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:
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; gene phylogeny; multilocus; interoperability; open science; reproducibility; public database; DNA alignment; Open Tree of Life; otol

Background

Phylogenies capture the shared history of organisms and provide key evolutionary context for our biological observations [1]. Updating existing phylogenies with publicly available molecular sequence data ~~that has never been incorporated into any phylogenetic estimate~~ provides the opportunity to simultaneously study the evolutionary history of many taxa in a reproducible and continuous manner. Here, we introduce Physcraper, a tool that establishes a data interoperability framework for biological databases to automate data connections across ~~databases, with the main goal of building on~~ biological databases. Physcraper's main goal is to build

upon published alignments and ~~extending~~ published phylogenies to extend existing phylogenetic inferences with more data and taxa, which improve phylogenetic reconstruction [2, 3], time of divergence estimates [4, 5], biogeographic analyses [6], as well as help in resolving phylogenetic conflict [7, 8, 3]. Physcraper updates a starting ~~tree and single locus alignments~~ single locus alignment and corresponding phylogeny with public DNA data from GenBank [9], and links the tips in ~~these~~ the updated trees to a unified, interoperable taxonomic resource [10].

Data such as geographical location, fossil ranges, and genetic and phenotypic information increasingly available in public biological databases constitute an amazing resource for ~~biological~~ scientific discovery [11]. One of the main challenges for automatic integration of ~~biological data across~~ data across biological databases are varying taxonomic idiosyncrasies. To address this challenge, the Open Tree of Life project (OpenTree) created a unified taxonomy for automatic taxonomic name standardization, by integrating taxonomic data from several databases [10], including the USA National Center for Biodiversity Information (NCBI) taxonomy [12, 13], and the Global Biodiversity Information Facility (GBIF) [14]~~among others. By using the existing OpenTree taxonomy programmatic tools to map tip names , Physcraper has~~ , among many others. All of OpenTree tools and methods are carried out via OpenTree's Application Programming Interfaces (APIs), that are open as services for use by the general public [15]. Physcraper leverages on existing OpenTree's taxonomic APIs to standardize taxon names in any phylogeny, providing a framework for ~~connecting~~ automatic connection of updated phylogenies with 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 [16]. 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 [16]. However, even thirteen years later, phylogenetic estimates for ~~many~~ most of these taxa are still not available [17]. ~~OpenTree's comprehensive~~ Using its unified taxonomy, OpenTree assembles a comprehensive synthetic tree of life ~~comprises~~ comprising 2.3 million tips, of which around only 90,000 are supported by ~~phylogenies~~ phylogenetic data uploaded to OpenTrees' database (the Phylesystem [18]) by curators - 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, but this data either has not been analysed or the analyses have not been made publicly available [17].

Assembling a DNA alignment from ~~such~~ a massive database such as GenBank can be done "by hand", but that is a ~~time consuming~~ time-consuming approach which is ~~not highly reproducible. A variety of~~ largely not reproducible. Various computational pipelines that mine DNA databases fast, efficiently, and reproducibly have been developed and used to infer phylogenetic relationships ~~in a variety of~~ of many organisms (e.g., [19, 20, 21, 22]). While genomics has, and will continue to, revolutionize phylogenetic inference, the diversity of alternative genomic sequencing approaches ~~implemented produce largely~~ that are implemented produce widely

non-overlapping homology hypotheses across taxa [23], creating challenges for phylogenetic reconstruction. Phylogenomics addresses this problem by focusing on targeted capture of informative regions [24]. However, fine-grained curated markers and alignments can significantly improve phylogenetic reconstructions, even in phylogenomic analyses [25].

Physcraper improves on previous work ~~in automating phylogenetic reconstructionby leveraging the power of~~ that automates phylogenetic reconstruction, by leveraging on the knowledge contained in existing homology hypotheses that phylogeneticists and 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 ~~[26]~~[26, 27, 28], which can be ~~leveraged~~ used as seeds to mine molecular databases, and as "jump-start" alignments for phylogenetic reconstructions [29] to continually enrich, update and compare ~~existing phylogenetic~~ new phylogenetic hypotheses to existing knowledge.

Physcraper is implemented as a Python pipeline that uses OpenTree's ~~programmatic access protocols (API's)~~ APIs to automatically link any phylogeny mapped to OpenTree's standardized taxonomy~~[18]~~, to alignments from TreeBASE, and data from GenBank. ~~Its utility~~ Physcraper's usage and functionalities are presented with a case-study analysis of a group of flowering plants, the hollies.

Implementation

Physcraper is implemented with Python and can be run on a Python interactive session, as a Python script, or using the command line interface we developed for it. It currently consists of 13 modules. For testing and improving Physcraper's Python code syntax quality, we used the Pylint software following instructions from its website [30] and manual [31], with a ".pylintrc" configuration file.

We improved code syntax of Physcraper's modules with low Pylint scores, and fixed code errors by following Pylint's recommendations. Based on Physcraper's software design choices, some of Pylint's recommendations were overruled by using its check-disabling system, and are explained along the code. As of now, all Physcraper modules have a Pylint score of 10/10.

The general Physcraper framework (Figure 1) consists of 4 steps: 1) identifying and processing a ~~tree~~ phylogeny 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~~ results to existing phylogenies.

The inputs: a ~~tree~~ phylogeny and an alignment

Taxon names in the input ~~tree~~ phylogeny must be standardized to ~~OpenTree taxonomy [10]~~ the OpenTree taxonomy [32] using OpenTree's bulk Taxonomic Name Resolution Service (TNRS) tool [33]. Users can upload their own ~~tree~~phylogeny, or choose from among the 2, 950 ~~standardized trees~~ curated phylogenies stored in OpenTree's Phylesystem ~~[34, 18]~~ database [34] that also have alignments available on ~~TreeBASE [26]~~the TreeBASE database [35, 36].

The input alignment is a single locus DNA dataset that was used in part or in whole to generate the input ~~tree~~phylogeny. Physcraper retrieves TreeBASE alignments automatically. Alternatively, users ~~must~~ can provide the path to a local copy of the alignment of their choosing. Only taxa that are both in the sequence alignment and in the ~~tree~~ phylogeny 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 [37] 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~~ which corresponds to the Most Recent Common Ancestor (MRCA) ~~API [38, 10], as the MRCA~~ of all ingroup taxa that is also a named clade in the NCBI taxonomy (Figure 1). The search taxon is identified using OpenTree's taxonomic API [38].

BLAST is performed using the blastn algorithm [39] implemented in BioPython ~~'s~~ 1.71 [40] NCBIWWW module [41] 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~~ minimum and maximum sequence length threshold, defined as ~~the~~ a proportion of the average sequence 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 OpenTree's 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 [40]. 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 [42] 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~~ Phylogenetic reconstruction and comparison

RAxML [43] is implemented to reconstruct a Maximum Likelihood (ML) gene ~~tree~~ phylogeny for each input alignment with default settings (GTRCAT model and 100 bootstrap replicates with default algorithm), using ~~input tree~~ the input phylogeny as starting tree for ML searches. Bootstrap results are summarized using DendroPy's SumTrees module [44].

Physcraper's main result is an updated phylogenetic hypothesis for the search taxon. Updated and original ~~tree~~ phylpogeny are compared with Robinson-Foulds weighted and unweighted metrics ~~estimated~~ calculated with Dendropy [44], and with a node by node comparison between the synthetic OpenTree and the original and updated ~~tree~~ phylogenies individually, using OpenTree's conflict API [45].

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~~ 400 [46] and 500 accepted living species [47], and is the only extant ~~clade~~ taxon within the family Aquifoliaceae, in the order Aquifoliales of flowering plants [48].

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 [49, 50, 51, 52], but only two have data openly available~~[46, 53]. [46] made original tree and alignment~~, the "Gottlieb2005" study [46] and the "Yao2020" study [53]. Gottlieb2005 made their original published phylogeny and alignment openly available in TreeBASE (study 1091 [54]). The ~~tree sampling~~ Gottlieb2005 phylogeny samples 41 species~~is also~~, is available from OpenTree's Phylesystem (study pg_2827 [55]), and has been integrated into OpenTree's synthetic tree [56]. The ~~most recent~~ Yao2020 *Ilex* ~~tree [53]is~~ phylogeny is the most recent one for the genus [53], and it is only available in OpenTree's Phylesystem (study ot_1984 [57]), and in the DRYAD repository [58]. With 175 tips, the ~~[53] tree~~ Yao2020 phylogeny [53] is the best sampled phylogeny ~~yet~~ available for the ~~hollies.~~ genus *Ilex*. In order to showcase Physcraper's performance, we used the Gottlieb2005 phylogeny and alignment to update relationships in the genus *Ilex*, with the expectation to use the Yao2020 phylogeny, the best sampled and most recent one, as a "gold" standard to compare and verify Physcraper's results.

We ran Physcraper on a laptop Linux computer to update an internal transcribed spacer DNA region (ITS) alignment that was used to construct the phylogeny from [46], using a local GenBank database. BLAST and RAxML analyses ran for 19hrs 45min, with bootstrap analyses taking an additional 13hrs. The ~~updated [46] tree~~ Gottlieb2005 phylogeny [46] updated using Physcraper (Figure 2; Physcraper updated phylogeny from now on) 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~~ ML phylogeny from the RAxML analysis 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 ,~~ As comparison with the Physcraper updated phylogeny, the Yao2020 phylogeny [53] also contains all 41 distinct taxa ~~from the original [46]study, and contributes~~ sampled in the Gottlieb2005 phylogeny [46], while contributing phylogenetic data to 134 additional *Ilex* taxa~~, from which~~. From these, 67 taxa are also in ~~updated [46]. While [53] also used~~ the Physcraper updated phylogeny. While the Yao2020 phylogeny [53] was also constructed using ITS as a marker, their GenBank data is not released yet~~, so~~. Hence, Physcraper was unable to incorporate 68 ~~additional taxa into the analysis. However,~~ taxa that are only on the Yao2020 phylogeny because the DNA data is unavailable. We also note that Physcraper incorporates 18 *Ilex* taxa that are not in the Yao2020 phylogeny [53]. These taxa appear nested among other *Ilex* species and visual inspection of the DNA sequences suggests they are correctly assigned as *Ilex*. The ITS alignment that underlies the Yao2020 phylogeny was constructed without any tool to scrape

GenBank [53], which could explain why Physcraper was able to incorporate these 18 ~~taxa that were not in [53].~~ additional *Ilex* taxa in the Physcraper updated phylogeny (Figure 2).

Verification test

To test the accuracy of Physcraper we designed a verification test. We pruned 9 out of the 41 original tips of the Gottlieb2005 phylogeny [46], corresponding to a 20% trim, excluding the outgroups. We then performed a Physcraper run to test if we would recover the dropped tips. We successfully recovered 6 out of 9 pruned tips in the Physcraper updated phylogeny. Closer examination of results revealed that sequences for the 3 missing tips were correctly retrieved with BLAST along with sequences from the other 6 tips recovered in the updated phylogeny. By following the GenBank accession numbers reported in the original publication belonging to the ITS sequences of the missing tips, we observed that these three sequences contain a 100 bp long gap of unidentified nucleotides (Ns) that is completely absent from any of the sequences in the original alignment. This caused these three GenBank sequences in particular to exceed Physcraper's default sequence length cutoff of 120%, being thus filtered and excluded from the alignment step onwards. These sequences do appear in the Physcraper results in the list of matches from GenBank which did not fit the sequence length cutoffs set in the configuration file. This "seqlen_mismatch.txt" file includes the accession number, taxon, and sequence length of all sequences filtered based on sequence length.

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.

Physcraper is a tool that builds upon previous knowledge stored in published alignments and phylogenies, taking advantage of OpenTree's services to facilitate comparison of phylogenies, with the main goal of extending our knowledge of phylogenetic relationships across the tree of life.

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 [20, 16, 59, 60, 22]. In particular, Phylota [16], and PHLAWD [19], 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 less manual downstream processing) and eases comparison with previous phylogenetic knowledge.

Results from the verification test highlight the importance of incorporating existing expertly curated homology statements to automatically update phylogenetic relationships, instead of ignoring the information they contain and building homology statements fully *de novo*.

We encourage users to look at the output files containing information about the filtered sequences, and potentially modify configuration parameters such as the sequence length cutoff parameter, based on the filtered sequences. Default filtering parameters are arbitrary, but we hope that by making the process of locating homologous sequences online reproducible, and tracking what filters are used, we make it easier for researchers to delve into the effect that different choices have on their inferences. This is in contrast to "manual" searches for taxa, where similarly arbitrary filters are applied, but are difficult to trace. As many studies have shown [61] the effect of missing data can be enigmatic, and interact with the true phylogenetic relationships for the data set at hand. There is not currently strong support in the literature for any particular cutoff value, and rather than prescribe specific approaches, we encourage users to explore the effects of different choices on their phylogenetic inferences. In addition, by providing the output files at each step of the analysis, it is straightforward to assess how changing parameter and software choices do or do not drive differences in phylogenetic inference. By gathering the sequences, and making the unaligned files easily available to users, researchers can compare if applying any alternate alignment tool of their choice affects inferences. Once sequences are aligned, they can apply and compare inferences from any phylogenetic software.

Organellar genome sequences, such as chloroplasts and mitochondria will also generally be excluded from automatic addition based on default Physcraper length cutoffs. Multiple sequence alignment of loci of drastically different lengths is unfeasible, and we have found in testing that it often returns incorrect results, splitting shorter sequences with many long gaps to align with exact matches across the entire longer locus. While it would be possible to directly extract the BLAST match from genomes, this would exclude potentially homologous flanking regions which are not matched by BLAST's local search algorithm, but that may be important for phylogenetic inference. Instead we list the accession numbers for these matches in the "seqlen_mismatch.txt" file, for users to assess and incorporate appropriate homologous regions to their alignment of interest.

Unlike phylogenetic placement approaches [62, 63], which add new taxa without modifying the input ~~tree~~phylogeny, Physcraper estimates all the relationships anew in the context of the new data. PUMPER [21] 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 [64]. ~~However, Physcraper facilitates~~ In addition, single gene phylogenies with very high numbers of taxa may lack sufficient signal for accurate phylogenetic resolution [65]. The Physcraper workflow avoids this challenge by focusing on ingroup taxa of an existing phylogeny, using markers that have been assessed and proven appropriate for that phylogenetic scope in past publications. Also, Physcraper thins alignments by removing sequences identical to

original and newly added sequences, and by setting a maximum number of sequences per taxon. Nonetheless, it is incumbent on users to assess their final inference with respect to statistical support and biological plausibility.

In the era of phylogenomics, rigorous analyses of multiple loci still allow for more complex evolutionary models than analyses of large genomic data sets, and in many cases can provide better evolutionary estimates. For example [66] show that when applying coalescent models, there is more information in two genes of 300 bp each than in 600 independent sites. Physcraper is designed to facilitate 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 [67] ~~, BEAST2 [68],~~ or SVD Quartets [69]. Physcraper's "multi_locus.py" module allows to automatically merge the outputs of Physcraper runs from different loci into input files for the two software mentioned above, or as concatenated alignments for supermatrix analyses.

Our example application of Physcraper to update a phylogeny of the genus *Ilex* is based on a single marker, so we expect for it to be not as well resolved as phylogenies resulting from analyses that used multiple markers. Although not perfect, we think the Physcraper updated *Ilex* phylogeny seems biologically reasonable in different ways. All samples corresponding to the ingroup cluster together forming a monophyletic group (Figure 2A), and samples belonging to the same *Ilex* species also form monophyletic groups (Figure 2B). ~~Rigorous analyses of multiple loci allows for more complex evolutionary models than analysesof large genomic data sets, and can provide better evolutionary estimates.~~ A notable exception is samples of the species *Ilex theezians*, which appear as non-monophyletic in the updated phylogeny as well as in the original Gottlieb2005 phylogeny. analyses should be conducted A visual comparison of the Yao2020 phylogeny and the original Gottlieb2005 phylogeny suggests that the relationships within the genus *Ilex* are still being actively determined, and that increased taxon sampling might be key to resolve them.

Physcraper has the added advantage of facilitating the linkage of taxonomic information about tips in the output phylogenies to data available in a variety of biological databases [10], such as geographical locations for taxa from the GBIF [14]. Taxonomic links, and comparisons to existing published phylogenies in the OpenTree data store can also help flag paralogous sequences. Accidentally including paralogs as homologs is a risk in phylogenetic analyses, and can be more prevalent in automated analyses than in manually curated analyses. We provide users with several tools to try to assess homology of their aligned sequences. The estimated gene phylogeny itself is an evolutionarily explicit way to visualize gene evolution, which in concert with taxonomic labelling can reveal paralogy. OpenTree's conflict analysis tool informs the users of whether their phylogeny contains major conflicts with established taxonomy and any phylogenetic context they wish to compare to. This tool also returns information on taxonomic and phylogenetic conflicts that exist in the original input phylogeny. Detected conflicts may be a sign that taxonomy needs to be updated, or may be a sign that non-homologous sequences have been included in the analysis. These taxonomic and phylogenetic conflicts flag regions of the phylogeny for the researcher to more closely examine and assess homology.

The Physcraper workflow can be used to rapidly (in a matter of hours) create phylogenies which can address challenges overarching both fields of ecology and evolution, such as phylogenetically placing newly discovered species [70], curating taxonomic assignments [71], and generating custom trees for ecological [72] and evolutionary downstream analyses [73].

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~~ , applications and education. 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~~for phylogenetics, 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

API: Application Programming Interface

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

~~The~~ Code and 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.

Funding
This 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.

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.

Acknowledgements
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. We also thank the valuable comments of David Posada, Rutger Vos, and an anonymous reviewer that greatly improved an earlier version of this manuscript.

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. Hillis, D.M.: Inferring complex phylogenies. Nature **383**(6596), 130–131 (1996)
3. Natsidis, P., Tsakogiannis, A., Pavlidis, P., Tsigenopoulos, C.S., Manousaki, T.: Phylogenomics investigation of sparids (Teleostei: Spariformes) using high-quality proteomes highlights the importance of taxon sampling. Communications biology **2**(1), 1–10 (2019)
4. Schulte, J.A.: Undersampling taxa will underestimate molecular divergence dates: an example from the South American lizard clade Liolaemini. International Journal of Evolutionary Biology **2013** (2013)
5. Soares, A.E., Schrago, C.G.: The influence of taxon sampling on bayesian divergence time inference under scenarios of rate heterogeneity among lineages. Journal of Theoretical Biology **364**, 31–39 (2015)
6. Kayaalp, P., Stevens, M.I., Schwarz, M.P.: Back to Africa: increased taxon sampling confirms a problematic Australia-to-Africa bee dispersal event in the Eocene. Systematic Entomology **42**(4), 724–733 (2017)
7. Hedtke, S.M., Townsend, T.M., Hillis, D.M.: Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Systematic Biology **55**(3), 522–529 (2006)
8. Townsend, J.P., Lopez-Giraldez, F.: Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Systematic Biology **59**(4), 446–457 (2010)
9. 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
10. 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
11. Baxevanis, A.D., Bateman, A.: The importance of biological databases in biological discovery. Current Protocols in Bioinformatics **50**(1), 1–1 (2015)
12. 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
13. 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)
14. GBIF Secretariat: GBIF Backbone Taxonomy. Checklist Dataset. doi:10.15468/39omei. Accessed via GBIF.org on April 2021. https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c
15. OpenTreeOfLife, Redelings, B., Cranston, K.A., Allman, J., Holder, M.T., McTavish, E.J.: Open Tree of Life APIs V. 3.0. https://github.com/OpenTreeOfLife/germinator/wiki/Open-Tree-of-Life-Web-APIs
16. 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
17. 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
18. 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
19. 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
20. 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
21. 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
22. Pearse, W.D., Purvis, A.: phylogenerator: an automated phylogeny generation tool for ecologists. Methods in Ecology and Evolution **4**(7), 692–698 (2013)

23. 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
24. 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
25. 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
26. Piel, W., Chan, L., Dominus, M., Ruan, J., Vos, R., Tannen, V.: Treebase v. 2: A database of phylogenetic knowledge. e-Biosphere. London (2009)
27. Vos, R.A., Balhoff, J.P., Caravas, J.A., Holder, M.T., Lapp, H., Maddison, W.P., Midford, P.E., Priyam, A., Sukumaran, J., Xia, X., *et al.*: NeXML: rich, extensible, and verifiable representation of comparative data and metadata. Systematic Biology **61**(4), 675–689 (2012). doi:10.1093/sysbio/sys025
28. Piel, W.H., Vos, R.A.: Treebasedmp: A toolkit for phyloinformatic research. bioRxiv, 399030 (2018)
29. Morrison, D.A.: Multiple sequence alignment for phylogenetic purposes. Australian Systematic Botany **19**(6), 479–539 (2006). doi:10.1071/SB06020
30. Thénault, Sylvain (Logilab S.A.): Pylint. Accessed: March 2021. https://www.pylint.org/
31. Thénault, Sylvain (Logilab S.A.), PyCQA, and contributors: Pylint User Manual. Accessed: March 2021. http://pylint.pycqa.org/en/latest/
32. OpenTreeOfLife, Redelings, B., Cranston, K.A., Allman, J., Holder, M.T., McTavish, E.J.: Open Tree of Life Taxonomy V. 3.2. https://tree.opentreeoflife.org/about/taxonomy-version/ott3.2
33. OpenTreeOfLife: Name Resolution (TNRS) bulk mapping tool. https://tree.opentreeoflife.org/curator/tnrs/
34. 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
35. Piel, W., Chan, L., Dominus, M., Ruan, J., Vos, R., Tannen, V.: TreeBASE: A Database of Phylogenetic Knowledge. https://treebase.org/treebase-web/home.html
36. Vos, R.: SuperTreeBASE: data dump and code to summarize TreeBASE. https://github.com/TreeBASE/supertreebase
37. 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
38. OpenTreeOfLife, Rees, J.A., Cranston, K.: OpenTree's taxonomic MRCA API. https://github.com/OpenTreeOfLife/germinator/wiki/Taxonomy-API-v3#mrca
39. 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
40. 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
41. The BioPython Contributors (1999-2018): BioPython 1.71, Module Bio.Blast.NCBIWWW. Accessed: April 19, 2018. https://biopython.org/DIST/docs/api/Bio.Blast.NCBIWWW-module.html
42. 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
43. 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
44. Sukumaran, J., Holder, M.T.: DendroPy: a Python library for phylogenetic computing. Bioinformatics **26**(12), 1569–1571 (2010). doi:10.1093/bioinformatics/btq228
45. 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
46. 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
47. The Plant List 2013. Version 1.1: List of Name Records for the Generic Epithet *Ilex*. http://www.theplantlist.org/tpl1.1/search?q=ilex
48. Chase, M.W., Christenhusz, M., Fay, M., Byng, J., Judd, W.S., Soltis, D., Mabberley, D., Sennikov, A., Soltis, P.S., Stevens, P.F.: An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society **181**(1), 1–20 (2016)
49. 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
50. 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
51. 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
52. 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
53. 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

54. Gottlieb, A.M., Giberti, G.C., Poggio, L.: TreeBASE study 1091. https://treebase.org/treebase-web/search/study/summary.html?id=1091
55. Gottlieb, A.M., Giberti, G.C., Poggio, L.: Phylesystem study pg_2827. https://tree.opentreeoflife.org/curator/study/edit/pg_2827/?tab=home
56. 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
57. 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
58. 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
59. 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
60. 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
61. Huang, H., Knowles, L.L.: What Is the Danger of the Anomaly Zone for Empirical Phylogenetics? Systematic Biology, 047 (2009). doi:10.1093/sysbio/syp047. Accessed 2015-07-24
62. 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
63. 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
64. 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
65. Morel, B., Barbera, P., Czech, L., Bettisworth, B., Hübner, L., Lutteropp, S., Serdari, D., Kostaki, E.-G., Mamais, I., Kozlov, A.M., Pavlidis, P., Paraskevis, D., Stamatakis, A.: Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. Molecular Biology and Evolution (msaa314) (2020). doi:10.1093/molbev/msaa314. Accessed 2021-03-30
66. Zhu, T., Yang, Z.: Complexity of the simplest species tree problem. Molecular Biology and Evolution (msab009) (2021). doi:10.1093/molbev/msab009. Accessed 2021-03-30
67. 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
68. 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
69. 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
70. 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
71. 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
72. Helmus, M.R., Ives, A.R.: Phylogenetic diversity–area curves. Ecology **93**(sp8), 31–43 (2012). doi:10.1890/11-0435.1
73. 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
74. OpenTreeOfLife, Redelings, B., Reyes, L.L.S., Cranston, K.A., Allman, J., Holder, M.T., McTavish, E.J.: Open Tree of Life Synthetic Subtree of the Genus *Ilex*, Node Id Ott727571. https://tree.opentreeoflife.org/opentree/opentree12.3@ott727571/Ilex

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 using a starting phylogeny and an alignment from ~~original~~ [46] ~~tree~~ (Gottlieb2005 data in ~~B~~text). 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~~ from the Gottlieb2005 internal transcribed spacer DNA region (ITS) original alignment~~that generated tree in B)~~, filtered them following criteria ~~from~~ described 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~~ B) Results of conflict ~~analyses~~ analysis performed ~~with~~ using OpenTree~~tools~~'s conflict tool [45]. The Physcraper updated Gottlieb2005 phylogeny in (A) was compared to an *Ilex* OpenTree synthetic subtree v. 12.3 [74] constructed using taxonomy of the genus as backbone and resolving branches based on phylogenetic data from the original Gottlieb2005 phylogeny.