

1 Physcraper: a python package for continual update of evolutionary
2 estimates using the Open Tree of Life

3
4 **1. Luna L. Sanchez Reyes**

5 School of Natural Sciences, University of California, Merced

6 email: sanchez.reyes.luna@gmail.com

7 **2. Martha Kandziora**

8 School of Natural Sciences, University of California, Merced

9 Department of Botany, Faculty of Science, Charles University, Prague, Czech Republic

10 email: kandziom@natur.cuni.cz

11 **3. Emily Jane McTavish**

12 School of Natural Sciences, University of California, Merced

13 email: ejmctavish@gmail.com

14 **Correspondence address:** Science and Engineering Building 1, University of California, Merced, 5200 N.
15 Lake Rd, Merced CA 95343

16 **Correspondence email:** sanchez.reyes.luna@gmail.com, ejmctavish@gmail.com

¹⁷ **Running title:** Updating gene trees with the Open Tree of Life

¹⁸ **Word count:** 2419

¹⁹ **Manuscript prepared for submission to Methods in Ecology and Evolution**

²⁰ **Article type:** Application

Abstract

1. Phylogenies are a key part of research in many areas of biology. Tools that automate some parts of the process of phylogenetic reconstruction, mainly molecular character matrix assembly, have been developed for the advantage of both specialists in the field of phylogenetics and nonspecialists. However, interpretation of results, comparison with previously available phylogenetic hypotheses, and choice of one phylogeny for downstream analyses and discussion still impose difficulties to one that is not a specialist either on phylogenetic methods or on a particular group of study.
2. Physcraper is a command-line Python program that automates the update of published phylogenies by adding public DNA sequences to underlying alignments of previously published phylogenies. It also provides a framework for straightforward comparison of published phylogenies with their updated versions, by leveraging upon tools from the Open Tree of Life project to link taxonomic information across databases.
3. Physcraper can be used by the nonspecialist, as a tool to generate phylogenetic hypotheses based on publicly available expert phylogenetic knowledge. Phylogeneticists and taxonomic group specialists will find it useful as a tool to facilitate molecular dataset gathering and comparison of alternative phylogenetic hypotheses (topologies).
4. The Physcraper workflow demonstrates the benefits of doing open science for phylogenetics, encouraging more researchers to strive for better sharing practices. Physcraper can be used with any OS and is released under an open-source license. Detailed instructions for installation and use are available at <https://physcraper.readthedocs>.

Keywords: gene tree, interoperability, open science, open tree of life, phylogeny, public database, python, reproducibility, taxonomy, updated alignment

1 Introduction

Phylogenies capture the shared history of organisms and provide key evolutionary context for our biological observations. Public biological databases constitute an amazing resource for evolutionary studies, but a large portion of public molecular data has never been incorporated into any phylogenetic estimate. Extending existing phylogenies with new DNA sequence data, geographical location, and other metadata in a reproducible and continuous manner is possible by automating connections between biological databases. Here, we introduce Physcraper, a tool to build upon DNA datasets that taxon specialists have assessed and deemed appropriate for a specific phylogenetic scope, by using data from public DNA databases to update a starting tree and single locus alignments.

Taxonomic idiosyncrasies across databases represent a huge challenge for automatic integration of data into phylogenies; challenge that can be addressed with a unified taxonomy for name standardization. The Open Tree of Life project (OpenTree) constructs a comprehensive tree of life by synthesizing published phylogenies along with taxonomic data. OpenTree’s “synthetic” tree comprises 2.3 million tips, of which around 90,000 are supported by phylogenies - the remaining 1.4 million taxa are placed in the tree based on taxonomy. To achieve this, OpenTree unifies taxonomy from various databases (Rees & Cranston 2017), including the USA National Center for Biodiversity Information (NCBI) molecular database GenBank (Benson *et al.* 2000), among others. This OpenTree taxonomy represents a key resource for connecting data from any biological database that has been integrated to it.

Another challenge for incorporating public molecular data into existing phylogenies is assembling high-quality homology hypotheses. While genomics has, and will continue to, revolutionize phylogenetic inference, the variety of alternative genomic sequencing approaches produce largely non-overlapping genomic datasets across taxa, creating challenges in wide scale phylogenetic reconstruction. Phylogenomics ameliorate this problem by focusing on targeted capture of informative characters from independent and single-copy genetic markers (Andermann *et al.* 2020). Yet, decades of single locus sequencing have generated massive amounts of homologous DNA datasets that can be used for phylogenetic reconstruction at many scales. Moreover, species tree reconstructions from multiple single locus datasets using the multispecies coalescent model are

considered the gold standard for inferring species relationships (Song *et al.* 2012).

More than a decade ago, GenBank release number 159 (April 15, 2007) already hosted 72 million DNA sequences that were gauged to have the potential to resolve phylogenetic relationships of 98.05% of the almost 241,000 distinct taxa in the NCBI taxonomy at the time (Sanderson *et al.* 2008). Assembling a DNA alignment from such a massive database can be done “by hand”, but it is a largely time consuming and mostly non-reproducible approach. Computational pipelines that make DNA sequence search faster and more efficient, as well as more reproducible, have been applied to infer phylogenetic relationships in a variety of organisms (e.g., Smith *et al.* 2009; Izquierdo-Carrasco *et al.* 2014; Antonelli *et al.* 2017). However, thoughtfully curated markers and alignments can improve phylogenetic reconstructions, even in phylogenomic reconstructions (Fragoso-Martínez *et al.* 2017).

A way to incorporate the benefits from massive amounts of newly released molecular data and fine-grained curation from human experts, is to rely on manually curated homology hypotheses as “jump-start” alignments (Morrison 2006). The TreeBASE database (Piel *et al.* 2009) hosts about 8,200 alignments from published peer-reviewed studies, providing information on evolutionary relationships of around 100,000 distinct taxa (see TreeBASE’s website about). Linking published alignments generated by experts with molecular data that has not yet been included in any public phylogenetic estimate has the potential to accelerate the enrichment and updating of phylogenetic relationships in many regions of the tree of life.

Physcraper relies on programmatic access protocols (API’s) and the OpenTree taxonomy to automatically link phylogenetic data to molecular data from GenBank, alignments from TreeBASE, and phylogenies from OpenTree’s Phylesystem, to continually update and enrich phylogenetic knowledge based on expertly-curated homology hypotheses. Physcraper also provides new types of access to various OpenTree tools for comparison of existing phylogenetic hypotheses with newly generated ones. Physcraper is coded as a Python pipeline that focuses on data interoperability, by using the OpenTree taxonomy as a way to link taxon data from different databases. This integration also allows users to rapidly place new data from a diverse range of biological databases in an evolutionary context, opening the possibility for a variety of comparative downstream analyses.

2 The Physcraper framework

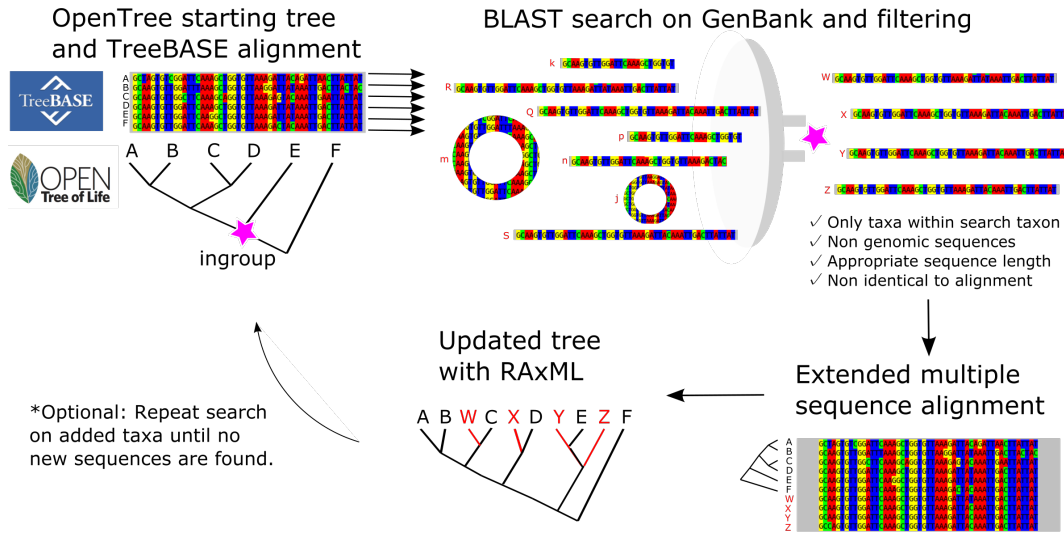


Figure 1: The Physcraper framework consists of 4 steps (see text). The software is fully described on its documentation website at physcraper.readthedocs.io, along with installation instructions, function usage descriptions, examples and tutorials.

The general Physcraper framework is depicted in Figure 1. It consists of 4 steps: 1) identifying and processing a tree and its underlying alignment; 2) performing a BLAST search of DNA sequences from the original alignment on GenBank, and filtering of new sequences; 3) profile-aligning new sequences to the original alignment; 4) performing a phylogenetic analysis and comparing the updated tree to previous phylogenetic estimates.

2.1 The inputs: a tree and an alignment

Taxon names in the input tree must be standardized to the OpenTree taxonomy (Rees & Cranston 2017) using OpenTree’s bulk Taxonomic Name Resolution Service TNRS tool. Users can upload their own tree, or choose from among the 2, 950 standardized trees stored in OpenTree’s Phylesystem that also have alignments available on TreeBASE (Piel *et al.* 2009).

The input alignment is a single locus alignment that was used in part or in whole to generate the input tree. If it is stored in TreeBASE, Physcraper retrieves it automatically. Otherwise, the user must provide the path to a local copy of the alignment. Only taxa that are both in the sequence alignment and in the tree are considered further for analysis; at least one taxon and its corresponding sequence are required.

2.2 DNA sequence search and filtering

The Basic Local Alignment Search Tool, BLAST (Altschul *et al.* 1990) is used for DNA sequence search on a remote or local GenBank database. It is constrained to a “search taxon”, a taxonomic group in the NCBI taxonomy that is automatically identified using the OpenTree API (Rees & Cranston 2017), as the most recent common ancestor of the ingroup taxa in the tree that is also a named clade in the NCBI taxonomy (Fig. 1). Alternatively, users can arbitrarily define a search taxon that is either a more or less inclusive clade relative to the ingroup of the input tree.

BLAST is implemented with the BLAST command line tools `blastn` function (Camacho *et al.* 2009) and the BioPython (Cock *et al.* 2009) BLAST function from the NCBIWWW module modified to accept an alternative BLAST address. Each sequence in the alignment is BLASTed once against all DNA sequences in GenBank. New sequences are excluded for analysis if 1) they do not belong to the search taxon; 2) they have an e-value above the cutoff (default to 0.00001); 3) they fall outside a min and max length threshold, defined as the proportion of the average length without gaps of all sequences in the input alignment (default values of 80% and 120%, respectively); 4) or if they are either identical to or shorter than an existing sequence in the input alignment and they represent the same taxon in the OpenTree taxonomy or the NCBI taxonomy. An arbitrary maximum number of randomly chosen sequences per taxon are allowed (default to 5).

Reverse, complement, and reverse-complement sequences are identified and translated using BioPython internal functions (Cock *et al.* 2009). Iterative cycles of BLAST searches can be performed, by blasting all new sequences until no new ones are found. By default only one BLAST cycle is performed.

2.3 New DNA sequence alignment

MUSCLE (Edgar 2004) is used to perform a profile alignment in which the original alignment is used as a template of homology criteria to align new sequences. The final alignment is not further automatically checked, and additional inspection and refinement are recommended.

2.4 Tree reconstruction and comparison

A Maximum Likelihood (ML) gene tree is reconstructed for each alignment provided, using RAxML (Stamatakis 2014) with default settings (GTRCAT model and 100 bootstrap replicates with the default algorithm) and using the input tree as a starting tree for the ML searches. Bootstrap results are summarized with DendroPy’s SumTrees module (current version 4.4.0; Sukumaran & Holder 2010).

Physcraper’s main result is an updated phylogenetic hypothesis for the search taxon. The updated tree is compared to the original tree using an automated conflict analysis which calculates Robinson Foulds weighted and unweighted metrics using Dendropy (Sukumaran & Holder 2010), and performs a node by node comparison between the synthetic OpenTree and the original and updated tree individually, using OpenTree’s conflict API (Redelings & Holder 2017). For the conflict analysis to be meaningful, the root of the tree needs to be accurately identified. The best way to currently do this is by hand, although a default rooting based on OpenTree’s taxonomy is also available.

3 Case Study: The hollies

A user is interested in the phylogenetic relationships within the genus *Ilex*. Commonly known as “hollies”, the genus encompasses between 400-700 living species, and is the only extant clade within the family Aquifoliaceae, order Aquifoliales of flowering plants.

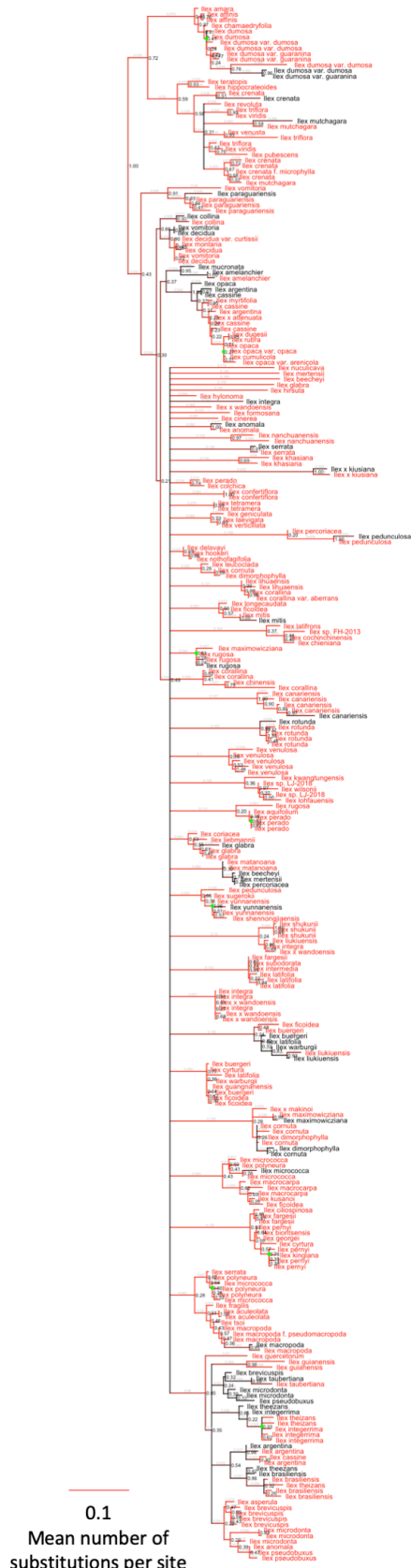
An online literature review in June 2020 (google scholar search for “ilex phylogeny”) reveals that there are several published phylogenetic trees showing relationships within the hollies (Cuénoud *et al.* 2000; Setoguchi & Watanabe 2000; Selbach-Schnadelbach *et al.* 2009; Manen *et al.* 2010), but only two have their data available publicly (Gottlieb *et al.* 2005; Yao *et al.* 2020). Gottlieb *et al.* (2005) made original tree and alignment data available in TreeBASE. The “Gottlieb2005” tree sampling 41 species was added to the OpenTree Phylesystem and its information has been integrated into OpenTree’s synthetic tree.

The most recent *Ilex* tree from Yao *et al.* (2020), has been made available in the OpenTree Phylesystem and in the DRYAD repository. The “Yao2020” tree, is the best sampled phylogenetic tree yet available for the

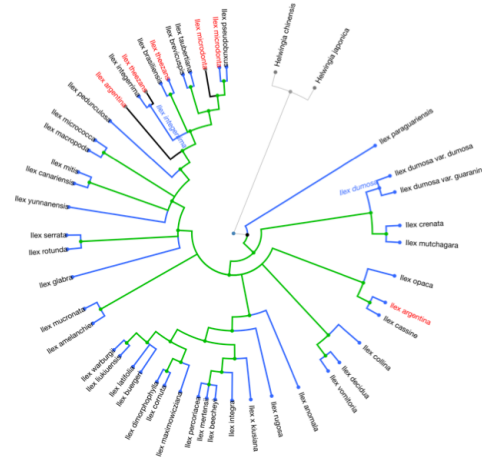
hollies, with 175 tips.

Figure 2 shows results from the Physcraper analysis of an alignment of the internal transcribed spacer DNA region (ITS) from Gottlieb *et al.* (2005). Physcraper ran on a local BLAST database, on a laptop Linux computer for 19hrs 45min to perform BLAST and RAxML analyses, with bootstrap analyses taking an additional 13hrs. The updated Gottlieb2005 tree contains all 41 distinct taxa from the original study plus 231 new tips, contributing phylogenetic data to 84 additional *Ilex* taxa. The best RaxML tree is 99% resolved, with 25% of nodes with bootstrap support < 0.1 and 48% nodes with bootstrap support > 0.75 . A large portion of internal branches are negligibly small, with 30 branches < 0.00001 substitution rate units, from which only 9 have a bootstrap support > 0.75 (Fig. 2). For comparison, the Yao2020 tree also contains all 41 distinct taxa from the original Gottlieb2005 study, and contributes phylogenetic data to 134 additional *Ilex* taxa, from which 67 are also in the Physcraper updated Gottlieb2005 tree. While Yao *et al.* (2020) also used ITS as a marker, their data in GenBank is not public yet, so Physcraper was unable to incorporate 68 additional taxa into the analysis. However, Physcraper was able to incorporate 18 taxa that were not in Yao2020. This might be caused by the method they used to download existing ITS *Ilex* sequences from GenBank, which is not fully explained in the publication, but seems to be a “manual” process.

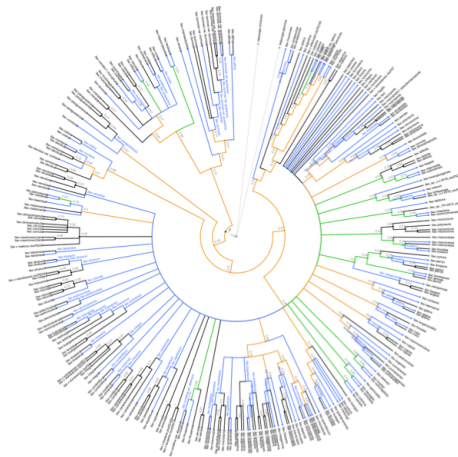
A) Updated Gottlieb2005 consensus tree



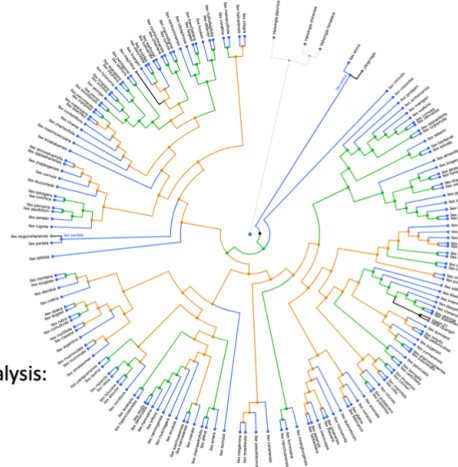
B) Original Gottlieb2005 tree conflict
48 tips, 41 taxa



C) Updated Gottlieb2005 tree conflict
231 new tips, 84 taxa not in B, 18 taxa not in D



D) Yao2020 tree conflict
134 new tips. and taxa not in B., 68 taxa not in C



Conflict analysis:
 • Resolves
 • Agrees
 • Conflicts

Figure 2: A) Phylogenetic tree obtained by updating the Gottlieb et al. 2005 tree in B) using Physcraper.

Figure 2 caption continued: Tips in original alignment and new tips added with Physcraper are depicted in black and red, respectively. Physcraper obtained sequences from the GenBank database via local BLAST of all sequences in the original alignment that generated tree in B), filtered them following criteria specified in section “DNA sequence search and filtering”, aligned them to the original alignment using MUSCLE and performed a phylogenetic reconstruction using RAxML with 100 bootstraps. B-D show results of the conflict analysis performed with OpenTree tools.

4 Discussion

Databases that preserve and democratize access to biological data are essential resources for our scientific endeavor. New molecular data keep accumulating and tools facilitating its integration into existent evolutionary knowledge are essential.

Phylogenetic pipelines designed to make evolutionary sense of the vast amount of public molecular data (e.g., Phylota (Sanderson *et al.* 2008), PHLAWD (Smith *et al.* 2009), SUPERSMART (Antonelli *et al.* 2017)) focus on generating full phylogenies *de novo*, i.e., inferring phylogenetic relationships from a newly generated homology hypothesis, as opposed to e.g., supertrees, that are generated by summarizing previous phylogenetic estimates. While Physcraper does not generate phylogenies *de novo* in a traditional sense, it successfully generates new phylogenetic knowledge, revealing the importance of open science in facilitating phylogenetic placement of public molecular data and accelerating the enrichment and updating of phylogenetic relationships in any region of the tree of life. The PUMPER pipeline (Izquierdo-Carrasco *et al.* 2014) also uses the concept of updating pre-existing alignments to incorporate public molecular data into phylogenies. Unfortunately, installation of the tool was unsuccessful following instructions from the author, and we were unable to benchmark a comparison.

Physcraper generates individual gene trees, which fail to capture the complexity of species’ evolutionary history (Song *et al.* 2012). Yet, Physcraper makes it straightforward to gather alignments and gene trees for multiple loci from a taxonomic group of interest, which can then be used to perform coalescent analysis to generate species trees with e.g., ASTRAL (Mirarab *et al.* 2014), BEAST2 (Bouckaert *et al.* 2019), or SVD

Quartets (Chifman & Kubatko 2014)).

PhyScraper has the potential to link phylogenies to data available in any of the taxonomies integrated by the OpenTree taxonomy (Rees & Cranston 2017), such as geographical locations from the Global Biodiversity Information Facility, or fossils from the Paleobiology Database. The PhyScraper workflow can be used to rapidly (in a matter of hours) address challenges overarching both fields of ecology and evolution, such as placing newly discovered species phylogenetically (Webb *et al.* 2010), obtaining trees for ecophylogenetic studies (Helmus & Ives 2012), systematizing molecular (and other) databases, i.e., curating taxonomic assignments (San Mauro & Agorreta 2010), and generating custom trees for ecological and evolutionary downstream analyses (Stoltzfus *et al.* 2013).

Data repositories hold more information than meets the eye. Besides the main data, they are rich sources of metadata that can be leveraged for the advantage of all areas of biology as well as the advancement of scientific policy and applications. Usually, initial ideas about the data are changed by new analyses. PhyScraper can provide context for these ideas by streamlining inferences integrating new data with existing knowledge.

5 Acknowledgements

Research was supported by the grant “Sustaining the Open Tree of Life”, National Science Foundation ABI No. 1759838, and ABI No. 1759846. Computer time was provided by the Multi-Environment Research Computer for Exploration and Discovery (MERCED) cluster from the University of California, Merced (UCM), supported by the NSF Grant No. ACI-1429783.

We thank the members of the OpenTree development team and the “short bar” Science and Engineering Building 1, UCM, joint lab paper discussion group for valuable comments on this manuscript.

The authors have no conflict of interest to declare.

6 Authors' Contributions

LLSR wrote the manuscript, alignment code, documentation, performed analyses and developed examples; MK wrote code for ncbidataparser module, filtering of sequences per OTU and using offline blast searches, wrote documentation and tests; EJM conceived study, wrote most of the code, documentation and tests. All authors contributed to the manuscript and gave final approval for publication.

7 Data Archiving

Physcraper source code available at <https://github.com/McTavishLab/physcraper>

Documentation available at <https://physcraper.readthedocs.io/en/latest/index.html>

Illustrated examples available at <https://github.com/McTavishLab/physcraperex>

This is a reproducible manuscript available at https://github.com/McTavishLab/physcraper_ms

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Andermann, T., Torres Jiménez, M.F., Matos-Maraví, P., Batista, R., Blanco-Pastor, J.L., Gustafsson, A.L.S., Kistler, L., Liberal, I.M., Oxelman, B., Bacon, C.D. & Antonelli, A. (2020). A guide to carrying out a phylogenomic target sequence capture project. *Frontiers in Genetics*, **10**, 1–20.
- Antonelli, A., Hettling, H., Condamine, F.L., Vos, K., Nilsson, R.H., Sanderson, M.J., Sauquet, H., Scharn, R., Silvestro, D., Töpel, M. & others. (2017). Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of taxa. *Systematic Biology*, **66**, 152–166.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. & Wheeler, D.L. (2000). GenBank. *Nucleic Acids Research*, **28**, 15–18.

238 Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J.,
239 Jones, G., Kühnert, D., Maio, N.D., Matschiner, M., Mendes, F.K., Müller, N.F., Ogilvie, H.A., Plessis, L.
240 du, Poppinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M.A., Wu, C.-H., Xie, D., Zhang, C.,
241 Stadler, T. & Drummond, A.J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary
242 analysis. *PLOS Computational Biology*, **15**, e1006650.

243 Camacho, C., George, C., Vahram, A., Ning, M., Jason, P., Kevin, B. & Thomas, L. (2009). BLAST+:
244 Architecture and applications. *BMC Bioinformatics*, **10**, 421.

245 Chifman, J. & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinform-*
246 *atics*, **30**, 3317–3324.

247 Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff,
248 F., Wilczynski, B. & others. (2009). Biopython: freely available Python tools for computational molecular
249 biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

250 Cuénoud, P., Martinez, M.A. del P., Loizeay, P.-A., Spichiger, R., Andrews, S. & Manen, J.-F. (2000).
251 Molecular phylogeny and biogeography of the genus *Ilex* L.(Aquifoliaceae). *Annals of Botany*, **85**, 111–122.

252 Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic*
253 *Acids Research*, **32**, 1792–1797.

254 Fragoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Lemmon, E.M.,
255 Lemmon, A.R., Sazatornil, F. & Mendoza, C.G. (2017). A pilot study applying the plant Anchored Hybrid
256 Enrichment method to New World sages (*Salvia* subgenus Calosphace; Lamiaceae). *Molecular Phylogenetics*
257 *and Evolution*, **117**, 124–134.

258 Gottlieb, A.M., Giberti, G.C. & Poggio, L. (2005). Molecular analyses of the genus *Ilex* (Aquifoliaceae) in
259 southern south america, evidence from afp and its sequence data. *American Journal of Botany*, **92**, 352–369.

260 Helmus, M.R. & Ives, A.R. (2012). Phylogenetic diversity–area curves. *Ecology*, **93**, S31–S43.

261 Izquierdo-Carrasco, F., Cazes, J., Smith, S.A. & Stamatakis, A. (2014). PUmPER: Phylogenies updated
 262 perpetually. *Bioinformatics*, **30**, 1476–1477.

263 Manen, J.-F., Barriera, G., Loizeau, P.-A. & Naciri, Y. (2010). The history of extant *Ilex* species (Aquifoli-
 264 aceae): evidence of hybridization within a Miocene radiation. *Molecular Phylogenetics and Evolution*, **57**,
 265 961–977.

266 Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S. & Warnow, T. (2014). ASTRAL:
 267 Genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548.

268 Morrison, D.A. (2006). Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*,
 269 **19**, 479–539.

270 Piel, W., Chan, L., Dominus, M., Ruan, J., Vos, R. & Tannen, V. (2009). Treebase v. 2: A database of
 271 phylogenetic knowledge. E-biosphere.

272 Redelings, B.D. & Holder, M.T. (2017). A supertree pipeline for summarizing phylogenetic and taxonomic
 273 information for millions of species. *PeerJ*, **5**, e3058.

274 Rees, J.A. & Cranston, K. (2017). Automated assembly of a reference taxonomy for phylogenetic data
 275 synthesis. *Biodiversity Data Journal*.

276 Sanderson, M.J., Boss, D., Chen, D., Cranston, K.A. & Wehe, A. (2008). The PhyLoTA Browser: Processing
 277 GenBank for Molecular Phylogenetics Research. *Systematic Biology*, **57**, 335–346.

278 San Mauro, D. & Agorreta, A. (2010). Molecular systematics: A synthesis of the common methods and the
 279 state of knowledge. *Cellular & Molecular Biology Letters*, **15**, 311.

280 Selbach-Schnadelbach, A., Cavalli, S.S., Manen, J.-F., Coelho, G.C. & De Souza-Chies, T.T. (2009). New
 281 information for *Ilex* phylogenetics based on the plastid psbA-trnH intergenic spacer (Aquifoliaceae). *Botanical*
 282 *Journal of the Linnean Society*, **159**, 182–193.

283 Setoguchi, H. & Watanabe, I. (2000). Intersectional gene flow between insular endemics of *Ilex* (Aquifoliaceae)
284 on the Bonin Islands and the Ryukyu Islands. *American Journal of Botany*, **87**, 793–810.

285 Smith, S.A., Beaulieu, J.M. & Donoghue, M.J. (2009). Mega-phylogeny approach for comparative biology:
286 An alternative to supertree and supermatrix approaches. *BMC Evolutionary Biology*, **9**, 37.

287 Song, S., Liu, L., Edwards, S.V. & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using
288 phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, **109**,
289 14942–14947.

290 Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
291 phylogenies. *Bioinformatics*, **30**, 1312–1313.

292 Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C.M., Vaidya, G., Pontelli, E.,
293 Cranston, K., Vos, R. & others. (2013). Phylotastic! Making tree-of-life knowledge accessible, reusable and
294 convenient. *BMC Bioinformatics*, **14**, 158.

295 Sukumaran, J. & Holder, M.T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinform-*
296 *atics*, **26**, 1569–1571.

297 Webb, C.O., Slik, J.F. & Triono, T. (2010). Biodiversity inventory and informatics in Southeast Asia.
298 *Biodiversity and Conservation*, **19**, 955–972.

299 Yao, X., Song, Y., Yang, J.-B., Tan, Y.-H. & Corlett, R.T. (2020). Phylogeny and biogeography of the hollies
300 (*Ilex* L., Aquifoliaceae). *Journal of Systematics and Evolution*, **58**, 1–10.