RESEARCH

Physcraper: A Python package for continually updated phylogenetic trees using the Open Tree of Life

Luna L. Sanchez Reyes[1*], Martha Kandziora[1,2] and Emily Jane McTavish[1*]

[*]Correspondence:
sanchez.reyes.luna@gmail.com;
ejmctavish@ucmerced.edu
[1]School of Natural Sciences,
University of California, Merced,
USA
Full list of author information is
available at the end of the article

Abstract

Background: 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 non-specialists. 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.

Results: 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. The program 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).

Conclusions: The Physcraper workflow showcases 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 usage are available at https://physcraper.readthedocs.io.

Keywords: gene tree; interoperability; open science; reproducibility; public database; DNA alignment

Background

Phylogenies capture the shared history of organisms and provide key evolutionary context for our biological observations [1]. Updating existing phylogenies with molecular sequence data that has never been incorporated into any phylogenetic estimate provides the opportunity to capture the evolutionary history of many taxa. Here, we introduce Physcraper, a tool that automates database connections to build on published alignments to extend phylogenetic inferences with more taxa. Physcraper can be used to update a starting tree and single locus alignments with public DNA data, and link the tips in these trees to a unified, interoperable taxonomic resource [2].

Data such as geographical location, fossil ranges, and genetic and phenotypic information increasingly available in public databases constitute an amazing resource for biological discovery [3]. One of the main challenges for automatic integration of biological data across databases are varying taxonomic idiosyncrasies. To address this challenge, the Open Tree of Life project (OpenTree) created a unified taxonomy for name standardization, by integrating taxonomic data from several databases [2], including the USA National Center for Biodiversity Information (NCBI) taxonomy [4, 5], and the Global Biodiversity Information Facility (GBIF) [6] among others. By using the existing OpenTree taxonomy programmatic tools to map tip names, Physcraper has a framework for connecting updated phylogenies data from any biological database.

Decades of single locus sequencing have generated massive amounts of homologous DNA datasets that have the potential to be used for phylogenetic reconstruction at many scales [7]. 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 [7]. However, even thirteen years later, phylogenetic estimates for many of these taxa are not available [8]. OpenTree's comprehensive tree of life 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. There is a considerable amount of phylogenetically informative data in GenBank with the potential to fill these phylogenetic gaps in the tree of life that, but this data either has not been analysed or the analyses have not been made publicly available [8].

Assembling a DNA alignment from such a massive database can be done "by hand", but that is a time consuming approach which is not highly reproducible. A variety of computational pipelines that mine DNA databases fast, efficiently, and reproducibly have been developed and used to infer phylogenetic relationships in a variety of organisms (e.g., [9, 10, 11]). While genomics has, and will continue to, revolutionize phylogenetic inference, the diversity of alternative genomic sequencing approaches implemented produce largely non-overlapping homology hypotheses across taxa [12], creating challenges for phylogenetic reconstruction. Phylogenomics addresses this problem by focusing on targeted capture of informative regions [13]. However, fine-grained curated markers and alignments can significantly improve phylogenetic reconstructions, even in phylogenomic analyses [14].

Physcraper improves on previous work in automating phylogenetics by leveraging the power of existing homology hypotheses that taxon specialists have assessed and deemed appropriate for a specific phylogenetic scope. There are almost 8,200 publicly available, peer-reviewed curated alignments, covering around 100,000 distinct taxa in the TreeBASE database [15], which can be leveraged as seeds to mine molecular databases, and as "jump-start" alignments for phylogenetic reconstructions [16] to continually enrich, update and compare existing phylogenetic knowledge.

Physcraper is a Python pipeline using OpenTree's programmatic access protocols (API's) to automatically link phylogenies mapped to a standardized taxonomy [17], to alignments from TreeBASE [18], data from GenBank [19]. Physcraper provides automated and reproducible workflows for updating existing phylogenetic estimates.

Implementation

The general Physcraper framework (Figure 1) consists of 4 steps: 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.

The inputs: a tree and an alignment

Taxon names in the input tree must be standardized to OpenTree taxonomy [2] using OpenTree's bulk Taxonomic Name Resolution Service (TNRS) tool [20]. Users can upload their own tree, or choose from among the 2, 950 standardized trees stored in OpenTree's Phylesystem [21, 17] that also have alignments available on TreeBASE [15].

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.

DNA sequence search and filtering

The Basic Local Alignment Search Tool, BLAST [22] 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's taxonomic Most Recent Common Ancestor (MRCA) API [23, 2], as the MRCA of all ingroup taxa that is also a named clade in the NCBI taxonomy (Figure 1).

BLAST is performed using the 'blastn' algorithm [24] implemented in BioPython's [25] NCBIWWW module [26] 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 80respectively); 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 [25]. 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.

New DNA sequence alignment

MUSCLE [27] 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.

Tree reconstruction and comparison

RAxML [28] 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 [29].

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 [29], and with a node by node comparison between the synthetic OpenTree and original and updated tree individually, using OpenTree's conflict API [30].

Results

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 [31, 32, 33, 34], but only two have data publicly available [35, 36]. [35] made original tree and alignment available in TreeBASE (study 1091 [37]). The tree sampling 41 species is also available from OpenTree's Phylesystem (study pg_2827 [38]), and has been integrated into OpenTree's synthetic tree [39]. The most recent *Ilex* tree [36] is available in OpenTree's Phylesystem (study ot_1984 [40]), and in the DRYAD repository [41]. With 175 tips, the [36] 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 [35], using a local GenBank database. BLAST and RAxML analyses ran for 19hrs 45min, with bootstrap analyses taking an additional 13hrs. The updated [35] tree (Figure 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 (Figure 2). For comparison, [36] also contains all 41 distinct taxa from the original [35] study, and contributes phylogenetic data to 134 additional *Ilex* taxa, from which 67 are also in updated [35]. While [36] 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 [36].

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 contribute to the acceleration of scientific discovery.

We believe this is a key step to successfully establish an open, reproducible workflow for phylogenetics, facilitating phylogenetic knowledge for ecologists and other non-specialists, effectively democratizing phylogenetic studies.

As a tool for automatizing phylogenetic reconstruction from molecular databases, Physcraper presents several advantages over existing phylogenetic pipelines designed to make evolutionary sense of the vast amount of public molecular data available.

Several analysis tools create full phylogenies *de novo* by mining of molecular databases [10, 7, 42, 43, 44]. In particular, Phylota [7], and PHLAWD [9], have been cited and used abundantly.

Physcraper builds on this automated database mining concept by incorporating prior phylogenetic work and existing taxonomic domain knowledge on appropriate markers and alignment construction. This decreases error (requiring more manual downstream processing) and eases comparison with previous phylogenetic knowledge. Unlike placement approaches [45, 46], which add new taxa without modifying the input tree, Physcraper estimates all the relationships in the context of the new data. PUMPER [11] shares these conceptual strengths, but is no longer under active development, is challenging to install and run, and has resulted in very few phylogenetic analyses since its publication.

Physcraper generates gene trees, which individually do not capture the full complexity of species' evolutionary history [47]. However, physcraper facilitates gathering alignments and gene trees for multiple loci from a group of interest, that together can be used to reconstruct species trees taking into account coalescent processes with ASTRAL [48], BEAST2 [49], or SVD Quartets [50]). Rigorous analyses of multiple loci allows for more complex evolutionary models than analyses of large genomic data sets, and can provide better evolutionary estimates.

Physcraper has the added advantage of linking taxonomic information about tips in the output phylogenies to data available in a variety of biological databases [2], such as geographical locations for taxa from the GBIF [6].

The Physcraper workflow can be used to rapidly (in a matter of hours) to create phylogenies which can address challenges overarching both fields of ecology and evolution, such as phylogenetically placing newly discovered species [51], curating taxonomic assignments [52], and generating custom trees for ecological [53] and evolutionary downstream analyses [54].

Conclusions

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.

Availability and requirements

Project name: Physcraper

Project home page: https://physcraper.readthedocs.io/en/latest/index.html
Operating System: Linux, Mac, Windows
Programming Language: Python
Other requirements: Dependencies
License: GNU
Any restrictions to use by non-academics: As specified by the License

Abbreviations

OpenTree: The Open Tree of Life project
  TNRS: Taxonomic Name Resolution Service
  MRCA: Most Recent Common Ancestor
  BLAST: Basic Local Alignment Search Tool
  NCBI: USA National Center for Biodiversity Information
  GBIF: Global Biodiversity Information Facility

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The datasets generated and analysed during the current study are available in the repositories "physcraper" containing the source code, https://github.com/McTavishLab/physcraper; "physcraperex" containing the examples, https://github.com/McTavishLab/physcraperex; and, "physcraper_ms" containing this reproducible manuscript, https://github.com/McTavishLab/physcraper_ms.

Competing interests
The authors declare that they have no competing interests.

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. . . .

Author details
[1]School of Natural Sciences, University of California, Merced, USA. [2]Department of Botany, Faculty of Science, Charles University, Prague , Czech Republic.

References
1. Dobzhansky, T.: Nothing in biology makes sense except in the light of evolution. The american biology teacher **35**(3), 125–129 (1973)
2. Rees, J.A., Cranston, K.: Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodiversity Data Journal (5) (2017). doi:10.3897/BDJ.5.e12581
3. Baxevanis, A.D., Bateman, A.: The importance of biological databases in biological discovery. Current Protocols in Bioinformatics **50**(1), 1–1 (2015)
4. Federhen, S.: The NCBI Taxonomy database. Nucleic Acids Research **40**(D1), 136–143 (2012). doi:10.1093/nar/gkr1178. Publisher: Oxford Academic. Accessed 2020-12-03
5. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I.: NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database **2020** (2020)
6. Secretariat, G.: GBIF Backbone Taxonomy (2019). doi:10.15468/39omei. Accessed 2020-12-03

7. Sanderson, M.J., Boss, D., Chen, D., Cranston, K.A., Wehe, A.: The PhyLoTA Browser: Processing GenBank for Molecular Phylogenetics Research. Systematic Biology **57**(3), 335–346 (2008). doi:10.1080/10635150802158688. https://academic.oup.com/sysbio/article-pdf/57/3/335/24203605/57-3-335.pdf

8. McTavish, E.J., Drew, B.T., Redelings, B., Cranston, K.A.: How and Why to Build a Unified Tree of Life. BioEssays **39**(11) (2017). doi:10.1002/bies.201700114. Number: 11. Accessed 2018-04-10

9. Smith, S.A., Beaulieu, J.M., Donoghue, M.J.: Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evolutionary Biology **9**(1), 37 (2009). doi:10.1186/1471-2148-9-37

10. Antonelli, A., Hettling, H., Condamine, F.L., Vos, K., Nilsson, R.H., Sanderson, M.J., Sauquet, H., Scharn, R., Silvestro, D., Töpel, M., *et al.*: Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of taxa. Systematic Biology **66**(2), 152–166 (2017). doi:10.1093/sysbio/syw066

11. Izquierdo-Carrasco, F., Cazes, J., Smith, S.A., Stamatakis, A.: Pumper: phylogenies updated perpetually. Bioinformatics **30**(10), 1476–1477 (2014). doi:10.1093/bioinformatics/btu053

12. Jones, M.R., Good, J.M.: Targeted capture in evolutionary and ecological genomics. Molecular Ecology **25**(1), 185–202 (2016). doi:10.1111/mec.13304

13. 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.: A guide to carrying out a phylogenomic target sequence capture project. Frontiers in Genetics **10**(1407), 1–20 (2020). doi:10.3389/fgene.2019.01407

14. Fragoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Lemmon, E.M., Lemmon, A.R., Sazatornil, F., Mendoza, C.G.: A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus Calosphace; Lamiaceae). Molecular Phylogenetics and Evolution **117**, 124–134 (2017). doi:10.1016/j.ympev.2017.02.006

15. Piel, W., Chan, L., Dominus, M., Ruan, J., Vos, R., Tannen, V.: Treebase v. 2: A database of phylogenetic knowledge. e-Biosphere. London (2009)

16. Morrison, D.A.: Multiple sequence alignment for phylogenetic purposes. Australian Systematic Botany **19**(6), 479–539 (2006). doi:10.1071/SB06020

17. McTavish, E.J., Hinchliff, C.E., Allman, J.F., Brown, J.W., Cranston, K.A., Holder, M.T., Rees, J.A., Smith, S.A.: Phylesystem: A git-based data store for community-curated phylogenetic estimates. Bioinformatics **31**(17), 2794–2800 (2015). doi:10.1093/bioinformatics/btv276

18. Sanderson, M., Donoghue, M., Piel, W., Eriksson, T.: Treebase: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of Botany **81**(6), 183 (1994). doi:publication does not exist

19. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L.: Genbank. Nucleic Acids Research **28**(1), 15–18 (2000). doi:10.1093/nar/28.1.15

20. OpenTreeOfLife: Name Resolution (TNRS) bulk mapping tool. https://tree.opentreeoflife.org/curator/tnrs/

21. OpenTreeOfLife, E.J. McTavish, Hinchliff, C.E., Allman, J.F., Brown, J.W., Cranston, K.A., Holder, M.T., Rees, J.A., Smith, S.A.: Phylesystem's top-level repository in the Open Tree of Life phylogenetic study document store. https://github.com/opentreeoflife/phylesystem

22. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2

23. OpenTreeOfLife, Rees, J.A., Cranston, K.: OpenTree's taxonomic MRCA API. https://github.com/OpenTreeOfLife/germinator/wiki/Taxonomy-API-v3#mrca

24. Camacho, C., George, C., Vahram, A., Ning, M., Jason, P., Kevin, B., Thomas, L.: BLAST+: Architecture and applications. BMC Bioinformatics **10**(1), 421 (2009). doi:10.1186/1471-2105-10-421

25. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.*: Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11), 1422–1423 (2009). doi:10.1093/bioinformatics/btp163

26. The BioPython Contributors, .-.: BioPython's Bio.Blast.NCBIWWW module. https://biopython.org/DIST/docs/api/Bio.Blast.NCBIWWW-module.html

27. Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32**(5), 1792–1797 (2004). doi:10.1093/nar/gkh340

28. Stamatakis, A.: Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9), 1312–1313 (2014). doi:10.1093/bioinformatics/btu033

29. Sukumaran, J., Holder, M.T.: DendroPy: a Python library for phylogenetic computing. Bioinformatics **26**(12), 1569–1571 (2010). doi:10.1093/bioinformatics/btq228

30. Redelings, B.D., Holder, M.T.: A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. PeerJ **5**, 3058 (2017). doi:10.7717/peerj.3058

31. Cuénoud, P., Martinez, M.A.d.P., Loizeay, P.-A., Spichiger, R., Andrews, S., Manen, J.-F.: Molecular phylogeny and biogeography of the genus *ilex* l.(aquifoliaceae). Annals of Botany **85**(1), 111–122 (2000). doi:10.1006/anbo.1999.1003

32. 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 **57**(3), 961–977 (2010). doi:10.1016/j.ympev.2010.09.006

33. 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 **87**(6), 793–810 (2000). doi:10.2307/2656887

34. Selbach-Schnadelbach, A., Cavalli, S.S., Manen, J.-F., Coelho, G.C., De Souza-Chies, T.T.: New information for *Ilex* phylogenetics based on the plastid psbA-trnH intergenic spacer (Aquifoliaceae). Botanical Journal of the Linnean Society **159**(1), 182–193 (2009). doi:10.1111/j.1095-8339.2008.00898.x

35. Gottlieb, A.M., Giberti, G.C., Poggio, L.: Molecular analyses of the genus *ilex* (aquifoliaceae) in southern south

america, evidence from aflp and its sequence data. American Journal of Botany **92**(2), 352–369 (2005). doi:10.3732/ajb.92.2.352

36. Yao, X., Song, Y., Yang, J.-B., Tan, Y.-H., Corlett, R.T.: Phylogeny and biogeography of the hollies (*ilex* l., aquifoliaceae). Journal of Systematics and Evolution **58**(5), 1–10 (2020). doi:10.1111/jse.12567

37. Gottlieb, A.M., Giberti, G.C., Poggio, L.: TreeBASE study 1091. https://treebase.org/treebase-web/search/study/summary.html?id=1091

38. Gottlieb, A.M., Giberti, G.C., Poggio, L.: Phylesystem study pg_2827. https://tree.opentreeoflife.org/curator/study/edit/pg_2827/?tab=home

39. OpenTreeOfLife, Redelings, B., Reyes, L.L.S., Cranston, K.A., Allman, J., Holder, M.T., McTavish, E.J.: Open Tree of Life Synthetic subtree, node id mrcaott68451ott89474. https://tree.opentreeoflife.org/opentree/opentree12.3@mrcaott68451ott89474/Ilex-theizans--Ilex-dumosa

40. Yao, X., Song, Y., Yang, J.-B., Tan, Y.-H., Corlett, R.T.: Phylesystem study ot_1984. https://tree.opentreeoflife.org/curator/study/view/ot_1984

41. Yao, X., Song, Y., Yang, J.-B., Tan, Y.-H., Corlett, R.T.: Phylogeny and biogeography of the hollies (*Ilex* L., Aquifoliaceae), Dryad, Dataset. https://datadryad.org/stash/dataset/doi:10.5061/dryad.k0p2ngf4x. Accessed April 2020

42. Smith, S.A., Walker, J.F.: Pyphlawd: A python tool for phylogenetic dataset construction. Methods in Ecology and Evolution **10**(1), 104–108 (2019). doi:10.1111/2041-210X.13096

43. Bennett, D.J., Hettling, H., Silvestro, D., Zizka, A., Bacon, C.D., Faurby, S., Vos, R.A., Antonelli, A.: phylotar: An automated pipeline for retrieving orthologous dna sequences from genbank in r. Life **8**(2), 20 (2018). doi:10.3390/life8020020

44. Pearse, W.D., Purvis, A.: phylogenerator: an automated phylogeny generation tool for ecologists. Methods in Ecology and Evolution **4**(7), 692–698 (2013)

45. Berger, S.A., Krompass, D., Stamatakis, A.: Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. Systematic Biology, 010 (2011). doi:10.1093/sysbio/syr010. Accessed 2015-01-12

46. Matsen, F., Kodner, R., Armbrust, E.V.: pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC bioinformatics **11**(1), 538 (2010). Number: 1. Accessed 2013-12-17

47. Song, S., Liu, L., Edwards, S.V., Wu, S.: Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences **109**(37), 14942–14947 (2012). doi:10.1073/pnas.1211733109

48. Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T.: ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics **30**(17), 541–548 (2014). doi:10.1093/bioinformatics/btu462

49. Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N.D., Matschiner, M., Mendes, F.K., Müller, N.F., Ogilvie, H.A., Plessis, L.d., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M.A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., Drummond, A.J.: BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology **15**(4), 1006650 (2019). doi:10.1371/journal.pcbi.1006650. Publisher: Public Library of Science

50. Chifman, J., Kubatko, L.: Quartet inference from SNP data under the coalescent model. Bioinformatics **30**(23), 3317–3324 (2014). doi:10.1093/bioinformatics/btu530. Publisher: Oxford Academic

51. Webb, C.O., Slik, J.F., Triono, T.: Biodiversity inventory and informatics in Southeast Asia. Biodiversity and Conservation **19**(4), 955–972 (2010). doi:10.1007/s10531-010-9817-x

52. San Mauro, D., Agorreta, A.: Molecular systematics: a synthesis of the common methods and the state of knowledge. Cellular & Molecular Biology Letters **15**(2), 311 (2010). doi:10.2478/s11658-010-0010-8

53. Helmus, M.R., Ives, A.R.: Phylogenetic diversity–area curves. Ecology **93**(sp8), 31–43 (2012). doi:10.1890/11-0435.1

54. Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C.M., Vaidya, G., Pontelli, E., Cranston, K., Vos, R., *et al.*: Phylotastic! making tree-of-life knowledge accessible, reusable and convenient. BMC Bioinformatics **14**(1), 158 (2013). doi:10.1186/1471-2105-14-158

Figures

Figure 1 The Physcraper framework consists of four general steps. The software is fully described on its documentation website at https://physcraper.readthedocs.io, along with installation instructions, function usage descriptions, examples and tutorials.

Figure 2 A) Phylogeny updated with Physcraper from original [35] tree in B. 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.