Molecular Phylogenetics and Evolution 87 (2015) 46–49

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Short Communication

RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical biogeography

Yan Yu [a,b,*], A.J. Harris [c], Christopher Blair [b,d], Xingjin He [a,*]

[a] Key Laboratory of Bio-Resources and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, Sichuan 610065, PR China
[b] Department of Biology, Duke University, Box 90338, BioSci 130 Science Drive, Durham, NC 27708, USA
[c] Department of Botany, Oklahoma State University, 301 Physical Science, Stillwater, OK 74078-3013, USA
[d] Department of Biological Sciences, New York City College of Technology, The City University of New York, 300 Jay Street, Brooklyn, NY 11201, USA

ARTICLE INFO

Article history:
Received 4 October 2014
Revised 9 March 2015
Accepted 12 March 2015
Available online 26 March 2015

Keywords:
Historical biogeography
Ancestral state reconstruction
Software

ABSTRACT

We announce the release of Reconstruct Ancestral State in Phylogenies (RASP), a user-friendly software package for inferring historical biogeography through reconstructing ancestral geographic distributions on phylogenetic trees. RASP utilizes the widely used Statistical-Dispersal Vicariance Analysis (S-DIVA), the Dispersal–Extinction–Cladogenesis (DEC) model (Lagrange), a Statistical DEC model (S-DEC) and BayArea. It provides a graphical user interface (GUI) to specify a phylogenetic tree or set of trees and geographic distribution constraints, draws pie charts on the nodes of a phylogenetic tree to indicate levels of uncertainty, and generates high-quality exportable graphical results. RASP can run on both Windows and Mac OS X platforms. All documentation and source code for RASP is freely available at http://mnh.scu.edu.cn/soft/blog/RASP.

1. Introduction

Historical biogeography can be defined as the study of species distributions over evolutionary time scales. Methods to reconstruct ancestral geographical distributions using a combination of phylogenetic and distributional information are increasing rapidly, and several new software packages are being developed. However, many of these packages are command-line only and can present a steep learning curve for researchers not familiar with the interface (Landis et al., 2013; Ree and Smith, 2008), limiting the widespread adoption of computational methods of historical biogeography.

To make historical biogeographic reconstructions using phylogenies more accessible, we introduce RASP (Reconstruct Ancestral State in Phylogenies), which provides a graphical user interface (GUI) for existing popular historical biogeographic software packages. Since its original inception, RASP has been in wide use for historical biogeographic research (e.g. Miraldo and Hanski, 2014; Moyle et al., 2012; Schenk et al., 2013), presumably because the program aggregates and enhances all methods from diverse software packages. In RASP 3.0, we have now improved the implementation of the Statistical-Dispersal Vicariance Analysis (S-DIVA; Yu et al., 2010), added the Dispersal–Extinction–Cladogenesis (DEC; Ree and Smith, 2008) and BayArea (Landis et al., 2013) (Fig. 1a) models, and written a Statistical DEC (S-DEC; Beaulieu et al., 2013) method. Here, we describe the implementations of S-DIVA, DEC, S-DEC, and BayArea, as well as two additional tools in RASP.

2. Description

2.1. Enhanced S-DIVA method

The most well-known and commonly used event-based method of biogeographic inference is Dispersal–Vicariance Analysis (DIVA) (Ronquist, 1996, 1997, 2001). Nylander et al. (2008) applied a Bayesian approach to DIVA (Bayes-DIVA) in which biogeographic reconstructions were averaged over a sample of highly probable Bayesian trees. The S-DIVA method is an expansion of Bayes-DIVA and has been previously described in detail (Yu et al., 2010). Briefly, in S-DIVA the occurrence of an ancestral range at a node could be calculated using the frequency of all of the alternative reconstructions generated by the DIVA algorithm for each tree in the data set (when "Allow Reconstruction" is checked), while Bayes-DIVA uses only the summary of the alternative reconstructions. In addition, trees from sources other than Bayesian analyses

* Corresponding authors at: Key Laboratory of Bio-Resources and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, Sichuan 610065, PR China (Y. Yu).

E-mail addresses: aj.harris@okstate.edu (A.J. Harris), xjhe@scu.edu.cn, sculab@gmail.com (X. He).

http://dx.doi.org/10.1016/j.ympev.2015.03.008
1055-7903/

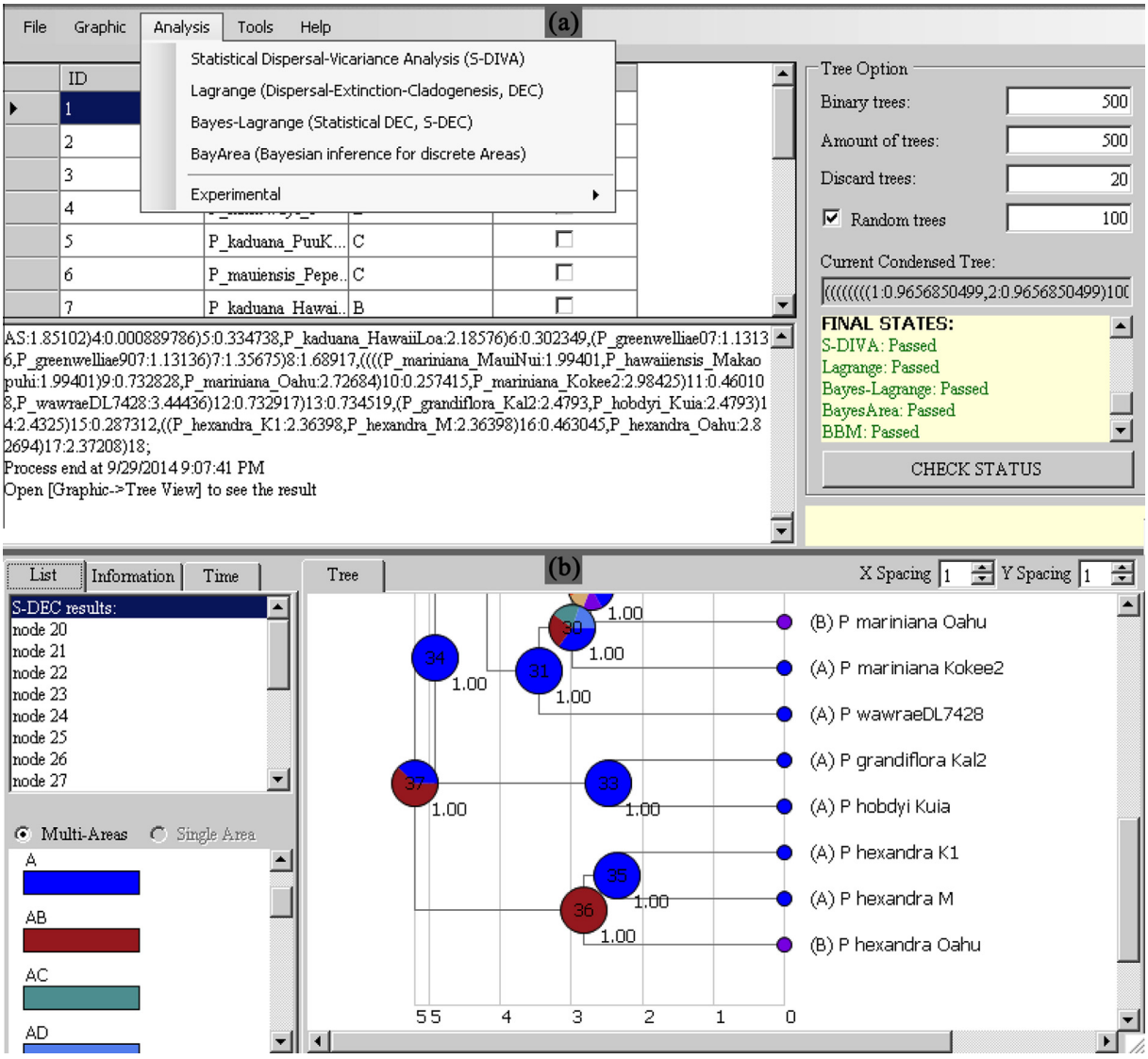

Fig. 1. (a) The main screen of RASP. The "Analysis" menu in the menu bar provides access to S-DIVA, DEC (Lagrange), S-DEC (Bayes–Lagrange) and BayArea methods. (b) The "Tree View" screen of RASP. The modern range for each taxon is color coded, and is drawn on the terminal lineages before each taxon's name. Pie charts at internal nodes represent the marginal probabilities for each alternative ancestral area.

are allowed. In contrast with many other biogeographic methods, S-DIVA can explicitly utilize an entire posterior distribution of trees to account for both phylogenetic uncertainty and uncertainty in ancestral states. S-DIVA is in wide use for estimating historical biogeography (e.g., Harris et al., 2013; Miraldo and Hanski, 2014) and has also been applied to testing the coevolutionary history of parasites (Razo-Mendivil et al., 2011) and bacteria (Comas et al., 2013) with their hosts.

The main feature that is new to S-DIVA in RASP is the ability to remove user-specified, widespread distributions from analyses. This option is particularly useful for excluding biologically unlikely widespread ranges and hypothesis testing. To allow removal of user-specified geographic ranges, we modified the source code of DIVA (Ronquist, 2001). The original DIVA algorithm encodes four different types of biogeographic events: dispersal, extinction, vicariance and duplication (Ronquist, 1997). As DIVA optimizes reconstructions across a phylogenetic tree, the algorithm follows a rule set in which an optimal distribution of an ancestral node cannot contain a unit area not occupied by any descendant. The outcome of this rule is that extinction events will never appear in dispersal–vicariance optimizations (Ronquist, 2001; Kodandaramaiah, 2010). Thus, if some user-specified ranges are excluded, a null (or empty) result may occur; namely if the only geographic ranges that are consistent with the rule have been eliminated. For example, suppose that the total distribution is {*A*, *B*, *C*}, $N_L = A$ and $N_R = B$. When geographic range *AB* is excluded, *ABC* should be proposed as the ancestral range, but *ABC* violates the rule set of the DIVA algorithm. Therefore, we have made the following modification in S-DIVA: Assume that the ancestral range of the node, *i*, is A_i, then the descendant nodes (terminal) are N_L and N_R. Let $|X|$ be the number of elements in *X*. Then the cost of an extinction event E_i could be calculated as $|A_i| - |N_L \cup N_R \cap A_i|$. When no ranges are excluded, the algorithm of S-DIVA is as same as DIVA.

Our enhanced S-DIVA algorithm has been tested using the simulated data set of 100 randomly sampled trees from Harris and Xiang (2009; data available from http://www.plant-biogeography.webs.com). Results obtained from S-DIVA were identical to those reported from manual calculations by Harris and Xiang (2009).

2.2. DEC and S-DEC model

The DEC and S-DEC model of geographic range evolution is implemented in RASP using the publically available source code for the C++ version of Lagrange (Smith, 2010). All basic options of Lagrange are available within the RASP GUI. In addition, the RASP version of DEC exports the likelihood of all possible biogeographic scenarios estimated at a given node. Akaike weights (AICw) for alternative models can then be calculated and interpreted as the relative probability of different ancestral ranges, which are displayed as pie charts on the nodes of a phylogenetic tree (Beaulieu et al., 2013). Thus, the additions to the DEC model in RASP allow for a more comprehensive evaluation of the degree of ancestral state uncertainty in biogeographic reconstructions.

In recent years, many researchers have become interested in combining Bayesian estimates of evolutionary relationships and divergence times with the DEC model to incorporate phylogenetic uncertainty in biogeographic inference (Beaulieu et al., 2013; Smith, 2009). The application of DEC across a distribution of phylogenetic trees has been tentatively termed "Bayes–Lagrange", and may be considered a non-parametric Bayesian method (similar to Bayes-DIVA in Harris and Xiang, 2009; Johns, 1957).

In RASP, we call the implementation of "Bayes–Lagrange" Statistical Dispersal–Extinction–Cladogenesis (S-DEC) to emphasize the similarity of the technique to S-DIVA; namely that it summarizes biogeographic reconstructions across all user-supplied trees. In S-DEC, the DEC model is applied to each ultrametric tree within a posterior distribution resulting from a Bayesian phylogenetic analysis. Subsequently, the probability (p) of an ancestral range x at node n on a summary tree (e.g., maximum clade credibility tree, majority rule consensus, etc.) is calculated as $p(x_n) = p_n * \sum_{t=1}^{m} w(x_n)t$ where t is the selected tree and m is the total number of sampled trees. At each node on the summary tree, a corresponding AICw is calculated for alternative ancestral states where $w(x_n)_t$ is the AICw of an ancestral range x at node n for tree t, and p_n is the support for the node.

Both the DEC and S-DEC models in RASP have been tested using a data set comprising the plant family Hyacinthaceae (Subfamily Urgineoideae), for which a biogeographic history was previously estimated with DEC in Lagrange (Ali et al., 2013). We obtained results from RASP that were identical to those estimated with DEC/S-DEC in Lagrange.

2.3. BayArea method

BayArea extends the application of biogeographic models to the analysis of realistic problems that involve a large number of areas (Landis et al., 2013).

The main feature that is new in RASP with regard to the original BayArea software is the ability to re-calculate ancestral state probabilities of each unit area with alternative burn-in values. This differs from the original implementation of BayArea, in which changing the burn-in value necessitated performing the entire analysis again. In addition, we have enhanced BayArea in RASP so that it calculates the probabilities of ancestral states at nodes from the estimated frequencies of alternative states during each cycle of the BayArea MCMC. In BayArea, binary character states are used to code the geographic range for a species as being present (1) or absent (0) in each unit area. The $P_1(x_i)$ and $P_0(x_i)$ is the average probability of the presence (1) and absence (0) over all sampled generations, respectively, of the ancestral species in area X_i. Assume that $D = \{X_1, X_2 \ldots X_n\}$ is the set of all unit areas and that $R = \{Y_1, Y_2 \ldots Y_2{}^n{}_{-1}\}$ is all possible combinations of unit areas; in other words, all possible ranges. The probability of an ancestral range in set R is calculated as:

$$P(Y_i)\left(\prod_{X_i \in Y_i} P_1(X_j)\right)\left(\prod_{X_i \in (D \cap Y_i)^c} P_0(X_k)\right)$$

We have tested the BayArea method in RASP using the data set from the original program package (Landis et al., 2013). Our results from BayArea in RASP were identical to those from the original program.

2.4. Two additional tools

In additional to providing users with a comprehensive friendly GUI for commonly used historical biogeographic software packages, RASP provides two additional and useful tools. First, the *results combine* tool facilitates combining multiple runs from the same data set using the same method of inference. This tool is particularly useful for combing the results of independent MCMC analyses in BayArea. We recommend running more than one analysis when using the BayArea method to increase effective sample sizes and as a comparative test for aberrant behaviors in any one Markov chain.

The second new tool in RASP is the *group remove* tool. The *group remove* tool enables users to prune trees prior to performing biogeographic analyses. Pruning is accomplished without derooting trees or altering branch lengths at other nodes. This tool is particularly useful for removing widely distributed outgroups. A widely distributed outgroup, particularly with a taxonomic rank above species, may provide little useful information for ingroup reconstructions and may lead to erroneous results; namely extremely large, improbable ranges on their parent nodes (Buerki et al., 2011; Ronquist, 1997; Xiang and Thomas, 2008; Link-Pérez et al., 2011). While an outgroup of this nature may be suitable for phylogeny estimation, it is not suitable for biogeographic analysis. In addition to widely distributed outgroups, the *group remove* tool can be used to remove any user-specified group.

3. Comparisons and recommendations

As the number of models for biogeographic inference has continued to increase, the challenge of selecting a method suitable for a particular data set has become greater. This is especially true because different methods may yield different results for the same data (Xiang and Thomas, 2008; Ali et al., 2013). We generally recommend utilizing all methods in RASP to compare the degree of congruence among models. We also recommend the R-package BioGeoBEARS (Matzke, 2013), which finds the model that best fits the data and implements DIVALIKE (a likelihood interpretation of DIVA), and BAYAREALIKE (a likelihood interpretation of BayArea, without the Bayesian sampling ability of BayArea) (Matzke, 2014).

In general, different methods make different assumptions and are optimal under different conditions. For example, the DEC model and its derivatives can accommodate differing dispersal probabilities among areas across different time-periods and can integrate branch lengths, divergence times, and geological information (Ree and Smith, 2008). Thus, DEC and S-DEC (and DECLIKE) models may be good choices for biogeographic reconstruction when relationships among distributional areas through time are known with relatively high certainty (Moreau and Bell, 2013). In contrast, the S-DIVA algorithm is hard-wired to favor models of vicariance and is generally biased against early dispersal (Ronquist, 1997; Nylander et al., 2008). As a consequence, S-DIVA tends to reconstruct wide ancestral ranges on deeper nodes, particularly the root node (Buerki et al., 2011; Kodandaramaiah, 2010). However, the model in S-DIVA is more suitable than other methods when testing the co-evolutionary history of parasites or

bacteria with their hosts (Comas et al., 2013; Razo-Mendivil et al., 2011), because the original DIVA algorithm was derived from a simple model of parasite-host interactions (Ronquist, 1996). Further, S-DIVA is much faster than DEC and its derivatives. One of the biggest current limitations of DEC and S-DEC is computational burden. For example, analyses on a personal computer are often constrained to two or three geographic regions when there are more than 50 species in the tree. Thus, S-DIVA may represent a good choice among methods when little information is available to inform models in DEC, for co-evolutionary studies, or when the computational demand of DEC is limiting.

While both S-DIVA and DEC assume a model where lineages bifurcate and never multifurcate, BayArea can accept trees with polytomies directly giving the researcher more flexibility in analysis. Thus, the method is particularly attractive if researchers are more interested in the ancestral distribution of key nodes and wish to utilize a Bayesian approach to biogeographic inference. However, BaysArea cannot define the dispersal rate, constrain the maximum number of areas at each node, or exclude widespread and unlikely ancestral areas before analysis. In practice, users may also need to repeat the BaysArea analysis many times with more than 10 million generations to get a stable result. Being Bayesian, users should also check for consistency of results using different burnin values. Finally, the program currently accepts only a single phylogenetic tree, thus excluding phylogenetic uncertainty into biogeographic inference. In RASP, we have limited the maximum number of areas and species in a tree for BayArea to 26 and 512, respectively. We recommend that users who want to calculate hundreds of areas and species to use the command version of BayArea on a High-Performance Computing Cluster (HPCC).

In summary, RASP combines four quantitative, phylogeny-based historical biogeographic methods in a single, user-friendly package. More specifically, RASP implements the S-DIVA, DEC, S-DEC and BayArea methods. The output of all four methods can be displayed and exported as high quality graphics that are directly comparable. We will continue to develop RASP with a focus on implementing new algorithms for biogeographic inference and integrating more third party tools. All versions of RASP and a comprehensive manual are available freely from http://mnh.scu.edu.cn/soft/blog/RASP.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31270241, 31100161), and the Specimen Platform of China, Teaching Specimen's sub-platform (Web, http://mnh.scu.edu.cn/).

References

Ali, S.S., Pfosser, M., Wetschnig, W., et al., 2013. Out of Africa: Miocene dispersal, vicariance, and extinction within Hyacinthaceae Subfamily Urgineoideae. J. Integr. Plant Biol. 55 (10), 950–964.

Beaulieu, J.M., Tank, D.C., Donoghue, M.J., 2013. A Southern Hemisphere origin for campanulid angiosperms, with traces of the break-up of Gondwana. BMC Evol. Biol. 13 (1), 80.

Buerki, S., Forest, F., Alvarez, N., et al., 2011. An evaluation of new parsimony-based versus parametric inference methods in biogeography: a case study using the globally distributed plant family Sapindaceae. J. Biogeogr. 38 (3), 531–550.

Comas, I., Coscolla, M., Luo, T., et al., 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat. Genet. 45 (10), 1176–1182.

Harris, A.J., Xiang, Q.Y., 2009. Estimating ancestral distributions of lineages with uncertain sister groups: a statistical approach to dispersal vicariance analysis and a case using *Aesculus* L. (Sapindaceae) including fossils. J. Syst. Evol. 47, 349–368.

Harris, A.J., Wen, J., Xiang, Q.J., 2013. Inferring the biogeographic origins of inter-continental disjunct endemics using a Bayes-DIVA Approach. J. Syst. Evol. 51, 117–133.

Johns, M.V., 1957. Non-parametric empirical Bayes procedures. Ann. Math. Statist. 28, 649–669.

Kodandaramaiah, U., 2010. Use of dispersal–vicariance analysis in biogeography–a critique. J. Biogeogr. 37 (1), 3–11.

Landis, M.J., Matzke, N.J., Moore, B.R., 2013. Bayesian analysis of biogeography when the number of areas is large. Syst. Biol. 62, 789–804.

Link-Pérez, M.A., Watson, L.E., Hickey, R.J., 2011. Redefinition of Adiantopsis Fée (Pteridaceae): systematics, diversification, and biogeography. Taxon 60 (5), 1255–1268.

Matzke, N.J., 2013. BioGeoBEARS: BioGeography with Bayesian (and Likelihood). Evolutionary Analysis in R Scripts. Release R package version 0.2.1. <http://CRAN.R-project.org/package=BioGeoBEARS>.

Matzke, N.J., 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. Systematic Biology. syu056.

Miraldo, A., Hanski, I., 2014. Competitive release leads to range expansion and rampant speciation in Malagasy dung beetles. Syst. Biol., syu011

Moreau, C.S., Bell, C.D., 2013. Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. Evolution 67 (8), 2240–2257.

Moyle, R.G., Andersen, M.J., Oliveros, C.H., et al., 2012. Phylogeny and biogeography of the core babblers (Aves: Timaliidae). Syst. Biol. 61 (4), 631–651.

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*). Syst. Biol. 57, 257–268.

Razo-Mendivil, U., de León, G. Pérez-Ponce, 2011. Testing the evolutionary and biogeographical history of *Glypthelmins* (Digenea: Plagiorchiida), a parasite of anurans, through a simultaneous analysis of molecular and morphological data. Mol. Phylogenet. Evol. 59 (2), 331–341.

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

Ronquist, F., 1996. Dispersal Vicariance Analysis (DIVA) 1.1. User's manual.

Ronquist, F., 1997. Dispersal–vicariance analysis: a new approach to the quantification of historical biogeography. Syst. Biol. 46, 195–203.

Ronquist, F., 2001. Dispersal Vicariance Analysis (DIVA), version v. 1.2.

Schenk, J.J., Rowe, K.C., Steppan, S.J., 2013. Ecological opportunity and incumbency in the diversification of repeated continental colonizations by muroid rodents. Syst. Biol. 62 (6), 837–864.

Smith, S.A., 2009. Taking into account phylogenetic and divergence-time uncertainty in a parametric biogeographical analysis of the Northern Hemisphere plant clade Caprifolieae. J. Biogeogr. 36 (12), 2324–2337.

Smith, S.A., 2010. Lagrange C++ Manual.

Xiang, Q.Y., Thomas, D.T., 2008. Tracking character evolution and biogeographic history through time in Cornaceae – does choice of methods matter? J. Syst. Evol. 46, 349–374.

Yu, Y., Harris, A.J., He, X.J., 2010. S-DIVA (statistical dispersal–vicariance analysis): a tool for inferring biogeographic histories. Mol. Phylogenet. Evol. 56, 848–850.