Journal of Biogeography

METHODOLOGICAL APPLICATION

An event model for phylogenetic biogeography using explicitly geographical ranges

J. Salvador Arias[1,2,*]

[1]*Unidad Ejecutora Lillo (CONICET-Fundación Miguel Lillo), Miguel Lillo 251, S.M. de Tucumán (04000) Tucumán, Argentina,* [2]*Cátedra de Biogeografía, Facultad de Ciencias Naturales, Universidad Nacional de Tucumán, Miguel Lillo 205, S.M. de Tucumán (04000) Tucumán, Argentina*

*Correspondence: J. Salvador Arias, Cátedra de Biogeografía, Facultad de Ciencias Naturales, Universidad Nacional de Tucumán, Miguel Lillo 205, S.M. de Tucumán (04000), Tucumán, Argentina.
E-mail: jsalarias@csnat.unt.edu.ar

ABSTRACT

Aim To develop and implement a method for phylogenetic biogeography that is both event based and geographically explicit, that is, that uses the geographical ranges observed in the terminals instead of 'predefined areas.'

Methods The method, GEM (Geographically explicit Event Model), attributes vicariance, sympatry (range copying), point sympatry (subset sympatry) or founder events, to the internal nodes of the tree. The cost of a reconstruction is calculated as the event cost plus the amount of range changes along a branch, and the best reconstruction is the combination of the event and range assignments that minimize the cost.

Results The approach was implemented in a computer program, EVS, using a geographical data model (a raster) in which range changes were measured by pixel counts. The program can be used in real-sized datasets, using an heuristic to find reasonable solutions in short times.

Main conclusion GEM provides a method for direct analysis of joint data on phylogeny and explicit distribution ranges, and proposes both the ancestral ranges and the biogeographical events connected with cladogenesis.

Keywords

ancestral ranges, dispersal, event-based biogeography, extinction, founder event, geographical data models, historical biogeography, phylogenetic biogeography, sympatry, vicariance

INTRODUCTION

The objective of phylogenetic biogeography (also known as 'taxon history biogeography' or 'lineage geohistory') is inferring the evolution of the distribution range in a particular clade given its phylogenetic relationships and geographical ranges of its terminals (Brundin, 1966; Hennig, 1966; Hovenkamp, 1997, 2002; Ronquist, 1997; Ree *et al.*, 2005). A major breakthrough in the field was the development of event-based methods, which include dispersal–vicariance analysis (DIVA, Ronquist, 1997) and the Dispersal-Extinction-Cladogenesis model (DEC, Ree *et al.*, 2005). The event-based methods have the advantage that they infer both the ancestral range and the biogeographical processes (event scenario) to be at work during cladogenesis. Range changes along a branch have either a given cost (DIVA) or probability (DEC). The optimal reconstruction is found by searching the ancestral range and cladogenetic event assignments that minimize the total cost (DIVA) or maximize the likelihood (DEC).

The simultaneous inference of the event scenario and the ancestral range made both DIVA and DEC as the preferred tool in phylogenetic biogeography studies. But these methods have a critical drawback: they discard the explicit geographical range of the terminals analysed. Instead, they use a set of 'areas' or 'units' defined prior to the analysis (in this article I will use 'units' to refer to this predefined areas, and 'area' or 'geographical area' to indicate a measure of a surface). How to define these units is often far from clear, hampering testing of alternatives or reuse of data. Because of computational constraints (Ronquist, 1997; Ree *et al.*, 2005; Ree & Smith, 2008; Ree & Sanmartín, 2009), the number of units allowed by different programs is usually small (e.g. 8–12). As a consequence, many of these units represent large geographical areas, with the ranges of most terminals matching the units poorly. This results in many allopatric taxa lumped into a

http://wileyonlinelibrary.com/journal/jbi
doi:10.1111/jbi.13024

single unit, whereas some other taxa, with relatively small ranges, are scored as 'widespread.' As these methods treat taxa on a single unit and on multiple units differently, the resulting reconstruction may drastically change when the definition of the units is slightly modified, even if the actual geographical information (i.e. geographical area covered by a range, and its location) remains exactly the same (Arias *et al.*, 2011).

Hovenkamp (1997, 2001, 2002) produced a second breakthrough in the field when he proposed the use of explicit geographical ranges. The first method that uses explicit geographical data was developed for phylogeography, in which geographical locations of each haplotype are used as points (a vector data model), to infer, as a point, the position of ancestral haplotypes (Lemmon & Lemmon, 2008; Lemey *et al.*, 2010). This method has been extended to handle species by modelling their geographical range as polygons, and using these either as source for sampling of geographical points (Nylinder *et al.*, 2014) or directly as geographical ranges (Quintero *et al.*, 2015), but both methods model ancestral ranges as a single point. Other methods use an arbitrary grid (a raster data model) to represent spatial data, then treating ranges in both terminal and internal nodes as geographical areas. These methods are the spatial analysis of vicariance (Arias *et al.*, 2011; a method based on the ideas of Hovenkamp, 1997, 2001) and the Dispersal-Extinction model (DE, Landis *et al.*, 2013). In the spatial analysis of vicariance, the geographical range in an internal node is calculated as the sum of the ranges of its descendants (OR assignment), nodes without allopatric descendants have an extra cost and the distribution of some taxa might be ignored (with a given extra cost) in attempting to increase the number of nodes with allopatric descendants. On the other hand, in DE each descendant inherits a copy of its ancestor range, which is modified by addition and deletion of individual pixels (grid cells) along the branch.

With current methods, the cost to be paid for using explicit geography is to discard multiple biogeographical events: in phylogeographic methods and DE, only sympatry is allowed at each internal node, whereas in the spatial analysis of vicariance, nodes that cannot be interpreted as allopatric are left unassigned (i.e. not associated with specific events) and have an extra cost. While both Arias *et al.* (2011, p. 625) and Landis *et al.* (2013, p. 803) acknowledge that a multiple event-based approach is desirable, they do not attempt any solution.

Here I describe GEM (Geographically explicit Event Model) that can be seen as an attempt to merge the ideas of Hovenkamp (1997, 2001) and Ronquist (1997), as it is a multiple event method that uses explicit geographical ranges.

THE GEOGRAPHICALLY EXPLICIT EVENT MODEL

As in other event-based methods, there are two different kinds of processes modelled. First, how ranges are inherited by descendants at the cladogenetic event, and second what happened to a range along a lineage (Ronquist, 1997; Ree *et al.*, 2005).

In contrast with methods based on predefined areas, instead of evaluating events on the basis of the number of units in the ancestral range (Ronquist, 1997; Ree *et al.*, 2005), in GEM any event can be assigned to an internal node, without assuming a particular size of the range assigned to the node. How each descendant inherits a part of the geographical range is determined by the type of event assigned to the node. In the first event, vicariance, each descendant inherits a mutually exclusive part of the ancestor's range (the only event considered in the spatial analysis of vicariance, Arias *et al.*, 2011) (Fig. 1a). In the second event, sympatry (or range copying, Matzke, 2014), each descendant inherits an identical copy of the ancestor's range (the only event considered in DE) (Fig. 1b). If one of the descendants has a small subset of its ancestor's range, the assignment of sympatry will require an extensive extinction on that descendant; in such cases it is better to model sympatry as an event in which one descendant inherits a point inside the ancestor's range, whereas the other inherits the full ancestral range; this event is called here point sympatry (called subset sympatry by Matzke, 2014; used in predefined units in DEC) (Fig. 1c). In the fourth event, founder event, one descendant inherits the whole ancestor's range, and the other descendant starts as a founder population outside the ancestral range (used in predefined units by Matzke, 2014) (Fig. 1d). Ideally, the cost of each event is inversely related to the likelihood of that event.

Once an event and ancestral range are assigned to a tree node, the range change along the lineage (i.e. the amount of area gained or lost) is used as an extra cost. Different combinations of events and ancestral ranges at each node will produce reconstructions with different costs; those combinations with minimum cost are considered to be the best reconstructions. If event costs are assigned appropriately, this is also the most likely explanation of the data.

GEM IMPLEMENTATION

Geographical data model

The geographical data model used to represent both terminal and internal node ranges is a boolean raster grid in which a pixel is assigned to a set (a range) if there is a presence in the area covered by that pixel. The selection of the geographical data model is a matter of convenience: most operations just require simple pixel counts, unions and intersections. Although the method could be implemented with a vector data model using polygons or prim networks to represent ranges, calculations of surface overlap, or other geographical operations, should produce similar results but with more complex calculations.

Incomplete ranges

Unless distribution range maps are used, ranges are expected to be derived from point locations (i.e. georeferenced

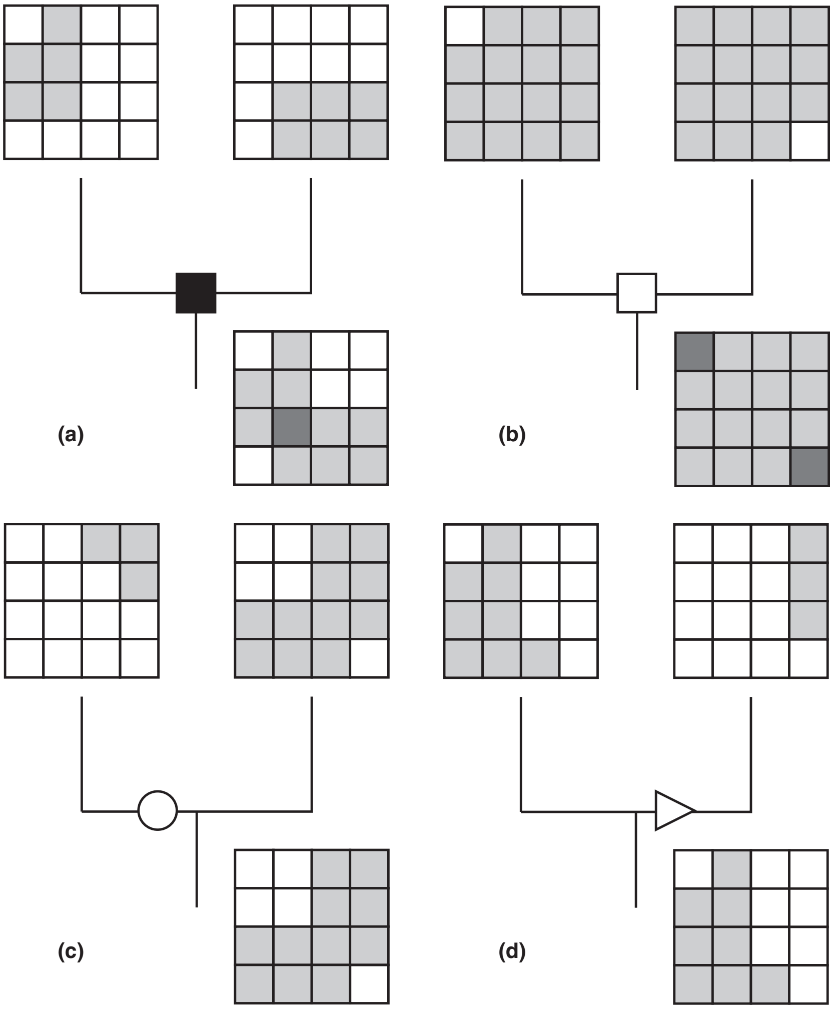

Figure 1 A graphical cartoon of the events implemented in GEM: (a) vicariance (the pixel in dark grey has an extra cost: it is an overlap); (b) sympatry (pixels in dark grey have an extra cost: they are absent in one of the descendants); (c) point sympatry and (d) founder event. Symbols for events were adapted from Page (1994) and Ronquist (2003).

specimens). This means that most ranges are only partially known, and increasing the raster resolution will increase the gaps among observations. To take this problem into account two pixel sets (rasters) are used for each node (either terminal or internal): the first containing only pixels with recorded observations, and the second using a simple filling (or buffering) algorithm: a given number of pixels around each pixel with an observation are set as filled. Note that if the filling is set as zero (e.g. when using range maps), both pixel sets will be identical. The main advantage of using a filling is that it allows the use of small pixels without introducing extensive gaps, as well as removing the sensitivity of point of origin of the raster (e.g. Aagesen *et al.*, 2009).

Cost assignment functions

As the cost of a reconstruction along a lineage is the range change, using a raster data model, this cost is expressed using the number of pixel changes (i.e. pixels as units of area). The main problem with any event-based method using a high-resolution raster is the large size of the data. A Sankoff-like optimization (Sankoff, 1975) would require a three-dimensional cost matrix (Ronquist, 1997) as some events, such as vicariance, depend on the range assignments in both descendant nodes. Fortunately, an optimal ancestral range is constrained by two rules (modified from Ronquist, 1997): (1) if an ancestor range includes a pixel, this pixel should be included in at least one of the descendants; and (2) if a pixel is shared by two descendants, this pixel should be present in the ancestor's range. Given these constraints, each event assignment can be calculated directly with the cost equations discussed below.

For notation, let D_i be the pixel set of the range assigned to node i, $\mathrm{D}_{f(i)}$ be the filled pixel set of i, and $|D_i|$ and $|\mathrm{D}_{f(i)}|$ the number of pixels in each set, respectively, a be the ancestral node and b and c the descendant nodes.

Vicariance

The cost of vicariance assignment is given by the sum of the overlap of each descendant against the other, plus the cost of a vicariance event (V):

$$C = |D_b \cap D_{f(c)}| + |D_c \cap D_{f(b)}| + V$$

This function strongly penalizes the overlap because it indicates that at least one descendant crossing a difficult to cross barrier.

Sympatry

The cost of assigning sympatry is just the cost of gained and lost pixels for each descendant relative to the ancestral distribution, plus the cost of a sympatry event (S):

$$C_b = \bigl(|D_b| - |D_b \cap D_{f(a)}|\bigr) + \bigl(|D_a| - |D_a \cap D_{f(b)}|\bigr)$$
$$C_c = \bigl(|D_c| - |D_c \cap D_{f(a)}|\bigr) + \bigl(|D_a| - |D_a \cap D_{f(c)}|\bigr)$$
$$C = C_b + C_c + S$$

Point sympatry

In point sympatry, the cost of the assignment is given by the size of the range on the point descendant minus one (so one pixel ranges will have no cost), plus the cost of point sympatry event (P). If the point descendant is c, the cost is as follows:

$$C = (|D_c| + \bigl(|D_c| - |D_c \cap D_{f(a)}|\bigr) - 1) + P$$

Also, in this function, the number of pixels outside the ancestral range is added as extra cost. This allows a differential cost with a founder event (see below).

Founder event

In a founder event the cost of the assignment is given by the size of the range of the founder descendant minus one (so one pixel ranges will have no cost), plus the cost of a founder event (J). If the founder descendant is c, the cost is as follows:

$$C = (|D_c| + |D_c \cap D_{f(a)}|\bigr) - 1) + J$$

Also, in this function, the cost of the pixels in overlap with the ancestral range is added as extra cost because they imply a backwards crossing over a barrier from the founder to the main stock. This also allows a cost differentiation with respect of point sympatry (see above).

Completing the model

Widespread ancestors

One of the most commonly observed problems with event-based methods is their tendency to produce widespread ancestors (e.g. Ronquist, 1997; Ree *et al.*, 2005). Proposed solutions include limiting the size of the ancestral area (Ronquist, 1997; Ree *et al.*, 2005), or suggesting events that reduce the ancestral range (e.g. Matzke, 2014). Instead of limiting the number of units (as used by DIVA and DEC), the approach taken here is to assign an additional cost based on size of the range assigned to the node minus one (so a single pixel range will be zero), modified by a weighting factor, Z. The extra cost for the range size of a node n is as follows:

$$C = (|D_n| - 1)/Z$$

The user can decide the value of factor Z, the greater the factor, the less influential the range size.

Event costs

As in GEM each event has a particular cost, this can be used to emulate previously proposed methods. For example, GEM-DE model (V=P=J=∞) emulates Landis *et al.*'s (2013) approach, and GEM-VIP (P=J=∞, V=S) emulates Arias *et al.*'s (2011). If a method limits a particular event to a single unit event, then that event is taken as not modelled geographically by the method, as it is dependent on the scale. For example, GEM-DEC only allows point sympatry (V=S=J=∞) because DEC only accepts vicariance for a single area, and GEM-DIVA only allows vicariance (S=P=J=∞) because DIVA only accepts sympatry for a single area.

In the computer implementation of GEM (see below), by default all events have the same cost, that is, they are considered equally likely. Under that weighting schema, reconstructions are discriminated only by the amount of added or deleted pixels in each lineage, that is, by the number of pixels that contradicts the event assignation. Of course, this cost can be changed by the user of the program.

Searching for optimal reconstructions

Although the use of cost assignment functions solves the problem of defining a three-dimensional cost matrix, it does not solve the problem of the ambiguity of assignments. Ambiguity consumes large amounts of computing resources (both time and memory) in nearly equivalent reconstructions, with small changes in ancestral ranges, and no changes at all in the events assigned to nodes. For example, suppose that two descendants share 90 pixels and each one has 10 exclusive pixels. If sympatry is assigned, a lot of different pixel assignments will be stored in memory, with the same cost, that just differ in how the 20 extra pixels are assigned to the ancestor (from 10 pixel extinctions to 10 pixel dispersals in each descendant).

Here I propose a heuristic approach that gives priority to event assignments, with conservative assignments of ancestral ranges. This approach is akin to the 'fixed state optimization' used in sequence data (Wheeler, 1999). Instead of lots of range assignments in ancestors, just three choices are attempted: the union of both descendant ranges (in vicariance and sympatry), or the range of one of the descendants (in sympatry, point sympatry, and founder event). That is, for each node, just eight different combinations of an ancestral range plus event will be evaluated.

Even with this simplification, searching for an exact solution is almost impossible, as eight alternative assignments per node imply 8^{t-1} possible reconstructions for a tree with t terminals. Therefore, I use the 'flip algorithm,' derived from Page's algorithm (Page, 1994), and previously implemented in vip (Arias *et al.*, 2011). This heuristic changes the ancestral range plus event assignment a single node at time and keeps those changes that reduce the cost, until no modification can improve the cost of the reconstruction (Table 1).

Although the method could be modified to include polytomous nodes, that is not attempted here. Instead, when a tree has a polytomous node, a conservative solution is used: the union of all descendants is assigned, and neither an extra cost, nor an event, is given to that node.

Computer implementation

The method presented here was implemented as a computer program evs, written in go (http://www.golang.org) as a command line application. To aid visualization of results a simple viewer was written in C using the gtk-2 library (http://www.gtk.org). The source of evs and its viewer is available at http://github.com/js-arias, and binaries for Linux and Windows are available at http://www.lillo.org/phylogeny/GEM.

Table 1 The flipping algorithm. In this algorithm the initial value of bestCost is calculated beforehand (e.g. with all nodes vicariant). The *eventsAssignment* list contains the eight valid *event Assignments* (see text). The method SetEvent changes the event assigned to a node and returns the new cost of that assignment.

Line Instruction

```
repeat := true
while repeat {
    repeat = false
    Randomize(nodeList)
    for n := range nodeList {
        Randomize(eventAssignment)
        ce = n.event
        for e := range eventAssignment {
            if ce == e {
                continue
            }
            cost := n.SetEvent(e)
            if cost < bestCost {
                bestCost = cost
                repeat = true
                break
            }
        }
        if repeat {
            break
        }
        n.SetEvent(ce)
    }
}
```

AN EMPIRICAL EXAMPLE

I apply the method proposed here to the 'Euvireya' clade of the plant genus *Rhododendron* L. This group is distributed in the Malesian archipelago, and their phylogenetic relationships, distribution and biogeography have recently received significant attention. Heads (2003) studied the group from a panbiogeographical perspective. Brown *et al.* (2006) provide explicit geographical data but, just like subsequent workers (Webb & Ree, 2012; Landis *et al.*, 2013), used predefined units instead of the compiled geographical data.

Materials and methods

I use the phylogeny reported by Webb & Ree (2012) with geographical data taken, by hand, from Brown *et al.*'s (2006) maps with explicit sampling points for 62 species included in the phylogeny. Data from *R. sarcodes* (geographical data not given by Brown *et al.*) was taken from a range map in Heads (2003; which is compatible with the predefined unit used by Brown *et al.*). As Brown *et al.* do not report geographical ranges of the outgroups, records from *R. maddenii* were taken from GBIF (http://www.gbif.org) and *R. lindleyi* was removed from the analysis as no records were available in that database (so the tree used here has 64 terminals instead of 65).

I ran the data with GEM as implemented in evs, using a raster grid with pixels of 1°×1° degrees, with a filling of 1. The cost of the four cladogenetic events is set to one. To penalize large ancestral ranges I use a $Z = 10$. The search was made with the flipping algorithm applying 10 independent runs each with 10,000 flip replicates (for a grand total of 10^5 flip replicates). For comparison, a search using the same parameters was performed without using the Z modification (so ancestral ranges are not penalized by their size). Several independent runs were also made, using the same raster and search parameters, in which one or more kind of events were not allowed.

For comparison I also ran the spatial analysis of vicariance (Arias *et al.*, 2011) with vip (available at: http://www.lillo.org.ar/phylogeny/VIP), using the same raster parameters, accepting an overlap up to 25%, using real values to measure overlap, an elimination cost of 2 and 0 (i.e. maximizing disjunctions), and a search with 10,000 flip iterations. I also ran the bayarea (available as source code at: http://github.com/mlandis/bayarea), the Bayesian implementation of the DE model (Landis *et al.*, 2013) using the geographical coordinates of the centre of each observed pixel. I made six independent runs using the m3 model (which takes into account the current distance between pixels), each using 1.6×10^8 generations, sampling parameters and histories every 1000 generations (default), and discarding the first 10^7 samples as burn-in. Convergence of MCMC was verified with the Gelman diagnostic (Gelman & Rubin, 1992) as implemented in the R package 'coda' (Plummer *et al.*, 2006). Ancestral ranges were taken from a single run (run 2, picked at random), and only pixels with a posterior probability > 0.5 taken into account.

Datasets, batch files used for runs, logs and output of the programs are available in Appendix S1 in Supporting Information.

Results

The search with GEM found two different reconstructions with a cost of 293.80. The reconstructions have the same event assignment in all nodes. Overall numbers of events are 11 vicariance events, 22 sympatry events, six point sympatry events, and 24 founder events (Table 2; Fig. 2). All ancestral areas assignments are available in the Fig. S2.1 in Appendix S2.

The ancestor of the 'Euvireya' clade (see Fig. 2a; Appendix S2) was reconstructed as widespread, present in Southern Malay Peninsula, Borneo, Sulawesi and Seram, with a founder event in one of its descendants to Sumatra and Java. If this reconstruction is correct, then Euvireya never 'crosses' the Wallace line (contra Webb & Ree, 2012; Landis *et al.*, 2013), as the group is already present in both parts of the Wallace line at the beginning of its history. Webb & Ree (2012) estimated the age of the group as *c.* 55–45 Ma, so this reconstruction is consistent with West Sulawesi attached to Borneo at the beginning of the Cenozoic (Hall, 1996; Lohman *et al.*, 2011; Zahirovic *et al.*, 2014). However, some specific clades within the group have crossed the line at later times. For the origin of Eastern Malesia clade, the reconstruction found a founder event (Fig. 2c,d; see Appendix S2) from Western and Middle Malesia (i.e. a crossing of the Lydekker's line), although this reconstruction is sensitive to *Z* (the ancestral area size penalization; results not shown).

Table 2 Results of GEM under different event and parameter combinations, for the geographical dataset of Brown *et al.* (2006) and the phylogeny of Webb & Ree (2012). The model names reflect their most similar found in literature (see text). Found indicates the number of found reconstructions. Columns V, S, P and J indicates the number of events found of vicariance, sympatry, point sympatry and founder event, respectively, using a n-dash to indicate events prohibited under that model. The cost column indicates the cost of the reconstruction(s). All models, except for GEM-noZ, use a range weight of 1/10 for the range size of each node, so GEM-noZ is the only cost value not comparable with other cost values. The ratio of the best solution against each model is indicated for quick comparison of the effect of prohibiting one or more events relatively to the full events used in GEM. Models are sorted by its cost.

Model	Found	V	S	P	J	Cost	Ratio
GEM-noZ	91	8–13	19–23	7–13	18–24	206.00	–
GEM	2	11	22	6	24	293.80	–
GEM-VIP+J	1	12	24	–	27	311.88	0.942
GEM-No vicariance	2	–	22–23	6–7	34	323.20	0.909
GEM-DE+J	1	–	26	–	37	337.80	0.870
GEM-No founder	3	24–25	26–28	11–12	–	342.00	0.859
GEM-VIP	1	32	31	–	–	366.00	0.803
GEM-No sympatry	2	4	–	12	47	421.90	0.696
GEM-DEC+J	5	–	–	9–14	49–54	437.90	0.671
GEM-DIVA+J	24	5	–	–	58	455.70	0.645
GEM-DE	20	–	63	–	–	481.70	0.610
GEM-DEC	48	–	–	63	–	630.50	0.466
GEM-DIVA	1	63	–	–	–	1111.70	0.264

Comparisons

For this dataset, the search without penalizing the size of the ancestral range produces more ambiguous results and larger ancestral areas. When comparing ancestral ranges, the most notable difference is that in some reconstructions, New Guinea forms part of the ancestral range of Euvireya.

Using different combinations of event costs it is possible to compare the full model to other models that have more restricted sets of events (Table 2). All combinations that prohibit one or more events are at least 5% more costly than the reconstructions using all events. This is expected, given the complex history of the region: all four events need to be invoked at some part of the tree to obtain the best account of current ranges.

When looking at the results of BAYAREA, when an ancestor is 'widespread', it contains few, but widespread, pixels. For example, the reconstruction of Euvireya has only five pixels: one in Sumatra, one in Borneo, two in Java and one in New Guinea. In cases in which two descendants are allopatric, most of the times the reconstruction is a single pixel in the middle of both descendant ranges, or one or two pixels in one of each descendant ranges, or an ancestral range without a pixel with a posterior larger than 0.5. When derived nodes in GEM are inferred as sympatric, both EVS and BAYAREA produce nearly identical results. Most significant are the differences in computing time: whereas all EVS analyses with all models took about 40–50 min, every single run of BAYAREA takes about 40 h in the same machine.

In contrast, VIP produces more widespread ancestral ranges than EVS. But as the main objective of VIP is to find disjunctions, it is worth to notice that most disjunctions are shared between both results. In fact, in different EVS results, most disjunctions (regardless of whether they are vicariance or founder event) remain the same (Table 2), which implies that disjunctions are a more stable result than ancestral ranges.

DISCUSSION

Predefined areas versus explicit geography

The event-based method presented here departs from previous methods such as DIVA or DEC in that it uses an explicit

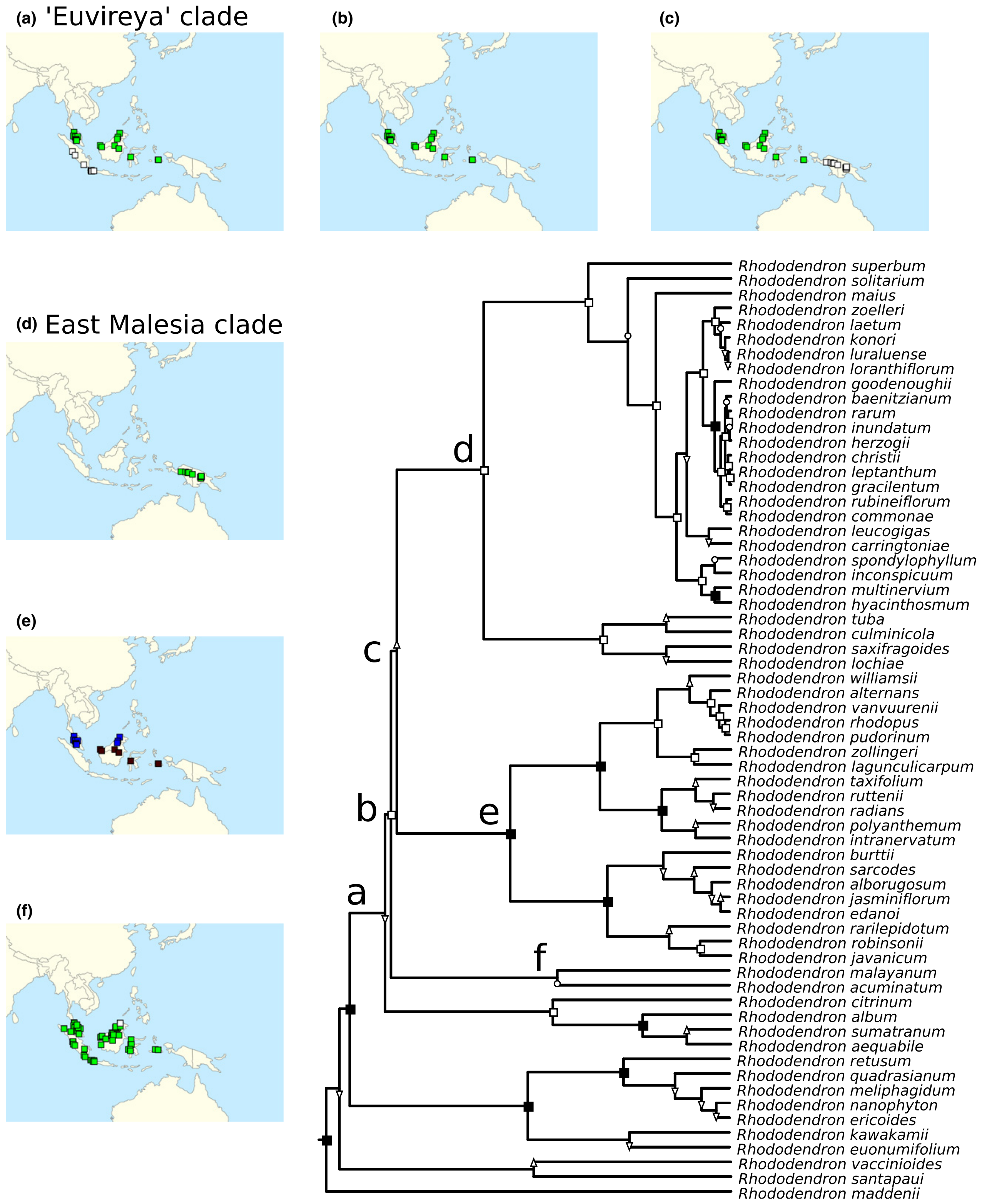

Figure 2 The optimal reconstructions found for the geographical data, the distribution of *Rhododendron* from Brown *et al.* (2006) and the phylogenetic tree of Webb & Ree (2012). Symbols for reconstruction are the same as used in the Fig. 1. (a–f) Maps with the ancestral range reconstructions for the nodes indicated in the reconstruction. In maps with sympatry (b, d), point sympatry (f) and founder event (a, c), the ancestral range is drawn in green (web version), whereas in vicariance (e) the ancestral range is the union of descendant ranges the different colours provide a graphical aid to indicate the disjunction. In founder event (a, c) and point sympatry (f), 'founder' and 'point' descendant are drawn in white squares. For all ancestral node reconstructions see the Appendix S2.

representation of the geography of the terminals analysed (i.e. a geographical data model). In this section I want to expand some particular aspects of the method with an emphasis on the geographical nature of the data.

Most recent discussion on phylogenetic biogeography is centred around the development of statistical-based methods (e.g. Lamm & Redelings, 2009; Ree & Sanmartín, 2009; Ronquist & Sanmartín, 2011; Wen *et al.*, 2013). Whereas this discussion is welcome, much of that discussion has the problem that the geographical context of the data is replaced by a 'dispersal matrix' (e.g. Ree *et al.*, 2005; Ree & Sanmartín, 2009). As a result, the development of such methods usually sacrifices geographical resolution (i.e. it requires fewer input units) in favour of model parametrization, for example, GeoSSE (Goldberg *et al.*, 2011) and Shiba (Webb & Ree, 2012) have complex and detailed process models but allow only a handful of units. Decreasing resolution has been justified both on computational grounds (Ree *et al.*, 2005; Ree & Smith, 2008; Ree & Sanmartín, 2009), but also with the argument that higher resolution reduces 'information content' (Ree & Sanmartín, 2009) by producing more gaps in the dataset. But, instead of using predefined units for this second case, a more fruitful approach is to use geographical methods to reduce those sampling gaps. For example, Aagesen *et al.* (2009) propose the use of buffering, which produces a filling around an observed pixel (implemented here), and Landis *et al.* (2013) propose the use of predictive niche models.

There is no clear definition of what a predefined unit is. They are justified on several grounds, such as areas of endemism (e.g. Nelson & Platnick, 1981; Parenti & Ebach, 2009; Wiley & Lieberman, 2011), geological areas (e.g. Wiley & Lieberman, 2011), hypothesis of 'spatial homology' (Morrone, 2001), biotic elements (e.g. Hausdorf, 2002) or 'by the question being asked' (Ree *et al.*, 2005; Ree & Sanmartín, 2009; Ronquist & Sanmartín, 2011). Whatever the definition, some methods make a distinction between ancestral ranges with a single unit and ancestral ranges with multiple units, without considering the geographical area, for example, in original DIVA description (Ronquist, 1997) sympatry in multiple areas is prohibited, and in original DEC description both vicariance and sympatry in multiple areas is prohibited (Ree *et al.*, 2005). When such restriction is enforced, results may depend more on the prior geographical partition than on actual distribution data. For example, suppose a tree with four terminals, all occupying the same area, a square of 100 km^2, that is labelled A, so the input tree on DIVA will be (A,(A,(A,A))) and the result will be sympatry on every node, with A as ancestral range, and no dispersal. If you split unit A in two units of equal size (say A1, and A2), the geographical content of the data does not change (it covers the same geographical area: each terminal has a geographical area of 100 km^2, in the same spatial location), but the reconstruction is different both in terms of the events (either vicariance or sympatry, dispersals in most terminals) and in terms of the ranges (in each node all possible ancestral range combinations are possible).

There are several advantages to using explicit geographical information. For example, it usually requires primary data. As a consequence testing or reproducing results, and sharing, curating, verifying and expanding the dataset, or even using it for any other kind of analysis, will be easier and more transparent: instead of a collection of terminals assigned to a poorly defined units on nebulous grounds, the dataset will be a list of actual sampling points (which might include specimen locality data and museum catalogue numbers, as in standard taxonomy revisions) or explicit range maps.

The use of explicit geographical ranges also removes some problems that are a consequence of predefined areas rather than a consequence of biogeographical methodology (Hovenkamp, 1997, 2002). Two clear cases are the 'biogeographical assumptions' on widespread taxa (Nelson & Platnick, 1981), of how to deal with overlapped terminal ranges (Axelius, 1991).

Disjunctions

The method presented here is a development of the spatial analysis of vicariance (Arias *et al.*, 2011). Although both methods have somewhat different objectives (inference of disjunctions vs. inference of events and ancestral ranges), they use the same algorithmic mechanics. Their main difference is in how they score the range assignment to internal nodes, which in turn may (obviously) change the optimal assignment of each ancestor. As assignments in GEM take more possible events into account, GEM allows a more detailed explanation of the nature of the disjunction (vicariance vs. founder event). Also, instead of *ad hoc* elimination of some node ranges to increase the number of disjunctions, explicit event and cost assignments in GEM allow an explanation of observed ranges with a strong biogeographical component. So, the method is, as defined by Hovenkamp (1997, 2002), a full taxon history method.

Some authors (e.g. Ree *et al.*, 2005; Lamm & Redelings, 2009; Kodandaramaiah, 2010) have argued that methods based on disjunctions, such as DIVA (Ronquist, 1997) or the spatial analysis of vicariance (Arias *et al.*, 2011), favour or overweight vicariance. At least for the dataset analysed here, this criticism does not hold, as results of GEM (which does not favour any event over others) are highly similar to those using GEM-VIP (which only allows vicariance and dispersal), at least in terms of disjunctions. Then both objectives, the detection of disjunction and the assignment of ancestral ranges, are better seen as two sides of the same coin: improving one will help improve the other (Hovenkamp, 2002).

Hovenkamp (1997, 2001, 2002) argued that instead of predefined areas, it is better to use explicit ranges and focus on disjunctions. This work (as Arias *et al.*, 2011) is based on Hovenkamp's ideas, and shows that they can be implemented in a formal way. Although it does not provide a complete computational implementation of the Hovenkamp analysis, it can be used as a starting point (see, e.g. Domínguez *et al.*, 2016).

Ancestral range assignments

For the present implementation I made a very simplistic assignment of ancestral ranges. This has the heuristic value of speeding up reconstruction when the data are slightly ambiguous. Such ambiguity is surely the most common situation in biogeography, in which even two species with very similar ranges can somewhat differ, then the inference machinery without such heuristics would easily stall in a large number of almost identical solutions with the same cost. I hope that future research will find algorithms that allow faster examination of some configurations (maybe as a refinement after an initial search using the algorithm proposed here) that can increase the goodness of a particular set of event assignments.

Despite its simplicity, the cost function (i.e. the optimality criterion) does not depend on how the ancestral range is assigned, so even if a better algorithm for assigning ancestral ranges is eventually developed, it would be possible to make the comparison between the results of the new algorithm with respect to the ones found with the current heuristics.

Limitations of the method

Although the method presented here has the advantage of being fully event based and geographically explicit, it has some limitations that cannot yet be solved. When possible, I will try to point to some potential solution, or how the research might be directed to solve these open problems.

The method has the flexibility of any event cost schema, but this leaves open the question of how to calculate these costs. One of the advantages of the explicitly probabilistic methods is that they provide a direct way to numerically compare several cost schemas. In GEM such comparison can only be done when comparing models in which one or more events are prohibited and all other events have the same cost (as is done in the empirical example). Fortunately, this is the most usual test done in probability-based studies (e.g. Matzke, 2014). As in general it seems preferable to take into account all the events (instead of prohibiting them), it is worth to note that a more critical test in the context of predefined units should be a comparison of alternative unit assignments (as suggested by Landis *et al.*, 2013).

Other tests to estimate the effect of the event costs can be performed, for example, using a range permutation (Page, 1994; Siddall, 2001; Ronquist, 2003): ranges in terminals are permuted at random, and an analysis is done, and then comparing the score of the original data against the permuted data. This is repeated several times, and cost schemas that produce the most significant results will be preferred. In this kind of study we are choosing the parameters that maximize our chances of detecting historical signal in the data (Ronquist, 2003). As EVS is relatively quick, this kind of tests (either prohibiting events or a permutation test) can be easily implemented. In fact procedures developed for DIVA (e.g. Bayes-DIVA, Nylander *et al.*, 2008), can be implemented in a straightforward fashion to EVS.

Another limitation of the method is that it does not include branch length information. Although a simple implementation is added to the model in EVS (in which branch lengths are used to down weight the cost of dispersal and extinction, and to up weight ancestral range size costs), this tentative solution has a problem: it will penalize vicariance, as this is the only event that adds a significant amount of pixels to an ancestral distribution, without the benefit of down weighting those pixels. Until a more appropriate proposal of how branch lengths can be used in the context of parsimony, it may be best that these data remain ignored. It can be argued that while anagenetic dispersal and extinction are not modelled with branch lengths, the most important impact on distributions happens at speciation events (e.g. vicariance, founder events), which is the main reason to look for an event-based approach.

The current GEM definition does not include distances (i.e. a factor for the length of dispersal). While this is an important subject, and I am working on several potential alternatives, I am not able to provide a solution in this moment. If a researcher believes that an analysis with branch lengths and/or distances is required she should consider using a method such as DE (Landis *et al.*, 2013), at the cost of not using event assignments at nodes.

CONCLUDING REMARKS

GEM provides a tool for reconstruction of ancestral ranges in a clade using explicit geographical ranges and a cost schema for biogeographical events. GEM can be seen as a part of a trend, started by Hovenkamp, of developing methods which explicitly incorporate the geographical information of distribution ranges. While the mechanics of the inference is important, so is the way in which the spatial data is represented. I hope that the framework presented here helps move the discussion of biogeographical methods towards its geographical aspect.

ACKNOWLEDGEMENTS

I thank C.A. Szumik and P.A. Goloboff for their continued support on the development of the ideas presented in this study. M.D. Casagranda has given encouragement and countless comments on this work. M. Landis, I. Sanmartín (A.E.) and two anonymous reviewers provided many helpful and detailed comments to improve this study. I also thank J. Grosso for her love and support. This work was made at the facilities provided by the Unidad Ejecutora Lillo (UEL), CONICET-Fundación Miguel Lillo. Funding was provided by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET, doctoral and post-doctoral fellowships, and PIP 0687 to Pablo Goloboff), and the Global Biodiversity Information Facility (GBIF, a Young Research Award).

REFERENCES

Aagesen, L., Szumik, C.A., Zuloaga, F.O. & Morrone, O. (2009) Quantitative biogeography in the South America highlands—recognizing the Altoandina, Puna and Prepuna through the study of Poaceae. *Cladistics*, **25**, 295–310.

Arias, J.S., Szumik, C.A. & Goloboff, P.A. (2011) Spatial analysis of vicariance: a method for using direct geographical information in historical biogeography. *Cladistics*, **27**, 617–628.

Axelius, B. (1991) Areas of distribution and areas of endemism. *Cladistics*, **7**, 197–199.

Brown, G.K., Nelson, G. & Ladiges, P.Y. (2006) Historical biogeography of *Rhododendron* section *Vireya* and the Malesian archipelago. *Journal of Biogeography*, **33**, 1929–1944.

Brundin, L. (1966) Transantartic relationships and their significance, as evidenced by Chironomid midges. *Kungliga Svenska Vetenskapsakademiens Handligar, Fjärde serien*, **11**, 1–142.

Domínguez, M.C., Agrain, F.A., Flores, G.E. & Roig-Juñent, S.A. (2016) Vicariance events shaping Southern South American insect distributions. *Zoologica Scripta*, **45**, 504–511.

Gelman, A. & Rubin, D.B. (1992) Inferences from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

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

Hall, R. (1996) Reconstructing Cenozoic SE Asia. *Tectonic evolution of Southeast Asia* (ed. by R. Hall and D. Blundell), pp. 153–184. Geological Society, London.

Hausdorf, B. (2002) Units in biogeography. *Systematic Biology*, **51**, 648–652.

Heads, M. (2003) Ericaceae in Malesia: vicariance biogeograpahy, terrane tectonics and ecology. *Telopea*, **10**, 311–450.

Hennig, W. (1966) *Phylogenetic systematics*. University of Illinois Press, Urbana.

Hovenkamp, P. (1997) Vicariance events, not areas, should be used in biogeographical analysis. *Cladistics*, **13**, 67–79.

Hovenkamp, P. (2001) A direct method for the analysis of vicariance patterns. *Cladistics*, **17**, 260–265.

Hovenkamp, P. (2002) Biogéographie de la vicariance: “mess’ ou message? *Biosystema 20, Systématique et biogéographie* (ed. by P. Deleporte, J.-F. Silvain and J.-P. Hugot), pp. 15–26. Société de Systématique, Paris.

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

Lamm, K.S. & Redelings, B.D. (2009) Reconstructing ancestral ranges in historical biogeography: properties and prospects. *Journal of Systematics and Evolution*, **47**, 369–382.

Landis, M.J., Matzke, N.J., Moore, B.R. & Huelsenbeck, J.P. (2013) Bayesian analysis of biogeography when the number of areas is large. *Systematic Biology*, **62**, 789–804.

Lemey, P., Rambaut, A., Welch, J.J. & Suchard, M.A. (2010) Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, **27**, 1877–1885.

Lemmon, A.R. & Lemmon, E.M. (2008) A likelihood framework for estimating phylogeographic history on a continuous landscape. *Systematic Biology*, **57**, 544–561.

Lohman, D.J., de Bruyn, M., Page, T., von Rintelen, K., Hall, R., Ng, P.K.L., Shih, H.-T., Carvalho, G.R. & von Rintelen, T. (2011) Biogeography of the Indo-Australian Archipelago. *Annual Review of Ecology, Evolution, and Systematics*, **42**, 205–226.

Matzke, N.J. (2014) Model selection in historical geography reveals that founder-event speciation is a crucial process in island clades. *Systematic Biology*, **63**, 951–970.

Morrone, J.J. (2001) Homology, biogeography and areas of endemism. *Diversity and Distributions*, **7**, 297–300.

Nelson, G. & Platnick, N. (1981) *Systematics and biogeography: cladistics and vicariance*. Columbia University Press, New York.

Nylander, J.A.A., Olsson, U., Alström, P. & Sanmartín, I. (2008) Accounting for phylogenetic uncertainty in biogeography: a Bayesian approach to dispersal-vicariance analysis of the thrushes (Aves: *Turdus*). *Systematic Biology*, **57**, 257–268.

Nylinder, S., Lemey, P., de Bruin, M., Suchard, M.A., Pfeif, B.E., Walsh, N. & Anderberg, A.A. (2014) On the biogeography of *Centipeda*: a species-tree diffusion approach. *Systematic Biology*, **63**, 178–191.

Page, R.D.M. (1994) Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics*, **10**, 155–173.

Parenti, L.R. & Ebach, M.C. (2009) *Comparative biogeography: discovering and classifying biogeographical patterns of a dynamic Earth*. University of California Press, Berkeley.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006) Coda: convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.

Quintero, I., Keil, P., Jetz, W. & Crawford, F.W. (2015) Historical biogeography using species geographical ranges. *Systematic Biology*, **64**, 1059–1073.

Ree, R.H. & Sanmartín, I. (2009) Prospects and challenges for parametric models in historical biogeographical inference. *Journal of Biogeography*, **36**, 1211–1220.

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.

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

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

Ronquist, F. (2003) Parsimony analysis of coevolving associations. *Tangled trees: phylogeny, cospeciation, and coevolution* (ed. by R.D.M. Page), pp. 22–64. University of Chicago Press, Chicago.

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

Sankoff, D. (1975) Minimal mutation tree sequences. *SIAM Journal on Applied Mathematics*, **28**, 35–42.

Siddall, M.E. (2001) Computer-Intensive randomization in systematics. *Cladistics*, **17**, S35–S52.

Webb, C.O. & Ree, R.H. (2012) Historical biogeography in Malesia. *Biotic evolution and environmental changes in Southeast Asia* (ed. by D. Gower, K. Johnson, J. Richardson, B. Rosen, L. Ruber and S. Williams), pp. 191–215. Cambridge University Press, Cambridge.

Wen, J., Ree, R.H., Ickert-Bond, S.M., Nie, Z. & Funk, V. (2013) Biogeography: where do we go from here? *Taxon*, **62**, 912–927.

Wheeler, W.C. (1999) Fixed character states and the optimization of molecular sequence data. *Cladistics*, **15**, 379–385.

Wiley, E.O. & Lieberman, B.S. (2011) *Phylogenetics: theory and practice of phylogenetic systematics*, 2nd edn. Wiley-Blackwell, Hoboken.

Zahirovic, S., Seton, M. & Müller, R.D. (2014) The Cretaceous and Cenozoic tectonic evolution of Southeast Asia. *Solid Earth*, **5**, 227–273.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Input data and output logs for *Rhodondendron* dataset.

Appendix S2 Optimal reconstructions for *Rhodondendron* dataset.

BIOSKETCH

J. Salvador Arias teaches biogeography at the Universidad Nacional de Tucumán (Argentina). He has a strong interest in theoretical, computational and geographical aspects of historical biogeography. He also works on phylogenetic methodology.

Editor: Isabel Sanmartín