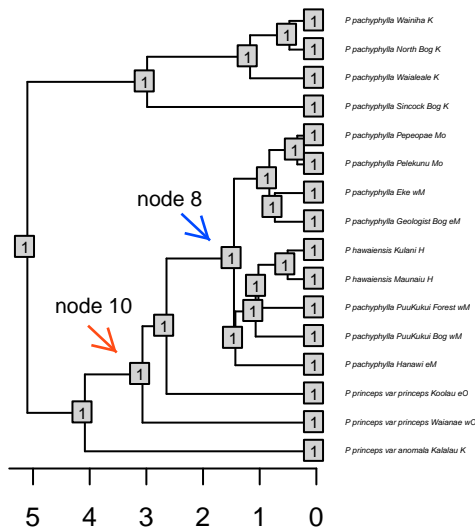
A)

DEC most probable number of areas reconstructed



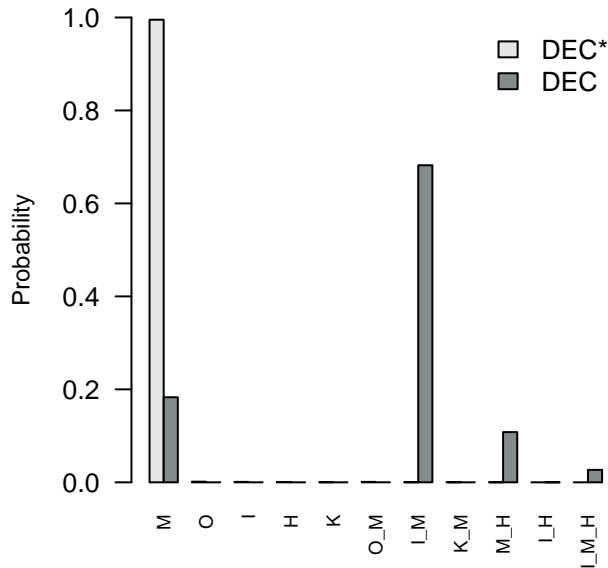
B)

DEC* most probable number of areas reconstructed



C)

Ancestral area reconstructions for node 8



D)

Ancestral area reconstructions for node 10

