FishPhyloMaker: An R package to generate phylogenies for ray-finned fishes

Authors: Gabriel Nakamura[1, 2, *], Aline Richter[1], Bruno E. Soares[3]

1 – Universidade Federal do Rio Grande do Sul, Departamento de Ecologia, Bento Gonçalves Avenue, 9500, CP

2 – INCT Ecology, Evolution, and Biodiversity Conservation

3 – Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação em Ecologia

*correspondence author: gabriel.nakamura.souza@gmail.com

Abstract

1 – Phylogenies summarize information for evolutionary and ecological studies. They allow investigating hypotheses from trait evolution to the relationship between evolutionary diversity and ecosystem functioning. However, obtaining a phylogenetic hypothesis for local fish assemblages can be difficult, hindering studies involving this group.

2 – We developed the FishPhyloMaker R package to facilitate the obtention of phylogenetic information for ray-finned fishes. FishPhyloMaker automates the insertion procedure of species in the most comprehensive phylogeny of ray-finned fishes following their taxonomic positions.

3 – The main functions of the FishPhyloMaker package, FishTaxaMaker() and FishPhyloMaker(), assess the validity of species names and generate dated phylogenies for a local pool of species, respectively.

4 – FishPhyloMaker facilitates the generation of phylogenetic trees through a reliable and reproducible way for the most diversified group of vertebrates. The package adopts well-known rules of insertion, which will expand the range of evolutionary and ecological questions that can be addressed using ray-finned fishes as study models.

Key-words: Community phylogenetics, phylogenetic tools, gap-analysis, Darwinian shortfall, Actinopterygii

Introduction

Phylogenies have been widely explored in ecology in the last decades due to the development of theoretical frameworks, numerical methods, and software (*e.g.,* Webb et al. 2008; Felsenstein 1985). The research agenda in ecology and evolution encompasses phylogenetic approaches from organismal to macroecological-scale, including trait evolution, invasion ecology, metacommunity ecology, and ecosystem functioning (Cavender-Bares et al., 2009). Hence, comprehensive phylogenetic trees must be available to address those topics.

Well-established phylogenies for most of known species are available for some groups, such as birds (Jetz et al., 2012) and plants (Magallón et al., 2015). Inversely, available phylogenies for bony fishes (Betancur et al., 2017; Rabosky et al., 2018) display issues related to the taxonomic position of some clades (e.g., non-monophyletic groups) and the lack of species representativeness. The latter issue hampers answering some questions on the ecology and evolution of bony fishes by generating inaccuracy to estimates of phylogenetic signal, trait evolution, and phylogenetic diversity (Seger et al. 2010; Boettiger et al., 2012a), or even impeding their calculation.

A short-term solution to tackle the Darwinian shortfall for ray-finned fishes (*i.e.*, the lack of phylogenetic information for species) would be coupling the phylogenetic information with cladistic classification to produce more comprehensive phylogenies (Diniz-Filho et al., 2013). This solution is laborious and lacks reproducibility when adding many species manually, and the specific steps are not precisely documented (Webb et al., 2008). On the other hand, molecular techniques generate comprehensive phylogenies but demand high expertise and financial costs (Roquet et al., 2013). Therefore, automatizing the procedures of constructing synthesis phylogenies "by hand" (Webb & Donoghue, 2005) provides a more reliable and short-term solution for evolutionary ecologists.

Ray-finned fishes (Actinopterygii) exhibit a complex evolutionary history and high ecological diversity (Albert et al., 2020), making them an interesting group to address questions in the interface of ecology and evolution (*e.g.*, Roa-Fuentes et al. 2019; Nakamura et al. 2020). Nonetheless, studies addressing those questions are scarce compared to clades that present specific tools to build local phylogenies (*e.g.,* Webb & Donoghue 2005 for mammals and plants; Jin & Qian 2019 for plants). This scenario suggests that the difficulty in obtaining phylogenetic information can hinder our efforts to understand fish ecology and evolution. Additionally, our knowledge about the Darwinian shortfalls for fishes is restricted to few lineages (*e.g.*, Freitas et al., 2021), which impedes the mapping of the relative demand of additional efforts in specific regions or clades.

Here, we present the package *FishPhyloMaker*, a tool in the R environment that automatizes the construction of phylogenetic trees for ray-finned fishes in local or regional pools of species. Our package overcomes the main problems associated with manually building phylogenies for ray-finned fishes by following a specific and documented procedure and reducing the manual labor in large phylogenies.

Inside the Fish(PhyloMaker): an overview of the package

FishPhyloMaker is a freely-available R package containing two main functions, FishTaxaMaker() and FishPhyloMaker(). Below, we describe these two functions highlighting the input data, intermediate steps, and output objects. Brief descriptions of the package functions are available in Table 1.

FishTaxaMaker()

FishTaxaMaker() checks the validity of species names provided by the user and prepares a formatted data frame to be used in the FishPhyloMaker() function.

The input data must be a string vector or a data frame containing a list of species from the regional pool or an occurrence matrix (sites x species). The genus and specific epithet (or subspecies) must be separated by underline (e.g., *Genus_epithet*). The function first classifies the provided species names as valid or synonymies based on Fishbase (Froese & Pauly, 2006) by using the *rfishbase* package (Boettiger et al., 2012b). A new column summarizes names initially valid and the current valid names substituting identified synonymies. Unknown species to Fishbase are printed in the command line, and the user must manually inform the Family and the Order of the species. The output of the function is a list containing three elements: 1) a data frame displaying the taxonomic information (Valid name, Subfamily, Family, Order, Class, and SuperClass) for each provided species; 2) a data frame displaying the taxonomic information (Species, Family, and Order) only for the valid species; 3) a character vector displaying the species names not found in Fishbase.

Table 1: Functions presented in the package FishPhyloMaker and their descriptions.

Function	Description
FishTaxaMaker()	Checks species names according to Fishbase and prepares the species list for the other functions in the package.
whichFishAdd()	Identifies the species already included in the mega-tree and in which taxonomic level each remaining species will be inserted.
FishPhyloMaker()	Builds the phylogeny and may return a data frame identifying step-by-step the performed insertions.
Darwinian_deficit()	Calculates the Darwinian shortfall for the provided species list through a Phylogenetic

Diversity (PD) ratio:

$$^{PD_{inserted}}\Big/_{PD_{inserted} + PD_{present\ in\ tree}}$$

FishPhyloMaker()

The function builds a phylogenetic hypothesis for the provided species list by inserting in and pruning species from the Rabosky's et al. (2020) phylogenetic tree (Figure 1) downloaded by the fishtreeoflife R package (Chang et al. 2019). This phylogeny is the most up-to-date and comprehensive phylogenetic hypothesis for ray-finned fishes.

The input for FishPhyloMaker() can be the second element in the list returned by FishTaxaMaker() or a manually constructed data frame with the same configuration (species, family, and order names for each taxon). The function contains three logical arguments: insert.base.node, return.insertions and progress.bar. These three arguments are set by default as FALSE, TRUE, and TRUE, respectively.

The function identifies which of the provided species are in the backbone phylogenetic tree. If all of them are already present in the backbone tree, the function returns a pruned one. If any of the provided species is not in the backbone tree, the function performs a four-level insertion routine. First, species from genera already included in the backbone tree are inserted as polytomies at the most recent ancestral node that links all congeneric species (or as the sister species of the only species representing a genus in the backbone tree, as shown in i in Figure 1). Second, species not inserted in the previous step are then inserted at the family level by an interactive procedure using a returned list of all the genera within the same family of the target species. The user has the option to insert the target species as a sister taxon to a genus (ii in Figure 1, option 1, *Loricaria* genus), between two genera (ii in Figure 1, option 2, between *Loricaria* and *Hypostomus*), or at the node of the family (ii in Figure 1, option 3). If the user enters a single genus from the list, the function splits its branch and inserts the target species as a sister taxon (option 1). If the user enters two genera separated by a blank space, the function inserts the target species as a polytomy at the most recent node that links the selected genera (option 2). If the user enters the family name, the function inserts the target species at the family node as a polytomy (option 3). Third, if any remaining species can now be inserted at the genus level, the function repeats the first procedure but records it as a Congeneric family-level insertion (iii in Figure 1). Fourth, remnant species are inserted at the order level following similar to the second step by an interactive procedure using a returned list of all the families within the order of the target species. Hence, the user may specify a family to insert the target species as sister taxon (option 1), two families to insert it as a polytomy at the most recent node linking them (option 2), or the order to insert it as a sister taxon (option 3). The function will not perform insertions steps beyond the order level because it would add too much uncertainty to the phylogenetic tree.

Setting the argument insert.base.node as TRUE automatically inserts the target species from the second and fourth steps in the family and order nodes, respectively. This setting facilitates the insertion of a large number of species or species with the phylogenetic position unknown. The default output is a list with two objects: (i) the pruned tree including only the provided species list (Final tree in Figure 1); (ii) a data frame identifying if each provided species was initially present in the backbone tree, in which step it was inserted, or not inserted at all.

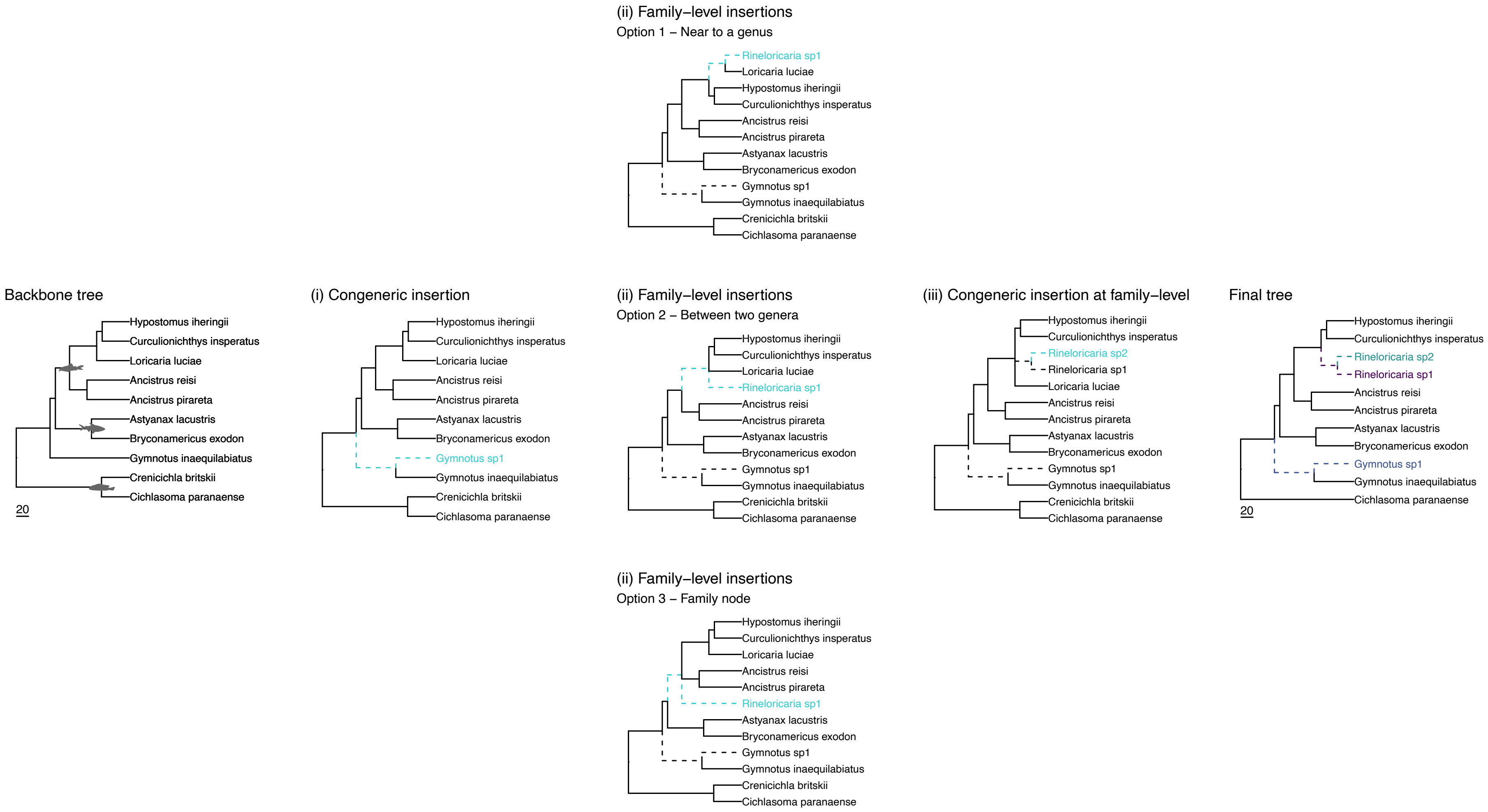

Figure 1: Schematic representation of insertion and subsetting procedure performed by the FishPhyloMaker() function. Here we used a hypothetical phylogeny containing ten species and four families (silhouettes inside the tree) as the backbone phylogeny. Step (i) represents the

congeneric level of insertion. Step (ii) represents the three options that the user may choose in the Family-level round of insertions (Option 1 – near to a genus; Option 2 – between two genera; Option 3 – at the family node). (iii) represents the congeneric insertions at the family level, and, finally, the final pruned tree containing only the species of interest.

Example analysis

We provide an example of the usage of the FishPhyloMaker package by creating a phylogenetic tree for a global dataset of freshwater fishes inhabiting four ecoregions: Afrotropic, Indo-Malay, Nearctic, and Neotropic (Tedesco et al., 2017). This dataset encompasses extensive occurrence data for freshwater fishes and allowed in-depth investigation on the global patterns of species distribution and their evolutionary determinants (*e.g.*, Miller & Román-Palácios, 2021).

We prepared the occurrence data using the function FishTaxaMaker(). The occurrence matrix encompassed 2478 species, from which 2477 were valid. We applied the FishPhyloMaker() function separately for ecoregions, thus building one phylogenetic tree for each (Figure 2). For simplicity, we set the argument insert.base.node as TRUE. The entire insertion procedure lasted approximately two hours using one core from a machine with an i5 processor. A total of 821 species were inserted, with the Afrotropics exhibiting the largest number of insertions (359 from 767).

Coding to reproduce this phylogenetic tree is provided at GitHub (GabrielNakamura/MS_FishPhyloMaker). Further exploratory analysis and a guide to package installation is available at https://gabrielnakamura.github.io/FishPhyloMaker/.

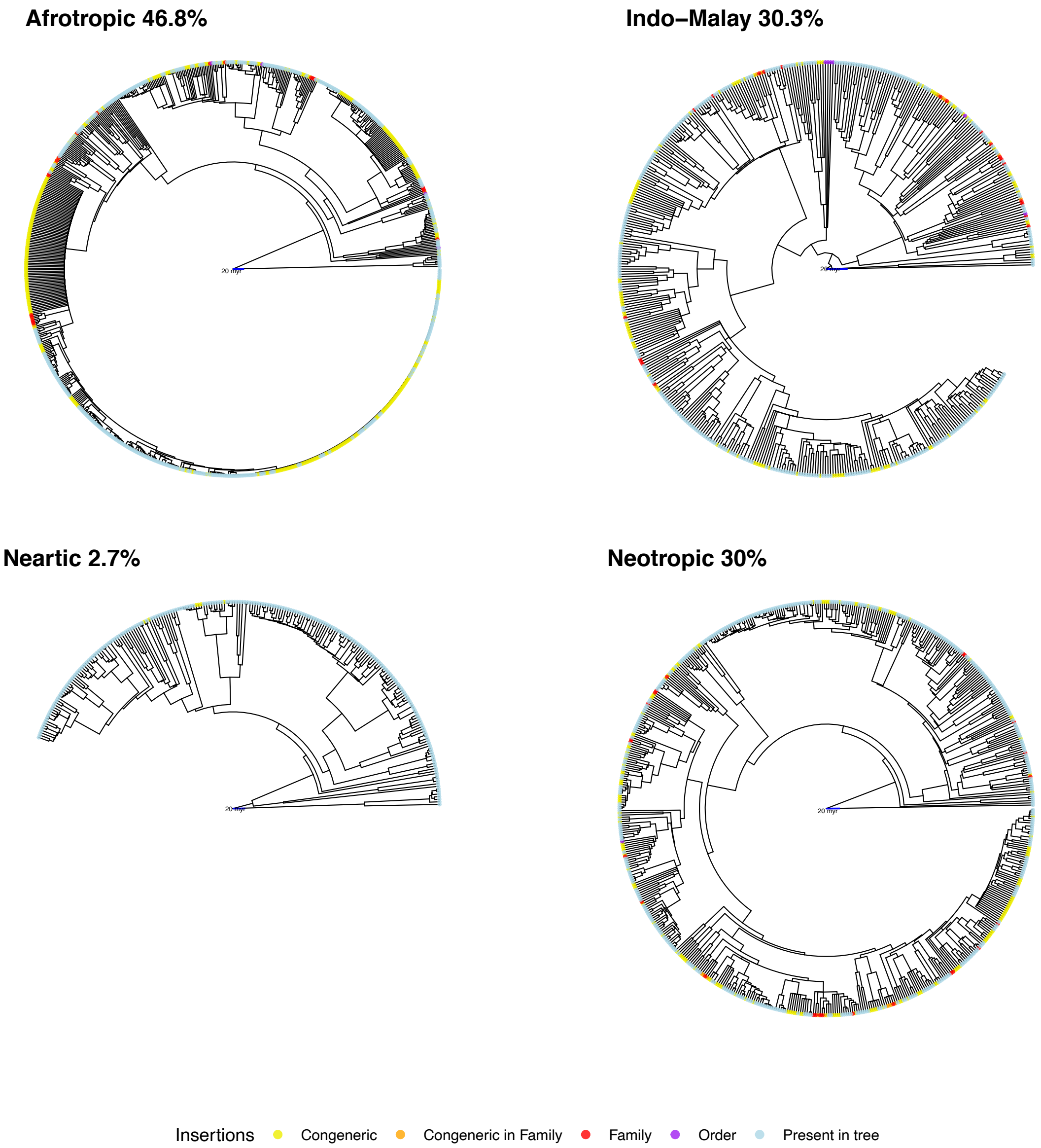

Figure 2: Phylogenetic trees generated with the FishPhyloMaker package for freshwater fishes inhabiting four ecoregions (Afrotropics, Indo-Malay, Nearctic, and Neotropic). The colored tip-points indicate if species were present in the backbone tree (Present in tree) or at which level they were inserted. The percentages of insertions over the total number of species for each ecoregion are shown.

Similarity and advances in relation with other approaches

We provided a user-friendly and reproducible way to construct a phylogenetic tree for a megadiverse group (Actinopterygii). The FishPhyloMaker package is in line with tools developed for plants, such as Phylomatic (C++ application) and V.PhyloMaker (R package) (Webb & Donoghue, 2005; Jin & Qian, 2019), but includes different features, such as new insertion options and records of performed insertions.

Limitations and possible applications

Future developments of the package should consider the Catalog of Fishes (Fricke & Eschemeyer, 2021) to improve the nomenclature checking procedures. Despite Fishbase being a widely used database to check for the taxonomic classification of fishes, it may present delays in updating taxonomic information because it is not its primary purpose. Inversely, the Catalog of Fishes is an authoritative taxonomic list frequently updated.

An inherent limitation of the phylogenetic hypothesis produced by FishPhyloMaker is the large number of polytomies resulting from the insertion procedures. We recommend that users directly assess how the phylogenetic uncertainty affects further analysis when not using a fully solved phylogenetic tree (Martins et al., 2013).

These limitations do not preclude the package applicability for studies in phylogenetic community ecology since synthesis phylogenies do not significantly impact phylogenetic diversity indices (Li et al., 2019). Moreover, this is the only automated tool able to provide a complete phylogenetic tree that can easily handle large datasets. FishPhyloMaker can be relevant for addressing several critical questions in ecology and evolution by facilitating the obtention of phylogenetic hypotheses for local pools of ray-finned fishes. This facilitation can be essential for regions with a large gap in the phylogenetic knowledge of fishes, such as the Neotropical region (Albert et al., 2020). Such phylogenetic hypotheses allow understanding how ecological traits evolved or how the current and past environmental conditions selected the lineages in different areas. At larger scales, biogeographical studies are usually restricted to one or a few lineages due to the availability of molecular phylogenies. The FishPhyloMaker package facilitates large-scale investigations on the biogeographic history of the most diverse group of vertebrates on Earth, the Actinopterygians, helping us understand the processes that drive this high diversity. Finally, we can map where the lack of phylogenetic information is the most critical once the function returns the insertion-level of species. This information can directly elucidate the patterns of the Darwinian shortfalls for ray-finned fishes. Therefore, we expect that the FishPhyloMaker package reduces the gaps and barriers to addressing ecological and evolutionary questions due to the difficulty or lack of a reliable phylogenetic hypothesis for local and regional pools of ray-finned fishes.

Acknowledgements

GN is member of the National Institutes for Science and Technology (INCT) in Ecology, Evolution and Biodiversity Conservation, supported by MCTIC/CNPq (proc. 465610/2014-5). BES and AR are grateful to FAPERJ and CAPES for their postdoctoral and doctoral grants, respectively. The Brazilian science resists.

Author´s contribution

GN design the study, GN and AR created the R package, GN, BES and AR wrote the manuscript.

Data availability

Data available from the Zenodo Digital Repository https://zenodo.org/record/4739032#.YJUyEGZKjoA

References

Albert, J. S., Tagliacollo, V. A., & Dagosta, F. (2020). Diversification of Neotropical Freshwater Fishes. *Annual Review of Ecology, Evolution, and Systematics*, *51*(1), 27–53. doi:10.1146/annurev-ecolsys-011620-031032

Betancur, R. R., Wiley, E. O., Arratia, G., Acero, A., Bailly, N., Miya, M., … Ortí, G. (2017). Phylogenetic classification of bony fishes. *BMC Evolutionary Biology*, *17*(1), 1–40. doi:10.1186/s12862-017-0958-3

Boettiger, C., Coop, G., & Ralph, P. (2012a). Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*, *66*(7), 2240–2251. doi:10.1111/j.1558-5646.2011.01574.x

Boettiger, C., Lang, D. T., & Wainwright, P. C. (2012b). rfishbase: exploring, manipulating and visualizing FishBase data from R. *Journal of Fish Biology*, *81*(6), 2030–2039. doi: 10.1111/j.1095-8649.2012.03464.x

Cavender-Bares, J., Kozak, K. H., Fine, P. V. a, & Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, *12*, 693–715. doi:10.1111/j.1461-0248.2009.01314.x

Chang, J., Rabosky, D. L., Smith, S. A., & Alfaro, M. E. (2019). An r package and online resource for macroevolutionary studies using the ray-finned fish tree of life. *Methods in Ecology and Evolution*, *10*(7), 1118–1124. doi:10.1111/2041-210X.13182

Diniz-Filho, J. A. F., Loyola, R. D., Raia, P., Mooers, A. O., & Bini, L. M. (2013). Darwinian shortfalls in biodiversity conservation. *Trends in Ecology and Evolution*, *28*(12), 689–695. doi:10.1016/j.tree.2013.09.003

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1–15. doi:0003-0147/85/2501-0001

Freitas, T. M. S., Stropp, J., Calegari, B. B., Calatayud, J., De Marco, P., Montag, L. F. de A.,

& Hortal, J. (2021). Quantifying shortfalls in the knowledge on Neotropical Auchenipteridae fishes. *Fish and Fisheries*, *22*(1), 87–104. doi:10.1111/faf.12507

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature*, *491*(7424), 444–448. doi:10.1038/nature11631

Jin, Y., & Qian, H. (2019). V.PhyloMaker: an R package that can generate very large phylogenies for vascular plants. *Ecography*, *42*(8), 1353–1359. doi:10.1111/ecog.04434

Li, D., Trotta, L., Marx, H. E., Allen, J. M., Sun, M., Soltis, D. E., … Baiser, B. (2019). For common community phylogenetic analyses, go ahead and use synthesis phylogenies. *Ecology*, *100*(9), 1–15. doi:10.1002/ecy.2788

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L., & Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist*, *207*(2), 437–453. doi:10.1111/nph.13264

Martins, W. S., Carmo, W. C., Longo, H. J., Rosa, T. C., & Rangel, T. F. (2013). SUNPLIN: Simulation with Uncertainty for Phylogenetic Investigations. *BMC Bioinformatics*, *14*(1). doi:10.1186/1471-2105-14-324

Nakamura, G., Vicentin, W., Súarez, Y. R., & Duarte, L. (2020). A multifaceted approach to analyzing taxonomic, functional, and phylogenetic β-diversity. *Ecology*. doi:10.1002/ecy.3122

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., … Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, *559*(7714), 392–395. doi:10.1038/s41586-018-0273-1

Roa-Fuentes, C. A., Heino, J., Cianciaruso, M. V., Ferraz, S., Zeni, J. O., & Casatti, L. (2019). Taxonomic, functional, and phylogenetic β-diversity patterns of stream fish assemblages in tropical agroecosystems. *Freshwater Biology*, *64*(3), 447–460.

doi:10.1111/fwb.13233

Roquet, C., Thuiller, W., & Lavergne, S. (2013). Building megaphylogenies for macroecology: Taking up the challenge. *Ecography*, *36*(1), 13–26. doi:10.1111/j.1600-0587.2012.07773.x

Tedesco, P. A., Beauchard, O., Bigorne, R., Blanchet, S., Buisson, L., Conti, L., … Oberdorff, T. (2017). Data Descriptor: A global database on freshwater fish species occurrence in drainage basins. *Scientific Data*, *4*, 1–6. doi:10.1038/sdata.2017.141

Webb, C. O., Ackerly, D. D., & Kembel, S. W. (2008). Phylocom: Software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, *24*(18), 2098–2100. doi:10.1093/bioinformatics/btn358

Webb, C. O., & Donoghue, M. J. (2005). Phylomatic: Tree assembly for applied phylogenetics. *Molecular Ecology Notes*, *5*(1), 181–183. doi:10.1111/j.1471-8286.2004.00829.x