

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

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

5 School of Natural Sciences, University of California, Merced

6 email: [sanchez.reyes.luna@gmail.com](mailto:sanchez.reyes.luna@gmail.com)

7 **2. Martha Kandziora**

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

9 email: [kandziom@natur.cuni.cz](mailto:kandziom@natur.cuni.cz)

10 **3. Emily Jane McTavish**

11 School of Natural Sciences, University of California, Merced

12 email: [ejmctavish@gmail.com](mailto:ejmctavish@gmail.com)

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

15 **Correspondence email:** [sanchez.reyes.luna@gmail.com](mailto:sanchez.reyes.luna@gmail.com), [ejmctavish@gmail.com](mailto:ejmctavish@gmail.com)

16 **Running title:** Continually updated gene trees with Physcraper

<sup>17</sup> **Word count:** 3646

<sup>18</sup> **Manuscript prepared for submission to Methods in Ecology and Evolution**

<sup>19</sup> **Article type:** Application

## Abstract

1. Phylogenies are a key part of research in many areas of biology. Tools that automatize 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 choosing 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 automatizes the update of published phylogenies by enriching underlying gene alignments with public DNA sequence data. It builds upon tools from the Open Tree of Life project to link taxonomic information across databases, providing a framework for straightforward comparison of published phylogenies with their updated versions.
3. Physcraper can be used by the nonspecialist, as a tool to generate phylogenetic hypotheses based on publicly available expert phylogenetic knowledge. Phylogeneticists and group specialists will find it useful as a tool to facilitate molecular dataset gathering and comparison of alternative phylogenetic hypotheses (topologies).
4. We hope that the Physcraper workflow demonstrates the benefits of doing open science for phylogenetics, encouraging more researchers to strive for better sharing practices. Physcraper can be used with any OS and is released under an open-source license. Detailed instructions for installation and use are available at <https://physcraper.readthedocs>.

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

# 1 Introduction

Phylogenetic estimates of evolutionary relationships capture the shared history of living organisms, and provide key context for all our biological observations. Public biological databases constitute an amazing resource for evolutionary estimation, but a large portion of molecular data publicly available has never been incorporated into any phylogenetic estimate. Extending existing phylogenetic estimates with new DNA sequence data, geographical location, fossil data, etc., in a reproducible and continuous manner is possible by automatizing connections between biological databases. Here, we propose Physcraper, a tool that uses existing phylogenetic research and data from public biological databases to update a starting tree and single locus alignments, to build upon molecular data that taxon specialists have assessed and deemed appropriate for a specific phylogenetic scope.

The prevalence of taxonomic idiosyncracies across databases represent a big challenge to automatically connect data from disparate biological databases in a phylogenetic context. To standardize taxonomic names, a unified taxonomy is needed. The main aim of the Open Tree of Life (OpenTree from now on) project is to construct a comprehensive, dynamic and digitally-available tree of life by synthesizing published phylogenetic trees along with taxonomic data. Currently, OpenTree’s “synthetic” tree comprises 2.3 million tips, of which around 90,000 are represented by phylogenetic estimates - the remaining 1.4 million taxa are placed in the tree based on their taxonomic assignment. To achieve this, OpenTree unifies taxonomy from various databases (Rees & Cranston 2017) such as the USA National Center for Biodiversity Information (NCBI) molecular database GenBank (Benson *et al.* 2000; Wheeler *et al.* 2000), the Global Biodiversity Information Facility (GBIF; Secretariat 2017), and the World Register of Marine Species [WoRMS; [www.marinespecies.org/](http://www.marinespecies.org/)], providing a key resource that can be used to connect data from virtually any biological database to phylogenetic data that has been standardized to OpenTree’s unified taxonomy.

Another challenge to incorporating molecular data from public databases to update phylogenetic knowledge is assembling high-quality homology hypotheses. Species tree reconstructions from multi-gene data sets taking into account the multispecies coalescent model are seen as the gold standard for inferring species relationships (Song *et al.* 2012). While genomics have allowed us to obtain large quantities of molecular

data from one organism, different research questions call for different genomic sequencing approaches, from whole genomes, to transcriptomes, restriction-site associated DNA sequencing (RADSeq), single nucleotide polymorphisms (SNPs), microsatellites, and ultra-conserved elements (UCEs). Having a variety of approaches has resulted in largely non-overlapping genomic data sets across taxa. Moreover, some types of genomic data are not suitable for phylogenetic inference at any scale. For example, phylogenetic distance can cause allelic dropout at deeper divergences with RADSeq data sets (Eaton *et al.* 2016). In contrast, decades of single locus sequencing have generated homologous DNA data that can be used for phylogenetic reconstruction at many scales. Indeed, more than a decade ago, GenBank release number 159 (April 15, 2007) already hosted 72 million non-genomic DNA sequences that were gauged to have the potential to resolve phylogenetic relationships of most of the 241, 000 distinct taxa in the NCBI taxonomy at the time (about 98.05% of taxa; Sanderson *et al.* 2008). Assembling a DNA alignment from such a massive database can be done “by hand”, but it requires huge amounts of time and it is mostly a non-reproducible method. Computer pipelines that make DNA sequence search faster and more efficient, as well as more reproducible have been utilized to study evolutionary relationships among a variety of organisms (e.g., Smith *et al.* 2009; Izquierdo-Carrasco *et al.* 2014; Antonelli *et al.* 2017). Still, there is a large tradition in phylogenetics of building molecular character matrices of few, thoughtfully curated markers, as manual curation of alignments often improves phylogenetic reconstructions (Fragoso-Martínez *et al.* 2017).

A way to incorporate the best of two worlds (massive amounts of newly released molecular data and fine-grained curation from human experts) is to rely on published manually curated homology hypotheses as “jump-start” alignments (Morrison 2006). The TreeBASE database (Piel *et al.* 2009) hosts about 8, 200 publicly accesible alignments, providing information on evolutionary relationships of around 100, 000 distinct taxa (see TreeBASE’s website about), representing an untapped source of valuable expert knowledge with the potential to continuously enrich and update phylogenetic relationships in several different regions of the tree of life, by incorporating data from molecular databases that has not been included in any public phylogenetic estimate. The phylogenies associated to all these alignments have been integrated to the OpenTree’s datastore, the Phylesystem (McTavish *et al.* 2015), and metadata linking these trees to their corresponding alignment repository in TreeBASE is available.

We propose Physcraper as a Python encoded pipeline that relies on programmatic access protocols to link molecular data from GenBank, to alignments from TreeBASE and phylogenies from OpenTree's Phylsystem, to automatically update and enrich phylogenetic knowledge based on expertly-curated homology hypotheses. Physcraper also provides new types of access to various general OpenTree programmatic tools for custom comparison of phylogenetic hypotheses. It focuses on database interoperability, by automatizing taxonomic name matching across biological databases, so users can perform any downstream analysis straightforwardly, as well as placing new data from virtually any biological database in an evolutionary context.

## 2 The Physcraper framework

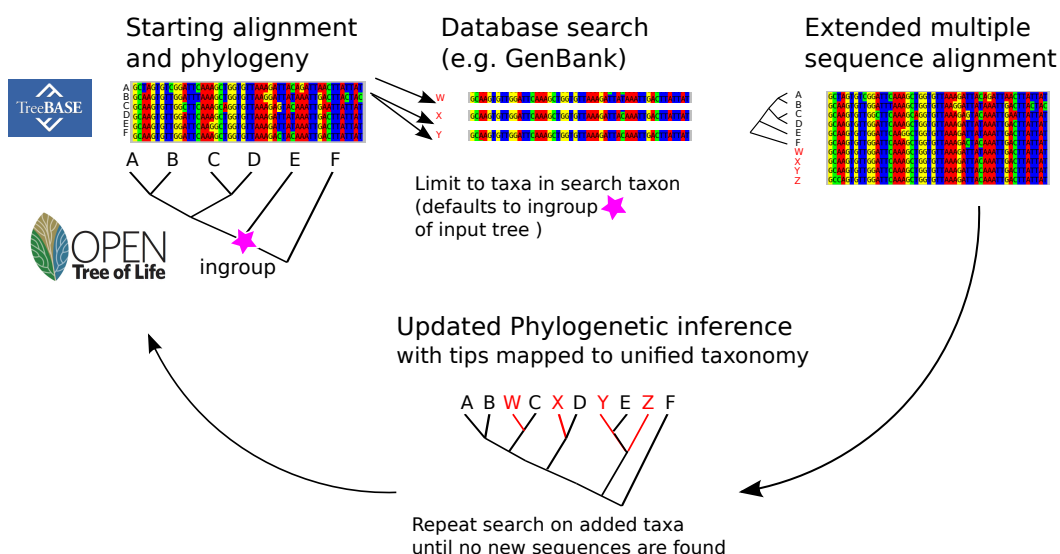


Figure 1: The Physcraper software is fully described on its documentation website at [physcraper.readthedocs.io](https://physcraper.readthedocs.io), along with installation instructions, tutorials, examples and function usage documentation.

The general Physcraper framework is shown in Figure 1. Next, we will describe the technical details of each step of the workflow.

### 2.1 The inputs: a tree and an alignment

- In order to take advantage of the OpenTree tools, it is recommended that the input tree is either stored in the OpenTree Phylsystem, or submitted via OpenTree's curator application (McTavish *et al.* 2015). If the user is not ready to make the input tree public, tree tip labels must be standardized

to the unified OpenTree taxonomy using the bulk Taxonomic Name Resolution Service TNRS tool. *Should we say that it is based on Boyle’s 2012 algorithm? or is it a TNRS algorithm independently developed by the OpenTree team?* This step is referred to as taxon name mapping. If taxonomic names can’t be mapped, their taxonomic information will be excluded from further analysis. Mapping tip names to OpenTree’s unified taxonomy saves a set of user defined characteristics that are essential for automatizing the phylogeny updating process. The most relevant of these is the taxonomic name mapping and the definition of ingroup and outgroup taxa, allowing to automatically set the root for the updated tree on the final steps of the pipeline. Currently, only trees connected to a published study can be stored in the Phylsystem. Users can choose from among the 2, 950 studies in OpenTree’s Phylsystem that have alignments on TreeBASE.

- The input alignment should be a single locus alignment that was used to generate the tree. Alignments are often stored in a public repository such as TreeBase (Piel *et al.* 2009; Vos *et al.* 2012), DRYAD ([www.datadryad.org](http://www.datadryad.org)), or a data repository associated with the journal where the tree was originally published. If the alignment is stored in TreeBase, Physcraper downloads it directly, either from the TreeBASE website ([www.treebase.org](http://www.treebase.org)) or through the TreeBASE GitHub repository (SuperTreeBASE; [github.com/TreeBASE/supertreebase](https://github.com/TreeBASE/supertreebase)). If the alignment is on another repository, or constitutes personal data, a path to a local copy of the alignment has to be provided.
- Single locus alignments sometimes have fewer taxa than the tree inferred from the full concatenated data, simply because a single molecular marker usually does not cover all the taxa sampled for the full phylogenetic analysis. Physcraper prunes the input tree to taxa found in the alignment, and verifies that all taxon names on the tips of the tree are in the DNA character matrix and vice versa. Technically, just one taxon name (and its corresponding sequence in the alignment) is needed to continue the algorithm. See next section.
- A summary “csv” file with results from the taxon name standardization and pruning is produced for the user.

- The standardized and pruned tree and alignment (checked tree and alignment from now on) are output as “newick” and “fasta” respectively in the “inputs” folder to be used in the following steps.

## 2.2 DNA sequence search and filtering

- Physcraper uses the GenBank DNA database as source to search for new sequences. The DNA sequence search can be performed on the GenBank remote database or in a GenBank local database set up by the user, which can speed up the search process. Detailed instructions to setup a local database are provided on Physcraper’s software documentation.

- The next step is to identify a “search taxon” to constrain the sequence search on the GenBank database within that taxonomic group. The search taxon can be chosen by the user from the NCBI taxonomy. If none is provided, then the search taxon is automatically set using the taxa in the input tree labeled as the “ingroup” (Fig. 1). The search taxon is The Most Recent Common Ancestor (MRCA) of the ingroup taxa in the OpenTree synthetic tree, that is also a named clade in the NCBI taxonomy. This is known in the OpenTree as the Most Recent Common Ancestral Taxon (MRCAT; also referred as the Least Inclusive Common Ancestral taxon - LICA). The MRCAT can be different from the phylogenetic MRCA when the latter is an unnamed clade in the synthetic tree. To identify the MRCAT of a group of taxon names, we use the OpenTree taxonomic tool v3 (Rees & Cranston 2017).

Users can provide a search taxon that is either a more or a less inclusive clade relative to the ingroup of the original phylogeny. If the search taxon is more inclusive, the sequence search will be performed outside the MRCAT of the matched taxa, e.g., including all taxa within the family or the order that the ingroup belongs to. If the search taxon is a less inclusive clade, the users can focus on enriching a particular clade/region within the ingroup of the phylogeny.

- The Basic Local Alignment Search Tool, BLAST (Altschul *et al.* 1990, 1997) is used to identify similarity between DNA sequences within the search taxon in a nucleotide database, and the sequences on the checked alignment. The `blastn` function from the BLAST command line tools (Camacho *et al.* 2009) is used for local database sequence searches. For remote database searches, we modified the BioPython



(Cock *et al.* 2009) BLAST function from the NCBIWWW module to accept an alternative BLAST address (URL). This is useful when a user has no access to the computer capacity needed to setup a local database, and a local blast database can be set up on a remote machine to BLAST avoiding NCBI's required waiting times, which slow down the searches markedly.

- A constrained BLAST search is performed, in which each sequence in the alignment is BLASTed once against all database DNA sequences belonging to the search taxon. All results from each BLAST run are stored, and sequences with match scores better than the e-value cutoff (default to 0.00001) are saved along with their corresponding metadata, i.e., their GenBank accession number. The full sequence for each match is downloaded from NCBI into a dedicated library within the “physcraper” folder, allowing for secondary analyses to run significantly faster.
- BLAST result sequences will be discarded if they fall outside the user set min and max length cutoffs, set as proportions of the average length without gaps of sequences in the input alignment (defaults values of 80% and 120%, respectively). This filtering guarantees the exclusion of whole genome sequences, which create problems in multiple sequence alignment. The GenBank accession numbers of sequences removed due to not meeting e-value or length cutoffs are stored in output files.
- All sequences accepted up to this point are assigned an internal identifier, and are further filtered.
- New sequences that are either identical or a subset of any existing sequence in the input alignment are discarded, unless they represent a different taxon in the unified taxonomy, or they are longer than the sequence in the input alignment.
- Among the filtered sequences, there are often several representatives per taxon. Although it can be useful to keep some of them, for example, to investigate monophyly within species, there can be hundreds of exemplar sequences per taxon for some markers. To control the number of sequences per taxon in downstream analyses, 5 sequences per taxon are chosen at random. This number is set by default but can be modified by the user.

- All BLAST and filtering parameters can be customized by the user.
- Reverse, complement, and reverse-complement BLAST result sequences are identified and translated using BioPython internal functions (Cock *et al.* 2009).
- Iterative cycles of sequence similarity search can be performed, by blasting the newly found sequences until no new sequences are found. By default only one BLAST search cycle is performed in which only sequences in the input alignment are blasted.
- New sequences passing all filtering steps are added to the “csv” taxon summary file.
- A “fasta” file containing all new filtered and processed sequences resulting from the BLAST search is generated for the user, and is used as an input for alignment.

## 2.3 New DNA sequence alignment

- By default, Physcraper uses the software MUSCLE (Edgar 2004) to perform DNA sequence alignments. Instructions on how to install all software dependencies used by Physcraper are provided in the documentation.
- The process to align new sequences consists of two steps. First, all new sequences are aligned using the default MUSCLE options.
- Second, a MUSCLE profile alignment is performed, in which the original alignment is used as a template to align the new sequences. This ensures that the final alignment follows the homology criteria established by the original alignment.
- The final alignment is not further processed by Physcraper. It is recommended that the alignment is checked by the user, by eye followed by manual refinement, or using a tool for automatic alignment processing (e.g., GBlocks; Castresana 2000, 2002).
- While curating the alignment is a critical step, it is not a reproducible one. The main reason for its lack of reproducibility might be that it is hard to track changes made on the alignment. A form of version control, to register the differences between the alignment that was produced by the software

and the manually curated alignment will be ideal.

- Users may also use Physcraper to only gather new GenBank sequences, to then apply their own preferred alignment and phylogenetic inference methods.

## 2.4 Tree reconstruction and comparison

- A Maximum Likelihood (ML) gene tree is reconstructed for each alignment provided, using the software RAxML (Stamatakis 2014) with default settings, such as a GTRCAT model of molecular evolution and 100 bootstrap replicates with the default algorithm. Currently only the number of bootstrap replicates can be specified by the user.
- By default, the original tree is used as a starting tree for the ML searches. Alternatively, users can set the original tree as a full topological constraint, or ignore it completely for the searches.
- Bootstrap results are summarized with the SumTrees module of DendroPy (current version 4.4.0; Sukumaran & Holder 2010).
- Physcraper’s final result is an updated phylogenetic hypothesis for the locus provided in the input alignment.
- Tips on all trees generated by Physcraper are defined by a taxon “name space”. The taxon metadata captures the NCBI accession information, as well as the taxon identifiers, allowing the user to perform comparisons and conflict analyses.
- Two ways to compare the updated tree with the original tree are implemented in Physcraper. First, Robinson Foulds weighted and unweighted metrics are estimated using Dendropy functions (Sukumaran & Holder 2010).
- Second, a conflict analysis is performed. This is a node by node comparison between the the synthetic OpenTree and the original and updated tree individually. This is performed with OpenTree’s conflict Application Programming Interface (Redelings & Holder 2017).
- For the conflict analysis to be meaningful, the root of the tree needs to be accurately defined.
- A suggested default rooting based on OpenTree’s taxonomy is implemented for now. This approach uses the taxon labels for all the tips in the updated tree, pulls an inferred subtree from OpenTree’s

taxonomy and then applies the same rooting to the inferred updated tree. However, if the updated tree changes expectations from taxonomy, the root may no longer be appropriate. Automatic identification of a phylogenetic tree root is indeed a difficult problem that has not been solved yet. The best way right now is for users to define outgroup directly on the updated tree, so trees are accurately rooted.

## 3 Examples

We will illustrate the utility of Physcraper in here with two use-case scenarios. One in which the user is interested in a particular group. Another one in which the user is interested in a particular phylogeny. A tutorial as well as illustrated examples of commands for every step needed to perform a Physcraper analysis are available elsewhere.

### 3.1 The hollies

A student is interested in the genus *Ilex*, the only extant clade within the family Aquifoliaceae, order Aquifoliales of flowering plants. The genus encompasses between 400-600 living species. A review of literature (google scholar search for “*ilex* phylogeny”) reveals that there are several published phylogenetic trees showing relationships within the hollies (Cuénoud *et al.* 2000; Setoguchi & Watanabe 2000; Selbach-Schnadelbach *et al.* 2009; Manen *et al.* 2010), but only two have their data available publicly (Gottlieb *et al.* 2005; Yao *et al.* 2020). Gottlieb *et al.* (2005) made tree and alignment data available in TreeBASE. The tree sampling 48 species was integrated to the OpenTree Phylesystem and is part of OpenTree’s synthetic tree. The most recent *Ilex* tree from Yao *et al.* (2020) has been made available in the OpenTree Phylesystem and in the DRYAD repository. It is the best sampled yet for the genus, with 200 species. However, it has not been added to OpenTree’s synthetic tree yet. This makes it a perfect case to test the basic functionalities of Physcraper: we know that the sequences of the most recently published tree have been made available on the GenBank database. Hence, we expect that updating the oldest tree should at least contain the same species sampled in the largest tree.

DESCRIBE RESULTS: SUMMARY OF NEW TAXA FOUND RELATIVE TO ORIGINAL TREE AND

256 RELATIVE TO OpenTree RF DISTANCE INTERPRETATION HOW MUCH TIME THE BLAST RUN  
257 TOOK ML ESTIMATES OF UPDATED TREE VS ORIGINAL TREE

258 FIGURE: FACE TO FACE ORIGINAL VS UPDATED PHYLOGENY, IN RED NEW TAXA NOT IN  
259 OpenTree.

## 260 **3.2 The Malvaceae**

261 A postdoc started working with a new reserach group. They are interested in solving relationships among  
262 lineages of the Malvaceae, a family of flowering plants with almost 6 000 known species, containing the  
263 relatives of cacao, cotton, durian and okra. A review of the literature shows them that there are many  
264 phylogenetic trees encompassing some of the linegaes in the group. However, the head of the research group  
265 wants to use a particular marker they believe to be the best one to be able to solve the relationships in the  
266 group. They have been working on the alignment for a long time and they want to incorporate new data into  
267 the hypothesis of homology that they have been curating and that they trust.

# Original tree

# Updated tree

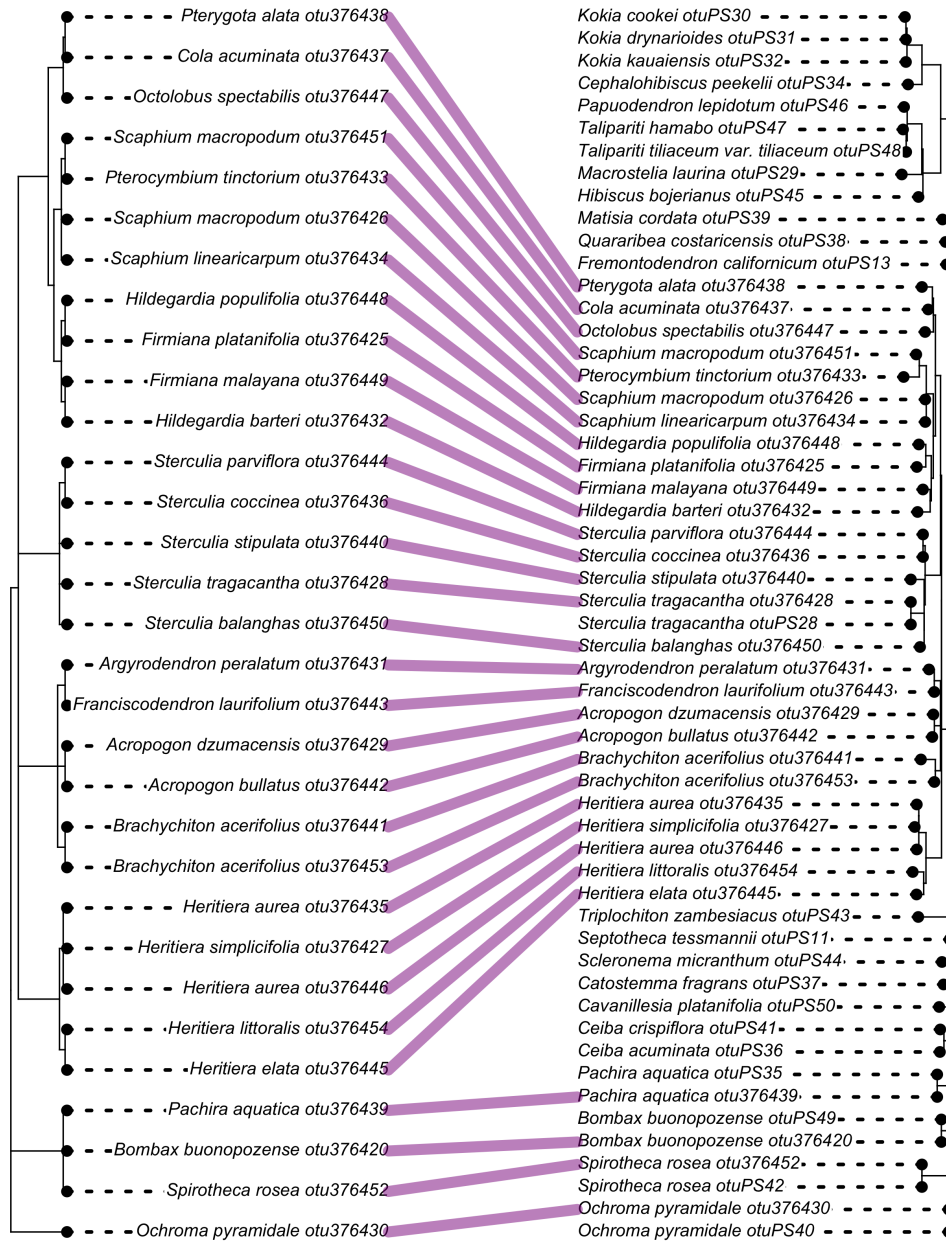


Figure 2: Comparison of original tree and tree updated with Physcraper, family Malvaceae.

## 4 Discussion

Data repositories hold even more information than meets the eye. Besides the actual data, they are rich sources of metadata that can be used for the advantage of all areas of biology as well as the advancement of scientific policy and applications.

COMPARE WITH PERFORMANCE OF OTHER PIPELINES FOR SEQUENCE SCRAPING WHY WE DID NOT MAKE A BENCHMARK COMPARISON

Many pipelines are making use of DNA data repositories in different ways. Most of them focus on efficient ways to mine the data – getting the most homologs. Some focus on accurate ways of mining the data – getting real and clean homologs. Others focus on refining the alignment. Most focus on generating full trees *de novo*, mainly for regions of the Tree of Life that have no phylogenetic assessment yet in published studies, but also for regions that have already been studied and which have phylogenetic data. However, expert phylogenetic knowledge is also an important source of data in public and open repositories that is not being used to its full potential.

All these tools are key efforts for advancing towards reproducibility in phylogenetics, a field that has relied on processes which are somewhat artisanal. Here, we highlight the potential of taking advantage of this careful curation work in previous phylogenetic estimates. By taking sources of information available from data repositories and present a method to link data from different repositories, while leveraging the knowledge and intuition of the expert community to build up our phylogenetic knowledge, we can use not only data accumulated in molecular data repositories, but phylogenetic knowledge accumulated in phylogenetic tree repositories.

While not generating full phylogenies *de novo*, Physcraper is still capable of generating new phylogenetic knowledge. Moreover, it can combine phylogenies with data from repositories other than molecular data. For example geographic locations (using GBIF ids), fossils (using PBDB ids), etc. *from Robert: I think you can sell the program more here. Why is it better than the other methods? You mentioned in lab meeting that its difficult to run other programs, talk about that here, talk about the speed and other advantages*

Physcraper has the potential to be applied for the advantage of the field to rapidly *HOW FAST IS “RAPID”*  
*mention it in results and then here again* place newly discovered species phylogenetically (Webb *et al.* 2010),  
obtain trees for ecophylogenetic studies (Helmus & Ives 2012), help to systematize molecular databases, i.e.,  
curate taxonomic assignments (San Mauro & Agorreta 2010), and rapidly generate custom species trees for  
downstream analyses (Stoltzfus *et al.* 2013).

## 5 Acknowledgements

Research was supported by the grant “Sustaining the Open Tree of Life”, National Science Foundation  
ABI No. 1759838, and ABI No. 1759846. Compute 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 members of the “short bar” Science and Engineering Building 1, University of California Merced,  
joint lab paper discussion meeting for valuable comments on a first version of this manuscript.

## 6 Authors’ Contributions

EJM: Conceived study, wrote most of the code, documentation and tests. MK: Wrote code for `ncbidataparser`  
module, filtering of sequences per OTU and using offline blast searches, wrote documentation and tests.  
LLSR: Wrote the manuscript, alignment code, documentation, performed analyses and developed examples.  
All authors contributed to the manuscript.

## 7 Data Availability

Physcraper source code available at <https://github.com/McTavishLab/physcraper> Documentation available at  
<https://physcraper.readthedocs.io/en/latest/index.html> Illustrated examples available at <https://github.com/McTavishLab/physcraperex> This is a reproducible manuscript available at [https://github.com/McTavishLab/physcraper\\_ms](https://github.com/McTavishLab/physcraper_ms)





- 317 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool.  
318 *Journal of molecular biology*, **215**, 403–410.
- 319 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped  
320 blast and psi-blast: A new generation of protein database search programs. *Nucleic acids research*, **25**,  
321 3389–3402.
- 322 Antonelli, A., Hettling, H., Condamine, F.L., Vos, K., Nilsson, R.H., Sanderson, M.J., Sauquet, H., Scharn, R.,  
323 Silvestro, D., Töpel, M. & others. (2017). Toward a self-updating platform for estimating rates of speciation  
324 and migration, ages, and relationships of taxa. *Systematic Biology*, **66**, 152–166.
- 325 Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. & Wheeler, D.L. (2000). GenBank.  
326 *Nucleic acids research*, **28**, 15–18.
- 327 Camacho, C., George, C., Vahram, A., Ning, M., Jason, P., Kevin, B. & Thomas, L. (2009). BLAST+:  
328 Architecture and applications. *BMC bioinformatics*, **10**, 421.
- 329 Castresana, J. (2002). GBLOCKS: Selection of conserved blocks from multiple alignments for their use in  
330 phylogenetic analysis. *Version 0.91 b. Copyrighted by J. Castresana, EMBL*.
- 331 Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic  
332 analysis. *Molecular biology and evolution*, **17**, 540–552.
- 333 Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff,  
334 F., Wilczynski, B. & others. (2009). Biopython: Freely available python tools for computational molecular  
335 biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
- 336 Cuénoud, P., Martinez, M.A. del P., Loizeay, P.-A., Spichiger, R., Andrews, S. & Manen, J.-F. (2000).  
337 Molecular phylogeny and biogeography of the genus *Ilex* L.(Aquifoliaceae). *Annals of Botany*, **85**, 111–122.

- Eaton, D.A.R., Spriggs, E.L., Park, B. & Donoghue, M.J. (2016). Misconceptions on Missing Data in RAD-seq  
Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology*, **66**, 399–412. Retrieved  
from <https://doi.org/10.1093/sysbio/syw092>
- Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic  
acids research*, **32**, 1792–1797.
- 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. (2017). 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.
- Gottlieb, A.M., Giberti, G.C. & Poggio, L. (2005). Molecular analyses of the genus *ilex* (aquifoliaceae) in  
southern south america, evidence from *atp* and its sequence data. *American Journal of Botany*, **92**, 352–369.
- Helmus, M.R. & Ives, A.R. (2012). Phylogenetic diversity–area curves. *Ecology*, **93**, S31–S43.
- Izquierdo-Carrasco, F., Cazes, J., Smith, S.A. & Stamatakis, A. (2014). PUmPER: Phylogenies updated  
perpetually. *Bioinformatics*, **30**, 1476–1477.
- Manen, J.-F., Barrier, G., Loizeau, P.-A. & Naciri, Y. (2010). The history of extant *ilex* species (aquifoliaceae):  
Evidence of hybridization within a miocene radiation. *Molecular Phylogenetics and Evolution*, **57**, 961–977.
- McTavish, E.J., Hinchliff, C.E., Allman, J.F., Brown, J.W., Cranston, K.A., Holder, M.T., Rees, J.A. &  
Smith, S.A. (2015). Phylsystem: A git-based data store for community-curated phylogenetic estimates.  
*Bioinformatics*, **31**, 2794–2800.
- Morrison, D.A. (2006). Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*,  
**19**, 479–539.
- Piel, W., Chan, L., Dominus, M., Ruan, J., Vos, R. & Tannen, V. (2009). Treebase v. 2: A database of  
phylogenetic knowledge. E-biosphere.

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

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

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

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

369 Secretariat, G. (2017). GBIF backbone taxonomy. *Checklist Dataset [cited 2017 Nov 14]*. doi, **10**.

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

373 Setoguchi, H. & Watanabe, I. (2000). Intersectional gene flow between insular endemics of ilex (aquifoliaceae)  
374 on the bonin islands and the ryukyu islands. *American Journal of Botany*, **87**, 793–810.

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

377 Song, S., Liu, L., Edwards, S.V. & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using  
378 phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, **109**,  
379 14942–14947.

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

382 Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C.M., Vaidya, G., Pontelli, E.,

383 Cranston, K., Vos, R. & others. (2013). Phylotastic! Making tree-of-life knowledge accessible, reusable and  
384 convenient. *BMC bioinformatics*, **14**, 158.

385 Sukumaran, J. & Holder, M.T. (2010). DendroPy: A python library for phylogenetic computing. *Bioinform-*  
386 *atics*, **26**, 1569–1571.

387 Vos, R.A., Balhoff, J.P., Caravas, J.A., Holder, M.T., Lapp, H., Maddison, W.P., Midford, P.E., Priyam,  
388 A., Sukumaran, J., Xia, X. & others. (2012). NeXML: Rich, extensible, and verifiable representation of  
389 comparative data and metadata. *Systematic biology*, **61**, 675–689.

390 Webb, C.O., Slik, J.F. & Triono, T. (2010). Biodiversity inventory and informatics in southeast asia.  
391 *Biodiversity and Conservation*, **19**, 955–972.

392 Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A. & Rapp,  
393 B.A. (2000). Database resources of the national center for biotechnology information. *Nucleic acids research*,  
394 **28**, 10–14.

395 Yao, X., Song, Y., Yang, J.-B., Tan, Y.-H. & Corlett, R.T. (2020). Phylogeny and biogeography of the hollies  
396 (*Ilex* l., *aquifoliaceae*). *Journal of Systematics and Evolution*.