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.

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

Highlights

- We provide the first automated procedure to check species names, construct phylogenetic trees and calculate Darwinian shortfalls for ray-finned fishes (Actinopterygii) by the R package FishPhyloMaker.
- This package provides functions to assemble phylogenies through a fast, reliable, and reproducible method, allowing its use and replicability by specialists and non-specialists in fish systematics.
- The package also provides an interactive procedure that gives more flexibility to the user when compared with other existing tools that construct phylogenetic trees for other highly speciose groups.
- The package includes a new method to compute Darwinian shortfalls for ray-finned fishes, but the rationale of the provided algorithm can be extended in future studies to be used in other groups of organisms

Abstract

Phylogenies summarize evolutionary information that is essential in the investigation of ecological and evolutionary causes of diversity patterns. They allow investigating hypotheses from trait evolution to the relationship between evolutionary diversity and ecosystem functioning. However, obtaining a comprehensive phylogenetic hypothesis can be difficult for some groups, especially those with a high number of species, that is the case for fishes, particularly tropical ones. The lack of species in phylogenetic hypotheses, called Darwinian shortfalls, can hinder ecological and evolutionary studies involving this group. To tackle this problem, we developed FishPhyloMaker, an R package that facilitates the generation of phylogenetic trees through a reliable and reproducible procedure, even for a large number of species. The package adopts well-known rules of insertion based on cladistic hierarchy, allowing its use by specialists and non-specialists in fish systematics. We tested the reliability of our algorithm in maintaining important properties of phylogenetic distances running a sensitivity analysis. We also exemplified the use of the FishPhyloMaker package by constructing complete phylogenies for fishes inhabiting the four richest freshwater ecoregions of the world. Furthermore, we proposed a new method to calculate Darwinian shortfalls and mapped this information for the major freshwater drainages of the world. FishPhyloMaker will expand the range of evolutionary and ecological questions that can be addressed using ray-finned fishes as study models, mainly in the field of community phylogenetics, by providing an easy and reliable way to obtain comprehensive phylogenies. Further, FishPhyloMaker presents the potential to be extended to other taxonomic groups that suffer from the same difficulty in the obtention of comprehensive phylogenetic hypothesis.

Keywords: Phylogenies, community phylogenetics, Darwinian shortfall, gap-analysis.

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. Large phylogenies were primarily developed by combining source-trees and published-trees (the supertree approach), by concatenating different data matrices of systematic phylogenetic characters to generate a single tree (the supermatrix approach), or by a mix of both approaches (Haeseler, 2012; Smith et al., 2009).

Well-established phylogenies for most of the known species are available for some groups, such as terrestrial vertebrates (birds (Jetz et al., 2012), mammals (Upham et al., 2019), amphibians (Jetz and Pyron, 2018), squamates (Tonini et al., 2016), sharks (Stein et al., 2018), and plants (Magallón et al., 2015), which also have powerful tools to generate phylogenetic trees for local/regional pools of species (*e.g.*, Webb & Donoghue 2005 for mammals and plants; Jin & Qian 2019 for plants, to the others see http://vertlife.org/phylosubsets/). 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 ray-finned fishes by generating inaccuracy in estimates of phylogenetic signal, trait evolution, and phylogenetic diversity (Seger et al., 2013; Boettiger et al., 2012a), or even impeding their calculation.

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). The difficulty in obtaining phylogenetic information can hinder our efforts to understand fish ecology and evolution. Additionally, the lack of phylogenetic information for species, *i.e.*, Darwinian shortfalls, is currently investigated in a few lineages (*e.g.*, Freitas et al., 2021), which impedes the mapping of the relative demand of additional efforts needed in entire regions or clades to uncover the phylogenetic history of fishes. This problem urges a rapid solution in the context of the accelerated loss of species (Chase et al., 2020).

A short-term solution to tackle the Darwinian shortfall for ray-finned fishes would be coupling the phylogenetic information with cladistic classification to produce comprehensive phylogenies (Diniz-Filho et al., 2013). This solution is laborious and lacks reproducibility when adding many species, and the specific steps are not precisely documented when did "by hand" procedures (Webb et al., 2008). An alternative would be using molecular techniques to generate comprehensive phylogenies (e.g. Pie et al., 2021). However, it demands high expertise and high financial investment (Roquet et al., 2013), limiting factors for several institutions. Therefore, automatizing the procedures of constructing comprehensive phylogenies using the information from cladistic hierarchy, as suggested by Diniz-Filho et al (2013), provides a more reliable, accessible, and short-term solution for evolutionary ecologists. The technique produces reliable phylogenetic information for community phylogenetics (Li et al., 2019).

In order to tackle the problem of obtention of comprehensive fish phylogenies in a reliable and reproducible way, we developed the FishPhyloMaker. This freely available R package facilitates the obtention of phylogenetic trees for ray-finned fishes. FishPhyloMaker automates the insertion procedure of species in the most comprehensive phylogeny (Rabosky et al., 2018) of ray-finned fishes following their taxonomic hierarchy. We illustrated how the FishPhyloMaker package solves the problem of obtaining comprehensive phylogenies by constructing phylogenetic trees for species inhabiting more than 3000 freshwater basins globally (Tedesco et al., 2017). Further, we developed a new method to quantify the Darwinian shortfalls, which we illustrate by mapping the Darwinian shortfalls for the abovementioned basins. Finally, we performed a sensitivity analysis to evaluate how our method preserves characteristics of the phylogenetic tree (pairwise distances among species and evolutionary distinctiveness), even with a varying number of inserted taxa. 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.

Methods

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

A stable version of FishPhyloMaker can be downloaded from the CRAN repository (https://cran.r-project.org/web/packages/FishPhyloMaker/index.html), and a development version is available at the GitHub repository (https://github.com/GabrielNakamura/FishPhyloMaker). All analyses shown here were performed using the development version of FishPhyloMaker.

FishPhyloMaker is a freely available R package containing three main functions, *FishTaxaMaker*, *FishPhyloMaker*, and *PD_deficit*. Below, we describe the functions to generate phylogenetic trees, highlighting the input data, intermediate steps, and output objects. Brief descriptions of the package functions are available in Table 1.

FishTaxaMaker

The *FishTaxaMaker* function checks the validity of species names provided by the user and prepares a formatted data frame for 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, 2000) 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 of these species. If the user types a Family not recognized in the FishBase, the user is asked to check the spelling and type the Order of this family. 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 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 backbone 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.
PD_deficit()	Calculates the Darwinian shortfall for the

	provided species list through a Phylogenetic Diversity (PD Faith (1992)) ratio

FishPhyloMaker

FishPhyloMaker is the core function of the package. This function builds a phylogenetic tree for the provided species list by inserting in and pruning species from the Rabosky et al., (2018) 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 the *FishPhyloMaker* function can be the second element in the list returned by the *FishTaxaMaker* (Taxon_data_FishPhyloMaker) function or a manually constructed data frame with the same configuration (species, family, and order names for each taxon). The function also 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, and allow the user to choose if the species must be at the base node of families/orders, if the insertions made by each species must be shown in the output and if a progress bar must be shown in the console.

The function works sequentially, first identifying which of the provided species are in the backbone phylogenetic tree (Rabosky et al., 2018). 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. In the case of *i* in Figure 1 the branch length is divided at half of its length and the species is inserted. 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, near to *Loricaria* genus), between two genera (*ii* in Figure 1, option 2, between genus *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 as a sister taxon of this genus (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 attaches 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 by splitting the branch length of the congeneric species at half of its length (*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 option facilitates the insertion of a large number of species or species with the unknown phylogenetic position. 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. This data frame will flag each species with one of the six classification based on the insertion procedure: 1 – Present in tree will indicate species that were already present in the backbone tree; 2 – Congeneric insertion will indicate species that present at least one species of the same genus in the backbone tree and was inserted as congeneric of this species; 3 – Family insertion will indicate inserted species that did not present any congeneric species at backbone tree, but had at least one species of the same family in backbone tree; 4 – Congeneric at Family-level will indicate species that was added as congeneric after another species of the same genus was inserted at the Family level; 5 – Order insertion will indicate inserted species that did not presented any species of the same family in the backbone tree and must be inserted near to an extant family or in node corresponding to the order root in the backbone tree; 6 – not inserted will indicate species that did not present any species of the same order in the backbone tree, therefore was not inserted due their high uncertainty in the phylogenetic position.

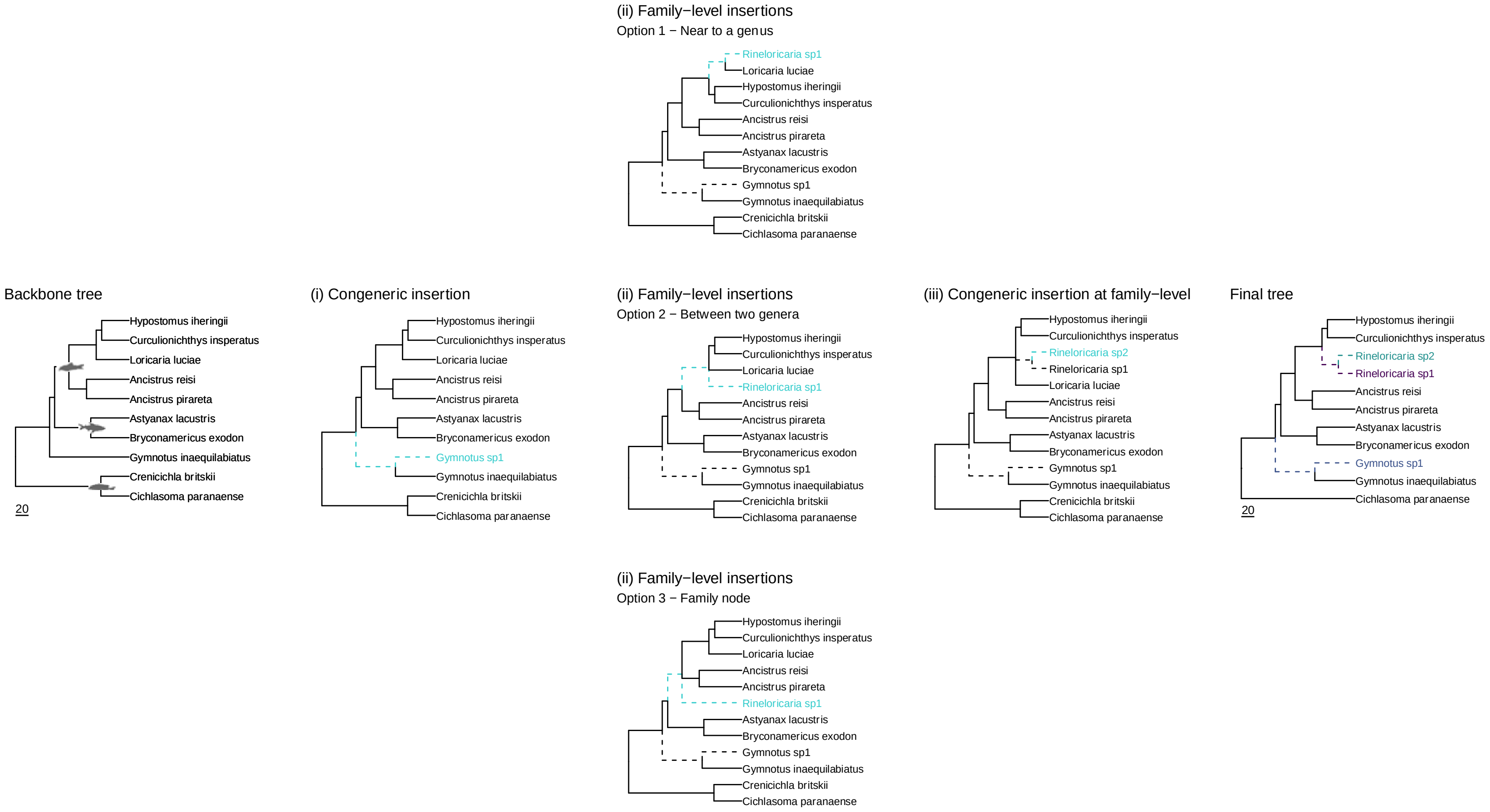

Figure 1: Schematic representation of insertion and subsetting procedure performed by the FishPhyloMaker() function. Here we used a hypothetical phylogeny containing ten species and four families (silhouettes inside the tree) as the backbone phylogeny. Step (i) represents the congeneric level of insertion.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

PD_deficit

The *PD_deficit* function calculates a measure of Darwinian shortfalls following Equation 1:

$$PD_{inserted} \Big/ PD_{inserted} + PD_{present\ in\ tree} \qquad (1)$$

In this function, PD_{inserted} is the sum of the branch lengths of species in the phylogenetic tree before the insertion procedure. $PD_{\text{present in tree}}$ is the sum of branch lengths of the species inserted in the tree. Therefore, the Darwinian deficit ranges from 0 (all species already present in the backbone tree before the insertion procedure) to 1 (all the species in the phylogenetic tree were inserted and were not presented in backbone phylogeny). PD_deficit function returns a vector with three values, the Darwinian shortfall (Equation 1), the total phylogenetic diversity calculated as the sum of branch lengths of the tree (PD_{total}) with all species provided by the user, the sum of branch lengths inserted ($PD_{inserted}$) in the tree and that was already present in the backbone tree ($PD_{present\ in\ tree}$). It is worth noting that the sum of $PD_{inserted}$ and $PD_{present}$ are complementary, summing up to PD_{total}. To calculate the Darwinian shortfall through the *PD_deficit* function, the user must provide a phylogenetic tree and a table of insertions, both obtained from the *FishPhyloMaker* function.

Sensitivity analysis

We performed a sensitivity analysis to assess how the insertion procedure implemented herein and the amount of inserted species affect two characteristics of phylogenetic trees: the mean pairwise distance among species and the phylogenetic distinctiveness.

We 1) randomly change the name of a subsample of species within Rabosky's phylogeny. Then, 2) we built a phylogeny for the species sampled with changed names in the previous step using the FishPhyloMaker function. Finally, we computed: 3) the matrix correlation (Pearson correlation) between the cophenetic distances of the subsampled species in Rabosky's phylogeny and the FishPhyloMaker phylogeny; and 4) the Pearson correlation between the phylogenetic distinctness values for the Rabosky's and FishPhyloMaker phylogenies. The evolutionary distinctness was calculated as the equal splits measure that is the sum of the contribution of all branches of a given lineage divided among its daughter branches (Redding and Mooers, 2006). Evolutionary distinctness measure was calculated using the phyloregion package (Daru et al., 2020).

The abovementioned steps (1, 2 and 3) were repeated 100 times for eleven different quantities (10%, 15%, 20%, 25%, 30%, 35%, 40%, 50%, 55%, 60%) of subsampled species from Rabosky's phylogeny and inserted by the FishPhyloMaker function.

Illustrating the use of FishPhyloMaker package

We provide an example of the usage of the *FishPhyloMaker* package by creating a phylogenetic tree using a global dataset of freshwater fishes inhabiting 3,119 freshwater drainage basins that cover more than 80% of the Earth surface and 14886 species (Tedesco et al., 2017). This dataset allowed in-depth investigation on the global patterns of species distribution and their evolutionary determinants (*e.g.*, Miller & Román-Palácios, 2021). We built a phylogenetic for all species presented in Tedesco´s et al. dataset and mapped all the insertions realized. Moreover, we used this same dataset to demonstrate how to map Darwinian shortfalls, calculated following Equation 1 through *PD_deficit* function for all the drainage basins in the Tedesco et al. (2017) dataset. All the analyses were performed using the development version of the FishPhyloMaker package, which can be downloaded using the following command line:

```
devtools::install_github("GabrielNakamura/FishPhyloMaker", ref = "main",
build_vignettes = TRUE)
```

We recommend that the user updates all the requested packages to avoid errors related to packages versions. We first prepared the fish occurrence by checking the validity of its names by using the function *FishTaxaMaker*. The occurrence matrix encompassed 14,886 species, from which 13,992 were valid names. The remaining 961 names were substituted by their corresponding valid names according to FishBase. We applied the *FishPhyloMaker* function to build a phylogenetic tree containing all the 14,886 species with valid names retrieved from *FishTaxaMaker* (Figure 2). For simplicity and reproducibility, we set the argument `insert.base.node` as TRUE, thus, inserting all species at the base node of its corresponding family and order when needed. We also set the argument `return.insertions = TRUE` for retrieving the insertion information of each species. Then, we applied the *PD_deficit* function to calculate the Darwinian shortfall for all the freshwater basins of the world harboring at least two species (Tedesco et al. 2017). The *PD_deficit* function was calculated considering congeneric insertions and insertions at the family level, however, the function may also include other levels of phylogenetic insertion, like order insertions. All the codes need to fully reproduce these analyses are provided at the GitHub repository (GabrielNakamura/MS_FishPhyloMaker). For further explanations and examples illustrating the usage of functions in the FishPhyloMaker package, the user can assess the package website https://gabrielnakamura.github.io/FishPhyloMaker/index.html and see the Articles section.

Results

The entire insertion procedure lasted approximately three hours using one core from a computer machine with an i5 processor. A total of 11,569 species were inserted, 6,418 species were already present in the backbone phylogeny, and 181 were not inserted at all, resulting in a phylogenetic tree containing 14,705 species (Figure 2). We also showed in Figure 2 all the insertions realized through the FishPhyloMaker function and the seven orders of ray-finned fishes with that present the highest number of species. We can see in Figure two that the insertions are evenly distributed throughout the phylogenetic tree.

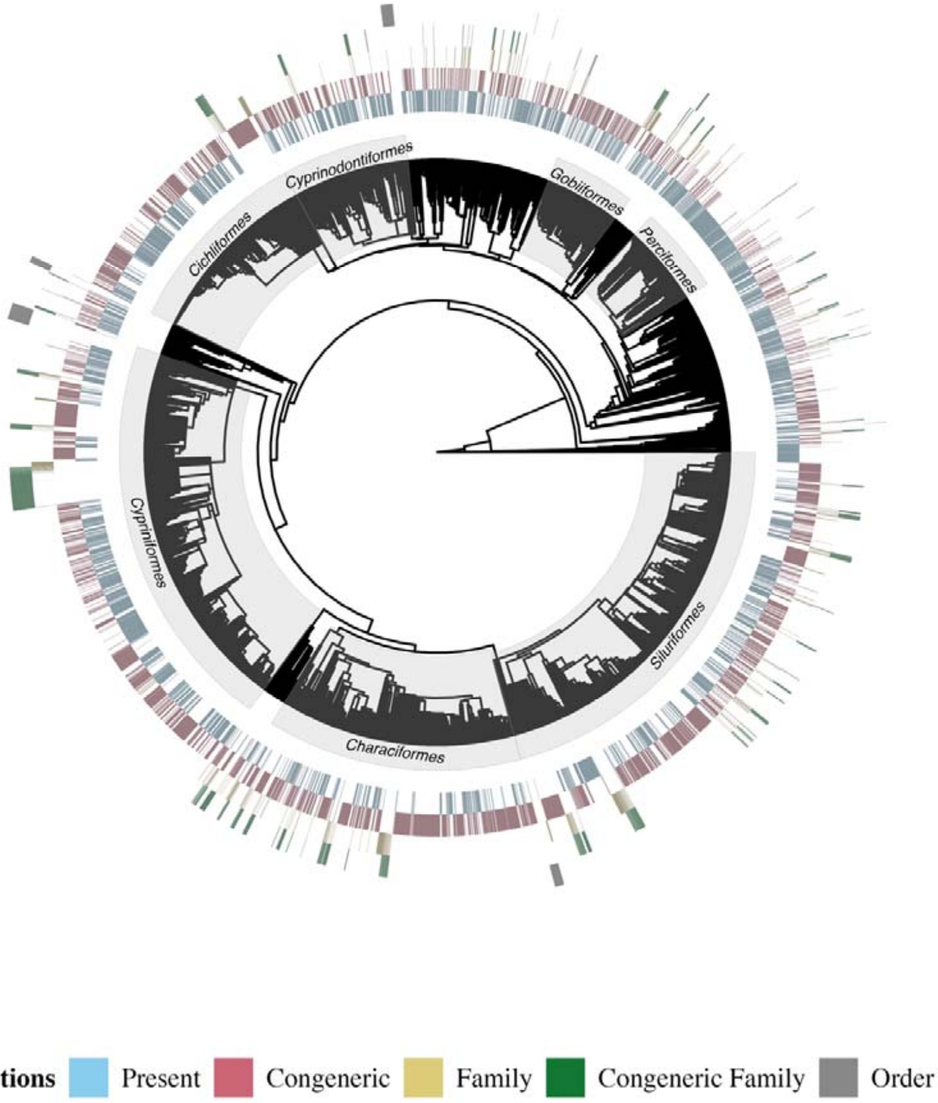

Figure 2: Phylogenetic tree obtained from FishPhyloMaker, containing 14,705 finned-ray species with their respective insertions. We also highlight in the gray rectangles the seven most speciose Orders.

We also depicted all the insertions made by FishPhyloMaker for all freshwater Ecoregion of the world. This was only possible because FishPhyloMaker flags all the insertions made during the insertion procedure. Figure 3 shows that Neotropics and Afrotropics regions exhibited the largest number of species inserted. On the contrary, despite the great area and number of basins, the Nearctic Ecoregion presented the smallest percentage of insertions, most of them congeneric. All Ecoregions and the percentage of species insertions per level are shown in Figure 3.

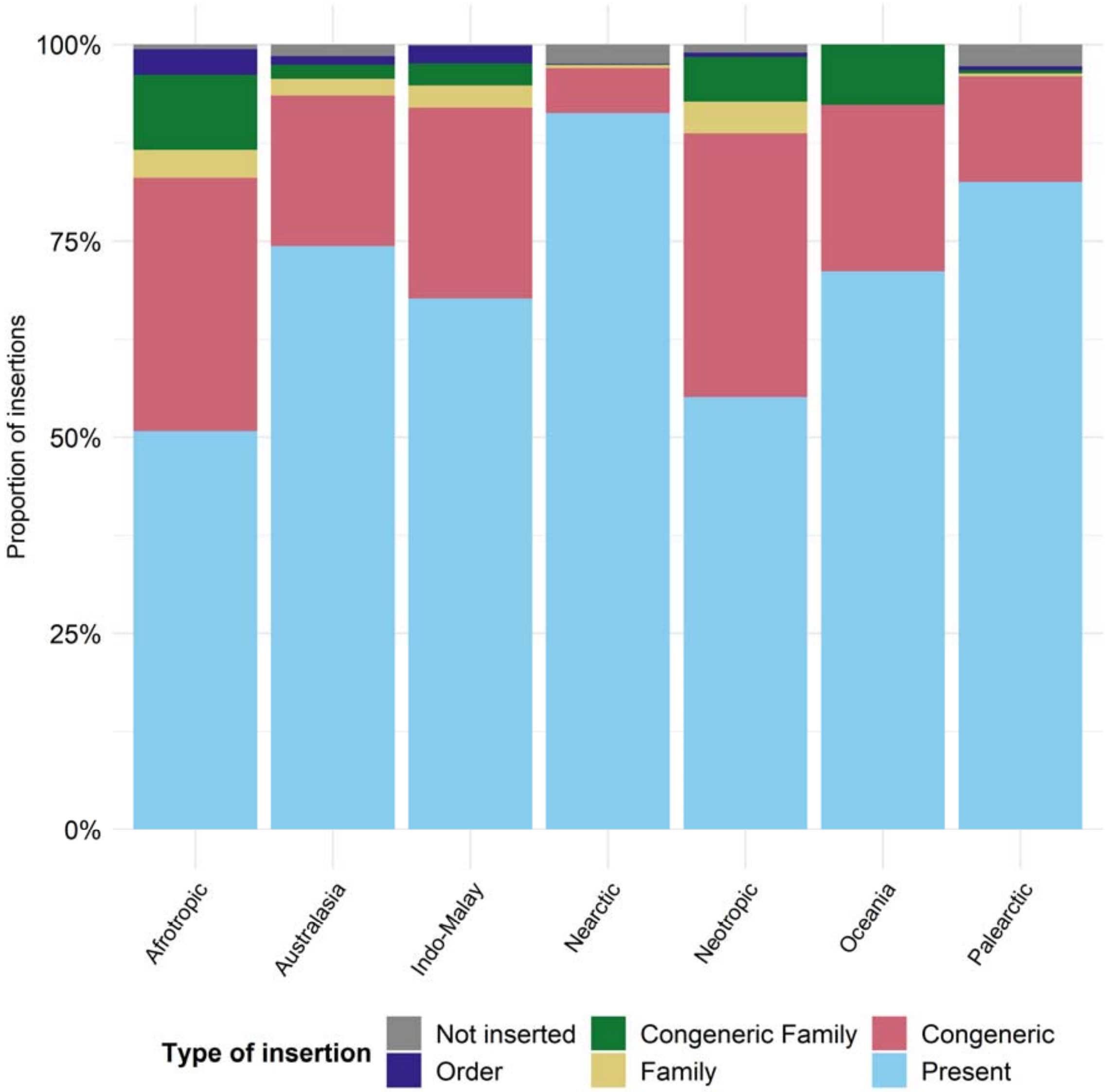

Figure 3: Barplot showing the percentage of species inserted in each one of the seven freshwater ecoregions of the world and their respective type of insertions mapped by FishPhyloMaker package.

We spatialized the Darwinian shortfalls per basin and observed that tropical regions exhibited the highest shortfalls, while northern sites had the lowest (Figure 4). The highest values of Darwinian shortfalls were found in Afrotropics and Neotropics, as some drainages did not harbor any (or only a few) species in the Rabosky's phylogeny. The grey areas correspond to sites that do not present species occurrences accordingly to Tedesco et al. (2017) or presented less than two occurrences for the Order considered. We also depicted the

Darwinian shortfalls for the four major orders in terms of species richness (bottom maps in Figure 4). For all the groups, the highest values of Darwinian shortfalls were found in the neotropical region, except for Cypriniformes, the group responsible for the highest values of Darwinian shortfalls in the watersheds in Asia and some basins in North America.

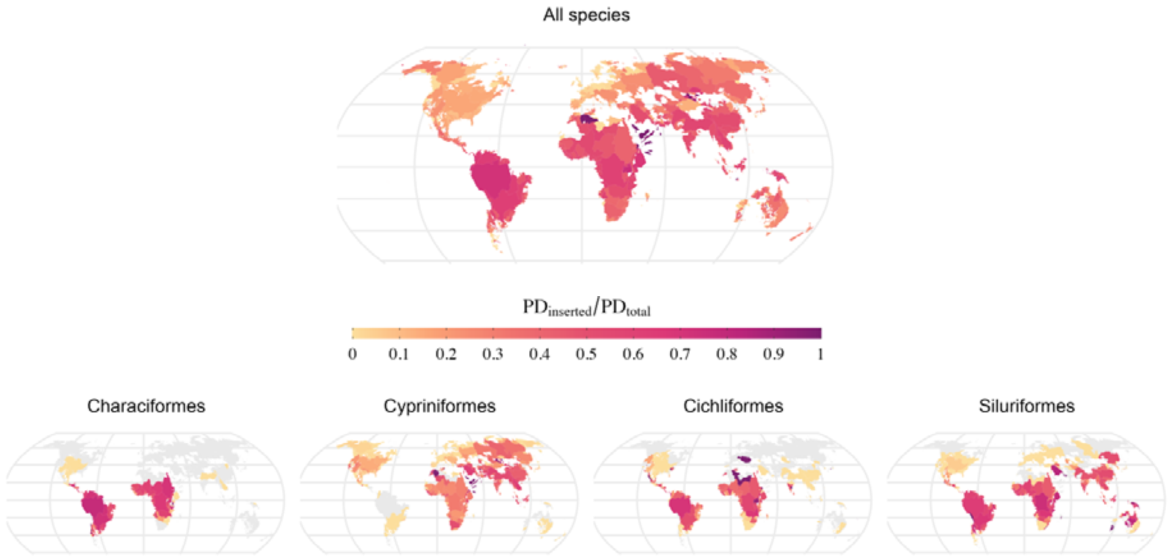

Figure 4: Global distribution of the Darwinian Shortfalls for ray-finned fishes, based on freshwater species occurrences in more than 3000 basins. Values near to 1 indicate a high Darwinian shortfall (a large number of congeneric insertions), while values near zero indicate low shortfalls. We depicted the Darwinian shortfall for the four major orders in terms of species richness (Characiformes, Cypriniformes, Cichliformes, and Siluriformes). Gray color indicates areas with no occurrence of species for a given order.

The sensitivity analysis highlights the strong correlation between the cophenetic distances of Rabosky's and FishPhyloMaker phylogenies (Figure 5 B) even in varying levels of taxa insertions. Inversely, an increasing number of insertions on Rabosky's phylogeny reduced the correlation between phylogenetic distinctness in the original phylogeny and that assembled by FishPhyloMaker (Figure 5 A).

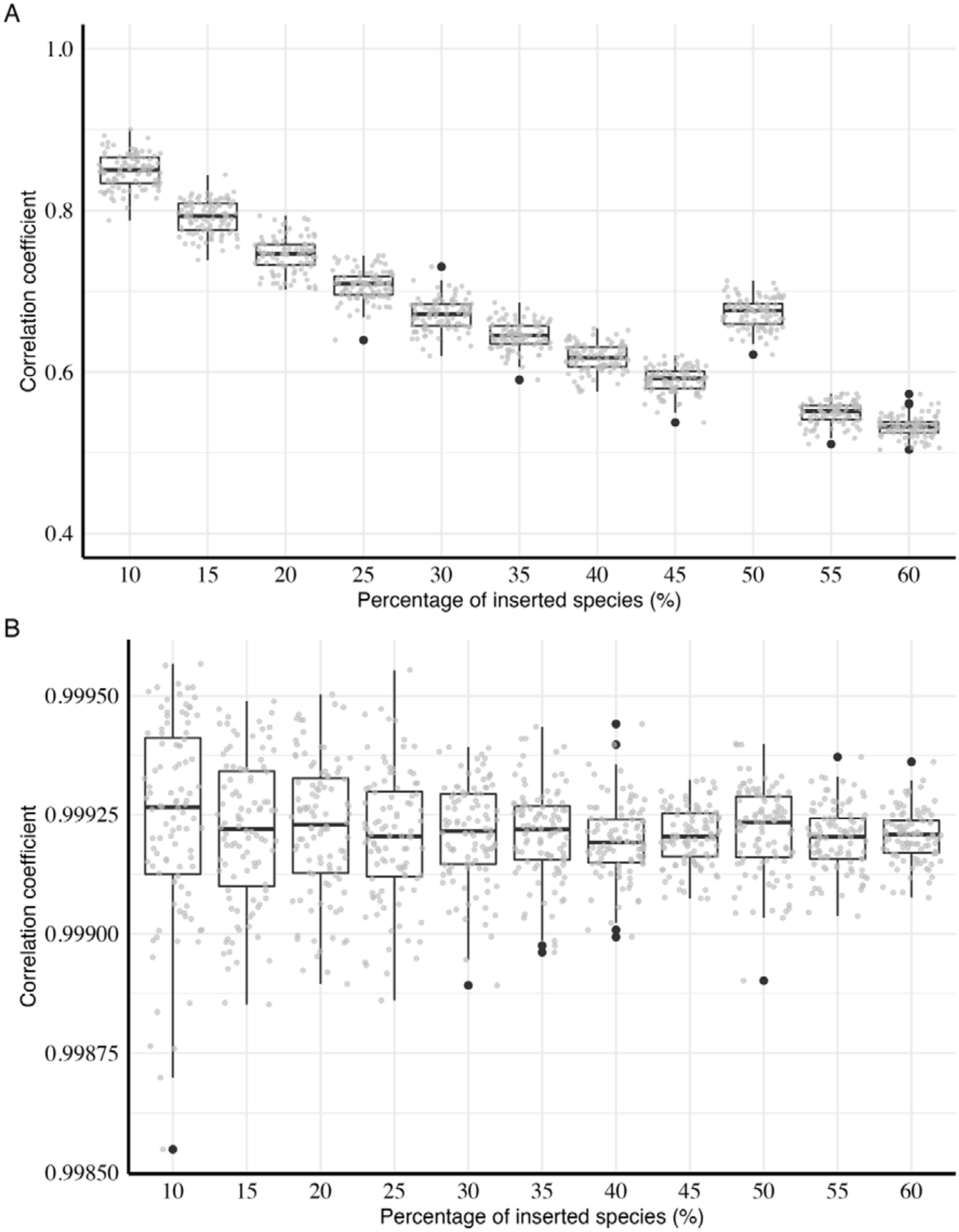

Figure 5: Barplots showing the correlation of evolutionary distinctness values (A) and between cophenetic distances (B) calculated from original phylogeny and inserted phylogenies with varying percentages of species inserted in the original phylogeny. Grey dots

represent individual correlation values. The lower and upper hinges in boxplots represent the first and third quantiles while the middle hinge represents the median.

Discussion

We provided a user-friendly, fast, reliable, and reproducible way to construct phylogenetic trees 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) (Jin and Qian, 2019; Webb and Donoghue, 2005), but includes different features. These features include new options for inserting species through an interactive procedure in phylogenies and recording insertions. The latter feature allows a better systematization of building supertrees and calculating the first, to our knowledge, quantitative measure of the Darwinian shortfall.

Whereas Phylomatic allows the insertion of absent species only as congeneric or at the node corresponding to the family of the focal species (Webb and Donoghue, 2005), the FishPhyloMaker package delivers options through an interactive procedure of insertion. The performed insertions can be easily recorded in an R script, providing flexibility and the same level of reproducibility as other algorithms designed for similar purposes (*e.g.*, Jin and Qian, 2019). This interactive option is a novelty when compared to similar insertion algorithms (*e.g.*, Phylomatic).

The spatial distribution of the Darwinian shortfall is paramount to guide our future efforts to understand the history of life. The phylogenetic gaps in the knowledge of ray-finned fishes are geographically biased, with tropical basins presenting higher Darwinian shortfalls levels, as evidenced in this study. This gap in evolutionary knowledge could lead to a bias in evaluating the effects of evolutionary history and the interpretation of macroecological patterns for fish assemblages in these regions, which can affect conservation decisions based on the phylogenetic dimension of diversity (Assis, 2018).

Several biological and sociological factors can explain the observed bias in Darwinian shortfalls. First, the regions exhibiting the most significant Darwinian gaps also exhibit the largest freshwater fish diversity, which we can not describe at the same speed as less biologically rich areas (Hortal et al., 2015). Second, on-ground accessibility, human occupation, and economic development constrain investments in biodiversity research (Moura et al., 2018; Moura and Jetz, 2021), which is probably more pronounced in tropical regions than temperate ones, which may hamper field sampling and phylogenetic analyses.

Despite being more simple when compared with other insertion methods (e.g., Pearse and Purvis, 2013), FishPhyloMaker provided reliable results by preserving important characteristics of the phylogenetic tree, as we showed through the sensitivity analysis. Commonly used measures of phylogenetic diversity are based on the pairwise distance of species from a phylogenetic tree (e.g., Kraft et al., 2007; Webb et al., 2002), and we showed that the algorithm implemented in FishPhyloMaker successfully preserve the distances among species in the phylogenetic tree even for a great number of insertions.

Limitations and possible applications

Future developments of the package should consider the Catalog of Fishes (van der Laan et al., 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). Furthermore, we recommend caution in the use of FishPhyloMaker phylogenies to compute measures that depend on speciation events (e.g., evolutionary distinctiveness and other split-based metrics) since the insertion procedure modifies the split events in the tree as shown in the sensitivity analysis.

These limitations do not preclude the package applicability for studies in phylogenetic community ecology since synthesis phylogenies do not significantly impact phylogenetic diversity indices as showed by previous studies (Li et al., 2019) and confirmed in ours (through sensitivity analysis). 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.

Biogeographical studies are usually restricted to one or a few lineages at larger scales due to the availability of molecular phylogenies (e.g. García-Andrade et al., 2021) or with phylogenies with a considerable number of absent species (Miller, 2021). 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, contributing not only to direct sampling and studying efforts but also to evidence the need for increased efforts to decolonize science (Trisos et al., 2021). 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.

Contributions

GN Conceptualization; Data curation; Formal Analysis; Methodology; Software; Writing – original draft. AR Data curation; Methodology; Software, Writing – review, and editing. BES Writing – original draft; Methodology.

Acknowledgments

GN is a 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 authors also thank valuable suggestions made by LDS Duarte and other two anonymous Reviewers. Brazilian science resists.

References

Albert, J.S., Tagliacollo, V.A., Dagosta, F., 2020. Diversification of Neotropical Freshwater Fishes. Annu. Rev. Ecol. Evol. Syst. 51, 27–53. https://doi.org/10.1146/annurev-ecolsys-011620-031032

Assis, L.C.S., 2018. Revisiting the Darwinian shortfall in biodiversity conservation. Biodivers. Conserv. 27, 2859–2875. https://doi.org/10.1007/s10531-018-1573-3

Betancur, R.R., Wiley, E.O., Arratia, G., Acero, A., Bailly, N., Miya, M., Lecointre, G., Ortí, G., 2017. Phylogenetic classification of bony fishes. BMC Evol. Biol. 17, 1–40. https://doi.org/10.1186/s12862-017-0958-3

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

Cavender-Bares, J., Kozak, K.H., Fine, P.V.A., Kembel, S.W., 2009. The merging of community ecology and phylogenetic biology. Ecol. Lett. 12, 693–715. https://doi.org/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 Ecol. Evol. 10, 1118–1124. https://doi.org/10.1111/2041-210X.13182

Chase, J.M., Blowes, S.A., Knight, T.M., Gerstner, K., May, F., 2020. Ecosystem decay exacerbates biodiversity loss with habitat loss. Nature 584, 238–243. https://doi.org/10.1038/s41586-020-2531-2

Daru, B.H., Karunarathne, P., Schliep, K., 2020. phyloregion: R package for biogeographical regionalization and macroecology. Methods Ecol. Evol. 11, 1483–1491. https://doi.org/10.1111/2041-210X.13478

Diniz-Filho, J.A.F., Loyola, R.D., Raia, P., Mooers, A.O., Bini, L.M., 2013. Darwinian shortfalls in biodiversity conservation. Trends Ecol. Evol. 28, 689–695. https://doi.org/10.1016/j.tree.2013.09.003

Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61, 1–10. https://doi.org/10.1016/0006-3207(92)91201-3

Felsenstein, J., 1985. Phylogenies and the comparative method. Am. Nat. 125, 1–15. https://doi.org/0003-0147/85/2501-0001

Freitas, T.M. da 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 Fish. 22, 87–104. https://doi.org/10.1111/faf.12507

van der Laan, R., Fricke, R. & Eschmeyer, W. N. (eds) 2021. ESCHMEYER'S CATALOG OF FISHES: CLASSIFICATION. (http://www.calacademy.org/scientists/catalog-of-fishes-classification/).Electronic version accessed dd mmm 2021.

García-Andrade, A.B., Carvajal-Quintero, J.D., Tedesco, P.A., Villalobos, F., 2021. Evolutionary and environmental drivers of species richness in poeciliid fishes across the Americas. Glob. Ecol. Biogeogr. 30, 1245–1257. https://doi.org/10.1111/geb.13299

Haeseler, A. V., 2012. Do we still need supertrees? BMC Biol. 10, 2–5. https://doi.org/10.1186/1741-7007-10-13

Hortal, J., De Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M., Ladle, R.J., 2015. Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. Annu. Rev. Ecol. Evol. Syst. 46, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Jetz, W., Pyron, R.A., 2018. The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. Nat. Ecol. Evol. 2, 850–858. https://doi.org/10.1038/s41559-018-0515-5

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, 444–448. https://doi.org/10.1038/nature11631

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

Kraft, N.J.B., Cornwell, W.K., Webb, C.O., Ackerly, D.D., 2007. Trait evolution, community assembly, and the phylogenetic structure of ecological communities. Am. Nat. 170, 271–283. https://doi.org/10.1086/519400

Li, D., Trotta, L., Marx, H.E., Allen, J.M., Sun, M., Soltis, D.E., Soltis, P.S., Guralnick, R.P., Baiser, B., 2019. For common community phylogenetic analyses, go ahead and use synthesis phylogenies. Ecology 100, 1–15. https://doi.org/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 Phytol. 207, 437–453. https://doi.org/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. https://doi.org/10.1186/1471-2105-14-324

Miller, E.C., 2021. Comparing diversification rates in lakes , rivers , and the sea 1–19. https://doi.org/10.1111/evo.14295

Moura, M.R., Costa, H.C., Peixoto, M.A., Carvalho, A.L.G., Santana, D.J., Vasconcelos, H.L., 2018. Geographical and socioeconomic determinants of species discovery trends in a biodiversity hotspot. Biol. Conserv. 220, 237–244. https://doi.org/10.1016/j.biocon.2018.01.024

Moura, M.R., Jetz, W., 2021. Shortfalls and opportunities in terrestrial vertebrate species discovery. Nat. Ecol. Evol. 5, 631–639. https://doi.org/10.1038/s41559-021-01411-5

Nakamura, G., Vicentin, W., Súarez, Y.R., Duarte, L., 2020. A multifaceted approach to analyzing taxonomic, functional, and phylogenetic β□diversity. Ecology.

https://doi.org/10.1002/ecy.3122

Pearse, W.D., Purvis, A., 2013. phyloGenerator: An automated phylogeny generation tool for ecologists. Methods Ecol. Evol. 4, 692–698. https://doi.org/10.1111/2041-210X.12055

Pie, M.R., Carrijo, T.F., Caron, F.S., 2021. The diversification of termites: Inferences from a complete species-level phylogeny. Zool. Scr. 1–11. https://doi.org/10.1111/zsc.12502

Rabosky, D.L., Chang, J., Title, P.O., Cowman, P.F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T.J., Coll, M., Alfaro, M.E., 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature 559, 392–395. https://doi.org/10.1038/s41586-018-0273-1

Redding, D.W., Mooers, A.O., 2006. Incorporating evolutionary measures into conservation prioritization. Conserv. Biol. 20, 1670–1678. https://doi.org/10.1111/j.1523-1739.2006.00555.x

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. Freshw. Biol. 64, 447–460. https://doi.org/10.1111/fwb.13233

Roa-Fuentes, C.A., Heino, J., Zeni, J.O., Ferraz, S., Cianciaruso, M.V., Casatti, L., 2020. Importance of local and landscape variables on multiple facets of stream fish biodiversity in a Neotropical agroecosystem. Hydrobiologia 7. https://doi.org/10.1007/s10750-020-04396-7

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

Seger, G.D.S., Duarte, L.D.S., Debastiani, V.J., Kindel, A., Jarenkow, J.A., 2013. Discriminating the effects of phylogenetic hypothesis, tree resolution and clade age estimates on phylogenetic signal measurements. Plant Biol. 15, 858–867. https://doi.org/10.1111/j.1438-8677.2012.00699.x

Smith, S.A., Beaulieu, J.M., Donoghue, M.J., 2009. Mega-phylogeny approach for comparative biology: An alternative to supertree and supermatrix approaches. BMC Evol. Biol. 9, 1–12. https://doi.org/10.1186/1471-2148-9-37

Stein, R.W., Mull, C.G., Kuhn, T.S., Aschliman, N.C., Davidson, L.N.K., Joy, J.B., Smith, G.J., Dulvy, N.K., Mooers, A.O., 2018. Global priorities for conserving the evolutionary history of sharks, rays and chimaeras. Nat. Ecol. Evol. 2, 288–298. https://doi.org/10.1038/s41559-017-0448-4

Tedesco, P.A., Beauchard, O., Bigorne, R., Blanchet, S., Buisson, L., Conti, L., Cornu, J.F., Dias, M.S., Grenouillet, G., Hugueny, B., Jézéquel, C., Leprieur, F., Brosse, S., Oberdorff, T., 2017. Data Descriptor: A global database on freshwater fish species occurrence in drainage basins. Sci. Data 4, 1–6. https://doi.org/10.1038/sdata.2017.141

Tonini, J.F.R., Beard, K.H., Ferreira, R.B., Jetz, W., Pyron, R.A., 2016. Fully-sampled phylogenies of squamates reveal evolutionary patterns in threat status. Biol. Conserv. 204, 23–31. https://doi.org/10.1016/j.biocon.2016.03.039

Trisos, C.H., Auerbach, J., Katti, M., 2021. Decoloniality and anti-oppressive practices for a more ethical ecology. Nat. Ecol. Evol. https://doi.org/10.1038/s41559-021-01460-w

Upham, N.S., Esselstyn, J.A., Jetz, W., 2019. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLOS Biol. 17, e3000494. https://doi.org/10.1371/journal.pbio.3000494

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

Webb, C.O., Ackerly, D.D., McPeek, M. a., Donoghue, M.J., 2002. Phylogenies and

Community Ecology. Annu. Rev. Ecol. Syst. 33, 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448

Webb, C.O., Donoghue, M.J., 2005. Phylomatic: Tree assembly for applied phylogenetics. Mol. Ecol. Notes 5, 181–183. https://doi.org/10.1111/j.1471-8286.2004.00829.x