A peer-reviewed version of this preprint was published in PeerJ on 30 March 2016.

MonoPhy: A simple R package to find and visualize monophyly issues

Orlando Schwery, Brian C O'Meara

Background. The monophyly of taxa is an important attribute of a phylogenetic tree, as a lack of it may hint at shortcomings of either the tree or the current taxonomy and can misguide subsequent analyses. While monophyly is conceptually simple, it is manually tedious and time consuming to assess on modern phylogenies of hundreds to thousands of species. **Results.** The R package *MonoPhy* allows assessment and exploration of monophyly of taxa in a phylogeny. It can assess the monophyly of genera using the phylogeny only, and with an additional input file any other desired higher taxa or unranked groups can be checked as well. **Conclusion.** Summary tables, easily subsettable results and several visualization options allow quick and convenient exploration of monophyly issues, thus making *MonoPhy* a valuable tool for any researcher working with phylogenies.

1   *MonoPhy*: A simple R package to find and visualize monophyly

2   issues

3   Orlando Schwery and Brian C. O'Meara

4   Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610, USA

5

6   Corresponding Author:

7   Orlando Schwery

8   425a Hesler, University of Tennessee, Knoxville, TN 37996-1610, USA

9   Email address: oschwery@vols.utk.edu

10

Abstract

Background. The monophyly of taxa is an important attribute of a phylogenetic tree, as a lack of it may hint at shortcomings of either the tree or the current taxonomy and can misguide subsequent analyses. While monophyly is conceptually simple, it is manually tedious and time consuming to assess on modern phylogenies of hundreds to thousands of species.

Results. The R package *MonoPhy* allows assessment and exploration of monophyly of taxa in a phylogeny. It can assess the monophyly of genera using the phylogeny only, and with an additional input file any other desired higher taxa or unranked groups can be checked as well.

Conclusion. Summary tables, easily subsettable results and several visualization options allow quick and convenient exploration of monophyly issues, thus making *MonoPhy* a valuable tool for any researcher working with phylogenies.

Introduction

Phylogenetic trees are undoubtedly crucial for most research in ecology or evolutionary biology. Whether one is studying trait evolution (e.g. Coddington 1988; Donoghue 1989), diversification (e.g. Gilinsky & Good 1991; Hey 1992), phylogeography (Avise et al. 1987), or simply relatedness within a group (e.g. Czelusniak et al. 1982; Shochat & Dessauer 1981; Sibley & Ahlquist 1981), bifurcating trees representing hierarchically nested relationships are central to the analysis. Exactly because phylogenies are so fundamental to the inferences we make, we need tools that enable us to examine how reconstructed relationships compare with existing assumptions, particularly taxonomy. We have computational approaches to estimate confidence for parts of a phylogeny (Felsenstein 1985; Larget & Simon 1999) or measuring distance between two phylogenies (Robinson 1971), but assessing agreement of a new phylogeny with existing taxonomy is often done manually. This does not scale to modern phylogenies of hundreds to thousands of taxa. Modern taxonomy seeks to name clades: an ancestor and all of its descendants (the descendants thus form a monophyletic group). Discrepancies between the new phylogenetic hypothesis and the current taxonomic classification may indicate that the phylogeny is wrong or poorly resolved. Alternatively, a well-supported phylogeny that conflicts with currently recognized groups might suggest that the taxonomy should be reformed. To identify such discrepancies, one can simply assess whether the established taxa are monophyletic. A lack of group monophyly signals a potential error that can affect downstream analysis and inference. For example, it will mislead ancestral trait or area reconstruction or introduce false signals when assigning unsampled diversity for diversification analyses (e.g. in *diversitree* (FitzJohn 2012) or BAMM (Rabosky 2014)). In general, a lack of monophyly can blur patterns we might see in the data otherwise.

The R package *MonoPhy* is a quick and user-friendly method for assessing monophyly of taxa in a given phylogeny. While the R package *ape* (Paradis et al. 2004) already contains the helpful function is.monophyletic, which also enables testing for monophyly, the functionality of *MonoPhy* is much broader. Apart from assessing monophyly for all groups and focal taxonomic

49   levels in a tree at once, *MonoPhy* is also not limited to providing a simple 'yes-or-no' output, but
50   rather enables the user to explore underlying causes of non-monophyly. In the following, we
51   outline the structure and usage of the package and provide examples to demonstrate its
52   functionality. For a more usage-focused and application-oriented treatment, one should refer to
53   the tutorial vignette (`vignette("MonoPhyVignette")`), which contains stepwise
54   instructions for the different functions and their options. For any other package details consult
55   the documentation (`help("MonoPhy")`).

Description

57   The package *MonoPhy* is written in R (R Development Core Team 2014, http://www.R-
58   project.org/), an increasingly important language for evolutionary biology. It builds on the
59   existing packages *ape* (Paradis et al. 2004), *phytools* (Revell 2012), *phangorn* (Schliep 2011),
60   *RColorBrewer* (Neuwirth 2014) and *taxize* (Chamberlain & Szocs 2013). A list of the currently
61   implemented commands is given in Table 1. Note that in the code and this paper, we distinguish
62   between tips: the organisms at the tip of the tree, and higher order taxa. Functions with 'taxa'
63   only return information about higher order taxa, not tips. The main function –
64   `AssessMonophyly` – evaluates the monophyly of each higher taxon by identifying the most
65   recent common ancestor (MRCA) of a collection of tips (e.g. all species in a genus), and then
66   returning all descendants of this node. The taxon is monophyletic if the number of its members
67   equals the number of descendants of its MRCA. If there are more descendants than taxon
68   members, the function will identify and list the tips that do not belong to the focal taxon, which
69   we then call 'intruders'.
70   Biologically, identifying a few intruders may suggest that the definition of a group should be
71   expanded; observing some group members in very different parts of the tree than the rest of their
72   taxon may instead suggest that these individuals were misidentified or that their placement is the
73   result of contaminated sequences. We thus implemented an option to specify a cutoff value,
74   which gives the minimal proportion of tips among the descendants of a taxon's MRCA that are
75   actual members of that taxon. If a given group falls below this value, the function will find the
76   'core clade' – a subclade for which the proportion matches or exceeds the cutoff value – and
77   label all taxon members outside of it as 'outliers'. As there is no objective criterion to decide at
78   what point individuals should be considered outliers, a reasonable cutoff value has to be chosen
79   by the user.
80   If the tree's tip labels are in the format '*Genus_speciesepithet*', the genus names will be extracted
81   and used as taxon assignments for the tips. If the tip labels are in another format, or other
82   taxonomic levels should be tested, taxon names can be assigned to the tips using an input file. To
83   avoid having to manually compose a taxonomy file for a taxon-rich phylogeny, *MonoPhy* can
84   automatically download desired taxonomic levels from ITIS or NCBI using *taxize* (Chamberlain
85   & Szocs 2013).
86

87   ***Table 1:*** *Functions of the package MonoPhy.*

Function name	Description
AssessMonophyly	Runs the main analysis to assess monophyly of groups on a tree
GetAncNodes	Returns MRCA nodes for taxa.
GetIntruderTaxa	Returns lists of taxa that cause monophyly issues for another taxon.
GetIntruderTips	Returns lists of tips that cause monophyly issues for a taxon.
GetOutlierTaxa	Returns lists of taxa that have monophyly issues due to outliers.
GetOutlierTips	Returns lists of tips that cause monophyly issues for their taxon by being outliers.
GetResultMonophyly	Returns an extended table of the results
GetSummaryMonophyly	Returns a summary table of the results
PlotMonophyly	Allows several visualizations of the result.

88

89   All inference results are stored in a solution object used for extracting information from by all
90   other commands (e.g. summary tables, intruder and outlier lists) for one or more higher-level
91   taxa of interest. PlotMonophyly reconstructs and plots the monophyly state of the tips using
92   *phytools* (Revell 2012). Apart from the basic monophyly plot (Fig.1), branches can be coloured
93   according to taxonomic groups or to highlight intruders and outliers. Monophyletic groups can
94   be collapsed and plots can be saved directly to PDF to facilitate the visualization of large trees.
95   It is important to remember that the results produced by the package are merely the product of
96   the used phylogeny and the available taxonomic information. It thus only makes the mismatches
97   between those accessible, but does not reveal any more than that. The decision of whether the
98   result suggests problems in the phylogeny or the taxonomy, or whether a tip should be
99   considered a rogue taxon and be removed, is entirely up to the user's judgment.
100   *MonoPhy* is available through CRAN (https://cran.r-project.org/package=MonoPhy/) and is
101   developed on GitHub (https://github.com/oschwery/MonoPhy). Intended extensions and fixes
102   can be seen in the issues list of the package's GitHub page. Among the planned extensions of the
103   package are: allowing trees with polytomies, multiple trees, displaying the result for specific
104   subtrees, proposing monophyletic subgroups, enabling formal tests for monophyly (incorporating
105   clade support) and providing increased plot customizability.

106   Examples

107   Our first example makes use of the example files contained in the package. They come from a
108   phylogeny of the plant family Ericaceae (Schwery et al. (2015) pruned to 77 species; original
109   data see Schwery et al. (2014)) and two taxon files assigning tribes and subfamilies to the tips (in
110   both files, errors have been introduced for demonstration purposes; see code and output for both
111   examples in Supplementary Data). Running the main analysis command AssessMonophyly
112   on genus level (i.e. tree only) and tribe level (i.e. tree plus taxonomy file) using standard settings

113   took 0.045 and 0.093 seconds respectively on a MacBook Pro with 2.4 GHz Intel Core i5 and
114   8GB Ram. We could now use the remaining commands to extract the information of interest
115   from the saved output object (e.g. summary tables, lists of problem taxa, etc.). The basic
116   monophyly plot for the genus level analysis is displayed for a subclade of the tree in Figure 1
117   (the figure of the full tree is shown in Fig. S1).
118   For the second example, we demonstrate the package's performance on a tree of 31,749
119   species of Embriophyta (Zanne et al. 2014; data see Zanne et al. 2013), using an outlier-cutoff of
120   0.9 this time. Just checking monophyly for genera took 1.78 hours, but revealed that 22% of
121   genera on the tree are not monophyletic, while around half of all genera are only represented by
122   one species each. Furthermore, we can see that the largest monophyletic genus is *Iris* (139 tips),
123   that *Justicia* had the most intruders (13 tips) and that *Acacia* produced the most outliers (99 tips).
124   Finally, with 2337 other tips as descendants of their MRCA, the 3 species of *Aldina* are most
125   spread throughout the tree.
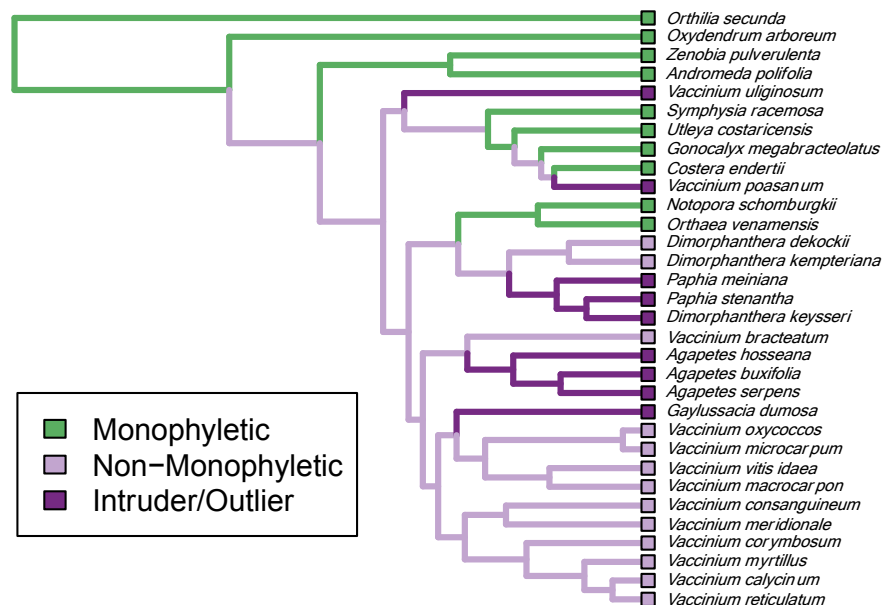
126   **Fig. 1. Monophyly plot of the genera of Ericaceae**. Close-up on subfamily Vaccinioideae only. Branches of the
127   tree coloured according to monophyly status. We can see that *Vaccinium* has two outliers and that its intruders are
128   *Paphia*, *Dimorphanthera*, *Agapetes* and *Gaylussacia*.

Citation

130   Researchers using *MonoPhy* in a published paper should cite this article and indicate the
131   used version of the package. The citation information for the current package version can
132   be obtained using `citation("MonoPhy")`.

133 Acknowledgements

138 References

139 Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, and Saunders NC.
140   1987. Intraspecific Phylogeography - The Mitochondrial-DNA Bridge Between
141   Population-Genetics and Systematics. *Annual Review of Ecology and Systematics*
142   18:489-522.
143 Chamberlain SA, and Szocs E. 2013. taxize: taxonomic search and retrieval in R.
144   *F1000Research* 2:191-191.
145 Coddington JA. 1988. Cladistic Tests Of Adaptational Hypotheses. *Cladistics-the*
146   *International Journal of the Willi Hennig Society* 4:3-22.
147 Czelusniak J, Goodman M, Hewettemmett D, Weiss ML, Venta PJ, and Tashian RE. 1982.
148   Phylogenetic Origins and Adaptive Evolution of Avian and Mammalian Hemoglobin
149   Genes. *Nature* 298:297-300.
150 Donoghue MJ. 1989. Phylogenies and the Analysis of Evolutionary Sequences, with
151   Examples from Seed Plants. *Evolution* 43:1137-1156.
152 Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap.
153   *Evolution* 39:783-791.
154 FitzJohn RG. 2012. Diversitree: comparative phylogenetic analyses of diversification in R.
155   *Methods in Ecology and Evolution* 3:1084-1092.
156 Gilinsky NL, and Good IJ. 1991. Probabilities of Origination, Persistence, and Extinction of
157   Families of Marine Invertebrate Life. *Paleobiology* 17:145-166.
158 Hey J. 1992. Using Phylogenetic Trees to Study Speciation and Extinction. *Evolution* 46:627-
159   640.
160 Larget B, and Simon DL. 1999. Markov chain Monte Carlo algorithms for the Bayesian
161   analysis of phylogenetic trees. *Molecular Biology and Evolution* 16:750-759.
162 Neuwirth E. 2014. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. ed.
163 Paradis E, Claude J, and Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in
164   R language. *Bioinformatics* 20:289-290.
165 R Development Core Team. 2014. R: A language and environment for statistical computing.
166   Vienna, Austria: R Foundation for Statistical Computing.
167 Rabosky DL. 2014. Automatic detection of key innovations, rate shifts, and diversity-
168   dependence on phylogenetic trees. *Plos One* 9:e89543.
169 Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other
170   things). *Methods in Ecology and Evolution* 3:217-223.
171 Robinson DF. 1971. Comparison of labeled trees with valency three. *Journal of*
172   *Combinatorial Theory, Series B* 11:105-119.
173 Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592-593.

174   Schwery O, Onstein RE, Bouchenak-Khelladi Y, Xing Y, Carter RJ, and Linder HP. 2014. Data
175   from: As old as the mountains: the radiations of the Ericaceae. Dryad Data
176   Repository.
177   Schwery O, Onstein RE, Bouchenak-Khelladi Y, Xing Y, Carter RJ, and Linder HP. 2015. As
178   old as the mountains: the radiations of the Ericaceae. *New Phytologist* 207:355-367.
179   Shochat D, and Dessauer HC. 1981. Comparative Immunological Study of Albumins of
180   Anolis Lizards of the Caribbean Islands. *Comparative Biochemistry and Physiology a-*
181   *Physiology* 68:67-73.
182   Sibley CG, and Ahlquist JE. 1981. *The phylogeny and relationships of the ratite birds as*
183   *indicated by DNA-DNA hybridization*.
184   Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara
185   BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ,
186   Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J,
187   Soltis PS, Swenson NG, Warman L, and Beaulieu JM. 2014. Three keys to the
188   radiation of angiosperms into freezing environments. *Nature* 506:89-+.
189   Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara
190   BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ,
191   Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J,
192   Soltis PS, Swenson NG, Warman L, Beaulieu JM, and Ordonez A. 2013. Data from:
193   Three keys to the radiation of angiosperms into freezing environments. Dryad Data
194   Repository.
195