`phyloregion`: R package for biogeographic regionalization and spatial conservation

Article type: Application

Barnabas H. Daru[1,*], Piyal Karunarathne[1], and Klaus Schliep[2]

[1]*Department of Life Science, Texas A&M University-Corpus Christi, Corpus Christi, 78412 TX, USA*

[2]*Department of Biology, University of Massachusetts Boston, Boston MA 02125, USA*

*Correspondence: Barnabas H. Daru (barnabas.daru@tamucc.edu)

Running headline: regionalization and spatial conservation in R

Summary

1. Biogeographical regionalization is the classification of regions in terms of their biotas and is key to understanding biodiversity patterns across the world. Previously, it was only possible to perform analysis of biogeographic regionalization on small datasets, often using tools that are difficult to replicate.
2. Here, we present `phyloregion`, a package for the analysis of biogeographic regionalization and spatial conservation 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 three to four orders of magnitude faster and memory efficient for cluster analysis, determining optimal number of clusters, evolutionary distinctiveness of regions, as well as analysis of more standard conservation measures of phylogenetic diversity, phylogenetic endemism, and evolutionary distinctiveness and global endangerment.
4. A case study of zoogeographic regionalization for 9574 species of squamate reptiles (amphisbaenians, lizards, and snakes) across the globe, reveals their evolutionary affinities, using visualization tools that allow rapid identification of patterns and underlying processes with user-friendly colours–for example–indicating the levels of differentiation of the taxa in different regions.
5. Ultimately, `phyloregion` would facilitate rapid biogeographic analyses that accommodates the ongoing mass-production of species occurrence records

and phylogenetic datasets at any scale and for any taxonomic group into completely reproducible R workflows.

Key-words: biogeography, bioinformatics, conservation, phylogenetics, regionalization, software

1.0 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 units of biogeographical regionalization i.e., “phyloregions” or “bioregions”, are key to our understanding of the ecological and historical drivers affecting species distribution in macroecological or large-scale conservation studies (Kreft & Jetz, 2010; Ladle & Whittaker, 2011; Moreno Saiz et al., 2013; Oikonomou et al., 2014; Ficetola et al., 2017; Morrone, 2018). When paired with phylogenetic information, biogeographical regionalization allows geographic regions that do not share any species in common to be quantified (Graham and Fine, 2008), and can identify patterns overlooked by species-level analyses (Holt et al. 2013; Daru et al. 2016). However, compared to the mass-production of species distribution and phylogenetic datasets, statistical and computational approaches necessary to analyze such data, and approaches that can incorporate efficient storage and manipulation of such data, are lacking.

A few open-source tools have recently become available and can provide infrastructural support for analysis of biogeographical regionalization. The *ape* package (Paradis and Schliep 2018) 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* focused on analysis of phylogenetic community structure and trait evolution (Kembel et al. 2010). The use of network methods to detect bioregions (Carstensen and Olesen 2009, Thébault 2013, Vilhena and 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 (Bloomfield et al. 2018). However, there is no consensus on which method is the most appropriate for biogeographical regionalization and spatial conservation at large scales (Dapporto et al., 2015; Bloomfield et al. 2018; Morrone, 2018). The most effective approach to biogeographical regionalization might therefore depend on the system under study and the research questions.

Here, we present `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 clustering substantially large-scale species assemblages, determining optimal number of clusters, quantifying evolutionary distinctiveness of phyloregions, and visualizing various facets of alpha and beta (differences in species composition between local communities) diversity. We illustrate the utility of the proposed

package with a simulated dataset and an empirical dataset on the flora of southern Africa that includes species distributions and their phylogenetic relationships. Moreover, we also present a case study for zoogeographic regionalization with the most comprehensive dataset on the phylogenetic relationships and geographic distributions for 9574 species of squamate reptiles (amphisbaenians, lizards, and snakes) across the globe, to demonstrate its potential for analysis at any scale and for any taxonomic group. Visualization tools allow rapid identification of phyloregions with colours in multidimensional scaling space indicating levels of differentiation of the taxa in different phyloregions.

2.0 Overview and general workflow of `phyloregion`

The `phyloregion` package interacts with several other R packages including *Matrix* (Bates and Maechler 2019), *ape* (Paradis & Schliep 2018), *betapart* (Baselga & Orme 2012), *raster* (Hijmans 2019), and *sp* (Bivand et al. 2013). We provide a workflow of the `phyloregion` package for biogeographical assessment of any selected taxa and region (**Fig. 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 development version of `phyloregion` is hosted on github at https://github.com/darunabas/phyloregion. To install `phyloregion`, in R, type:

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

2.1. Raw Data

2.1.1 Distribution data input

The `phyloregion` package ships with functions for manipulating at least three categories of distribution data at varying spatial grains and extents: point records, extent-of-occurrence polygons and raster layers. Extent-of-occurrence range maps can be derived from the IUCN Redlist spatial database (https://www.iucnredlist.org/resources/spatial-data-download), published monographs or field guides validated by taxonomic experts. Point records are commonly derived from GBIF, iDigBio, or CIESIN and typically have columns of geographic coordinates for each observation. Raster layers are typically derived from analysis of species distribution modeling, 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.

2.1.2 Phylogenetic data

Phylogenies are often derived from DNA sequences or supertree approaches, however, they tend to be prevalent with missing taxa for most non-charismatic groups e.g. plants or insects. When paired with distribution data, phylogenies can aid the discovery of common patterns and processes that underlie the formation of biogeographic regions (Wiley 1988, Daru et al. 2017). 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.

3.0 Data preparation and analyses

3.1. 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 analyzing 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 > 2^31 elements. One approach to overcome this limitation is to utilize sparse matrix, a matrix with a high proportion of zero entries (Duff 1977). Because a sparse matrix is comprised

159 mostly of 0s, it only stores the non-zero entries, from which several measures of
160 biodiversity including biogeographical regionalization can be calculated. Our
161 `sampl2sparse` function allows conversion of community data from either long or
162 wide formats to a condensed sparse matrix (**Fig. 2**) to ease downstream
163 analyses such as compositional dissimilarity and avoid the exhaustion of
164 computer memory capacities.
165
166 *3.2. Matching phylogeny and community composition data*
167 In community ecology and biogeographic analyses, it is sometimes desirable to
168 make sure that different datasets (e.g. community, phylogeny and trait) match
169 one another (Kembel et al. 2010). However, existing tools are not tailored for
170 comparing taxa in mega phylogenies spanning thousands of taxa with community
171 composition datasets at large scales. We present `match_phylo_comm` that
172 compares a sparse community matrix against a phylogenetic tree and adds
173 missing species to the tree at the genus or higher taxonomic levels.
174
175 *3.3. Generating beta diversity (phylogenetic and non-phylogenetic)*
176 The three commonly used methods for quantifying β-diversity, the variation in
177 species composition among sites, – Simpson, Sorenson and Jaccard (Laffan et
178 al. 2016) – are included in the package as a comparative and optimal selection
179 tool. The `phyloregion`'s functions `beta_diss` and `phylobeta` compute
180 efficiently pairwise dissimilarities matrices for large sparse community matrices

and phylogenetic trees for taxonomic and phylogenetic turnover, respectively. The results are stored as distance objects for later use.

3.4. 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 nine widely used hierarchical clustering algorithms (including UPGMA, and single linkage) on the (phylogenetic) beta diversity matrix for degree of data distortion using Sokal & 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.

3.5. Determining the optimal number of clusters

The function `optimal.phyloregion` utilizes the efficiency of the so-called "elbow" (also "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 NMDS

ordination, and are assessed for spatial coherence by mapping and/or quantifying their evolutionary distinctiveness.

3.6. Evolutionary distinctiveness of phyloregions

The function `ed_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 in the mechanisms that are responsible for generating ecological diversity such as speciation, niche conservatism, extinction and dispersal (Holt et al. 2013; Daru et al. 2017).

4.0. Visual representation and assessment of biogeographic regions

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

- `plot_phyloregion` can display clusters of cells (i.e. 'phyloregions' or 'bioregions') 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 (**Fig. 1**).

- `plot_evoldistinct` quantifies evolutionary distinctiveness of phyloregions in geographic space as the mean of pairwise beta diversity values between each phyloregion and all other phyloregions and displays them in HCL colour space (default is "`YlOrBr`" palette; **Fig. 1**). Darker regions indicate regions of higher evolutionary distinctiveness.
- `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.

5.0 Case study of biogeographical regionalization of squamate reptiles

We validated the application of the `phyloregion` package on the geographic distributions and phylogenetic data for all 9574 species of squamate reptiles across the globe (data: Tonini et al. 2016). Despite the fact that reptiles were part of the dataset used in Wallace's original zoogeographic regionalization along with birds, mammals and insects (Wallace 1876), they have been largely neglected in modern regionalization schemes (Kreft & Jetz 2010; Holt et al., 2013; Meiri & Chapple 2016). Nevertheless, squamate reptiles are one of the most diverse and widely distributed animal 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 biogeographic regionalization and spatial conservation at large scales.

We used updated extent-of-occurrence polygons representing the maximum geographical extent of each squamate reptile species (Roll et al. 2017). We ran the `polys2comm, sampl2sparse,` and `match_phylo_comm` wrapper functions to generate the community data at a resolution of 1°× 1°. 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), and took the mean across grid cells using `mean_matrix`. 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; review function `phylobeta`.

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 (**Fig. 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 show substantial congruence to Holt et al.'s (2013) updates of Wallace's original zoogeographic regions including

Oceanian, Australian, Madagascan, Palearctic and Nearctic (**Fig. 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 focal group being reptiles whereas Holt et al. present results for birds, mammals and amphibians; or differences in spatial grain size (1°× 1° in our study vs 2°× 2° in Holt et al. (2013)). Phylogenetic beta diversity and environmental correlates are systematically grain (spatial resolution) dependent (e.g. Keil et al. 2012).

Notably, geographically proximal phyloregions tend to have low levels of faunal similarity (**Fig. 3b**), suggesting spatial patterns of species diversity can have different phylogenetic structures (Hawkins et al. 2012). Mean phylogenetic turnover of squamate reptiles between a phyloregion and all other phyloregions (function `ed_phyloregion`) indicates a latitudinal gradient in evolutionary distinctiveness, with higher evolutionary distinctiveness in the tropics than in temperate phyloregions (**Fig. 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; **Fig. 3c**), reflecting strong niche conservatism or 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 brings a new component of evolutionary importance of each region to the biogeographic regionalization as well as for conservation prioritization. Most of the phyloregions found here spanned multiple ecoregions and biogeographic realms, suggesting that conservation planning should be adjusted to cover these larger phyloregions.

6.0. `phyloregion` as tool for spatial conservation

We demonstrate the utility of `phyloregion` in mapping standard conservation metrics of species richness, weighted endemism (`weighted.endemism`) and threat (`mapTraits`) as well as fast computations of phylodiversity measures such as phylogenetic diversity (`PD`), phylogenetic endemism (`PE`), and evolutionary distinctiveness and global endangerment (`EDGE`). The major advantage of these functions compared to available tools e.g. biodiverse (Laffan et al. 2010), is the ability to utilize sparse matrix that speeds up the analyses without exhausting computer memories, making it ideal for handling any data from small local scales to large regional and global scales.

6.0. Benchmarking `phyloregion`

We compared the execution time of `phyloregion`'s functions with available packages using exactly the same datasets (R code for benchmarking

317 `phyloregion` with available packages is provided as Data S1). Regardless of
318 the size of the distribution data and phylogenetic tree, `phyloregion` is 3 or 4
319 orders of magnitude faster and memory efficient (**Fig. 4**).
320
321 **7.0. Concluding Remarks**
322 Despite the few other tools that have provided support for biogeographic
323 regionalization and spatial conservation e.g. *ape* (Paradis & Schliep 2018),
324 betapart (Baselga & Orme 2012), or *vegan* (Oksanen et al. 2019) among many
325 others, `phyloregion` adds the following novelties compared to available
326 packages: 1) ability to utilize sparse matrix and large-scale phylogenies for
327 analysis of biogeographical regionalization and spatial conservation, allowing
328 normal operations of a typical matrix in base R to be done on the sparse matrix,
329 2) novel functions for speedy raw data conversion to sparse community matrix as
330 well as a user-friendly analysis of biogeographical regionalization into completely
331 reproducible R workflows, 3) although the functionality of the package has been
332 developed with biogeographical regionalization in mind, it can accommodate
333 analysis of spatial conservation at large scales such as mapping various
334 biodiversity metrics for conservation ranging from mapping biodiversity hotspots
335 of species richness, endemism, or threat. Other implementations of
336 `phyloregion` include the addition of phylogenetic information and sparse
337 community matrix to map evolutionary diversity including phylogenetic diversity,
338 phylogenetic endemism, and evolutionary distinctiveness and global
339 endangerment.

Overall, no hard rule exists on how to perform analysis of biogeographic regionalization or spatial conservation - the choice of approach will ultimately depend on the goal of the study, questions, hypotheses or the taxonomic group. The goal of `phyloregion` is to facilitate analysis of biogeographic regionalization and spatial conservation 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.

Authors' contributions

B.H.D. conceived the project. B.H.D. and K.S. developed the method. B.H.D., K.S., and P.K. tested the method. B.H.D. analyzed the data and led the writing with help from P.K. All co-authors assisted with edits and approve publication.

Data accessibility

The `phyloregion` R package and documentation are hosted at https://github.com/darunabas/phyloregion. 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 et al. 2019).

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: Second Edition. Springer New York.

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.

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.

Carstensen, D. W., & Olesen, J. M. (2009). Wallacea and its nectarivorous birds: nestedness and modules. Journal of Biogeography, 36(8), 1540–1550.

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 and Evolution, 32(11), 845–860.

386 Daru, B. H., Karunarathne, P., & Schliep, K. (2019). phyloregion: R package for
387 biogeographic regionalization and spatial conservation. Dryad, Dataset
388 DOI https://doi.org/10.5061/dryad.tdz08kpw6.
389 Daru, B. H., Van der Bank, M., Maurin, O., Yessoufou, K., Schaefer, H., Slingsby,
390 J. A., & Davies, T. J. (2016). A novel phylogenetic regionalization of the
391 phytogeographic zones of southern Africa reveals their hidden
392 evolutionary affinities. Journal of Biogeography, 43(1), 155-166.
393 Duff, I. S. (1977). A survey of sparse matrix research. Proceedings of the IEEE,
394 65(4), 500–535.
395 Ficetola, G. F., Mazel, F., & Thuiller, W. (2017). Global determinants of
396 zoogeographical boundaries. Nature Ecology and Evolution, 1(4), 0089.
397 Graham, C. H., & Fine, P. V. A. (2008). Phylogenetic beta diversity: linking
398 ecological and evolutionary processes across space in time. Ecology
399 Letters, 11, 1265–1277.
400 Hawkins, B.A. et al. (2012) Different evolutionary histories underlie congruent
401 species richness gradients of birds and mammals. J. Biogeogr. 39, 825–
402 841.
403 Hijmans, R. J. (2019). raster: Geographic Data Analysis and Modeling. R
404 package version 3.0-7. Retrieved from https://cran.r-
405 project.org/package=raster%0A
406 Holt, B. G., Lessard, J. P., Borregaard, M. K., Fritz, S. A., Araújo, M. B., Dimitrov,
407 D., … Rahbek, C. (2013). An update of Wallace's zoogeographic regions
408 of the world. Science, 339(6115), 74–78.

409 Kaschner, K., Ready, J. S., Agbayani, E., Rius, J., Kesner-Reyes, K., Eastwood,
410 P. D., … Close, C. H. (2016). AquaMaps: Predicted range maps for
411 aquatic species. Retrieved from www.aquamaps.org
412 Keil, P., Schweiger, O., Kühn, I., Kunin, W. E., Kuussaari, M., Settele, J., …
413 Storch, D. (2012). Patterns of beta diversity in Europe: the role of climate,
414 land cover and distance across scales. Journal of Biogeography, 39(8),
415 1473–1486.
416 Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H.,
417 Ackerly, D. D., … Webb, C. O. (2010). Picante: R tools for integrating
418 phylogenies and ecology. Bioinformatics, 26(11), 1463–1464.
419 Kreft, H., & Jetz, W. (2010). A framework for delineating biogeographical regions
420 based on species distributions. Journal of Biogeography, 37(11), 2029–
421 2053.
422 Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method.
423 Psychometrika, 29(2), 115–129.
424 Ladle, R., & Whittaker, R. J. (2011). Conservation biogeography. John Wiley &
425 Sons.
426 Laffan, S. W., Lubarsky, E., & Rosauer, D. F. (2010). Biodiverse, a tool for the
427 spatial analysis of biological and related diversity. Ecography, 33(4), 643–
428 647.
429 Laffan, S. W., Rosauer, D. F., Di Virgilio, G., Miller, J. T., González-Orozco, C.
430 E., Knerr, N., Thornhill, A. H., & Mishler, B. D. (2016). Range-weighted
431 metrics of species and phylogenetic turnover can better resolve

biogeographic transition zones. Methods in Ecology and Evolution, 7(5), 580-588.

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.

Moreno Saiz, J. C., Donato, M., Katinas, L., Crisci, J. V., & Posadas, P. (2013). New insights into the biogeography of south-western Europe: spatial patterns from vascular plants using cluster analysis and parsimony. Journal of Biogeography, 40(1), 90–104.

Morrone, J. J. (2018). The spectre of biogeographical regionalization. Journal of Biogeography, 45(2), 282–288.

Oikonomou, A., Leprieur, F., & Leonardos, I. D. (2014). Biogeography of freshwater fishes of the Balkan Peninsula. Hydrobiologia, 738(1), 205–220.

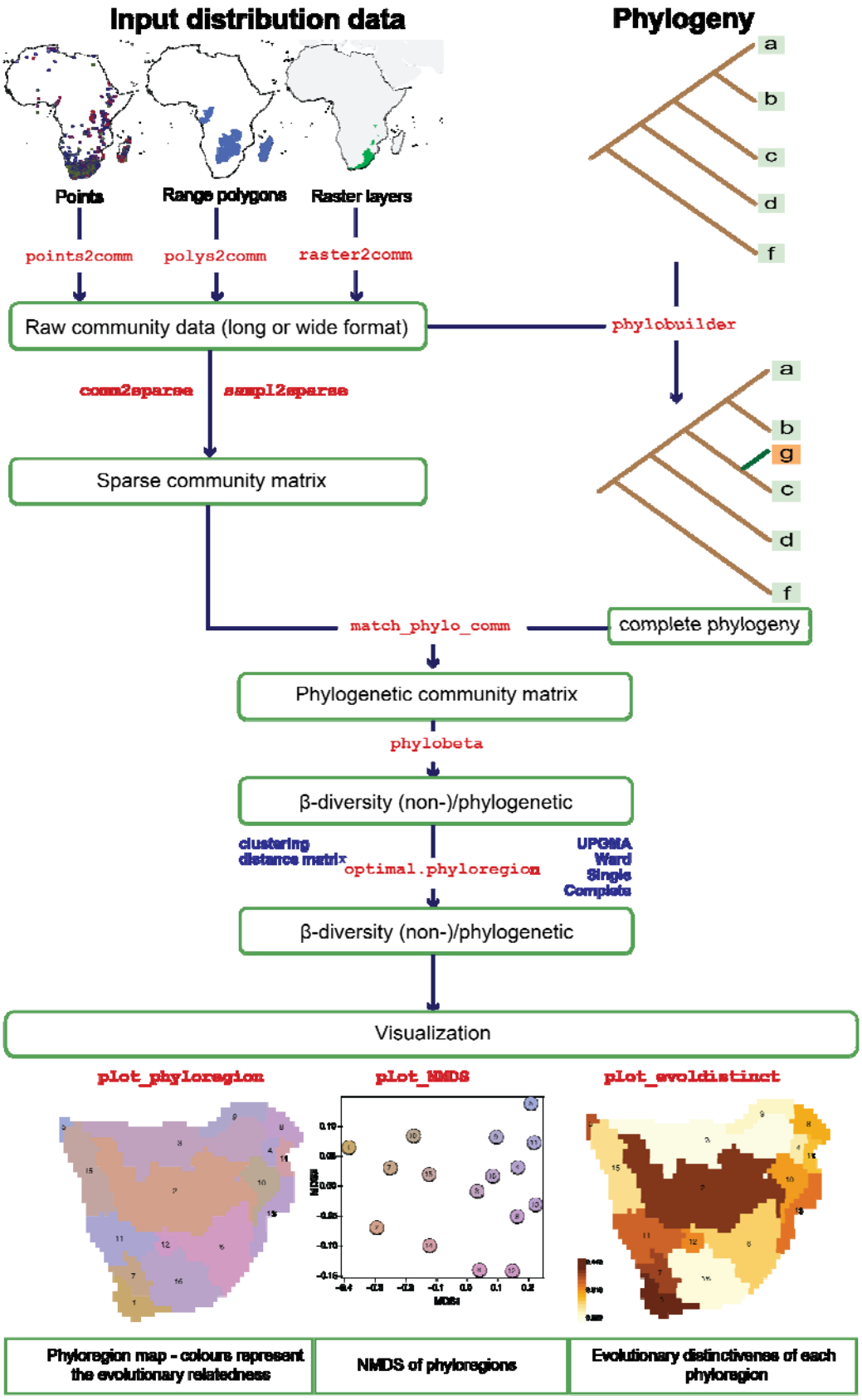
Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Solymos, P. (2019). vegan: community ecology package. R package version 2.5.6. Retrieved from https://cran.r-project.org/package=vegan

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

454 targeted reptile conservation. Nature Ecology and Evolution, 1(11), 1677–
455 1682.
456 Salvador, S., & Chan, P. (2004). Determining the number of clusters/segments in
457 hierarchical clustering/segmentation algorithms. In Proceedings -
458 International Conference on Tools with Artificial Intelligence, ICTAI (pp.
459 576–584).
460 Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective
461 methods. Taxon, 11(2), 33–40.
462 Ter Braak, C. J. F., & Prentice, I. C. (1988). A theory of gradient analysis.
463 Advances in Ecological Research, 18(C), 271–317.
464 Thébault, E. (2013). Identifying compartments in presence-absence matrices and
465 bipartite networks: insights into modularity measures. Journal of
466 Biogeography, 40(4), 759–768.
467 Tonini, J. F. R., Beard, K. H., Ferreira, R. B., Jetz, W., & Pyron, R. A. (2016).
468 Fully-sampled phylogenies of squamates reveal evolutionary patterns in
469 threat status. Biological Conservation, 204, 23–31.
470 Vilhena, D. A., & Antonelli, A. (2015). A network approach for identifying and
471 delimiting biogeographical regions. Nature Communications, 6, 6848.
472 Wallace, A. R. (1876). The geographical distribution of animals. Cambridge, UK:
473 Cambridge University Press Cambridge.
474 Wiley, E. O. (1988). Vicariance Biogeography. Annual Review of Ecology and
475 Systematics, 19, 513–542.
476

477 **Figures**



478

Fig. 1. Typical workflow for analysis of biogeographical regionalization using `phyloregion`. a) Distribution data (point records, polygons, and raster layers) is converted to a long community data frame format before b) conversion to a sparse community matrix. When paired with phylogenetic data, `phylobuilder` creates a subtree with largest overlap from a species list, thereby ensuring complete representation of missing data. c) phylocommunity matrix to visualization of results.

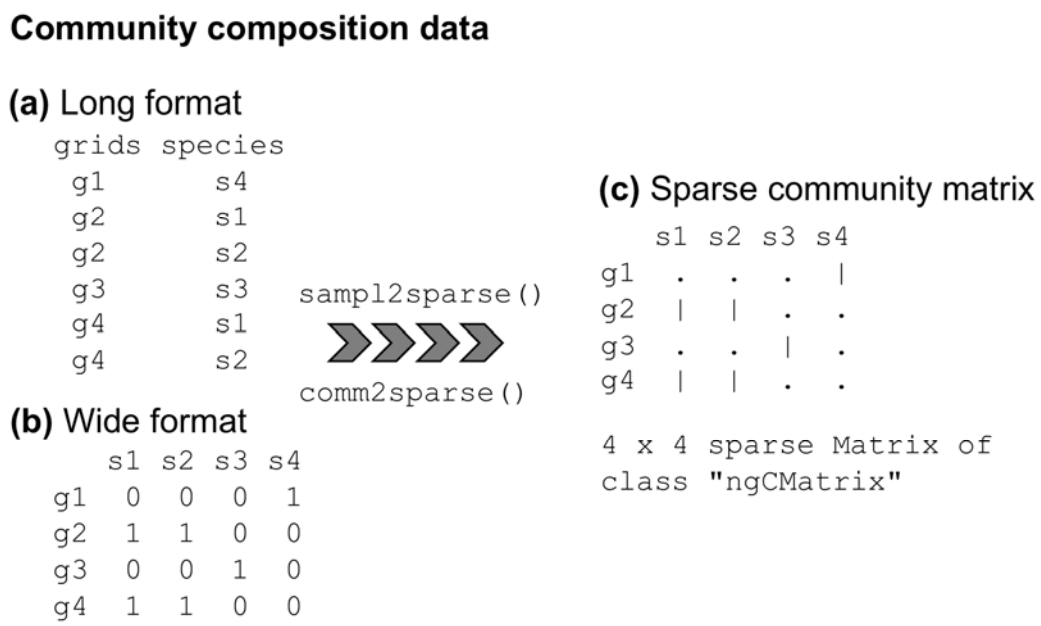
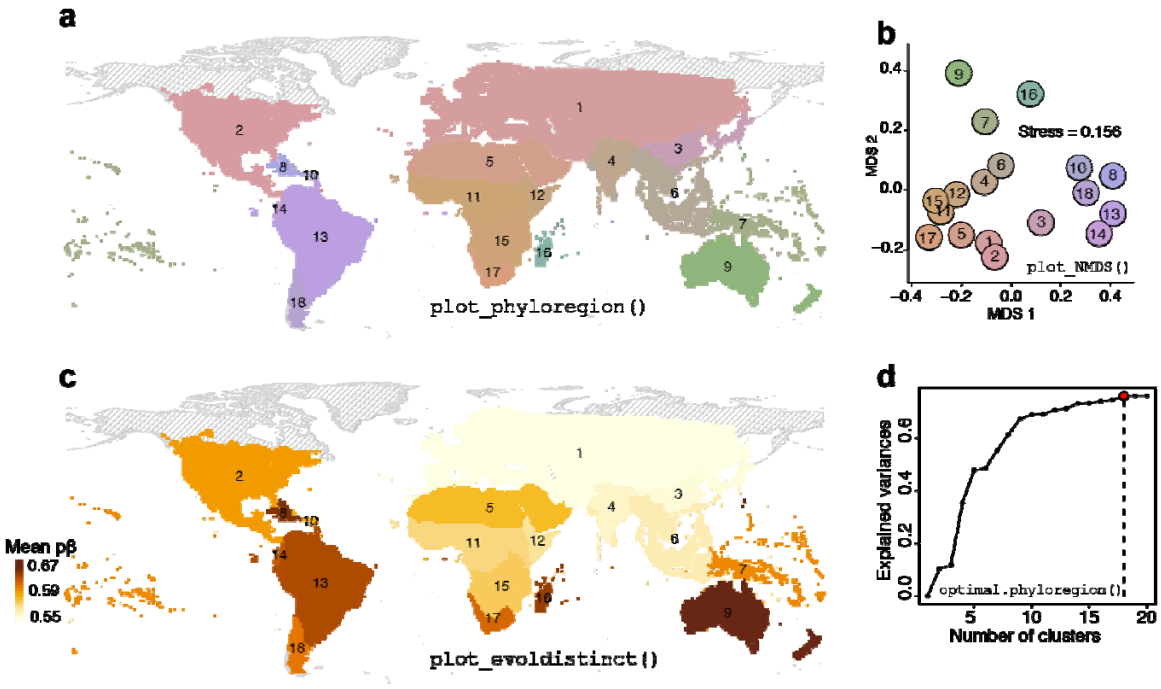


Fig. 2. Illustration showing community data conversion to sparse community matrix by (a) `sampl2sparse` function when the raw data is in long community data format, or (b) `com2sparse` for wide community data format. The result is (c) a sparse community matrix for downstream analysis.



494

495 **Fig. 3.** A global phylogenetic regionalization of 9574 species of squamate reptiles reveals their
496 evolutionary affinities. **a**, Map of phyloregions shows evolutionary affinities among disjunct
497 assemblages (function `plot_phyloregion`). **b**, The ordination of phyloregions in NMDS space
498 shows a tropical-temperate divide (function `plot_NMDS`). **c**, evolutionary distinctiveness is high in
499 the tropics than temperate bioregions (function `plot_evoldistinct`). **d**, the optimal number of
500 phyloregions (function `optimal.phyloregion`). Colours differentiating between phyloregions in
501 the map (**a**) and NMDS plot (**b**) are identical.

502

503

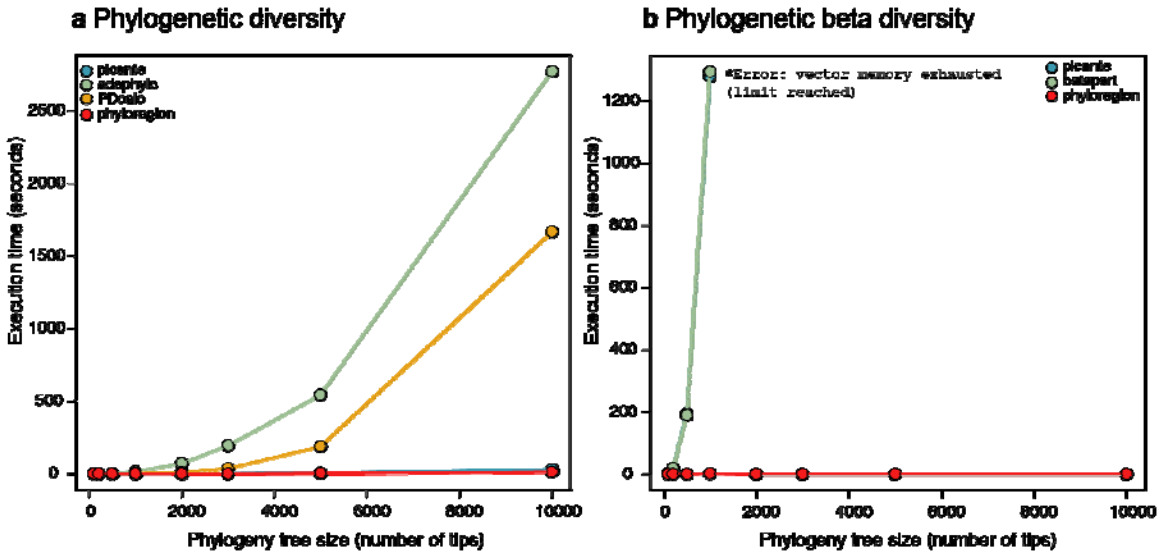


Fig. 4. Benchmarking `phyloregion` with available packages in analysis of **a**, phylogenetic diversity, and **b**, phylogenetic beta diversity.