

1 **phyloregion: R package for biogeographic regionalization and spatial**
2 **conservation**

3

4 **Article type:** Application

5

6 Barnabas H. Daru^{1,*}, Piyal Karunaratne¹, and Klaus Schliep²

7

8 *¹Department of Life Science, Texas A&M University-Corpus Christi, Corpus*

9 *Christi, 78412 TX, USA*

10 *²Department of Biology, University of Massachusetts Boston, Boston MA 02125,*

11 *USA*

12

13

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

15 **Running headline:** regionalization and spatial conservation in R

16

17 **Summary**

- 18 1. Biogeographical regionalization is the classification of regions in terms of their
19 biotas and is key to understanding biodiversity patterns across the world.
20 Previously, it was only possible to perform analysis of biogeographic
21 regionalization on small datasets, often using tools that are difficult to
22 replicate.
- 23 2. Here, we present `phyloregion`, a package for the analysis of biogeographic
24 regionalization and spatial conservation in the R computing environment,
25 tailored for mega phylogenies and macroecological datasets of ever-
26 increasing size and complexity.
- 27 3. Compared to available packages, `phyloregion` is three to four orders of
28 magnitude faster and memory efficient for cluster analysis, determining
29 optimal number of clusters, evolutionary distinctiveness of regions, as well as
30 analysis of more standard conservation measures of phylogenetic diversity,
31 phylogenetic endemism, and evolutionary distinctiveness and global
32 endangerment.
- 33 4. A case study of zoogeographic regionalization for 9574 species of squamate
34 reptiles (amphisbaenians, lizards, and snakes) across the globe, reveals their
35 evolutionary affinities, using visualization tools that allow rapid identification of
36 patterns and underlying processes with user-friendly colours—for example—
37 indicating the levels of differentiation of the taxa in different regions.
- 38 5. Ultimately, `phyloregion` would facilitate rapid biogeographic analyses that
39 accommodates the ongoing mass-production of species occurrence records

40 and phylogenetic datasets at any scale and for any taxonomic group into

41 completely reproducible R workflows.

42 **Key-words:** biogeography, bioinformatics, conservation, phylogenetics,

43 regionalization, software

44 **1.0 Introduction**

45 In biogeography, there is growing interest in the analysis of datasets of ever-
46 increasing size and complexity to explain biodiversity patterns and underlying
47 processes. A common approach is biogeographical regionalization, the grouping
48 of organisms based on shared features and how they respond to past or current
49 physical and biological determinants (Kreft & Jetz, 2010; Morrone, 2018). The
50 units of biogeographical regionalization i.e., “phyloregions” or “bioregions”, are
51 key to our understanding of the ecological and historical drivers affecting species
52 distribution in macroecological or large-scale conservation studies (Kreft & Jetz,
53 2010; Ladle & Whittaker, 2011; Moreno Saiz et al., 2013; Oikonomou et al.,
54 2014; Ficetola et al., 2017; Morrone, 2018). When paired with phylogenetic
55 information, biogeographical regionalization allows geographic regions that do
56 not share any species in common to be quantified (Graham and Fine, 2008), and
57 can identify patterns overlooked by species-level analyses (Holt et al. 2013; Daru
58 et al. 2016). However, compared to the mass-production of species distribution
59 and phylogenetic datasets, statistical and computational approaches necessary
60 to analyze such data, and approaches that can incorporate efficient storage and
61 manipulation of such data, are lacking.

62

63 A few open-source tools have recently become available and can provide
64 infrastructural support for analysis of biogeographical regionalization. The *ape*
65 package (Paradis and Schliep 2018) contains a comprehensive collection of tools
66 for analyses of phylogenetics and evolution and is useful for reading, writing, and

67 manipulating phylogenetic trees, among many other functions. The *betapart*
68 package (Baselga & Orme 2012) performs computations of total dissimilarity in
69 species composition along with their respective turnover and nestedness
70 components. *picante* focused on analysis of phylogenetic community structure
71 and trait evolution (Kembel et al. 2010). The use of network methods to detect
72 bioregions (Carstensen and Olesen 2009, Thébault 2013, Vilhena and Antonelli
73 2015), while not yet implemented in the R computing environment, provides an
74 alternative clustering method based on bipartite networks, and performs well at
75 identifying interzones between regions (Bloomfield et al. 2018). However, there is
76 no consensus on which method is the most appropriate for biogeographical
77 regionalization and spatial conservation at large scales (Dapporto et al., 2015;
78 Bloomfield et al. 2018; Morrone, 2018). The most effective approach to
79 biogeographical regionalization might therefore depend on the system under
80 study and the research questions.

81

82 Here, we present `phyloregion` R package that permits the integration of
83 phylogenetic relationships and species distributions for identifying
84 biogeographical regions of different lineages to elucidate the spatial and temporal
85 evolution of biota in a region. Specifically, `phyloregion` provides functions for
86 clustering substantially large-scale species assemblages, determining optimal
87 number of clusters, quantifying evolutionary distinctiveness of phyloregions, and
88 visualizing various facets of alpha and beta (differences in species composition
89 between local communities) diversity. We illustrate the utility of the proposed

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

100

101 **2.0 Overview and general workflow of `phyloregion`**

102 The `phyloregion` package interacts with several other R packages including
103 *Matrix* (Bates and Maechler 2019), *ape* (Paradis & Schliep 2018), *betapart*
104 (Baselga & Orme 2012), *raster* (Hijmans 2019), and *sp* (Bivand et al. 2013). We
105 provide a workflow of the `phyloregion` package for biogeographical
106 assessment of any selected taxa and region (**Fig. 1**). The workflow demonstrates
107 steps from preparation of different types of data to visualizing the results of
108 biogeographical regionalization, together with tips on selecting the optimal
109 method for achieving the best output, depending on the types of data used and
110 research questions. The development version of `phyloregion` is hosted on
111 github at <https://github.com/darunabas/phyloregion>. To install `phyloregion`, in
112 R, type:

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

117

118 **2.1. Raw Data**

119 *2.1.1 Distribution data input*

120 The `phyloregion` package ships with functions for manipulating at least three
121 categories of distribution data at varying spatial grains and extents: point records,
122 extent-of-occurrence polygons and raster layers. Extent-of-occurrence range
123 maps can be derived from the IUCN Redlist spatial database
124 (<https://www.iucnredlist.org/resources/spatial-data-download>), published
125 monographs or field guides validated by taxonomic experts. Point records are
126 commonly derived from GBIF, iDigBio, or CIESIN and typically have columns of
127 geographic coordinates for each observation. Raster layers are typically derived
128 from analysis of species distribution modeling, such as *aquamaps* (Kaschner et
129 al. 2016). An overview can be easily obtained with the functions `points2comm`,
130 `polys2comm` and `raster2comm` for point records, polygons, or raster layers,
131 respectively. Depending on the data source, all three functions ultimately provide
132 convenient interfaces to convert the distribution data to a community data frame
133 at varying spatial grains and extents for downstream analyses.

134

135 *2.1.2 Phylogenetic data*

136 Phylogenies are often derived from DNA sequences or supertree approaches,
137 however, they tend to be prevalent with missing taxa for most non-charismatic
138 groups e.g. plants or insects. When paired with distribution data, phylogenies can
139 aid the discovery of common patterns and processes that underlie the formation
140 of biogeographic regions (Wiley 1988, Daru et al. 2017). The function
141 `phylobuilder` appends missing taxa to a supertree. Unlike other tree-building
142 algorithms that manually graft missing taxa into a working supertree,
143 `phylobuilder` creates a subtree with the largest overlap from a species list at a
144 fast speed. If species in the taxon list are not in the tree (tip label), species will be
145 added at the most recent common ancestor at the genus or family level when
146 possible.

147

148 **3.0 Data preparation and analyses**

149 *3.1. Sparse community matrix*

150 A community composition dataset is commonly represented as a matrix of 1s and
151 0s with species as columns and rows as spatial cells or communities. In practice,
152 such a matrix can contain many zero values because species are known to
153 generally have unimodal distributions along environmental gradients (ter Braak &
154 Prentice, 1988), and storing and analyzing every single element of that matrix
155 can be computationally challenging and expensive. Indeed, for large matrices,
156 most base R functions cannot make a table with $> 2^{31}$ elements. One approach
157 to overcome this limitation is to utilize sparse matrix, a matrix with a high
158 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 `samp12sparse` 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

181 and phylogenetic trees for taxonomic and phylogenetic turnover, respectively.

182 The results are stored as distance objects for later use.

183

184 *3.4. Cluster algorithm selection and validation*

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

194

195 *3.5. Determining the optimal number of clusters*

196 The function `optimal.phyloregion` utilizes the efficiency of the so-called
197 "elbow" (also "knee") method corresponding to the point of maximum curvature
198 (Salvador & Chan, 2004), to determine the optimal number of clusters that best
199 describes the observed (phylogenetic) beta diversity matrix. Depending on the
200 research question, the scale of the cutting depth or clustering algorithm `method`
201 can be varied systematically. The output is used to visualize relationships among
202 phyloregions using hierarchical dendrograms of dissimilarity and NMDS

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

205

206 *3.6. Evolutionary distinctiveness of phyloregions*

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

215

216 **4.0. Visual representation and assessment of biogeographic regions**

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

- 219 • `plot_phyloregion` can display clusters of cells (i.e. ‘phyloregions’ or
220 ‘bioregions’) in multidimensional scaling colour space matching the colour
221 vision of the human visual system (Kruskal 1964). The colours indicate the
222 levels of differentiation of clades in different phyloregions. Phyloregions
223 with similar colours have similar clades and those with different colours
224 differ in the clades they enclose (**Fig. 1**).

- 225 • `plot_evoldistinct` quantifies evolutionary distinctiveness of
226 phyloregions in geographic space as the mean of pairwise beta diversity
227 values between each phyloregion and all other phyloregions and displays
228 them in HCL colour space (default is “YlOrBr” palette; **Fig. 1**). Darker
229 regions indicate regions of higher evolutionary distinctiveness.
- 230 • `plot_swatch` maps discretized values of a quantity based on
231 continuous numerical variables of their cells or sites for visualization as
232 heatmap in sequential colour palettes.

233

234 **5.0 Case study of biogeographical regionalization of squamate reptiles**

235 We validated the application of the `phyloregion` package on the geographic
236 distributions and phylogenetic data for all 9574 species of squamate reptiles
237 across the globe (data: Tonini et al. 2016). Despite the fact that reptiles were part
238 of the dataset used in Wallace’s original zoogeographic regionalization along with
239 birds, mammals and insects (Wallace 1876), they have been largely neglected in
240 modern regionalization schemes (Kreft & Jetz 2010; Holt et al., 2013; Meiri &
241 Chapple 2016). Nevertheless, squamate reptiles are one of the most diverse and
242 widely distributed animal groups in the world (Böhm et al. 2013). Most notably,
243 due to the high extinction rates they are facing, the distribution data, phylogeny,
244 and evolutionary relatedness of squamates have recently been well documented
245 (Tonini et al. 2016 and references therein). These make squamate reptiles an
246 ideal system to test the robustness and implementation of `phyloregion` for
247 biogeographic regionalization and spatial conservation at large scales.

248

249 We used updated extent-of-occurrence polygons representing the maximum
250 geographical extent of each squamate reptile species (Roll et al. 2017). We ran
251 the `polys2comm`, `sampl2sparse`, and `match_phylo_comm` wrapper
252 functions to generate the community data at a resolution of $1^\circ \times 1^\circ$. Note that this
253 resolution can be adjusted by varying the `res` argument in the function
254 `fishnet(mask, res = 0.5)`. We accounted for phylogenetic uncertainty in
255 our analyses by drawing 100 trees at random from a posterior distribution of fully
256 resolved trees (Tonini et al. 2016) to generate phylogenetic dissimilarity matrices
257 (with Simpson's pairwise phylogenetic dissimilarities as default), and took the
258 mean across grid cells using `mean_matrix`. Note that other dissimilarity indices
259 such as "Jaccard" and "Sorensen" can be used as desired (Laffan et al. 2016),
260 depending on the data used and research questions; review function
261 `phylobeta`.

262

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

268

269 The resulting phyloregions for squamate reptiles show substantial congruence to
270 Holt et al.'s (2013) updates of Wallace's original zoogeographic regions including

271 Oceanian, Australian, Madagascan, Palearctic and Nearctic (**Fig. 3a**). However,
272 we also identified some discrepancies. For example, the Afrotropical realm
273 (*sensu* Holt et al. 2013) was divided into four phyloregions in our study
274 corresponding to West and Central Africa (11), Horn of Africa (12), Zambezi
275 (15), and South African (17). We also identified a new phyloregion overlapping
276 Chile-Patagonian in temperate South America. This discrepancy might be due to
277 the focal group being reptiles whereas Holt et al. present results for birds,
278 mammals and amphibians; or differences in spatial grain size ($1^\circ \times 1^\circ$ in our study
279 vs $2^\circ \times 2^\circ$ in Holt et al. (2013)). Phylogenetic beta diversity and environmental
280 correlates are systematically grain (spatial resolution) dependent (e.g. Keil et al.
281 2012).

282

283 Notably, geographically proximal phyloregions tend to have low levels of faunal
284 similarity (**Fig. 3b**), suggesting spatial patterns of species diversity can have
285 different phylogenetic structures (Hawkins et al. 2012). Mean phylogenetic
286 turnover of squamate reptiles between a phyloregion and all other phyloregions
287 (function `ed_phyloregion`) indicates a latitudinal gradient in evolutionary
288 distinctiveness, with higher evolutionary distinctiveness in the tropics than in
289 temperate phyloregions (**Fig. 3c**), a similar observation to Tonini et al. (2016).

290 Notably, the Australian phyloregion has the highest mean phylogenetic turnover
291 (mean phylogenetic turnover between Australian and all other phyloregions =
292 0.67; **Fig. 3c**), reflecting strong niche conservatism or limited dispersal of
293 lineages in this phyloregion.

294

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

302

303 **6.0. `phyloregion` as tool for spatial conservation**

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

313

314 **6.0. Benchmarking `phyloregion`**

315 We compared the execution time of `phyloregion`'s functions with available
316 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.

340

341 Overall, no hard rule exists on how to perform analysis of biogeographic
342 regionalization or spatial conservation - the choice of approach will ultimately
343 depend on the goal of the study, questions, hypotheses or the taxonomic group.
344 The goal of `phyloregion` is to facilitate analysis of biogeographic
345 regionalization and spatial conservation at any scale and for any taxonomic
346 group, tailored to accommodate the ongoing mass-production of species
347 occurrence data and phylogenetic datasets.

348

349 **Acknowledgements**

350 B.H.D. is supported by Texas A&M University at Corpus Christi.

351

352 **Authors' contributions**

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

356

357 **Data accessibility**

358 The `phyloregion` R package and documentation are hosted at
359 <https://github.com/darunabas/phyloregion>. All data and scripts necessary to
360 repeat the analyses for the squamate reptiles described here have been made
361 available through the Dryad Digital Data Repository
362 <https://doi.org/10.5061/dryad.tdz08kpw6> (Daru et al. 2019).

363 **References**

- 364 Baselga, A., & Orme, C. D. L. (2012). betapart: an R package for the study of
365 beta diversity. *Methods in Ecology and Evolution*, 3(5), 808–812.
- 366 Bates, D., & Maechler, M. (2019). Matrix: sparse and dense matrix classes and
367 methods. R Package Version 1.2-17. Retrieved from [https://cran.r-](https://cran.r-project.org/package=Matrix)
368 [project.org/package=Matrix](https://cran.r-project.org/package=Matrix)
- 369 Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). Applied spatial data
370 analysis with R: Second Edition. Springer New York.
- 371 Bloomfield, N. J., Knerr, N., & Encinas-Viso, F. (2018). A comparison of network
372 and clustering methods to detect biogeographical regions. *Ecography*,
373 41(1), 1–10.
- 374 Böhm, M., Collen, B., Baillie, J. E. M., Bowles, P., Chanson, J., Cox, N., ... Zug,
375 G. (2013). The conservation status of the world's reptiles. *Biological*
376 *Conservation*, 157, 372–385.
- 377 Carstensen, D. W., & Olesen, J. M. (2009). Wallacea and its nectarivorous birds:
378 nestedness and modules. *Journal of Biogeography*, 36(8), 1540–1550.
- 379 Dapporto, L., Ciolli, G., Dennis, R. L. H., Fox, R., & Shreeve, T. G. (2015). A new
380 procedure for extrapolating turnover regionalization at mid-small spatial
381 scales, tested on British butterflies. *Methods in Ecology and Evolution*,
382 6(11), 1287–1297.
- 383 Daru, B. H., Elliott, T. L., Park, D. S., & Davies, T. J. (2017). Understanding the
384 processes underpinning patterns of phylogenetic regionalization. *Trends in*
385 *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-](https://cran.r-project.org/package=raster%0A)
405 [project.org/package=raster%0A](https://cran.r-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

- 432 biogeographic transition zones. *Methods in Ecology and Evolution*, 7(5),
433 580-588.
- 434 Meiri, S., & Chapple, D. G. (2016). Biases in the current knowledge of threat
435 status in lizards, and bridging the 'assessment gap'. *Biological
436 Conservation*, 204, 6–15.
- 437 Moreno Saiz, J. C., Donato, M., Katinas, L., Crisci, J. V., & Posadas, P. (2013).
438 New insights into the biogeography of south-western Europe: spatial
439 patterns from vascular plants using cluster analysis and parsimony.
440 *Journal of Biogeography*, 40(1), 90–104.
- 441 Morrone, J. J. (2018). The spectre of biogeographical regionalization. *Journal of
442 Biogeography*, 45(2), 282–288.
- 443 Oikonomou, A., Leprieur, F., & Leonardos, I. D. (2014). Biogeography of
444 freshwater fishes of the Balkan Peninsula. *Hydrobiologia*, 738(1), 205–
445 220.
- 446 Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D.,
447 ... Solymos, P. (2019). *vegan: community ecology package*. R package
448 version 2.5.6. Retrieved from <https://cran.r-project.org/package=vegan>
- 449 Paradis, E., & Schliep, K. (2019). *Ape 5.0: An environment for modern
450 phylogenetics and evolutionary analyses in R*. *Bioinformatics*, 35(3), 526–
451 528.
- 452 Roll, U., Feldman, A., Novosolov, M., Allison, A., Bauer, A. M., Bernard, R., ...
453 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.

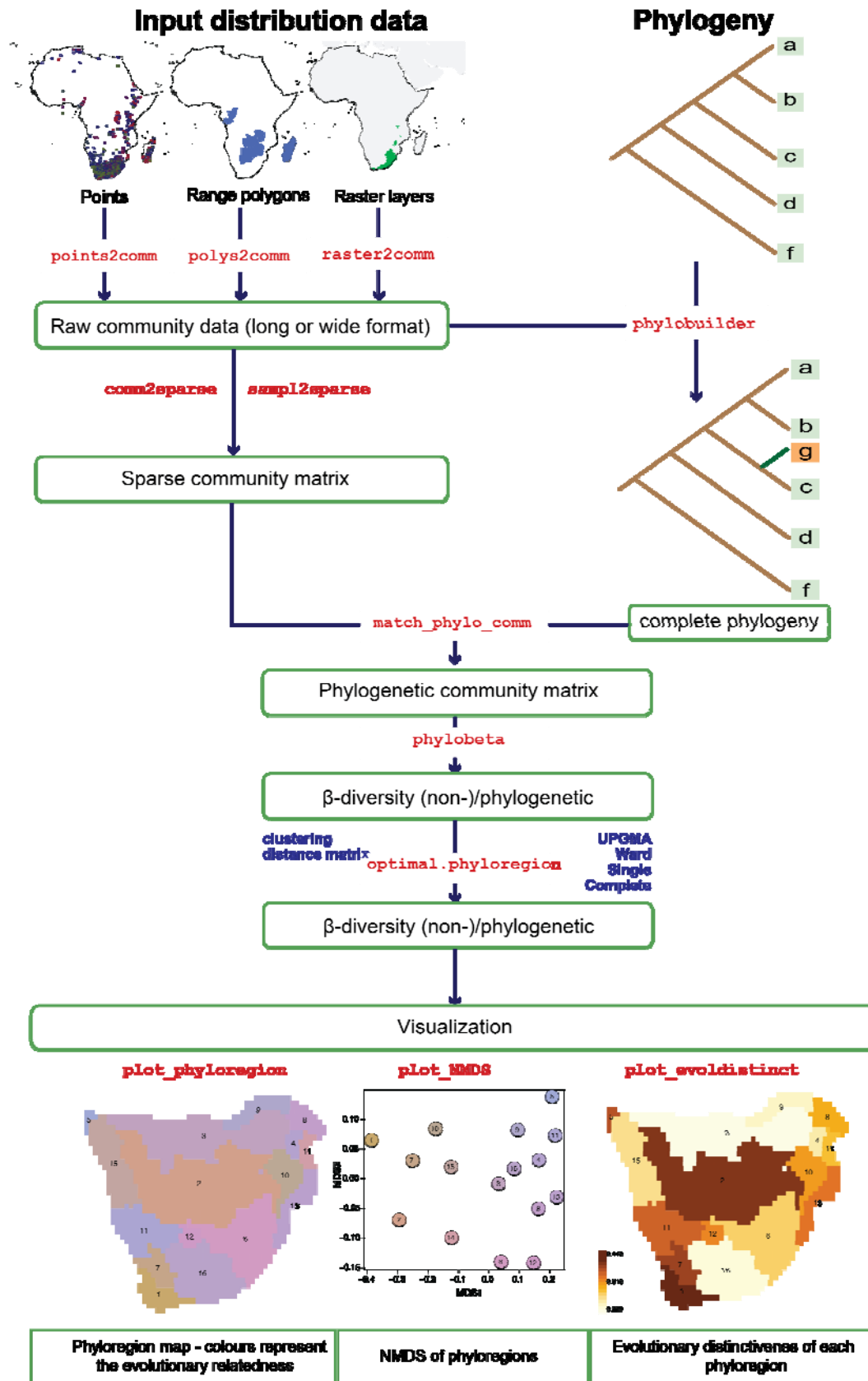
470 Vilhena, D. A., & Antonelli, A. (2015). A network approach for identifying and
471 delimiting biogeographical regions. *Nature Communications*, 6, 6848.

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**




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

485

Community composition data

(a) Long format

```
grids species
g1      s4
g2      s1
g2      s2
g3      s3
g4      s1
g4      s2
```

```
samp12sparse()

comm2sparse()
```

(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"
```

486

487 **Fig. 2.** Illustration showing community data conversion to sparse community matrix by (a)

488 `samp12sparse` function when the raw data is in long community data format, or (b)

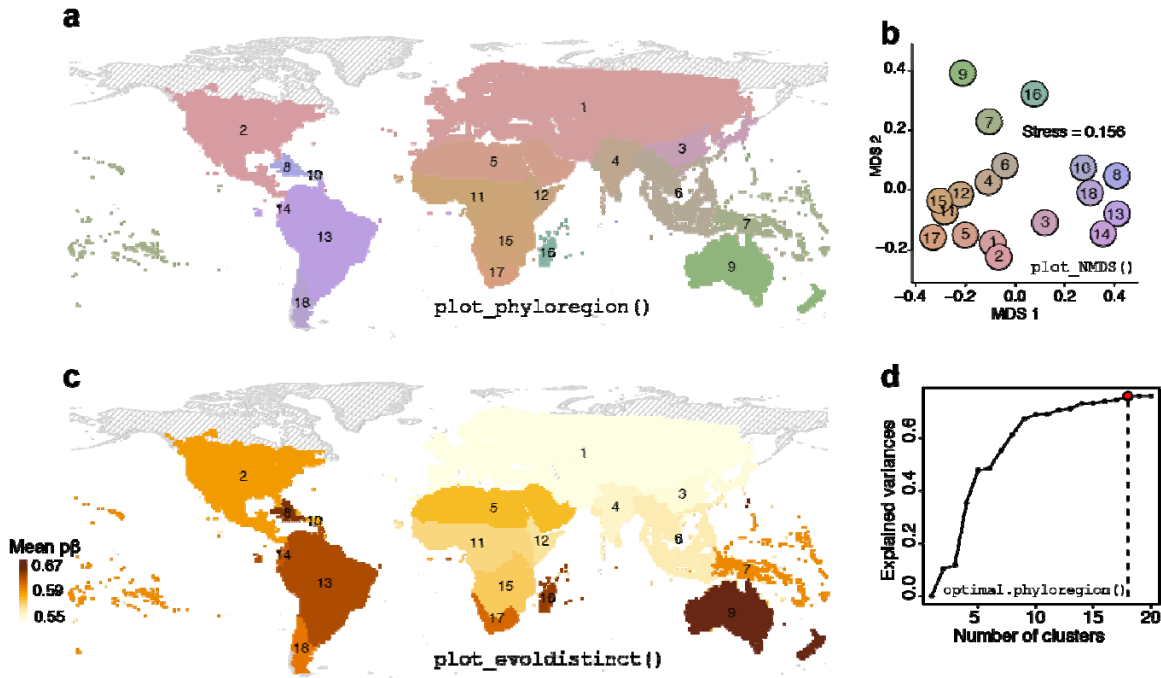
489 `com2sparse` for wide community data format. The result is (c) a sparse community matrix for

490 downstream analysis.

491

492

493

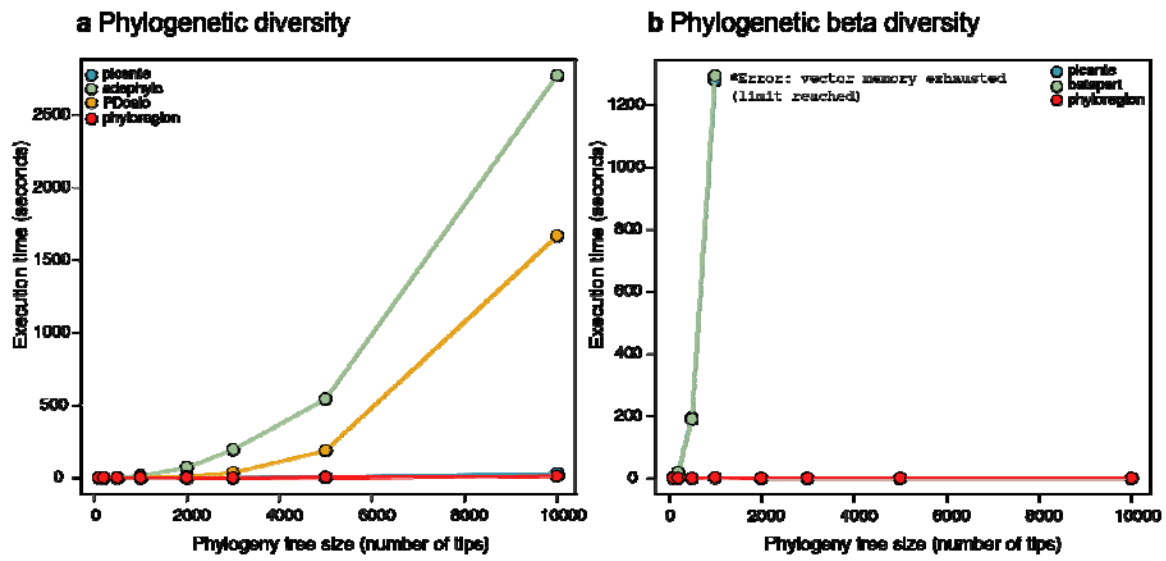


494

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



504

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

507