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

View the peer-reviewed version (peerj.com/articles/cs-56), which is the preferred citable publication unless you specifically need to cite this preprint.

Schwery O, O'Meara BC. (2016) *MonoPhy*: a simple R package to find and visualize monophyly issues. PeerJ Computer Science 2:e56 https://doi.org/10.7717/peerj-cs.56

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.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1600v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2015, publ: 19 Dec 2015

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

Orlando Schwery and Brian C. O'Meara

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

Corresponding Author:

Orlando Schwery

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

Email address: oschwery@vols.utk.edu

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1600v1 | CC-BY 4.0 Open Access | rec: 19 Dec 2015, publ: 19 Dec 2015

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

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

Description

The package *MonoPhy* is written in R (R Development Core Team 2014, http://www.R-project.org/), an increasingly important language for evolutionary biology. It builds on the existing packages *ape* (Paradis et al. 2004), *phytools* (Revell 2012), *phangorn* (Schliep 2011), *RColorBrewer* (Neuwirth 2014) and *taxize* (Chamberlain & Szocs 2013). A list of the currently implemented commands is given in Table 1. Note that in the code and this paper, we distinguish between tips: the organisms at the tip of the tree, and higher order taxa. Functions with 'taxa' only return information about higher order taxa, not tips. The main function – `AssessMonophyly` – evaluates the monophyly of each higher taxon by identifying the most recent common ancestor (MRCA) of a collection of tips (e.g. all species in a genus), and then returning all descendants of this node. The taxon is monophyletic if the number of its members equals the number of descendants of its MRCA. If there are more descendants than taxon members, the function will identify and list the tips that do not belong to the focal taxon, which we then call 'intruders'.

Biologically, identifying a few intruders may suggest that the definition of a group should be expanded; observing some group members in very different parts of the tree than the rest of their taxon may instead suggest that these individuals were misidentified or that their placement is the result of contaminated sequences. We thus implemented an option to specify a cutoff value, which gives the minimal proportion of tips among the descendants of a taxon's MRCA that are actual members of that taxon. If a given group falls below this value, the function will find the 'core clade' – a subclade for which the proportion matches or exceeds the cutoff value – and label all taxon members outside of it as 'outliers'. As there is no objective criterion to decide at what point individuals should be considered outliers, a reasonable cutoff value has to be chosen by the user.

If the tree's tip labels are in the format '*Genus_speciesepithet*', the genus names will be extracted and used as taxon assignments for the tips. If the tip labels are in another format, or other taxonomic levels should be tested, taxon names can be assigned to the tips using an input file. To avoid having to manually compose a taxonomy file for a taxon-rich phylogeny, *MonoPhy* can automatically download desired taxonomic levels from ITIS or NCBI using *taxize* (Chamberlain & Szocs 2013).

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.

All inference results are stored in a solution object used for extracting information from by all other commands (e.g. summary tables, intruder and outlier lists) for one or more higher-level taxa of interest. `PlotMonophyly` reconstructs and plots the monophyly state of the tips using *phytools* (Revell 2012). Apart from the basic monophyly plot (Fig.1), branches can be coloured according to taxonomic groups or to highlight intruders and outliers. Monophyletic groups can be collapsed and plots can be saved directly to PDF to facilitate the visualization of large trees.

It is important to remember that the results produced by the package are merely the product of the used phylogeny and the available taxonomic information. It thus only makes the mismatches between those accessible, but does not reveal any more than that. The decision of whether the result suggests problems in the phylogeny or the taxonomy, or whether a tip should be considered a rogue taxon and be removed, is entirely up to the user's judgment.

MonoPhy is available through CRAN (https://cran.r-project.org/package=MonoPhy/) and is developed on GitHub (https://github.com/oschwery/MonoPhy). Intended extensions and fixes can be seen in the issues list of the package's GitHub page. Among the planned extensions of the package are: allowing trees with polytomies, multiple trees, displaying the result for specific subtrees, proposing monophyletic subgroups, enabling formal tests for monophyly (incorporating clade support) and providing increased plot customizability.

Examples

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

took 0.045 and 0.093 seconds respectively on a MacBook Pro with 2.4 GHz Intel Core i5 and 8GB Ram. We could now use the remaining commands to extract the information of interest from the saved output object (e.g. summary tables, lists of problem taxa, etc.). The basic monophyly plot for the genus level analysis is displayed for a subclade of the tree in Figure 1 (the figure of the full tree is shown in Fig. S1).

For the second example, we demonstrate the package's performance on a tree of 31,749 species of Embriophyta (Zanne et al. 2014; data see Zanne et al. 2013), using an outlier-cutoff of 0.9 this time. Just checking monophyly for genera took 1.78 hours, but revealed that 22% of genera on the tree are not monophyletic, while around half of all genera are only represented by one species each. Furthermore, we can see that the largest monophyletic genus is *Iris* (139 tips), that *Justicia* had the most intruders (13 tips) and that *Acacia* produced the most outliers (99 tips). Finally, with 2337 other tips as descendants of their MRCA, the 3 species of *Aldina* are most spread throughout the tree.

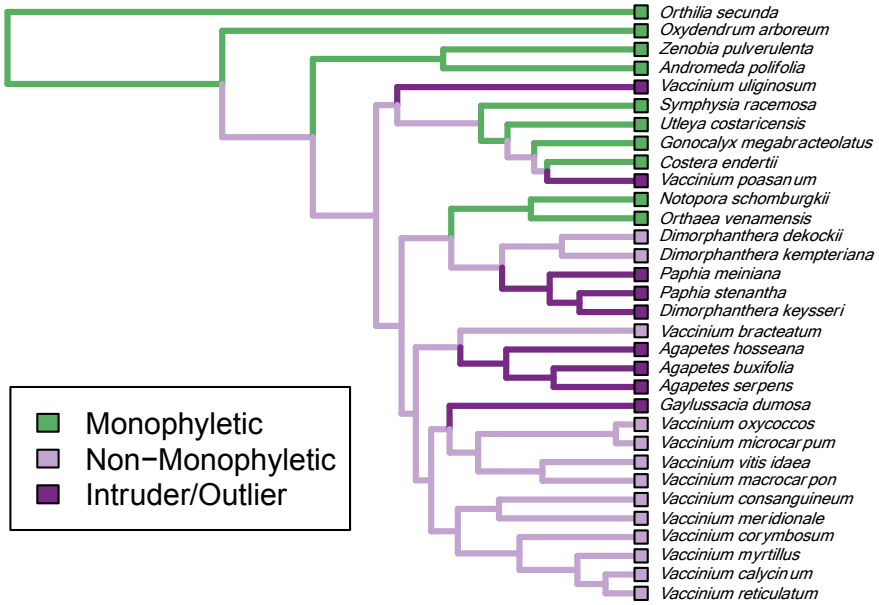

Fig. 1. Monophyly plot of the genera of Ericaceae. Close-up on subfamily Vaccinioideae only. Branches of the tree coloured according to monophyly status. We can see that *Vaccinium* has two outliers and that its intruders are *Paphia*, *Dimorphanthera*, *Agapetes* and *Gaylussacia*.

Citation

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

Acknowledgements

We want to thank the members of the O'Meara lab for helpful discussions, Brian Looney and Sam Borstein for beta testing, and the members of the Tank lab, Arne Mooers, Karen Cranston, Bruce Cochrane and Daniel Gates for great ideas on increasing the usefulness of this package.

References

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, and Saunders NC. 1987. Intraspecific Phylogeography - The Mitochondrial-DNA Bridge Between Population-Genetics and Systematics. *Annual Review of Ecology and Systematics* 18:489-522.

Chamberlain SA, and Szocs E. 2013. taxize: taxonomic search and retrieval in R. *F1000Research* 2:191-191.

Coddington JA. 1988. Cladistic Tests Of Adaptational Hypotheses. *Cladistics-the International Journal of the Willi Hennig Society* 4:3-22.

Czelusniak J, Goodman M, Hewettemmett D, Weiss ML, Venta PJ, and Tashian RE. 1982. Phylogenetic Origins and Adaptive Evolution of Avian and Mammalian Hemoglobin Genes. *Nature* 298:297-300.

Donoghue MJ. 1989. Phylogenies and the Analysis of Evolutionary Sequences, with Examples from Seed Plants. *Evolution* 43:1137-1156.

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* 39:783-791.

FitzJohn RG. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution* 3:1084-1092.

Gilinsky NL, and Good IJ. 1991. Probabilities of Origination, Persistence, and Extinction of Families of Marine Invertebrate Life. *Paleobiology* 17:145-166.

Hey J. 1992. Using Phylogenetic Trees to Study Speciation and Extinction. *Evolution* 46:627-640.

Larget B, and Simon DL. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16:750-759.

Neuwirth E. 2014. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. ed.

Paradis E, Claude J, and Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289-290.

R Development Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rabosky DL. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *Plos One* 9:e89543.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3:217-223.

Robinson DF. 1971. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory, Series B* 11:105-119.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592-593.

Schwery O, Onstein RE, Bouchenak-Khelladi Y, Xing Y, Carter RJ, and Linder HP. 2014. Data from: As old as the mountains: the radiations of the Ericaceae. Dryad Data Repository.

Schwery O, Onstein RE, Bouchenak-Khelladi Y, Xing Y, Carter RJ, and Linder HP. 2015. As old as the mountains: the radiations of the Ericaceae. *New Phytologist* 207:355-367.

Shochat D, and Dessauer HC. 1981. Comparative Immunological Study of Albumins of Anolis Lizards of the Caribbean Islands. *Comparative Biochemistry and Physiology a-Physiology* 68:67-73.

Sibley CG, and Ahlquist JE. 1981. *The phylogeny and relationships of the ratite birds as indicated by DNA-DNA hybridization*.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ, Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J, Soltis PS, Swenson NG, Warman L, and Beaulieu JM. 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506:89-+.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ, Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J, Soltis PS, Swenson NG, Warman L, Beaulieu JM, and Ordonez A. 2013. Data from: Three keys to the radiation of angiosperms into freezing environments. Dryad Data Repository.