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

Authors: Gabriel Nakamura1, 2, \*, Aline Richter1, Bruno E. Soares3

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

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

# Abstract

1 – Phylogenies comprises an essential source of information for evolutionary and ecological studies. They allow to investigate hypothesis from trait patterns evolution to the relationship among evolutionary diversity and ecosystem functioning. However, for some groups, like fishes, obtain a phylogenetic hypothesis that include all species of interest in a local assemblage can be a difficult task, hindering studies involving this group, mainly in tropical region.

2 – In order to facilitate the obtention of phylogenetic information for bony fish in a standardized and reliable way we developed the R package FishPhyloMaker. FishPhyloMaker automatize the insertion procedure of absent species in phylogeny using the most comprehensive phylogeny of finned-ray fishes as a backbone. The insertion is made sequentially, following the taxonomy hierarchy of species and the user can choose if the insertion will be done through an interactive procedure or automatically, facilitating the use of FishPhyloMaker package by non-specialists in the group.

3 – We presented the main functions of FishPhyloMaker package, FishTaxaMaker() and FishPhyloMaker(), that together allows for the assessment of validity of species names and generate dated phylogenies for a local pool of species, respectively. We also show how the insertions can be mapped in the phylogenetic tree, allowing to identify phylogenetic gaps in the relationships of species of the local assemblages.

4- FishPhyloMaker facilitates the generation of phylogenetic trees through a reliable and reproducible way for the most diversified group of vertebrates. The package adopt well known rules of insertion which will expand the range of evolutionary and ecological questions that can be addressed using bony fishes as a study model (mainly in community phylogenetic studies). Also, will allow to identify shortfalls in the evolutionary relationships of fish species.

# Introduction

Phylogenies have been widely used in ecology, particularly in the last 20 years, mainly due to the development of theoretical frameworks (*e.g.* Webb et al. 2008, Felsenstein 1985), numerical methods and software (Webb et al., 2008), advancing our knowledge at different scales of investigation. Currently, research agenda in ecology and evolution almost always involves the use of a phylogenetic approach, from species level to a macroecological scale, which makes mandatory the presence of a comprehensive phylogenetic tree (for traits (e.g. Maestri et al. 2017); invasion ecology (Pinto-Ledezma et al., 2020); metacommunity ecology (Pillar & Duarte, 2010); biodiversity and ecosystem functioning (Molina-Venegas et al. 2021))

The availability of phylogenetic trees has grown in the last decade, with some groups presenting well-established phylogenies, e.g. birds (Jetz, Thomas, Joy, Hartmann, & Mooers, 2012a), plants (Magallón, Gómez-Acevedo, Sánchez-Reyes, & Hernández-Hernández, 2015), butterflies (Chazot et al., 2019), bony fishes (Betancur et al., 2017; Rabosky et al., 2018). However, they differ in an important characteristic. While certain groups, present comprehensive phylogenies including all known species (e.g. birds), other groups, like fishes, species are absent in the most comprehensive phylogeny available. This imposes a problem since the lack of species in the phylogeny affect the calculation of metrics that indicates evolutionary process (*e.g.* diversification measures), generate inaccurate estimates of phylogenetic signal (Seger et al. 2010), trait evolution (Boettiger, Coop, & Ralph, 2012), underestimate phylogenetic diversity in local communities, or even impede the realization of an entire study.

The lack of phylogenetic information of species, specially from megadiverse groups and understudied regions of the world (e.g. the tropics), is one of the components of a pattern known as Darwinian shortfall (Diniz-Filho, Loyola, Raia, Mooers, & Bini, 2013). In an extensive revision of the components of Darwinian shortfalls, Diniz-Filho et al (2013) proposes some short-term solutions to tackle this problem. Particularly, regarding the lack of phylogenetic information for species, the authors suggest coupling the information present in available phylogenies with cladistic classification (from taxonomy hierarchy) to produce approximations of comprehensive phylogenies, a strategy that was already adopted by other groups (e.g. Phylomatic for plants and mammals Webb et al. 2002). However, this solution can become very laborious in the absence of an automatized procedure, particularly when many species must be added manually (Webb et al., 2008), besides lacking reproducibility due to specific decisions that are not precisely documented. On the other hand, obtention of comprehensive phylogenies can be done through the use of molecular techniques (Jetz, Thomas, Joy, Hartmann, & Mooers, 2012b; Pearse, Purvis, Cavender-bares, & Helmus, 2014), however, it requires expertise and present high financial costs (Roquet, Thuiller, & Lavergne, 2013). Therefore, assembly synthesis phylogenies “by-hand” with the assistance of a automatized procedure (Webb & Donoghue, 2005) can be a reliable solution in most cases.

Finned-ray fishes are an example of a megadiverse group in which the shortfall in phylogenetic information is pervasive, since most of its diversity is located at tropical region (Albert, Tagliacollo, & Dagosta, 2020). The complex evolutionary history, joint with heterogeneity of life history forms (Albert et al., 2020) make them an interesting group to address a range of questions in the interface of ecology and evolution (e.g. Roa-Fuentes et al. 2019, 2020; Nakamura et al. 2020). However, studies that used phylogenetic information in fishes are rare when compared with other groups that present specific tools to deal with the construction of local phylogenies (*e.g.* Webb and Donoghue 2005 for mammals and plants; Jin and Qian 2019 for plants), suggesting that the obtention of phylogenetic information can hinder our knowledge regarding fish ecology and evolution in megadiverse regions. Besides the operational problems to construct phylogenetic trees for local fish assemblages, our knowledge regarding the phylogenetic shortfalls for this group is restricted only to few lineages (e.g. Freitas et al., 2021) making difficult to map the regions that must be further studied.

In this study, we presented an R package FishPhyloMaker, that overcomes the main problems in the obtention of phylogenies for finned-ray fishes: lack of reproducibility and the labor intensiveness process for construct large phylogenies for specific local/regional pool of species. The package FishPhyloMaker allows for the construction of a phylogenetic tree for Actynopterygii species containing local/regional species informed by the user. Furthermore, we implemented in the package a new method to map Darwinian shortfalls in the distribution of finned-ray fishes.

# Inside the fish: overview of FishPhyloMaker

FishPhyloMaker is a freely R package that has two main functions, FishTaxaMaker() and FishPhyloMaker() (Table 1 present a list of all functions present in the package). The first generate a data frame from fish species names provided by the user, checking the validity of the species names by using the information presented in Fishbase database (Froese and Pauly, 2006) through the package rfishbase (Boettiger et al., 2012). Species that were not found in Fishbase are printed in the command line, and the user can inform manually the Family and the Order of these species. The output of FishTaxaMaker function is a list containing three information: 1) a data frame with a complete taxonomic information of the species provided by the user, joint with the valid name of the species, when present; 2) a data frame containing three columns with the valid names of species (s), the Family (f) and the Order (o). This data frame can be directly used in FishTaxaMaker() function to construct the phylogenetic tree; 3) a character vector containing all names of species informed by the user that was not found in Fishbase.

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

|  |  |
| --- | --- |
| **Function** | **Description** |
| FishTaxaMaker() | Check species names according with Fishbase and prepares the data of species name provided by the user |
| whichFishAdd() | Auxiliar function; can be used to know at which level each species will be inserted, or if the species are already presented in the mega-tree |
| FishPhyloMaker() | Core function; make the phylogeny and alternatively return a data frame indicating the order of insertions of species |
| Darwinian\_deficit | Calculates the Darwinian deficit for an assemblage |

FishPhyloMaker() function will use the information of the taxonomic hierarchy contained in the data frame returned from FishTaxaMaker(), joint with the information present in the fishtree of life project (Chang, Rabosky, Smith, & Alfaro, 2019) through the fishtree package (Chang et al., 2019) to assemble the phylogenetic tree with the valid species informed by the user.

FishPhyloMaker() works in a sequential way, with 4 levels of insertions that are explained in detailed next. First the function finds for all species of local data that are already presented in the mega-tree. If any species remains for insertion, then the function finds for congeneric species that are already present in tree. If any, the congeneric species are added to the tree as a sister species of the respective genus by dividing the branch of congeneric species by two and inserting the new species. If there are more than one congeneric species in the tree the species will be added as a polytomy at the node that corresponds to the most recent ancestral (MRCA) that links all congeneric species. This round of insertion is what we call Congeneric insertion.

After congeneric insertions, if there are species that already must be inserted in the tree, FishPhyloMaker finds for the all the genus in the tree that are of the same family of the species that must be inserted. This comprises the Family level round of insertions. In this round, the user has three insertion options: 1- the species can be added near to a genus of the same family, by dividing the branch length that links all species in this genus by two; 2- the species can be added between two different genus of the same family, by adding a new node between these two genus or 3- the species can be added in the node that unites all genus of the family (root family node) by adding a new branch at the family node. The user chooses one of these three options through an interactive procedure, in which all the genus of the same family already present in tree are printed in the console. The user must type the name of a given genus for the first option, the name of two genus separated by a blank space for the second option and the name of the family for the third option.

After all species with representative Genus of their families be inserted, if any species remains to be inserted in the tree, the function found for all families present in the tree that are of the same Order of the species that must be inserted, and an interactive procedure similar of that described previously for the family level is performed, but, instead to choose among Genus, one of the same three options previously presented must be chosen by typing the families inside a given Order, or the name of the Order to insert at the node that unites all families inside a given Order. This level of insertion is called Order insertion level. If any species remains to be inserted after these three levels, the function will inform the user that the species was not inserted in the tree. We adopt this decision since the insertions beyond Order will carry a lot of uncertainty regarding the phylogenetic position of this species. Finally, if the user do not want to realize the interactive procedure of insertion, for example, the case in which a great number of species that must be added, an automatic insertion at the family and order nodes can be performed by setting the argument insert.base.node equals TRUE.

Another point that is important to mention is that at each level of insertion a checking procedure is performed to check if any genus of the species that must be inserted was included in the tree. If this occur the species will be automatically inserted following the same procedure described in the Congeneric level of insertions. A scheme illustrating the whole procedure of insertions is showed in Figure 1.

Diagrama, Esquemático

Descrição gerada automaticamente

Figure 1: Schematic representation of insertion and subsetting procedure realized in FishPhyloMaker() function. Here we used a hypothetical phylogeny containing only 10 species and four families as the backbone phylogeny (A). B represents the Congeneric level of insertion. C represents the three options that the user may choose in the Family level round of insertions (C1 – at family node; C2 – near to a specific genus; C3 – between two genus). D represents the congeneric insertions at family level and finally, E represents the final tree containing only the species of interest

The last step (Figure 1 D colored lineages) consists in subset the tree to maintain only those species of interesting of the user. This is done by dropping all species that are not present at the names informed by the user. This returns a dated phylogenetic tree, and, optionally, a data frame with the same format of that the user obtained with FishTaxaMaker(), but with a new column indicating at which level the species was added in the procedure. If the species was already present in the tree it will labeled with “Present\_in\_tree” label, if the species was added as a congeneric species it will be labeled as “Congeneric\_insertion”, if species was added in at Family level be labeled as “Family\_insertion”, if species was added near to a family in a given order, i.e. Order insertion level, the species will be labeled as “Order\_insertion”, at last, if after all these steps the species was not added to the mega-tree it will appear as “Not\_inserted”.

The labels attributed to each species can give an idea of the uncertainty in the insertion process. Species inserted in the Order level carry much more uncertainty regarding its phylogenetic position than species added as congeneric species in the Congeneric level of insertions.

It is worth to highlight that the insertion procedures in FishPhyloMaker package carry similarities with other well accepted methodologies implemented for other groups, for example, Phylomatic (Webb & Donoghue, 2005) software. The procedures are similar in Phylomatic and FishPhyloMaker, being that the last present more options regarding the insertion procedures that the former (insertion between genus and near to a specific genus, and insertion between families and near to a given family).

# Example analysis

In order to show the functioning of FishPhyloMaker package we provided here an example of how to obtain a phylogenetic tree for fish species in the four main ecoregions of freshwater fishes: Afrotropic, Indo-Malay, Neartic and Neotropic (Tedesco et al., 2017). This dataset comprises the biggest occurrence data for freshwater species of the world and it has been widely used for investigating global patterns of species distribution and evolutionary determinants of freshwater fishes (e.g. Miller and Román-Palácios, 2021; Garcia-Andrade et al 2021).

We first checked and prepared the occurrence data using function FishTaxaMaker(). The FishTaxaMaker() function find for at least one valid name for all species in each of the ecoregions analyzed.

From data returned from FishTaxaMaker we used FishPhyloMaker() function to generated the phylogenies for each ecoregion, including all the species that lack the phylogenetic position for each ecoregion. For sake of simplicity and illustration we opt to insert species that lack congeneric species in the Rabosky´s phylogeny at the base node of the family or the order (when the species did not present any representative species from the same family), by setting argument return.insertions = TRUE. However, as highlighted earlier, the user can opt for manual insertions. The complete phylogenetic tree containing all the insertions returned from FishPhyloMaker()are showed in Figure 2.



Figure 2: Phylogenetic trees generated with FishPhyloMaker package. The colored tip-points show the level of insertion at which each species was inserted and correspond to the codes described in the main text. We generated phylogenetic trees for the four principal ecoregions in the world (Afrotropics, Indo-Malay, Neartic and Neotropic).

The whole procedure to insert all absent species in the Rabosky´s phylogeny, considering the four ecoregions, spend approximately two hours (using one core in a machine with i7 processor). A total of XXX species was inserted, with the Afrotropics being the region that presented the greater number of insertions (XXXX).

All the code needed to fully reproduce this phylogenetic tree is provided at (https://github.com/GabrielNakamura/MS\_FishPhyloMaker). Further analysis exploring the potential of FishPhyloMaker package can be accessed at https://gabrielnakamura.github.io/FishPhyloMaker/index.html.

# Similarity with other approaches, future developments and possible applications

The functionalities implemented in FishPhyloMaker package has consonance with other procedure developed for plant species, like Phylomatic (C++ aplication) and V.PhyloMaker (R package) (Webb & Donoghue, 2005; Jin & Qian, 2019). Some important differences here relative to these programs is that FishPhyloMaker expanded the insertion options that the user can made, at the same time that preserve the user friendliness by non-specialists in the group.

Future developments of FishPhyloMaker package consist in the inclusion of other online data bases in the functions that check for the names of species. Despite Fishbase be the most used online databases to check for taxonomic classification of fishes, other important sources were not contemplated in this version of the package, for example, the Catalog of Fishes (Fricke and Eschemeyer, 2021). We hope that in future versions this data base can be included, improving the scope of FishPhyloMaker package in taxonomic nomenclature checking.

Another possible drawback is that the phylogenetic hypothesis generated through FishPhyloMaker may present many polytomies (for example Figure 2, particularly in Afrotropics). To overcome this drawback we suggest that the phylogenetic tree be submitted to a procedure like that proposed by Martins et al. (2013) in order to assess how the phylogenetic uncertainty may affect downstream analysis that use not fully solved phylogenetic tree (e.g. Nakamura et al. 2021).

PhyloFishMaker will facilitate the obtention of phylogenetic hypothesis for local pool of finned-ray fishes, mainly in Neotropical region, that probably is the region that present the major gap regarding the knowledge of phylogenetic position of species (Albert et al., 2020). The ease to produce local fish phylogenies with all species will allow to better understand how fish traits evolved, how the current and past environmental conditions selected the lineages through different Basins, boosting our understanding of patterns of evolution of bony fish assemblages and species.

Much of the biogeographical studies using fishes are restricted only to a specific families that present molecular phylogenies, FishPhyloMaker will facilitate the obtention of phylogenetic information for local communities that generally involves a great number of fish lineages, especially in Neotropical region (Albert et al., 2020). This will allow to amplify the investigations of historical biogeography studies that uses Neotropical fishes as models, allowing for broaden the knowledge regarding the biogeographical understanding of the most diverse group of vertebrates on earth. Therefore, we expect that FishPhyloMaker reduce the gaps and barriers found in evolutionary and ecological studies due to the difficulty or absence of a reliable phylogenetic hypothesis for bony fishes.

# 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. (2012). Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*, *66*(7), 2240–2251. doi:10.1111/j.1558-5646.2011.01574.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

Chazot, N., Willmott, K. R., Lamas, G., Freitas, A. V. L., Piron-Prunier, F., Arias, C. F., … Elias, M. (2019). Renewed diversification following Miocene landscape turnover in a Neotropical butterfly radiation. *Global Ecology and Biogeography*, *28*(8), 1118–1132. doi:10.1111/geb.12919

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

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

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1–15. doi: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 and Fisheries*, *22*(1), 87–104. doi:10.1111/faf.12507

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

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, A. O. (2012b). 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

Maestri, R., Monteiro, L. R., Fornel, R., Upham, N. S., Patterson, B. D., & de Freitas, T. R. O. (2017). The ecology of a continental evolutionary radiation: Is the radiation of sigmodontine rodents adaptive? *Evolution*, *71*(3), 610–632. doi:10.1111/evo.13155

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. (2021). Taxonomic and phylogenetic beta diversity in headwater stream fish communities of the Paraná and Paraguai River basins. *Neotropical Ichthyology*, *19*(1), 1–17. doi:10.1590/1982-0224-2020-0126

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

Pearse, W. D., Purvis, A., Cavender-bares, J., & Helmus, M. R. (2014). Metrics and Models of Community Phylogenetics, 451–464. doi:10.1007/978-3-662-43550-2

Pillar, V. D., & Duarte, L. D. S. (2010). A framework for metacommunity analysis of phylogenetic structure. *Ecology Letters*, *13*, 587–596. doi:10.1111/j.1461-0248.2010.01456.x

Pinto-Ledezma, J. N., Villalobos, F., Reich, P. B., Catford, J. A., Larkin, D. J., & Cavender-Bares, J. (2020). Testing Darwin’s naturalization conundrum based on taxonomic, phylogenetic, and functional dimensions of vascular plants. *Ecological Monographs*. doi:10.1002/ecm.1420

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

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*. doi:10.1007/s10750-020-04396-7

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