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

1. Luna L. Sanchez Reyes

School of Natural Sciences, University of California, Merced

email: sanchez.reyes.luna@gmail.com

2. Martha Kandziora116*0.57

School of Natural Sciences, University of California, Merced

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

email: kandziom@natur.cuni.cz

3. Emily Jane McTavish

School of Natural Sciences, University of California, Merced

email: ejmctavish@ucmerced.edu

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

Correspondence email: sanchez.reyes.luna@gmail.com, ejmctavish@ucmerced.edu

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

Word count: 2969

Manuscript prepared for submission to Methods in Ecology and Evolution

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 selection 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 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. Updating existing phylogenies with molecular data that has never been incorporated into any phylogenetic estimate, geographical location, fossils, and other data in a reproducible and continuous manner is possible by establishing a data interoperability framework for biological databases. Here, we introduce Physcraper, a tool that automates database connections to build upon homology hypotheses that taxon specialists have assessed and deemed appropriate for a specific phylogenetic scope to update a starting tree and single locus alignments with public DNA data.

Taxonomic idiosyncrasies across databases represent a huge challenge for automatic integration of data into phylogenies, which 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 and taxonomy. 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 taxonomic data from various databases [1], including the USA National Center for Biodiversity Information (NCBI) molecular database GenBank [2], among others. The 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 it uses 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 loci [3]. Yet, decades of single locus sequencing have generated massive amounts of homologous DNA datasets that can be used for phylogenetic reconstruction at many scales.

More than a decade ago, GenBank release 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 [4]. 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 mine DNA databases fast, efficiently, and reproducibly, have been applied to infer phylogenetic relationships in a variety of organisms [5–7]. However, fine-grained curated markers and alignments can improve phylogenetic reconstructions, even in phylogenomic analyses [8].

There are almost 8,200 publicly available, peer-reviewed alignments, covering around 100,000 distinct taxa in the TreeBASE database [9], which can be used as seeds to mine molecular databases, and as "jump-start" alignments for phylogenetic reconstructions [10] to continually enrich, update and compare existing phylogenetic knowledge.

Physcraper is a Python pipeline using OpenTree's taxonomy and programmatic access protocols (API's) to implement a database interoperability framework that automatically links phylogenies that have been standardized to OpenTree taxonomy, to alignments from TreeBASE, data from GenBank, and phylogenies from OpenTree's Phylesystem. Physcraper aims to demonstrate the benefits of reproducible workflows and open science in phylogenetics, and encourage better data sharing practices in the community.

2 The Physcraper framework

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

2.1 The inputs: a tree and an alignment

Taxon names in the input tree must be standardized to OpenTree taxonomy [1] 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

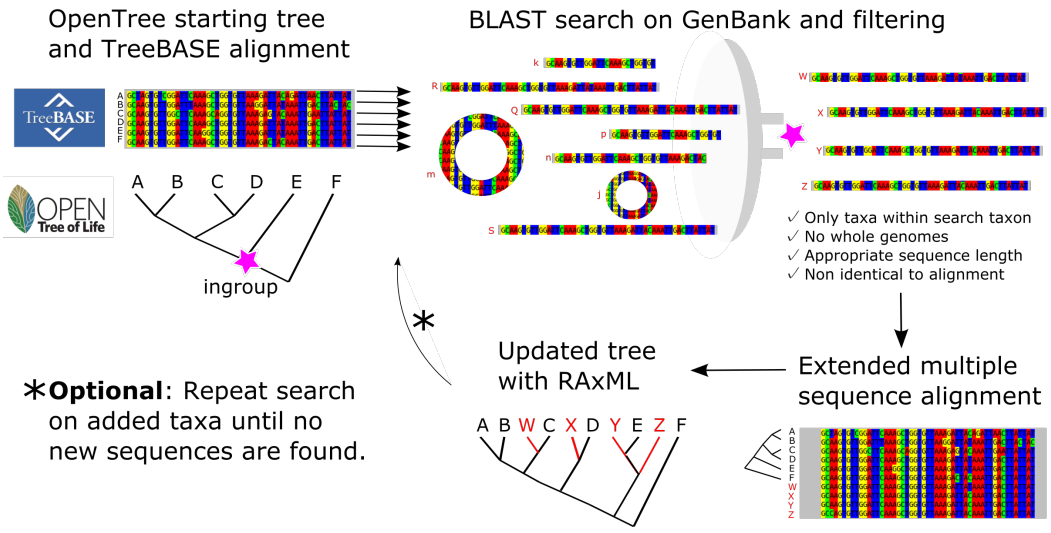

Figure 1: The Physcraper framework consists of 4 steps (see text). The software is fully described on its documentation website at physcraper.readthedocs.io, along with installation instructions, function usage descriptions, examples and tutorials.

[9].

The input alignment is a single locus DNA dataset that was used in part or in whole to generate the input tree. Physcraper retrieves TreeBASE alignments automatically. Alternatively, users 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 [11] 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 [1], as the most recent common ancestor of ingroup taxa 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 taxa.

BLAST is implemented with the `blastn` function [12] and the BioPython [13] BLAST function from 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 they 1)

are not in the search taxon; 2) have an e-value above the cutoff (default to 0.00001); 3) fall outside a min and max length threshold, defined as the proportion of the average length without gaps of all sequences in 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 OpenTree or 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 [13]. 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 [14] 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

RAxML [15] is implemented to reconstruct a Maximum Likelihood (ML) gene tree for each input alignment with default settings (GTRCAT model and 100 bootstrap replicates with default algorithm), using input tree as starting tree for ML searches. Bootstrap results are summarized using DendroPy's SumTrees module [16].

Physcraper's main result is an updated phylogenetic hypothesis for the search taxon. Updated and original tree are compared with Robinson-Foulds weighted and unweighted metrics estimated with Dendropy [16], and with a node by node comparison between the synthetic OpenTree and original and updated tree individually, using OpenTree's conflict API [17].

3 Case Study: The hollies

A user is interested in 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 phylogenies showing relationships within the hollies [18–21], but only two have data publicly available [22, 23]. [22] made original tree and alignment available in TreeBASE. The "Gottlieb2005" tree sampling 41 species was added to OpenTree Phylesystem and it has been integrated into OpenTree's synthetic tree.

The most recent *Ilex* tree [23] is available in OpenTree Phylesystem and in the DRYAD repository. With 175 tips, the "Yao2020" tree is the best sampled phylogeny yet available for the hollies.

We ran Physcraper on a laptop Linux computer to update an internal transcribed spacer DNA region (ITS) alignment from [22], using a local GenBank database. BLAST and RAxML analyses ran for 19hrs 45min, with bootstrap analyses taking an additional 13hrs. The updated Gottlieb2005 tree (Fig. 2) displays 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, Yao2020 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 updated Gottlieb2005. While [23] also used ITS as a marker, their GenBank data is not released 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.

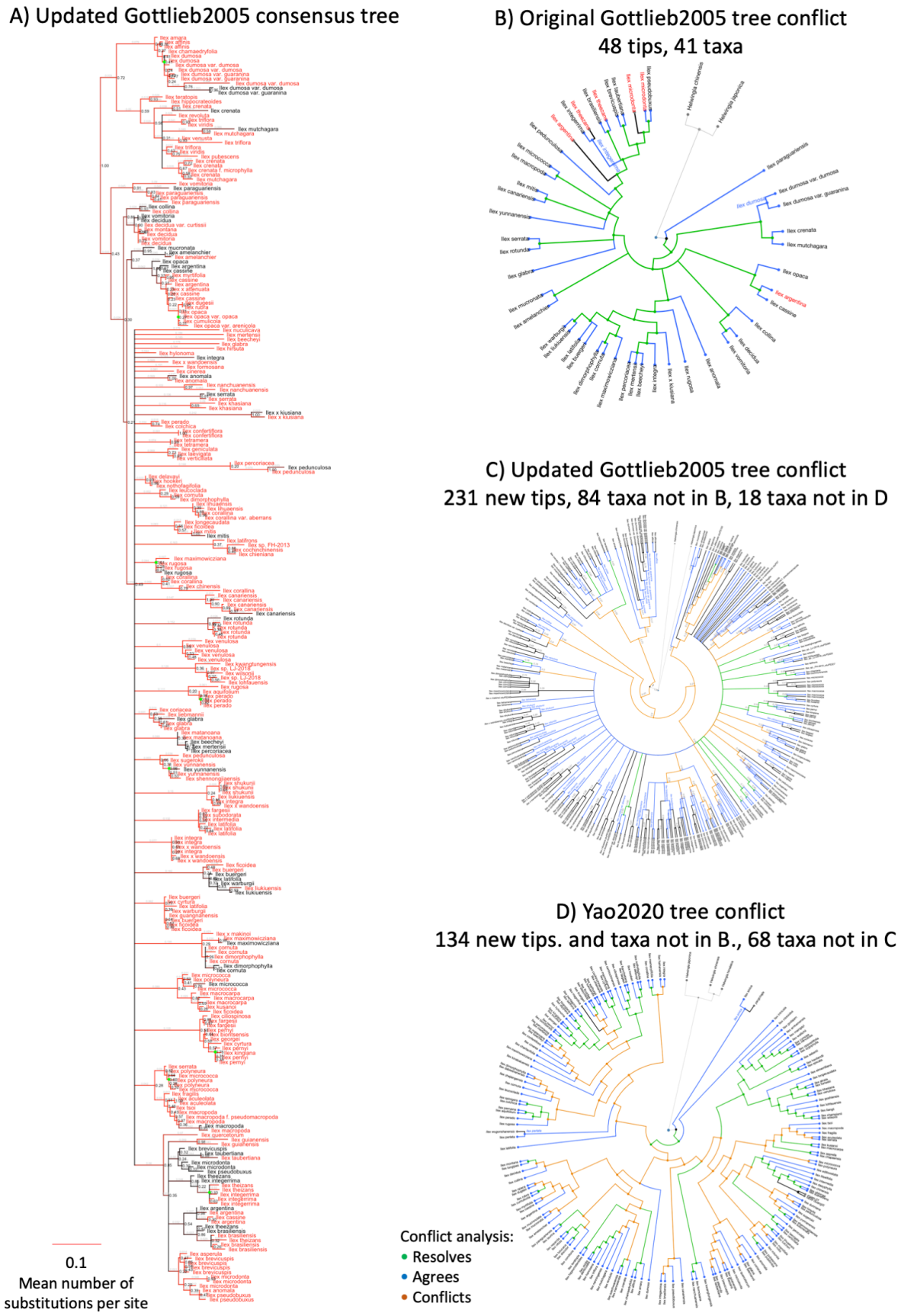

Figure 2: A) Phylogeny updated with Physcraper from Gottlieb et al. 2005 tree in B.

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 from section "DNA sequence search and filtering", aligned them to original alignment using MUSCLE and performed a phylogenetic reconstruction using RAxML with 100 bootstraps. B-D conflict analyses performed with OpenTree tools.

4 Discussion

Databases preserving and democratizing access to biological data have become essential resources for science. New molecular data keep accumulating and tools facilitating its integration into existent evolutionary knowledge are needed.

Phylogenetic pipelines designed to make evolutionary sense of the vast amount of public molecular data (e.g., Phylota [4], PHLAWD [5], SUPERSMART [6]) 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 enrichment and updating of phylogenetic relationships in any region of the tree of life. The PUMPER pipeline [7] also uses the concept of updating pre-existing alignments to incorporate public molecular data into phylogenies. Unfortunately, installation was unsuccessful following instructions from the author, and a comparison analysis was unfeasible.

Physcraper generates individual gene trees, failing to capture the complexity of species' evolutionary history [24]. Yet, Physcraper facilitates gathering alignments and gene trees for multiple loci from a group of interest, that can be used to reconstruct species trees with ASTRAL [25], BEAST2 [26], or SVD Quartets [27]).

Physcraper can potentially link phylogenies to data available in any of the taxonomies integrated in the

OpenTree taxonomy [1], 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 [28], systematizing molecular (and other) databases, i.e., curating taxonomic assignations [29], and generating custom trees for ecological [30] and evolutionary downstream analyses [31].

Data repositories hold more information than meets the eye. Beyond 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. Initial ideas about the data are constantly changed by results from new analyses. Physcraper provides a framework for reproducible phylogenetics that has the potential to consistently provide context for these ideas, highlighting the importance of data sharing and open science in the field, biology and science.

5 Acknowledgements

Research was supported by the grant "Sustaining the Open Tree of Life", NSF 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 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: https://github.com/McTavishLab/physcraper

Documentation: https://physcraper.readthedocs.io/en/latest/index.html

Examples: https://github.com/McTavishLab/physcraperex

Reproducible manuscript: https://github.com/McTavishLab/physcraper_ms

References

1. Rees JA, Cranston K. Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodiversity Data Journal. 2017.

2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Research. 2000;28:15–8.

3. Andermann T, Torres Jiménez MF, Matos-Maraví P, Batista R, Blanco-Pastor JL, Gustafsson ALS, et al. A guide to carrying out a phylogenomic target sequence capture project. Frontiers in Genetics. 2020;10:1–20.

4. Sanderson MJ, Boss D, Chen D, Cranston KA, Wehe A. The PhyLoTA Browser: Processing GenBank for Molecular Phylogenetics Research. Systematic Biology. 2008;57:335–46.

5. Smith SA, Beaulieu JM, Donoghue MJ. Mega-phylogeny approach for comparative biology: An alternative to supertree and supermatrix approaches. BMC Evolutionary Biology. 2009;9:37.

6. Antonelli A, Hettling H, Condamine FL, Vos K, Nilsson RH, Sanderson MJ, et al. Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of taxa. Systematic Biology. 2017;66:152–66.

7. Izquierdo-Carrasco F, Cazes J, Smith SA, Stamatakis A. PUmPER: Phylogenies updated perpetually. Bioinformatics. 2014;30:1476–7.

8. Fragoso-Martínez I, Salazar GA, Martínez-Gordillo M, Magallón S, Sánchez-Reyes L, Lemmon EM, et al. A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus Calosphace; Lamiaceae). Molecular Phylogenetics and Evolution. 2017;117:124–34.

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

10. Morrison DA. Multiple sequence alignment for phylogenetic purposes. Australian Systematic Botany. 2006;19:479–539.

11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215:403–10.

12. Camacho C, George C, Vahram A, Ning M, Jason P, Kevin B, et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009;10:421.

13. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.

14. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32:1792–7.

15. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

16. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. Bioinformatics. 2010;26:1569–71.

17. Redelings BD, Holder MT. A supertree pipeline for summarizing phylogenetic and taxonomic information

for millions of species. PeerJ. 2017;5:e3058.

18. Cuénoud P, Martinez MA del P, Loizeay P-A, Spichiger R, Andrews S, Manen J-F. Molecular phylogeny and biogeography of the genus *Ilex* L.(Aquifoliaceae). Annals of Botany. 2000;85:111–22.

19. Manen J-F, Barriera G, Loizeau P-A, Naciri Y. The history of extant *Ilex* species (Aquifoliaceae): evidence of hybridization within a Miocene radiation. Molecular Phylogenetics and Evolution. 2010;57:961–77.

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

21. Selbach-Schnadelbach A, Cavalli SS, Manen J-F, Coelho GC, De Souza-Chies TT. New information for *Ilex* phylogenetics based on the plastid psbA-trnH intergenic spacer (Aquifoliaceae). Botanical Journal of the Linnean Society. 2009;159:182–93.

22. Gottlieb AM, Giberti GC, Poggio L. Molecular analyses of the genus *Ilex* (Aquifoliaceae) in southern south america, evidence from aflp and its sequence data. American Journal of Botany. 2005;92:352–69.

23. Yao X, Song Y, Yang J-B, Tan Y-H, Corlett RT. Phylogeny and biogeography of the hollies (*Ilex* L., Aquifoliaceae). Journal of Systematics and Evolution. 2020;58:1–10.

24. Song S, Liu L, Edwards SV, Wu S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences. 2012;109:14942–7.

25. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics. 2014;30:i541–8.

26. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology. 2019;15:e1006650.

27. Chifman J, Kubatko L. Quartet inference from SNP data under the coalescent model. Bioinformatics.

2014;30:3317–24.

28. Webb CO, Slik JF, Triono T. Biodiversity inventory and informatics in Southeast Asia. Biodiversity and Conservation. 2010;19:955–72.

29. San Mauro D, Agorreta A. Molecular systematics: A synthesis of the common methods and the state of knowledge. Cellular & Molecular Biology Letters. 2010;15:311.

30. Helmus MR, Ives AR. Phylogenetic diversity–area curves. Ecology. 2012;93:S31–43.

31. Stoltzfus A, Lapp H, Matasci N, Deus H, Sidlauskas B, Zmasek CM, et al. Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. BMC Bioinformatics. 2013;14:158.