Available online at www.sciencedirect.com
ScienceDirect

SoftwareX

SoftwareX 3–4 (2015) 22–26

www.elsevier.com/locate/softx

Mapping phylogenetic endemism in R using georeferenced branch extents

Greg R. Guerin[a,*], Andrew J. Lowe[a,b]

[a] *Terrestrial Ecosystem Research Network, Adelaide-node, The Environment Institute, School of Biological Sciences, University of Adelaide, North Terrace, Adelaide, South Australia 5005, Australia*
[b] *Science, Monitoring and Knowledge, Department of Environment, Water and Natural Resources, Adelaide, South Australia, Australia*

Received 13 May 2015; received in revised form 10 September 2015; accepted 20 October 2015

Highlights

- *Phylogenetic endemism* (PE) measures range-restricted *phylogenetic diversity* (PD).
- Previous PE implementations measured range restriction via area of occupancy (AOO).
- The R functions map PE using AOO or extent of occurrence (EOO) with simple inputs.
- Range sizes and PE scores are poorly correlated between AOO and EOO methods.
- The functions provide new spatial information on biodiversity with R functionality.

Abstract

Applications are needed to map biodiversity from large-scale species occurrence datasets whilst seamlessly integrating with existing functions in R. *Phylogenetic endemism* (PE) is a biodiversity measure based on range-restricted *phylogenetic diversity* (PD). Current implementations use area of occupancy (AOO) or frequency to estimate the spatial range of branch-length (i.e. *phylogenetic range-rarity*), rather than extent of occurrence (EOO; i.e. *georeferenced phylogenetic endemism*), which is known to produce different range estimates. We present R functions to map PD or PE weighted by AOO or EOO (new georeferenced implementation), taking as inputs georeferenced species occurrences and a phylogeny. Non-parametric statistics distinguish PD/PE from trivial correlates of species richness and sampling intensity.

Keywords: Phylogenetic diversity; Phylogenetic endemism; Area of occupancy; Extent of occurrence; Minimum convex polygon; R function; Non-parametric statistics; Interquartile range

Code metadata

Current code version	1.0
Permanent link to code/repository used of this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-15-00013
Legal Code License	GNU GPL-3
Code versioning system used	Git
Software code languages, tools, and services used	R
Compilation requirements, operating environments & dependencies	R package dependencies: simba, geosphere, adehabitat, raster, ape
If available Link to developer documentation/manual	n/a
Support email for questions	greg.guerin@adelaide.edu.au

Abbreviations: PE, phylogenetic endemism; PD, phylogenetic diversity; AOO, area of occupancy; EOO, extent of occurrence; SIR, sum of inverse range-sizes; PRR, phylogenetic range-rarity; GPE, georeferenced phylogenetic endemism.

[*] Corresponding author.
E-mail address: greg.guerin@adelaide.edu.au (G.R. Guerin).

http://dx.doi.org/10.1016/j.softx.2015.10.002

1. Motivation and significance

The geographic restriction of biodiversity is of interest to the fields of biogeography and conservation biology [1,2], and in particular the historical development and conservation value of concentrations of endemic species [3,4]. Increasingly sophisticated and numerical methods have been developed to measure range-restricted biodiversity. For example, the sum of inverse range-sizes or 'SIR' [5,6] of a set of species in a community sample, is a numerically continuous alternative biodiversity measure to counts of species that have been categorically assigned as endemic to a pre-defined area.

Rosauer et al. [7] extended the concept of SIR metrics to Faith's *phylogenetic diversity* (PD), which is a measure of the evolutionary history represented among a set of species, calculated as the sum of the branch lengths of a phylogenetic tree containing species in a particular community sample [8–10]. This *phylogenetic endemism* [7] (PE) was defined as (1):

$$\text{PE} = \sum_{\{c \in C\}} \frac{L_c}{R_c}$$

read as the sum of the lengths of each branch in a tree containing a community sample of species divided by the geographic range of each branch (i.e. based on the species terminating from that branch), where L is branch length, R is range size, C is the tree and c is a particular branch of the tree.

While a number of methods exist for estimating range sizes [11,12], previously only the number of occupied map grid cells (based on recorded or modelled species occurrences) has been used to estimate ranges that weight this metric. Guerin et al. [13] showed that species ranges, and resulting SIR for map grid cells, estimated with alternative methods were poorly correlated. Specifically, the number of occupied cells (equivalent to frequency or area of occupancy—AOO) was not rank equivalent to measures of extent of occurrence (EOO), leading to recognition of different SIR measures, *range-rarity richness* (RRR) and *georeferenced weighted endemism* (GWE), respectively [5].

Since the concepts of range-restricted PD and SIR metrics are linked, we extend here the georeferenced implementation of SIR [13] to its PD equivalent and present new self-contained R [14] functions for calculating and mapping PD and PE, based on species records and relevant phylogenetic trees. The functions can calculate either the existing implementation of PE, where branch length is weighted by its spatial range in terms of the number of occupied grid cells (*phylogenetic range-rarity*; PRR), or our novel georeferenced implementation, where branch length is weighted by the 'span' of constituent species occurrences (*georeferenced phylogenetic endemism*; GPE). Alternatively, unweighted PD can be mapped. Non-parametric statistics are used to detect outlying grid cells (explained below).

These functions address two gaps in current research software: (1) georeferenced calculation of EOO as a weight for calculating PE, to provide different information on the range-restriction of biodiversity than current AOO implementations; (2) functionality in the R environment to map biodiversity metrics including PD/PE from large-scale species occurrence datasets, and to seamlessly integrate inputs and outputs with existing analysis packages. The functions are currently used by loading source into R and calling the functions on simple input data.

The functions are principally suited to mapping regional biodiversity to identify conservation priorities. An example of this application would be to convert georeferenced species inventory data into gridded biodiversity heat maps. The outputs are also useful for ecological models in situations where coarse resolution (i.e. map cells rather than field plots) is relevant, or that are based on existing regional inventory data.

The existing implementation of PE (and other biodiversity metrics) is available within the *perl*-based '*biodiverse*' software [15] with mapping functionality, as well as in the 'phylo.endemism.R' function of David Nipperess (released under the GNU GPL: http://davidnipperess.blogspot.com.au/2012/07/phyloendemism-r-function-for.html, accessed 4/2/2015), which calculates a numeric vector of PE for sites in an occurrence matrix but does not have mapping functionality. Our intention, therefore, is to provide a novel implementation with alternative branch weights and to make these functions available in the R environment with automated integration of point data with maps, without requiring sophisticated custom programming from the user. We modified coding for the conversion of phylogenetic data to matrix representation [16] from David Nipperess' function, while all other coding is new.

2. Software description

2.1. Overview

The software consists of two functions with separate source code, 'phylogenetic.endemism.R' and 'pe.null.test.R'. Once source code and desired input data are loaded into R, function 'phylogenetic.endemism.R' can be called on the input data and with arguments adjusted for desired settings, and 'pe.null.test.R' can subsequently be called on the returned object. Both functions automatically produce plots, mainly rasters (see example in Fig. 1): 'phylogenetic.endemism.R' returns a 'list' containing a numeric vector and a raster map of PD/PE scores, the weights used, binary matrices of species occurrences against grid cells, branches against species and branches against grid cells, and a vector of branch lengths; 'pe.null.test.R' returns a list containing numeric vectors of expected interquartile ranges for each species richness value, a vector and raster each for categorical and continuous outlier scores for cells and their statistical significance (p-value), and finally a numeric vector of species richness values. With these returned outputs, the user can then, if desired, reproduce plots with customised formatting, such as alternative colour schemes and axis labels, or use the items returned in downstream analysis.

2.2. Mapping raw phylogenetic endemism

The function 'phylogenetic.endemism.R' calculates and maps PE based on two alternative range weights, 'cell' (the

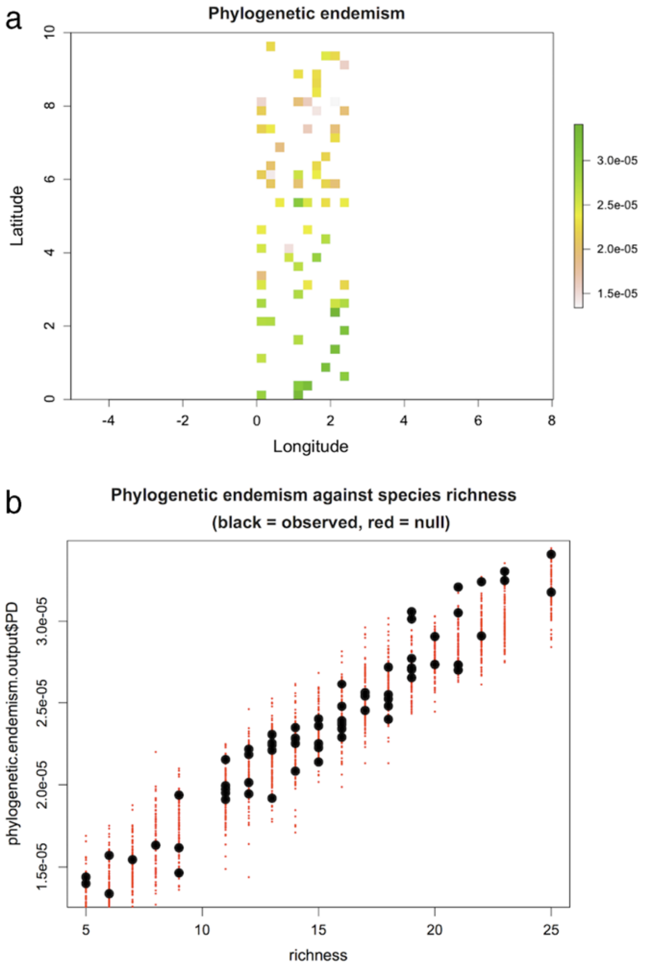

Fig. 1. Examples of default outputs from the *phylogenetic endemism* (PE) functions presented in this paper, based on the mite dataset (see main text). (a) Raw PE from example in main text, displayed on a raster map. Each grid cell in which species were recorded receives a score. (b) Observed PE scores from example in main text plotted against the observed species richness for the same cells (black) over a null distribution (red) representing PE from random draws from the species pool [observed values were re-plotted over red points using outputs from the functions (PE and richness scores) to make them clearer]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

number of cells occupied) and ‘span’ (greatest distance across the range), which are equivalent to PRR and GPE, respectively. The function can also accept user-supplied range weights or calculate and map unweighted PD. Occurrence input data can either be individual species records with fields for species name, longitude and latitude, or a species by sites matrix combined with an associated coordinates (longlat) by sites matrix. The phylogenetic data input is a tree with branch lengths of class ‘phylo’ [17]. Although the tree should obviously cover as many of the species in the incidence data as possible, the function strips species that do not occur in both datasets. Some difference in species name formatting (genus–species separator) is catered for between occurrence and phylogenetic datasets to deal with common input formats.

For GPE, species presences, georeferenced by the centroid coordinates for cells in which they have records and matched to a phylogenetic tree, are used to map the minimum convex polygon (hull) spanning occurrences of each branch, with the option of excluding geographical outliers. This process involves several computational steps: (1) converting the phylogenetic tree to a binary matrix [16]; (2) generating a species by grid cells occurrence matrix from the input point data; (3) intersecting the matrix-representation tree with the species occurrence matrix to generate a binary matrix of branch occurrences within grid cells [18]; (4) finally reconciling presences for particular branches with cell centroids and exclusion of geographical outliers; (5) a convex polygon is then drawn around the resulting set of coordinates.

Range ‘span’ is determined by calculating the maximum pairwise (great circle) distance between vertices of the convex polygon spanning occurrences of particular phylogenetic branches. Calculating pairwise distances between vertices, rather than all coordinates, significantly improves efficiency and computation time. If there are too few locations to form a polygon, the maximum distance is calculated from all pairwise comparisons, instead of polygon vertices. Using the calculated range weights for branches, PE is calculated and mapped for each grid cell that has data.

By default, the function automatically generates a raster map with the same geographical extent as the input data at a resolution of 0.25 degrees. The user can either specify an extent and resolution or define these by providing a reference raster object. Point records lying outside of the frame raster are excluded, so that a smaller extent can be analysed than that of the point records, if desired.

2.3. Non-parametric statistics

The function ‘pe.null.test.R’, takes the output of the ‘phylogenetic.endemism.R’ function and performs downstream tests for significant deviance from null expectations of PD/PE based on observed species richness (e.g. see [2,13,19]), an alternative to correcting for richness directly. The rationale is to determine whether observed PE represents an ‘enrichment’ of range-restricted branch length, or simply a trivial correlate of observed species richness, and therefore suffering the same sampling biases [20]. The expected distribution of PD/PE for a given species richness is determined by taking replicate (100 default, 1000 recommended) random draws of that number of species from the available pool without replacement, and comparing this to observed PD/PE. p-values are calculated based on the proportion of random replicates that are higher or lower that observed. Outlying PD/PE is also identified based on either a categorical cut-off (user-defined, with a default of more than 1.5 times outside the interquartile range of the null distribution) or a continuous measure [the factor of the interquartile range by which the score differs from the 50% quantile (≡median of the null distribution)]. These calculations are outputted as numeric vectors and gridded maps.

3. Illustrative examples

We present two examples of the use and application, respectively, of the software. The first example demonstrates the basic use of the functions in R and uses the Oribatid mite dataset [21] for convenience because it is available via the commonly used community ecology R package ‘vegan’ [22]. The dataset consists of a table (‘data frame’) of species occurrences against sites and a second table of map coordinates for the sites. Note the site coordinates in this example are metres along a Cartesian plane and are not geographic (longlat) coordinates as assumed by the function. Below is a simple introduction to the use of the functions in R (see also Fig. 1). Comments are preceded below by the # character.

```
####Preparation for this example:
library(vegan) #We load the vegan package
data(mite)
data(mite.xy) #We load the datasets from vegan
library(ape) #This package is required for the
phylogenetic functions to follow:
mite.tree<- rtree(n=ncol(mite),
tip.label=colnames(mite)) #Here, for this
example, we generate a phylogenetic tree of the
species in the mite dataset with random
relationships and branch lengths

###Using the functions to calculate and
test GPE:
source(‘‘phylogenetic.endemism.R’’)
mite.PE<- phylogenetic.endemism(mite,
records=‘‘site’’, site.coords=mite.xy,
sep.comm.spp=‘‘none’’, phylo.tree=mite.tree,
sep.phylo.spp=‘‘none’’, weight.type=‘‘geo’’)

source(‘‘pe.null.test.R’’)
pe.null.test(mite.PE)
```

The second example illustrates an empirical application of the functions to plant species recorded in systematically surveyed field plots from the Biological Survey of South Australia and *AusPlots* (Terrestrial Ecosystem Research Network) programs [23,24], a combined dataset of some 14,355 plots covering the state of South Australia. A phylogenetic tree of all species in the dataset was generated from Phylomatic Version 3 (http://phylodiversity.net/phylomatic/) and tree R20120829 [25], with node age constraints and branch-length adjustment [26,27]. We calculated PE for this dataset using, alternatively, cell-occupancy weights (i.e. PRR) and range-span weights (i.e. GPE). Specifically, we ran the phylogenetic.endemism function with weight.type set to either “cell” or “geo”. We then calculated correlation coefficients between the methods using Kendall’s τ, while partialling out the correlation between species richness and PE scores, because we would expect high rank correlation if the range estimation methods were equivalent (Fig. 2).

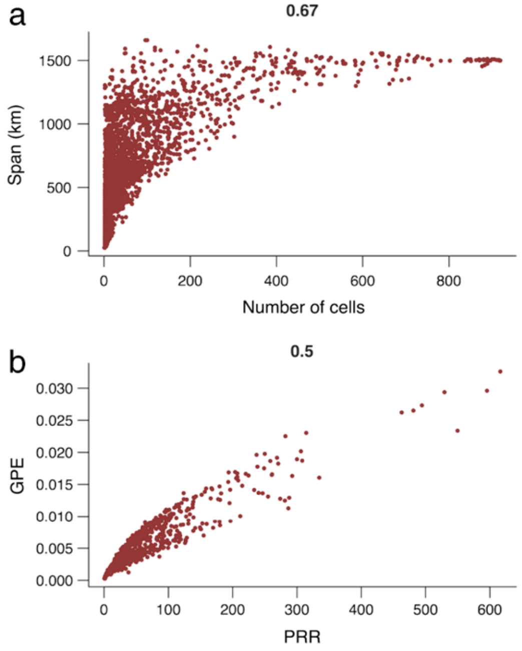

Fig. 2. Scatterplots of phylogenetic range estimates (a; $n = 2885$) and resulting *phylogenetic endemism* scores (b; for 919, 0.25° map grid cells) to compare the use of cell frequency (*phylogenetic range-rarity*; PRR) to range span (*georeferenced phylogenetic endemism*; GPE), as implemented in the reported software. Numbers above plots represent partial Kendall’s tau rank correlation coefficients, given species richness. Data are vegetation survey plots from the Biological Survey of South Australia and *AusPlots* combined with a Phylomatic tree of the species.

4. Impact

Interpretations of endemism as georeferenced restriction in EOO versus restriction in frequency or AOO are neither conceptually nor numerically equivalent [5,13]. Our second illustrative example analysis of data from an extensive network of vegetation survey plots in South Australia clearly shows that PE scores are poorly correlated between the two range estimation methods and highlight different aspects of the range-restriction of biodiversity. For example, not all phylogenetic branch-length that has a low AOO has a corresponding narrow EOO (Fig. 2). This is important because such measures may be used to rank areas for conservation value or assess biogeographic properties [5].

The software provides new functionality and metrics for pursuing a common research question in biogeography: *Are there concentrations of biodiversity in a landscape?* It also enables empirical testing of the more nuanced research question: *Is PE weighted by range extent equivalent to that weighted by range area and what are their best predictors?* The software therefore potentially informs the process of discovering hotspots of different aspects of biodiversity, which relate to different historical and ecological processes. The software has been used successfully as part of a regional biodiversity assessment (G.R. Guerin et al. unpubl. data).

The functions presented here enhance the pursuit of such biogeographical research questions by allowing the user to calculate and map PE onto raster maps from simple point data, using either range-weighting approach, as well as user-supplied weights or unweighted PD, providing alternative information on the distribution of biodiversity. This is a novel implementation of PE and contributes to much-needed self-contained functionality in R for mapping out biodiversity metrics from inventory type data and to allow seamless integration of such analyses with a suite of other functions in the R environment.

The functions were primarily designed for coarse-scale mapping of PE from inventory data for conservation and biodiversity management purposes, but can also provide outputs relevant for downstream modelling where it is useful to estimate PD/PE for map grid cells. For example, using this software it would be possible to calculate PE for map grid cells at a relevant resolution and immediately analyse the scores against predictor variables for the same grid cells, for example topographic heterogeneity within the cell or climatic parameters. This takes full advantage of the suite of analyses available in R (modelling, plotting etc.) with no need for importing and exporting files. For large projects and datasets, this workflow minimises data duplication and handling.

The functions differ from existing software through a combination of (1) automatically integrating point (species incidence) data with gridded map outputs; (2) the metrics calculated, including the novel GPE implementation, and; (3) integration with other data handling and analysis functions in the R environment to minimise data import and export steps. This new functionality enables users to map out (and subsequently analyse) PD/PE without custom programming the complex underlying calculations, making the process more feasible and less time-consuming.

5. Conclusions

Large species incidence datasets have become routinely available for analysis, for example through the Global Biodiversity Information Facility (GBIF) [28]. This information on species distribution, along with advances in modelling and phylogenetics, is enabling spatially explicit mapping and modelling of evolutionary history [6]. The implementations of PD and PE presented here for the R environment progress the suite of biodiversity metrics and software functionality available for this field of research including user-friendly access to existing and novel, spatially explicit metrics.

Acknowledgements

This research was supported by the Terrestrial Ecosystem Research Network (TERN), the South Australian Department of Environment, Water and Natural Resources (DEWNR) and the Australian Research Council. We thank Ed Biffin, Ben Sparrow, Susan Sweeney, Sandy Carruthers and Dan Rogers.

References

[1] Kier G, Barthlott W. Measuring and mapping endemism and species richness: a new methodological approach and its application on the flora of Africa. Biodivers Conserv 2001;10:1513–29.

[2] Slatyer C, Rosauer D, Lemckert F. An assessment of endemism and species richness patterns in the Australian Anura. J Biogeography 2007; 34:583–96.

[3] Beard JS, Chapman AR, Gioia P. Species richness and endemism in the Western Australian flora. J Biogeography 2000;27:1257–68.

[4] Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J. Biodiversity hotspots for conservation priorities. Nature 2000;403:853–8.

[5] Guerin GR, Lowe AJ. Sum of inverse range sizes (SIR), a biodiversity metric with many names and interpretations. Biodivers Conserv 2015; http://dx.doi.org/10.1007/s10531-015-0977-6.

[6] Crisp MD, Laffan S, Linder HP, Monro A. Endemism in the Australian flora. J Biogeography 2001;28:183–98.

[7] Rosauer D, Laffan SW, Crisp MD, Donnellan SC, Cook LG. Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. Mol Ecol 2009;18:4061–72.

[8] Faith DP. Conservation evaluation and phylogenetic diversity. Biol Conserv 1992;61:1–10.

[9] Asmyhr MG, Linke S, Hose G, Nipperess DA. Systematic conservation planning for groundwater ecosystems using phylogenetic diversity. PloS One 2014;9:e115132.

[10] Cardoso P, Rigal F, Borges PAV, Carvalho JC. A new frontier in biodiversity inventory: a proposal for estimators of phylogenetic and functional diversity. Methods Ecol Evol 2014;5:452–61.

[11] Gaston KJ, Quinn RM, Wood S, Arnold HR. Measures of geographic range size: the effects of sample size. Ecography 1996;19:259–68.

[12] Burgman MA, Fox JC. Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. Anim Conserv 2003;6:19–28.

[13] Guerin GR, Rukolainen L, Lowe AJ. A georeferenced implementation of weighted endemism. Methods Ecol Evol 2015;6:845–52.

[14] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2014. http://www.R-project.org/.

[15] Laffan SW, Lubarsky E, Rosauer DF. Biodiverse, a tool for the spatial analysis of biological and related diversity. Ecography 2010;33:643–7.

[16] Ragan MA. Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol 1992;1:53–8.

[17] Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 2004;20:289–90.

[18] Rodrigues AS, Gaston KJ. Maximising phylogenetic diversity in the selection of networks of conservation areas. Biol Conserv 2002;105: 103–11.

[19] Mishler BD, Knerr N, González-Orozco CE, Thornhill AH, Laffan SW, Miller JT. Phylogenetic measures of biodiversity and neo-and paleo-endemism in Australian Acacia. Nat Commun 2014;5:4473.

[20] Chao A, Chiu C-H, Hsieh TC, Davis T, Nipperess DA, Faith DP. Rarefaction and extrapolation of phylogenetic diversity. Methods Ecol Evol 2014;6:380–8.

[21] Borcard D, Legendre P, Drapeau P. Partialling out the spatial component of ecological variation. Ecology 1992;73:1045–55.

[22] Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Henry M, Stevens H, Wagner H, Vegan: Community Ecology Package. R package version 2.0-10. 2013; http://CRAN.R-project.org/package=vegan.

[23] Department of Environment, Water and Natural Resources. Biological Survey of South Australia—Vegetation Survey, Biological Database of South Australia (endemism mapping subset). Version 1.0 2014; http://dx.doi.org/10.4227/05/5444BAF32E94C. Obtained from Australian Ecological Knowledge and Observation System Data Portal (ÆKOS, http://www.portal.aekos.org.au/), made available by State of South Australia (Department of Environment, Water and Natural Resources), Adelaide, South Australia. Accessed 20 October 2014 (http://portal.aekos.org.au/dataset/168004).

[24] TERN AusPlots. Terrestrial Ecosystem Research Network AusPlots—Ausplots Rangelands Survey Program (biodiversity mapping supplement/subset), Version 1.0. 2015; http://dx.doi.org/10.4227/05/54C1B45A4CF2F. Obtained from Australian Ecological Knowledge and Observation System Data Portal (ÆKOS, http://www.portal.aekos.org.au/), made available by University of Adelaide. Accessed 23 January 2015. (http://portal.aekos.org.au/dataset/172373).

[25] Webb CO, Donoghue MJ. Phylomatic: tree assembly for applied phylogenetics. Mol Ecol Notes 2005;5:181–3.

[26] Wikström N, Savolainen V, Chase MW. Evolution of the angiosperms: calibrating the family tree. Proc R Soc B 2001;268:2211–20.

[27] Webb CO. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am Nat 2000;156:145–55.

[28] Flemons P, Guralnick R, Krieger J, Ranipeta A, Neufeld D. A web-based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA). Ecol Inform 2007;2:49–60.