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, beside their interface user-friendly. The package adopts 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). This tool allows us to advance 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. For this, is mandatory the presence of a comprehensive phylogenetic tree, even the main question be the traits evolution (e.g. Maestri et al. 2017); invasion ecology (Pinto-Ledezma et al., 2020); metacommunity ecology (Pillar & Duarte, 2010) or 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) and bony fishes (Betancur et al., 2017; Rabosky et al., 2018). However, they differ in an important characteristic, while certain groups present robust phylogenies that comprises all known species (e.g. birds), other groups, like fishes, have a lot of issues in relation with the taxonomic position (eg. non-monophyletic groups) and absence of species for the most comprehensive phylogeny available. The last issue 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, 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 (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, comprehensive phylogenies can be obtained 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 for most researchers.

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 compare an index of phylogenetic diversity (PD) calculated with the most complete phylogenetic tree available with a tree constructed with the FishPhyloMaker. This comparison makes clear the importance of taking into account all species/lineages of a local/regional pool and for identify the regions when this gap is more pronounced.

# Inside the fish: overview of FishPhyloMaker package

FishPhyloMaker is a freely R package that has two main functions, FishTaxaMaker() and FishPhyloMaker(). Below, we describe these two functions highlighting the input data, intermediate steps and output objects. A briefly description of all function of the package is presented in Table 1.

*FishTaxaMaker()*

The input data could be a string or a data frame with species list from the regional pool or an occurrence matrix, whit sites in rows and species in columns. The only requirement is that genus and specific epithet (or even the subspecies) must be separated by underline (eg. *Genus*\_*epithet*). The function thus search the user data in the Fishbase database (Froese and Pauly, 2006) employing functions available into rfishbase package (Boettiger et al., 2012). This search consists in find and classify the species list provided by user as valid names or synonymies. If the provided name is a synonymy, so the functions put in a new column the valid name from the species according with Fishbase. Species that were not found in Fishbase are printed in the command line, and the user must manually inform the Family and the Order which the species belong. The output of function is a list containing three information: 1) a data frame when the first column is the species names provided by used, a column with the valid names for each species and the subsequent columns are the hierarchical taxonomic information (Subfamily, Family, Order, Class and SuperClass); 2) a data frame containing the valid names of species (s), the Family (f) and the Order (o) information. This data frame could be low rows than the species list provided by user, because some species provided by user could be synonymies from a unique valid name. 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()*

Is a function to insert species and/or prune the backbone phylogenetic tree according with the species pool informed by user (Figure 1). The input of the function can be the second data frame generated by the *FishTaxaMaker*() function, or a data frame provided by user with the same information, and three logical arguments: insert.base.node, return.insertions and progress.bar. These three arguments are defined by default as FALSE, TRUE and TRUE, respectively.

First at all, the function finds for all species of local data that are already presented in the mega-tree. If all species are present, so the function return the pruned tree. If any of the species provided by the user is not found, then the FishPhyloMaker() works in a sequential way, with 4 levels of insertions: Congeneric insertion, Family-level insertions, Congeneric at Family-level insertions and Order insertions. (i) the Congeneric insertion consist in looking for genus in the tree that is the same genus of species not found. Once found, if there are only one species of the genus in the tree, the branch is split in two and the species is added as sister of the respective genus. When there are two or more species of the genus in the tree, the species is inserted as polytomy at the node that corresponds to the most recent ancestral (MRCA) that links all congeneric species. (ii) After this, if there are species that already must be inserted in the tree, FishPhyloMaker finds for all genus in the tree that are of the same family of the remaining species, comprising the Family-level of insertions. A list with all genus is printed, and interactively, the user can inform if the species must be inserted 1 – near to a genus, 2 – between two genera or 3 – at the node of the family. If option 1 is chosen, the user must provide the name of the genus in the list to which the species to be inserted is closest, the branch of the target genus is split in two and the species is added as sister. If option 2 is chosen, the user must provide two names, separated by a blank space. The species is inserted as a polytomy at the node that unites the two informed genera. If option 3 is chosen, the user must inform the family name, then the species is inserted at family node as a polytomy. (iii) After this procedure, the function checks if any remained species to be inserted could be inserted now as congeneric. If yes, the insertion follows the same procedure described at (i), but is discriminated in the output data as a Congeneric family-level insertion. iv) if any species remains to be inserted in the tree after the family-level insertion, the function works in a similar way to that already described in (ii), but instead a list of genera, a list of all families of the order to which the species belong is printed to user. Thus, the user can choose to insert the species 1 – near to a specified family, 2 – between two families or 3 – at the node of order. This is the step of Order insertion. If after this four steps there are still species to be inserted, 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.

These steps may be simplified if the user setting the argument insert.base.node = TRUE. This argument breaks the interactive procedure (options 1, 2 and 3 for (ii) and (iv)) and automatically insert all species not find in the congeneric insertions (steps (i) and (iii)) in the family or order nodes. This setting is recommendable if there are a large number of species that must be added by user or if the position on tree of the species that must be inserted is unknown or difficult to access.

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 pruned tree containing only the species of interest.

By default, the output of the FishPhyloMaker() function is a list with two objects. The first is the pruned tree, that maintain only those species informed by user in the data argument. The second is a copy of the informed data with the addition of a new column that discriminate for each species if it is already present in the tree (Present\_in\_Tree), inserted by step (i) “Congeneric\_insertion”, inserted in step (ii) by any choices “Family\_insertion”, inserted in step (iii) “Congeneric\_familiy\_level”, or inserted in step (iv) “Order\_insertion”. At last, if after all these steps the species was not added to the mega-tree, it will appear as “Not\_inserted”.

# 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(). For this dataset, from xx names provided in the occurrence matrix, xx were synonymies, remaining xx species to perform the FishPhyloMaker function. For each ecoregion, one phylogenetic tree was build, accounting with the all species inserted (Figure 2). For sake of simplicity and illustration, we opt to insert species that are not inserted at congeneric level (steps (ii) or (iii)) at the basal node of the family or the order (when the species did not present any representative species from the same family), by setting argument insert.base.node = TRUE. However, as highlighted earlier, the user can opt for manual insertions.



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 and advances in relation with other approaches

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). Beside the focal groups, there are some important differences in the FishPhyloMaker relative to these programs. First, we expanded the insertion options that the user can made, at the same time that preserve the user friendliness by non-specialists in the group. Further, we provide a reproductible way to construct phylogenetic tree for megadiverse group, like fishes.

**Limitations and possible applications**

Future developments of FishPhyloMaker package consist in the inclusion of other online databases in the FishTaxaMaker function, aiming to improve the nomenclature checking. 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).

An inherent limitation of the phylogenetic hypothesis produced by FishPhyloMaker is the large number of polytomies, in some cases (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, when a not fully solved phylogenetic tree is used (e.g. Nakamura et al. 2021).

However, these limitations do not preclude the package functioning or its applicability for studies in phylogenetic community ecology. Moreover, until now, this is the unique tool able to provide a complete phylogenetic tree that is totally automated, and that can handle easily for large datasets. The use of the FishPhyloMaker can be relevant for accessing several important issues in ecology and evolution. In a local-scale, the FhyloFishMaker could 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). Such phylogenetic hypothesis comprising all species allow to better understand how fish traits evolved or how the current and past environmental conditions selected the lineages through different Basins. On the other hand, large-scale studies with fishes, as the biogeographical studies, are restricted to one or a few specific families, due the availability of molecular phylogenies. With the FishPhyloMaker, all lineages can be used to evaluate the biogeographic history of the most diverse group of vertebrates on earth, and help us to understand the processes that drive this high diversity. Also, we can mapping where the lack of phylogenetic information is most critical, once that function returns the insertion-level of species. This could be very helpful to elucidate questions as the Darwinian shortfalls.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 finned-ray 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