Received: 27 December 2019 | Accepted: 14 August 2020

DOI: 10.1111/2041-210X.13478

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

phyloregion: R package for biogeographical regionalization and macroecology

Barnabas H. Daru[1] | Piyal Karunarathne[1] | Klaus Schliep[2]

[1]Department of Life Sciences, Texas A&M University-Corpus Christi, Corpus Christi, TX, USA

[2]Institute of Computational Biotechnology, Graz University of Technology, Graz, Styria, Austria

Correspondence
Barnabas H. Daru
Email: barnabas.daru@tamucc.edu

Funding information
Texas A&M University-Corpus Christi

Handling Editor: Daniele Silvestro

Abstract

1. Biogeographical regionalization is the classification of regions in terms of their biota and is key to our understanding of the ecological and historical drivers affecting species distribution in macroecological or large-scale conservation studies. However, despite the mass production of species distributions and phylogenetic data, statistical and computational infrastructure to successfully incorporate, manipulate and analyse such massive amounts of data had not been fully developed.
2. Here, we present phyloregion, a statistical package for the analysis of biogeographical regionalization and macroecology in the R computing environment, tailored for mega phylogenies and macroecological datasets of ever-increasing size and complexity.
3. Compared to available packages, phyloregion is several times faster and allocates less memory than other packages for analysis of alpha diversity (including phylogenetic diversity, phylogenetic endemism and evolutionary distinctiveness and global endangerment) and beta diversity (including cluster analysis, determining optimal number of clusters and evolutionary distinctiveness of regions).
4. We demonstrate the scalability of the package to large datasets with comprehensive phylogenies and global distribution maps of squamate reptiles (amphisbaenians, lizards and snakes), and show that different phyloregions differ strongly in evolutionary distinctiveness across scales. Visualization tools allow graphical exploration of the generated patterns of biogeographical regionalization and macroecology in geographical space.
5. Ultimately, phyloregion will facilitate rapid biogeographical analyses that will accommodate the ongoing mass production of species occurrence records and phylogenetic datasets at any scale and for any taxonomic group into completely reproducible R workflows.

KEYWORDS

biogeography, bioinformatics, biomes, conservation, phylogenetics, regionalization, software

1 | INTRODUCTION

In biogeography, there is growing interest in the analysis of datasets of ever-increasing size and complexity to explain biodiversity patterns and underlying processes. A common approach is biogeographical regionalization, the grouping of organisms based on shared features and how they respond to past or current physical and biological determinants (Kreft & Jetz, 2010; Morrone, 2018). The composition of species in biogeographical units (i.e. 'phyloregions' or 'bioregions') can reflect the historical processes such as extinction, speciation or dispersal that have shaped present-day distribution of biological diversity (Daru, Elliott, Park, & Davies, 2017; Ficetola, Mazel, & Thuiller, 2017; Kreft & Jetz, 2010; Morrone, 2018). When paired with phylogenetic information, biogeographical regionalization allows geographical regions that do not share any species in common to be quantified (Graham & Fine, 2008), and can identify patterns overlooked by species-level analyses (Daru et al., 2016; Edler, Guedes, Zizka, Rosvall, & Antonelli, 2017; Holt et al., 2013; Vilhena & Antonelli, 2015). However, compared to the mass production of species distribution and phylogenetic datasets, statistical and computational approaches necessary to analyse such data, and approaches that can incorporate efficient storage and manipulation of such data, are lacking.

A few open-source tools are available and can provide infrastructural support for analysis of biogeographical regionalization. For instance, the APE package (Paradis & Schliep, 2019) contains a comprehensive collection of tools for analyses of phylogenetics and evolution and is useful for reading, writing and manipulating phylogenetic trees, among many other functions. The BETAPART package (Baselga & Orme, 2012) performs computations of total dissimilarity in species composition along with their respective turnover and nestedness components. PICANTE focuses on analysis of phylogenetic community structure and trait evolution (Kembel et al., 2010). The use of network methods to detect bioregions (Bloomfield, Knerr, & Encinas-Viso, 2018; Carstensen & Olesen, 2009; Rosvall & Bergstrom, 2008; Thébault, 2013; Vilhena & Antonelli, 2015), while not yet implemented in the R computing environment, provides an alternative clustering method based on bipartite networks, and performs well at identifying interzones between regions (see Edler et al., 2017 for a simplified and accessible implementation). However, there is no consensus on which method is the most appropriate for biogeographical regionalization at large scales (Bloomfield et al., 2018; Dapporto, Ciolli, Dennis, Fox, & Shreeve, 2015; Morrone, 2018). The most effective approach to biogeographical regionalization might therefore depend on the system under study and the research questions.

Here, we present the phyloregion R package that permits the integration of phylogenetic relationships and species distributions for identifying biogeographical regions of different lineages to elucidate the spatial and temporal evolution of biota in a region. Specifically, phyloregion provides functions for analyses of standard alpha diversity metrics (such as phylogenetic diversity and phylogenetic endemism) as well as metrics for analysing spatial compositional turnover between communities (e.g. beta diversity, phylogenetic beta diversity and evolutionary distinctiveness of regions). We benchmark phyloregion against other packages for speed and memory allocation with an empirical dataset of the flora of southern Africa that includes species distributions and phylogenetic relationships for 1,400 taxa (data from Daru et al., 2016). Moreover, we also demonstrate the scalability of the package to big datasets using a case study of biogeographical regionalization with comprehensive phylogenies and distribution maps of 9,574 species of squamate reptiles (amphisbaenians, lizards and snakes) across the globe. Visualization tools allow graphical exploration of the generated patterns of biogeographical regionalization and macroecology in geographical space.

2 | OVERVIEW AND GENERAL WORKFLOW OF PHYLOREGION

The phyloregion package interacts with few other R packages including MATRIX (Bates & Maechler, 2019), APE (Paradis & Schliep, 2019), BETAPART (Baselga & Orme, 2012), RASTER (Hijmans, 2019) and SP (Bivand, Pebesma, & Gómez-Rubio, 2013). We provide a workflow of the phyloregion package for biogeographical assessment of any selected taxa and region (Figure 1). The workflow demonstrates steps from preparation of different types of data to visualizing the results of biogeographical regionalization, together with tips on selecting the optimal method for achieving the best output, depending on the types of data used and research questions. The package is available for direct installation through R from the Comprehensive R Archive Network (CRAN, https://CRAN.R-project.org/package=phyloregion), while the development version is hosted on GitHub at https://github.com/darunabas/phyloregion. To install phyloregion directly from CRAN, in R, type:

```
install.packages("phyloregion")
```

An alternative is to install the development version of phyloregion hosted on GitHub as follows:

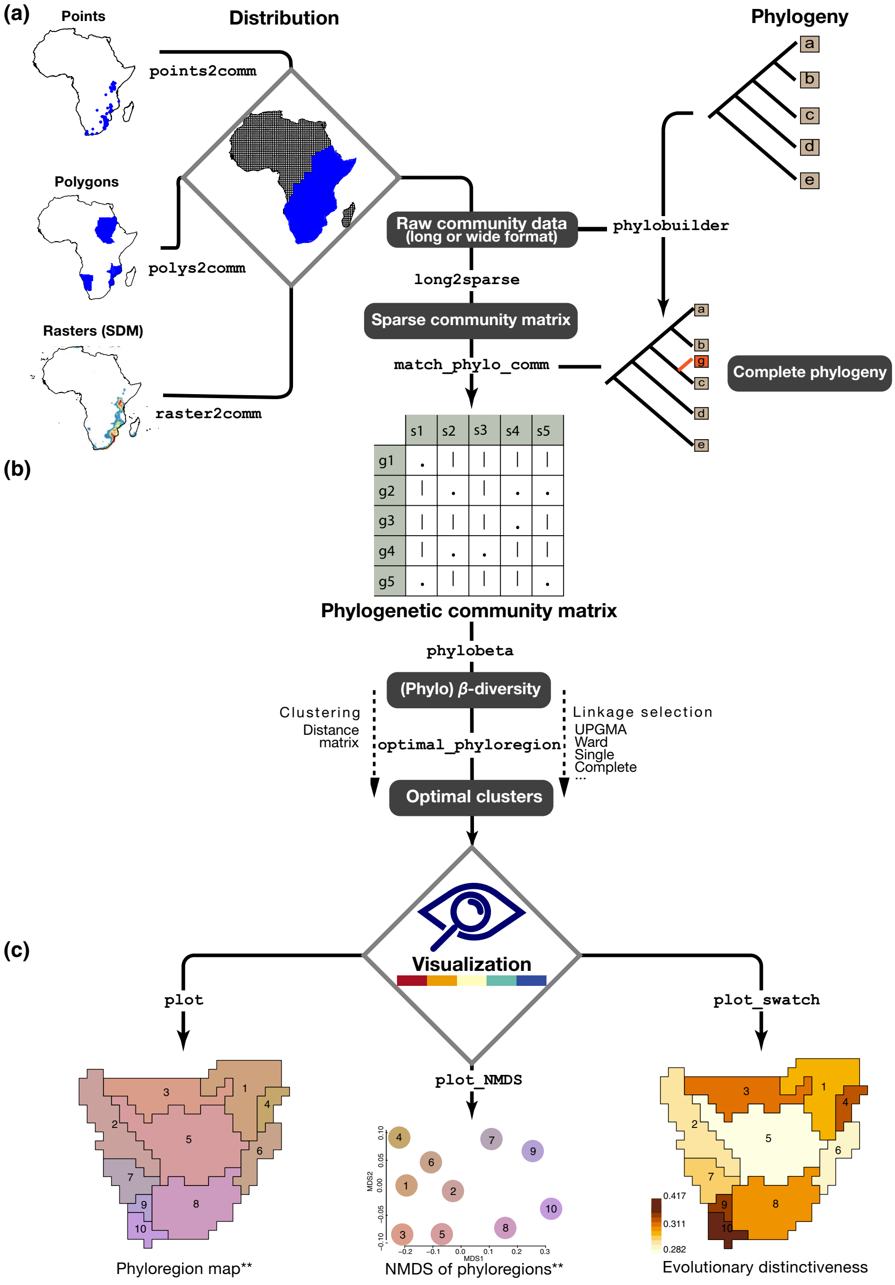
```
if (!requireNamespace("devtools", quietly = TRUE))
  install.packages("devtools")
devtools::install_github("darunabas/phyloregion")
library(phyloregion)
```

3 | RAW DATA

3.1 | Distribution data input

The phyloregion package provides functions for manipulating at least three categories of distribution data at varying spatial grains and extents: point records, polygons and raster layers. Polygons can be derived from the International Union for the Conservation of Nature's spatial database (https://www.iucnredlist.org/resources/spatial-data-download), published monographs or field guides that have been validated by taxonomic experts. Point records are commonly derived from major data hubs such as the Global Biodiversity Information Facility (Edwards, Lane, & Nielsen, 2000), Integrated

FIGURE 1 Typical workflow for analysis of biogeographical regionalization and macroecology using phyloregion. (a) Input data: distribution data (point records, polygons and raster layers) are converted to a long community data frame format before conversion to a sparse community matrix. When paired with phylogenetic data, the function phylobuilder creates a subtree with largest overlap from a species list, thereby ensuring complete representation of missing data. (b) Analysis: phyloregion allows analysis of standard alpha diversity metrics commonly used in conservation, such as phylogenetic diversity and phylogenetic endemism as well as metrics for analysing compositional turnover (e.g. beta diversity and phylogenetic beta diversity). (c) Visualization: efficient tools allow graphical exploration of the generated patterns of biogeographical regionalization and macroecology in geographical space. In the phylogenetic community matrix in (b), the zeros are represented with dots for clarity, whereas the non-zero elements are represented with vertical bars. Numbers on the maps are arbitrary and indicate the regions that have been delimited



	s1	s2	s3	s4	s5
g1	.	|	|	|	|
g2	|	.	|	.	.
g3	|	|	|	.	|
g4	|	.	.	|	|
g5	.	|	|	|	.

**Colour differences depict the amount of phylogenetic turnover among phyloregions

Digitized Biocollections (www.idigbio.org) or Botanical Information and Ecology Network (Enquist, Condit, Peet, Schildhauer, & Thiers, 2016), and typically have columns of geographical coordinates for each observation. We note, however, that these major data hubs show strong overlap in their collections. Raster layers are typically derived from species distribution modelling, such as *aquamaps* (Kaschner et al., 2016). An overview can be easily obtained with the functions points2comm, polys2comm and raster2comm for point records, polygons or raster layers respectively. Depending on the data source, all three functions ultimately provide convenient interfaces to convert the distribution data to a community data frame at varying spatial grains and extents for downstream analyses.

3.2 | Phylogenetic data

Phylogenies are often derived from DNA sequences; however, the issue of missing data is a significant obstacle in reconstructing phylogenetic relationships for most non-charismatic groups, for example, plants or insects. When paired with distribution data, phylogenies can aid the discovery of common patterns and processes that underlie the formation of biogeographical regions (Daru et al., 2017; Wiley, 1988). The function phylobuilder appends missing taxa to a supertree. Unlike other tree-building algorithms that manually graft missing taxa into a working supertree, phylobuilder creates a subtree with the largest overlap from a species list at a fast speed. If species in the taxon list are not in the tree (tip label), species will be added at the most recent common ancestor at the genus or family level when possible.

4 | DATA PREPARATION AND ANALYSES

4.1 | Sparse community matrix

A community composition dataset is commonly represented as a matrix of 1s and 0s, with species as columns and rows as spatial cells or communities. In practice, such a matrix can contain many zero values because species are known to generally have unimodal distributions along environmental gradients (ter Braak & Prentice, 1988), and storing and analysing every single element of that matrix can be computationally challenging and expensive. Indeed, for large matrices, most base R functions cannot make a table with more than 2^{31} elements. One approach to overcome this limitation is to utilize a sparse matrix, a matrix with a high proportion of zero entries (Duff, 1977). Because a sparse matrix is comprised mostly of 0s, it only stores the non-zero entries, from which several measures of biodiversity including biogeographical regionalization can be calculated. Our long2sparse function allows conversion of community data from either long or wide (dense2sparse) formats to a condensed sparse matrix (Figure 2) to ease downstream analyses such as compositional dissimilarity and avoid the exhaustion of computer memory capacities.

4.2 | Matching phylogeny and community composition data

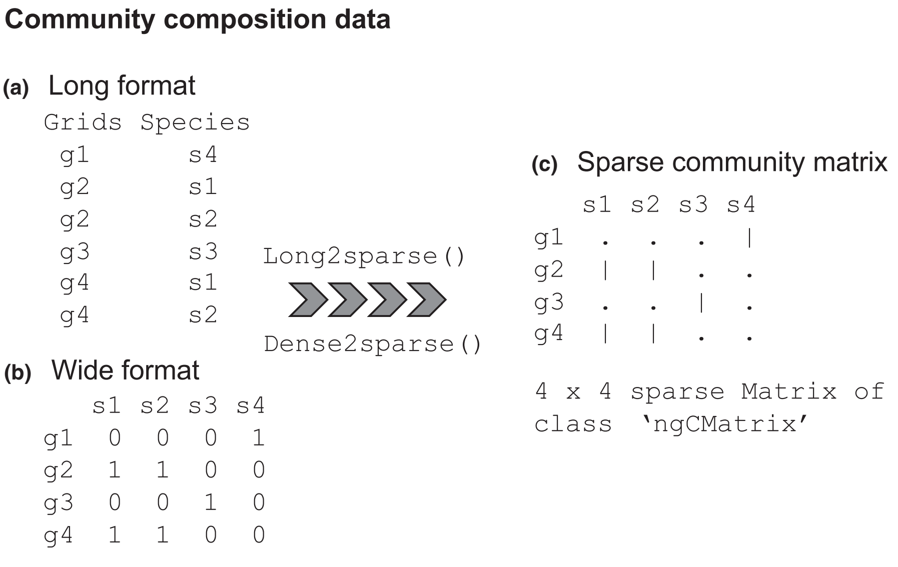
In community ecology and biogeographical analyses, it is sometimes desirable to make sure that the taxa in different datasets match each other (Kembel et al., 2010). However, existing tools are not tailored for comparing taxa in mega phylogenies spanning thousands of taxa with community composition datasets at large scales. We present match_phylo_comm that compares a sparse community matrix against a phylogenetic tree and adds missing species to the tree at the genus or higher taxonomic levels.

4.3 | Generating beta diversity (phylogenetic and non-phylogenetic)

The phyloregion package provides functions for analysis of compositional turnover (beta diversity) based on widely used dissimilarity indices such as Simpson, Sorensen and Jaccard (Laffan et al., 2016). The phyloregion's functions beta_diss and phylobeta compute efficiently the pairwise dissimilarity matrices for large sparse community matrices and phylogenetic trees for taxonomic and phylogenetic turnover respectively. The results are stored as distance objects for downstream analyses.

4.4 | Cluster algorithm selection and validation

To overcome the lack of a priori justification for using a particular method for identifying phyloregions, the function select_linkage can contrast eight widely used hierarchical clustering algorithms (including UPGMA and single linkage) on the (phylogenetic) beta diversity matrix for degree of data distortion using Sokal and Rohlf's (1962) cophenetic correlation coefficient. The cophenetic correlation coefficient measures how faithfully the original pairwise distance matrix is represented by the dendrogram (Sokal & Rohlf, 1962). Thus, the best method is indicated by higher correlation values, resulting in regions with a maximum internal similarity but with maximum differences from other regions.



Community composition data

(a) Long format

```
Grids Species
 g1      s4
 g2      s1
 g2      s2
 g3      s3
 g4      s1
 g4      s2
```

(b) Wide format

```
    s1 s2 s3 s4
g1   0  0  0  1
g2   1  1  0  0
g3   0  0  1  0
g4   1  1  0  0
```

(c) Sparse community matrix

```
   s1 s2 s3 s4
g1  .  .  .  |
g2  |  |  .  .
g3  .  .  |  .
g4  |  |  .  .

4 x 4 sparse Matrix of
class  'ngCMatrix'
```

FIGURE 2 Illustration showing community data conversion to a sparse community matrix by (a) long2sparse function when the raw data are in long community data format or (b) dense2sparse for wide community data format—typically 1s and 0s—with species as columns and rows as spatial cells or communities. The result is (c) a sparse community matrix, which holds only the non-zero data for downstream analysis. For this illustration, the zeros are represented with dots for clarity, whereas the non-zero elements are represented with vertical bars

4.5 | Determining the optimal number of clusters

The function optimal_phyloregion utilizes the efficiency of the so-called 'elbow' (also known as 'knee') method corresponding to the point of maximum curvature (Salvador & Chan, 2004), to determine the optimal number of clusters that best describes the observed (phylogenetic) beta diversity matrix. Depending on the research question, the scale of the cutting depth or clustering algorithm method can be varied systematically. The output is used to visualize relationships among phyloregions using hierarchical dendrograms of dissimilarity and non-metric multidimensional scaling (NMDS) ordination, and are assessed for spatial coherence by mapping and/or quantifying their evolutionary distinctiveness.

4.6 | Evolutionary distinctiveness of phyloregions

The function phyloregion estimates evolutionary distinctiveness of each phyloregion by computing the mean value of (phylogenetic) beta diversity between a focal phyloregion and all other phyloregions in the study area. It takes a distance matrix and returns a 'phyloregion' object containing a phyloregion × phyloregion distance object. Areas of high evolutionary distinctiveness can provide new insights on the mechanisms that are responsible for generating ecological diversity such as speciation, niche conservatism, extinction and dispersal (Daru et al., 2017; Holt et al., 2013).

5 | VISUAL REPRESENTATION AND ASSESSMENT OF BIOGEOGRAPHICAL REGIONS

The phyloregion package also provides a number of functions that aid visualization and assessment of biogeographical regions.

- plot.phyloregion can display clusters of cells (i.e. 'phyloregions' for phylogenetic approaches or 'bioregions' for non-phylogenetic approaches) in multidimensional scaling colour space matching the colour vision of the human visual system (Kruskal, 1964). The colours indicate the levels of differentiation of clades in different phyloregions. Phyloregions with similar colours have similar clades and those with different colours differ in the clades they enclose (Figure 1).
- plot_swatch maps discretized values of a quantity based on continuous numerical variables of their cells or sites for visualization as heatmap in sequential colour palettes. This function can also be used to quantify the evolutionary distinctiveness of phyloregions, defined as the mean of pairwise beta diversity values between each phyloregion and all other phyloregions and displays them in hue-chroma-luminance colours that are much more suitable for capturing human colour perception (Figure 1).

6 | CASE STUDY OF BIOGEOGRAPHICAL REGIONALIZATION OF SQUAMATE REPTILES

We validated the application of the phyloregion package on the geographical distributions and phylogenetic data for all 9,574 species of squamate reptiles across the globe (data from Tonini, Beard, Ferreira, Jetz, & Pyron, 2016). Despite the fact that reptiles were part of the dataset used in Wallace's original zoogeographical regionalization along with birds, mammals and insects (Wallace, 1876), they have been largely neglected in modern regionalization schemes (Edler et al., 2017; Holt et al., 2013; Kreft & Jetz, 2010; Meiri & Chapple, 2016). Nevertheless, squamate reptiles are one of the most diverse and widely distributed vertebrate groups in the world (Böhm et al., 2013). Most notably, due to the high extinction rates they are facing, the distribution data, phylogeny and evolutionary relatedness of squamates have recently been well-documented (Tonini et al., 2016 and references therein). These make squamate reptiles an ideal system to test the robustness and implementation of phyloregion for biogeographical regionalization at large scales.

We used updated polygons representing the maximum geographical extent of each squamate reptile species (Roll et al., 2017). We ran the polys2comm, long2sparse and match_phylo_comm wrapper functions to generate the community data at a resolution of 1° × 1° (see Figure S1 for an example of varying spatial extents). Note that this resolution can be adjusted by varying the res argument in the function fishnet(mask, res = 0.5). We accounted for phylogenetic uncertainty in our analyses by drawing 100 trees at random from a posterior distribution of fully resolved trees (Tonini et al., 2016) to generate phylogenetic dissimilarity matrices (with Simpson's pairwise phylogenetic dissimilarities as default because of its independence to differences in species richness among sites, Koleff, Gaston, & Lennon, 2003; Kreft & Jetz, 2010), and took the mean across grid cells using mean_dist. Note that other dissimilarity indices such as 'Jaccard' and 'Sorensen' can be used as desired (Laffan et al., 2016), depending on the data used and research questions.

Using the 'elbow method' (function optimal_phyloregion), we identified 18 optimal phyloregions (i.e. maximum explained variance of 0.72 for clustering achieved at $k = 18$) of squamate reptiles (Figure 3). UPGMA was identified as the best clustering algorithm (cophenetic correlation coefficient = 0.8; selected using function select_linkage).

The resulting phyloregions for squamate reptiles at the global extent show substantial congruence to Holt et al.'s (2013) updates of Wallace's original zoogeographical regions, including Oceanian, Australian, Madagascan, Palearctic and Nearctic (Figure 3a). However, we also identified some discrepancies. For example, the Afrotropical realm (sensu Holt et al., 2013) was divided into four phyloregions in our study corresponding to West and Central Africa (11), Horn of Africa (12), Zambezian (15) and South African (17). We also identified a new phyloregion overlapping Chile-Patagonian in temperate South America. This discrepancy might be due to the

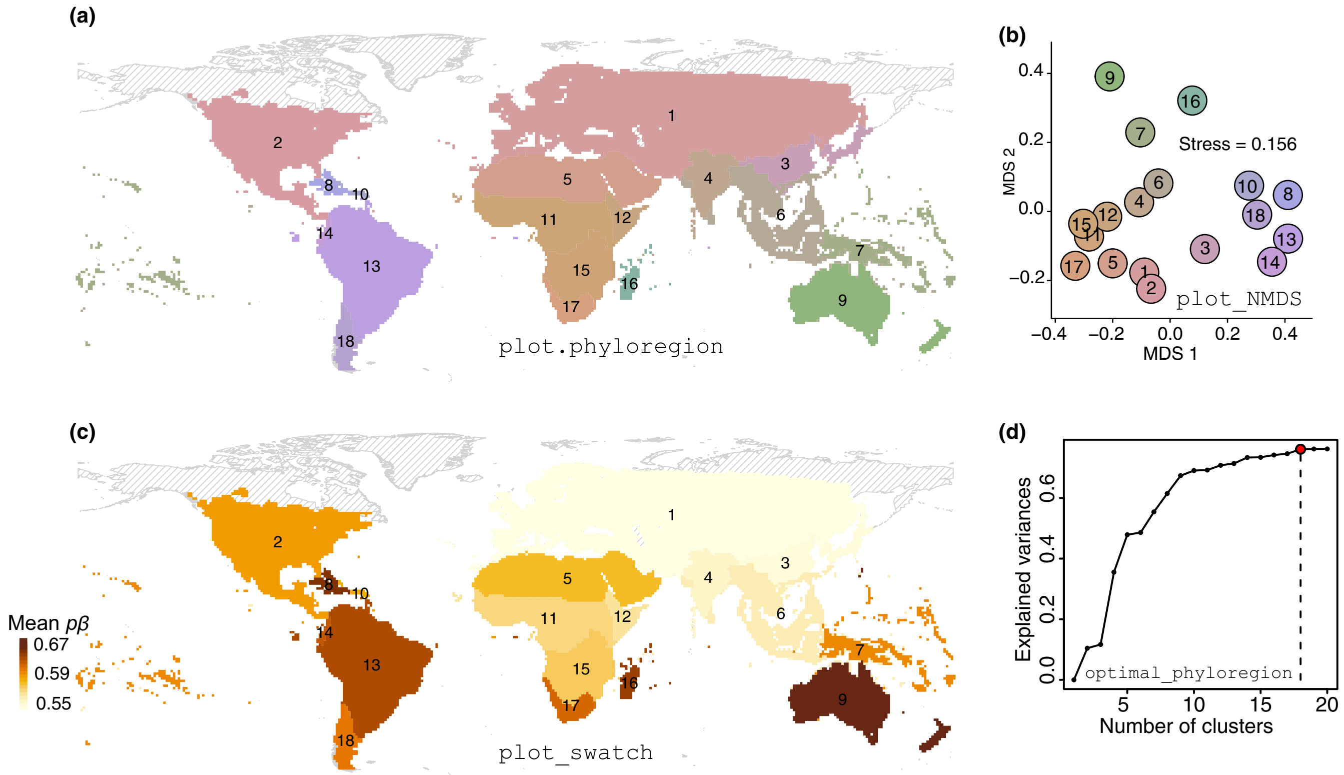


FIGURE 3 A global phylogenetic regionalization of 9,574 species of squamate reptiles reveals their evolutionary affinities. (a) Map of phyloregions shows evolutionary affinities among disjunct assemblages (function plot.phyloregion). (b) The ordination of phyloregions in NMDS space shows that different phyloregions differ strongly in evolutionary uniqueness (function plot_NMDS). (c) Map of evolutionary distinctiveness for squamate reptiles of the world, calculated as the mean value of phylogenetic beta diversity between a phyloregion and all other phyloregions in the study area (function plot_swatch). The colours indicate the degree to which each phyloregion differs from all other phyloregions based on mean pairwise phylogenetic beta diversity values, with darker colours indicating high evolutionary distinctiveness. Colour differences in the map (a) and NMDS plot (b) depict the amount of phylogenetic turnover among phyloregions. (d) The threshold of explained variances to identify the optimal number of phyloregions. The 'elbow' (optimal phyloregion) of the graph is indicated by the red circle. Numbers on the maps are arbitrary and indicate the regions that have been delimited

focal group being reptiles, whereas Holt et al. (2013) present results for birds, mammals and amphibians. However, as there is increasing evidence that patterns of biodiversity are scale-dependent (Daru, Farooq, Antonelli, & Faurby, 2020; Jarzyna & Jetz, 2018), it is likely that the mechanisms underlying patterns of biogeographical regionalization are sensitive to differences in spatial grain and extent (Daru et al., 2017; Holt et al., 2013; Keil et al., 2012). We found that these effects are lost at the continental to regional/local scales. At the continental extent, for instance, phyloregions are less spatially clumped and more differentiated from each other; for example, Africa split into 15 phyloregions at the continental extent compared to six phyloregions at the global extent (Figure S1). At the regional/local extent, spatial patterns of phylogenetic regionalization became more scattered across regions also with a loss of distinction (Figure S1c).

Box 1 Glossary of terms

Biogeographical regionalization: the partitioning of the biotic world into distinct geographical units.

Biogeographical realm: large biogeographical divisions within which ecosystems share a broadly similar evolutionary history.

Ecoregion: geographical regions that are defined by specific ecological patterns, including soil, flora and fauna, climatic conditions, among other factors.

Phyloregions: association of species into distinct phylogenetically delimited biogeographical units.

Geographically disjunct assemblages such as Panamanian and Temperate South America harbour closely related assemblages (Figure 3b), whereas some geographically proximal bioregions have low levels of faunistic similarity, suggesting spatial patterns of species diversity can have different phylogenetic structures (Hawkins et al., 2012). However, we also found that assemblages from different phyloregions tend to cluster with each other, for example, the Neotropics and Palaeotropics, suggesting that more spatially close phyloregions might also be more similar. Mean phylogenetic turnover of squamate reptiles between a phyloregion and all other phyloregions (function phyloregion) indicates that different phyloregions differ most strongly in evolutionary distinctiveness, with higher evolutionary distinctiveness in the tropics and southern hemisphere (Figure 3c), a similar observation to Tonini et al. (2016). Notably, the Australian phyloregion has the highest mean phylogenetic turnover (mean phylogenetic turnover between Australian and all other phyloregions = 0.67; Figure 3c), reflecting limited dispersal of lineages in this phyloregion.

The use of phylogenetic information and species distributions allows a deeper understanding of the mechanisms determining current patterns of biodiversity. Our evolutionary distinctiveness analysis in the recognized phyloregions (Box 1) brings a new component of evolutionary importance of each region to the biogeographical regionalization as well as for conservation prioritization. Most of the phyloregions found here spanned multiple ecoregions and biogeographical realms, suggesting that conservation planning should be adjusted to cover these larger phyloregions.

7 | BENCHMARKING PHYLOREGION

To benchmark phyloregion's functions with available packages for speed and memory allocation, we used an empirical dataset of the flora of southern Africa that includes species distributions and phylogenetic relationships for 1,400 plant taxa (data from Daru et al., 2016). This dataset is included in the phyloregion package directly as one of the example data in a helpfile: data(africa). We compared phyloregion to other packages in terms of analyses of alpha diversity metrics that are commonly used in biodiversity conservation, such as phylogenetic diversity and phylogenetic endemism as well as metrics for analysing compositional turnover (e.g. beta diversity and phylogenetic beta diversity; R code for benchmarking phyloregion with available packages is included in the benchmarking vignette; https://darunabas.github.io/phyloregion/articles/Benchmark.html). These tests indicate that phyloregion is faster and more efficient in memory allocation than other packages (Table 1; Figure S2).

TABLE 1 Comparison of phyloregion against comparable packages for analysis of alpha and beta diversity for woody plant species of southern Africa. The magnitude of change in memory allocation and run time are shown in parentheses

	Packages	Memory allocation (MB)	Speed (s)
Phylogenetic diversity	phyloregion	1.83	0.00287
	hilldiv	170.22 (×93)	1.09 (×380)
	pez	60.79 (×33)	0.13 (×45)
	picante	59.5 (×33)	0.12 (×42)
Phylogenetic endemism	phyloregion	1.06	0.0036
	pez	498.93 (×471)	1.89 (×525)
Beta diversity	phyloregion	0.40	0.0010
	betapart	0.60 (×1.5)	0.0011 (×1.1)
	vegan	1.02 (×2.6)	0.0013 (×1.3)
	BAT	31.76 (×79)	0.044 (×44)
Phylogenetic beta diversity	phyloregion	1.07	0.0051
	betapart	1,240 (×1,159)	2.14 (×420)
	picante	1,240 (×1,159)	4.38 (×859)
	BAT	207.39 (×194)	1.61 (×316)

Abbreviation: MB, megabyte.

8 | DISSIMILARITY-BASED REGIONALIZATION VERSUS NETWORK APPROACHES

Although phyloregion implements a dissimilarity-based algorithm to quantify spatial patterns of biodiversity, we recognize that such clustering methods could be sensitive to sampling biases in species occurrence data. This could potentially influence the delineation of bioregions, especially when the goal is to identify transition zones compared to network methods (e.g. Bloomfield et al., 2018; Vilhena & Antonelli, 2015). We did not explicitly evaluate the performance of our phyloregion package against network methods such as mapequation (Rosvall & Bergstrom, 2008) or modularity metrics (Guimera & Amaral, 2005); however, we believe that any difference between the two methods might be complementary and offer additional biogeographical insights where phyloregion is more informative at detecting bioregions at continental to global scales (Daru et al., 2017; Holt et al., 2013; Kreft & Jetz, 2010), whereas network methods are efficient at identifying interzones (Bloomfield et al., 2018). Depending on the goal of the study, users can integrate a dissimilarity-based approach (as implemented in phyloregion) with network approaches to understand patterns and processes of biogeographical regionalization. In addition to computations of compositional turnover, phyloregion can run other standard spatial biodiversity analyses such as phylogenetic diversity, phylogenetic endemism and evolutionary distinctiveness that are commonly used in conservation. It can also handle analysis for any form of organism at any grain size or spatial extent and does not suffer from a resolution limit unlike standard network-based approaches (Fortunato & Barthélemy, 2007; Kawamoto & Rosvall, 2015).

9 | CONCLUDING REMARKS

Although there are other packages such as APE (Paradis & Schliep, 2019), BETAPART (Baselga & Orme, 2012) or VEGAN (Oksanen et al., 2019) that can be used for analysis of biogeographical regionalization, phyloregion adds three novelties. First, it can utilize a sparse matrix by holding only the non-zero elements of a matrix thereby taking up significantly less memory and making it possible to handle larger datasets efficiently and rapidly. Second, it has novel functions for speedy raw data conversion from traditional data tables (e.g. a dense matrix of 1s and 0s) to a sparse community matrix as well as a user-friendly analysis of biogeographical regionalization into completely reproducible R workflows. Third, the functionality of the package can be extended for analysis of alpha diversity such as mapping hotspots of species richness, endemism or threat (e.g. Daru et al., 2020).

The goal of phyloregion is to facilitate the analysis of biogeographical regionalization and macroecology at any scale and for any taxonomic group, tailored to accommodate the ongoing mass production of species occurrence data and phylogenetic datasets.

ACKNOWLEDGEMENTS

B.H.D. is supported by Texas A&M University at Corpus Christi. We are grateful to Kristen Ruggles for helping with English language editing, and Abubakar Bello and Danmallam Bello Adamu for translating the abstract to Hausa language.

AUTHORS' CONTRIBUTIONS

B.H.D. conceived the project, analysed the data and led the writing with help from P.K.; B.H.D. and K.S. developed the method; B.H.D., K.S. and P.K. tested the method. All the co-authors assisted with editing and approved for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13478.

DATA AVAILABILITY STATEMENT

The phyloregion R package and documentation are hosted on CRAN (https://CRAN.R-project.org/package=phyloregion) and GitHub (https://github.com/darunabas/phyloregion), whereas a website with the major vignettes is available at https://darunabas.github.io/phyloregion/index.html. All data and scripts necessary to repeat the analyses for the squamate reptiles described here have been made available through the Dryad Digital Data Repository https://doi.org/10.5061/dryad.tdz08kpw6 (Daru, Karunarathne, & Schliep, 2019).

ORCID

Barnabas H. Daru https://orcid.org/0000-0002-2115-0257
Piyal Karunarathne https://orcid.org/0000-0002-1934-145X
Klaus Schliep https://orcid.org/0000-0003-2941-0161

REFERENCES

Baselga, A., & Orme, C. D. L. (2012). betapart: An R package for the study of beta diversity. *Methods in Ecology and Evolution*, *3*(5), 808–812.

Bates, D., & Maechler, M. (2019). *Matrix: sparse and dense matrix classes and methods*. R package version 1.2-17. Retrieved from https://cran.r-project.org/package=Matrix

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied spatial data analysis with R* (2nd ed.). New York, NY: Springer.

Bloomfield, N. J., Knerr, N., & Encinas-Viso, F. (2018). A comparison of network and clustering methods to detect biogeographical regions. *Ecography*, *41*(1), 1–10. https://doi.org/10.1111/ecog.02596

Böhm, M., Collen, B., Baillie, J. E. M., Bowles, P., Chanson, J., Cox, N., … Zug, G. (2013). The conservation status of the world's reptiles. *Biological Conservation*, *157*, 372–385. https://doi.org/10.1016/j.biocon.2012.07.015

Carstensen, D. W., & Olesen, J. M. (2009). Wallacea and its nectarivorous birds: Nestedness and modules. *Journal of Biogeography*, *36*(8), 1540–1550. https://doi.org/10.1111/j.1365-2699.2009.02098.x

Dapporto, L., Ciolli, G., Dennis, R. L. H., Fox, R., & Shreeve, T. G. (2015). A new procedure for extrapolating turnover regionalization at mid-small spatial scales, tested on British butterflies. *Methods in Ecology and Evolution*, *6*(11), 1287–1297.

Daru, B. H., Elliott, T. L., Park, D. S., & Davies, T. J. (2017). Understanding the processes underpinning patterns of phylogenetic regionalization. *Trends in Ecology & Evolution*, *32*(11), 845–860. https://doi.org/10.1016/j.tree.2017.08.013

Daru, B. H., Farooq, H., Antonelli, A., & Faurby, S. (2020). Endemism patterns are scale dependent. *Nature Communications*, *11*, 2115. https://doi.org/10.1038/s41467-020-15921-6

Daru, B. H., Karunarathne, P., & Schliep, K. (2019). phyloregion: R package for biogeographic regionalization and spatial conservation. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.tdz08kpw6

Daru, B. H., Van der Bank, M., Maurin, O., Yessoufou, K., Schaefer, H., Slingsby, J. A., & Davies, T. J. (2016). A novel phylogenetic regionalization of the phytogeographic zones of southern Africa reveals their hidden evolutionary affinities. *Journal of Biogeography*, *43*(1), 155–166.

Duff, I. S. (1977). A survey of sparse matrix research. *Proceedings of the IEEE*, *65*(4), 500–535. https://doi.org/10.1109/PROC.1977.10514

Edler, D., Guedes, T., Zizka, A., Rosvall, M., & Antonelli, A. (2017). Infomap Bioregions: Interactive mapping of biogeographical regions from species distributions. *Systematic Biology*, *66*(2), 197–204.

Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science*, *289*(5488), 2312–2314.

Enquist, B. J., Condit, R., Peet, R. K., Schildhauer, M., & Thiers, B. M. (2016). Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints*, *4*, e2615v2.

Ficetola, G. F., Mazel, F., & Thuiller, W. (2017). Global determinants of zoogeographical boundaries. *Nature Ecology & Evolution*, *1*(4). https://doi.org/10.1038/s41559-017-0089

Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(1), 36–41. https://doi.org/10.1073/pnas.0605965104

Graham, C. H., & Fine, P. V. A. (2008). Phylogenetic beta diversity: Linking ecological and evolutionary processes across space in time. *Ecology Letters*, *11*, 1265–1277. https://doi.org/10.1111/j.1461-0248.2008.01256.x

Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, *433*, 895–900. https://doi.org/10.1038/nature03288

Hawkins, B. A., McCain, C. M., Davies, T. J., Buckley, L. B., Anacker, B. L., Cornell, H. V., … Stephens, P. R. (2012). Different evolutionary histories underlie congruent species richness gradients of birds and mammals. *Journal of Biogeography*, *39*, 825–841. https://doi.org/10.1111/j.1365-2699.2011.02655.x

Hijmans, R. J. (2019). *raster: Geographic data analysis and modeling*. R package version 3.0-7. Retrieved from https://cran.r-project.org/package=raster%0A

Holt, B. G., Lessard, J. P., Borregaard, M. K., Fritz, S. A., Araújo, M. B., Dimitrov, D., … Rahbek, C. (2013). An update of Wallace's zoogeographic regions of the world. *Science*, *339*(6115), 74–78.

Jarzyna, M. A., & Jetz, W. (2018). Taxonomic and functional diversity change is scale dependent. *Nature Communications*, *9*, 2565. https://doi.org/10.1038/s41467-018-04889-z

Kaschner, K., Ready, J. S., Agbayani, E., Rius, J., Kesner-Reyes, K., Eastwood, P. D., & Close, C. H. (2016). *AquaMaps: Predicted range maps for aquatic species*. Retrieved from www.aquamaps.org

Kawamoto, T., & Rosvall, M. (2015). Estimating the resolution limit of the map equation in community detection. *Physical Review E*, *91*(1). https://doi.org/10.1103/PhysRevE.91.012809

Keil, P., Schweiger, O., Kühn, I., Kunin, W. E., Kuussaari, M., Settele, J., … Storch, D. (2012). Patterns of beta diversity in Europe: The role of climate, land cover and distance across scales. *Journal of Biogeography*, *39*(8), 1473–1486. https://doi.org/10.1111/j.1365-2699.2012.02701.x

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., … Webb, C. O. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, *26*(11), 1463–1464. https://doi.org/10.1093/bioinformatics/btq166

Koleff, P., Gaston, K. J., & Lennon, J. J. (2003). Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*, *72*, 367–382. https://doi.org/10.1046/j.1365-2656.2003.00710.x

Kreft, H., & Jetz, W. (2010). A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, *37*(11), 2029–2053. https://doi.org/10.1111/j.1365-2699.2010.02375.x

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129. https://doi.org/10.1007/BF02289694

Laffan, S. W., Rosauer, D. F., Di Virgilio, G., Miller, J. T., González-Orozco, C. E., Knerr, N., … Mishler, B. D. (2016). Range-weighted metrics of species and phylogenetic turnover can better resolve biogeographic transition zones. *Methods in Ecology and Evolution*, *7*(5), 580–588. https://doi.org/10.1111/2041-210X.12513

Meiri, S., & Chapple, D. G. (2016). Biases in the current knowledge of threat status in lizards, and bridging the 'assessment gap'. *Biological Conservation*, *204*, 6–15. https://doi.org/10.1016/j.biocon.2016.03.009

Morrone, J. J. (2018). The spectre of biogeographical regionalization. *Journal of Biogeography*, *45*(2), 282–288. https://doi.org/10.1111/jbi.13135

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Solymos, P. (2019). *vegan: Community ecology package*. R package version 2.5.6. Retrieved from https://cran.r-project.org/package=vegan

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528.

Roll, U., Feldman, A., Novosolov, M., Allison, A., Bauer, A. M., Bernard, R., … Meiri, S. (2017). The global distribution of tetrapods reveals a need for targeted reptile conservation. *Nature Ecology and Evolution*, *1*(11), 1677–1682. https://doi.org/10.1038/s41559-017-0332-2

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–1123. https://doi.org/10.1073/pnas.0706851105

Salvador, S., & Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Proceedings of the Sixteenth IEEE International Conference on Tools with Artificial Intelligence* (pp. 576–584). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*(2), 33–40. https://doi.org/10.2307/1217208

Ter Braak, C. J. F., & Prentice, I. C. (1988). A theory of gradient analysis. *Advances in Ecological Research*, *18*(C), 271–317.

Thébault, E. (2013). Identifying compartments in presence–absence matrices and bipartite networks: Insights into modularity measures. *Journal of Biogeography*, *40*(4), 759–768. https://doi.org/10.1111/jbi.12015

Tonini, J. F. R., Beard, K. H., Ferreira, R. B., Jetz, W., & Pyron, R. A. (2016). Fully-sampled phylogenies of squamates reveal evolutionary patterns in threat status. *Biological Conservation*, *204*, 23–31. https://doi.org/10.1016/j.biocon.2016.03.039

Vilhena, D. A., & Antonelli, A. (2015). A network approach for identifying and delimiting biogeographical regions. *Nature Communications*, *6*, 6848. https://doi.org/10.1038/ncomms7848

Wallace, A. R. (1876). *The geographical distribution of animals*. Cambridge, UK: Cambridge University Press.

Wiley, E. O. (1988). Vicariance biogeography. *Annual Review of Ecology and Systematics*, *19*, 513–542. https://doi.org/10.1146/annurev.es.19.110188.002501

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Daru BH, Karunarathne P, Schliep K. phyloregion: R package for biogeographical regionalization and macroecology. *Methods Ecol Evol*. 2020;11:1483–1491. https://doi.org/10.1111/2041-210X.13478