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

1. Luna L. Sanchez Reyes

School of Natural Sciences, University of California, Merced

email: sanchez.reyes.luna@gmail.com

2. Martha Kandziora

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

email: kandziom@natur.cuni.cz

3. Emily Jane McTavish

School of Natural Sciences, University of California, Merced

email: ejmctavish@gmail.com

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

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

Running title: Continually updated gene trees with Physcraper

Word count: 4198

Manuscript prepared for submission to Methods in Ecology and Evolution

Article type: Application

Abstract

1. Phylogenies are a key part of research in many areas of biology. Tools that 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 an open-source, command-line Python program that automatizes the update of published phylogenies by enriching underlying gene alignments with public DNA sequence data, and linking taxonomic information across databases. This provides a framework for comparison of published phylogenies with their updated versions, by using the conflict Application Programming Interface (API) function of the Open Tree of Life project.

3. Physcraper can be used by the nonspecialist, as a tool to generate phylogenetic hypotheses based on already 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.

GenBank, the USA National Center for Biodiversity Information (NCBI) molecular database, release number 159 (April 15, 2007) hosted 72 million DNA sequences that were gauged to have the potential to resolve phylogenetic relationships of most of its 241 000 distinct taxa (about 98.05% of taxa in the NCBI taxonomy release 159; Sanderson *et al.* 2008). Currently, estimates of phylogenetic relationships are publicly available for about 100 000 taxa only (Piel *et al.* 2009; Hinchliff *et al.* 2015; OpenTreeOfLife *et al.* 2019), representing less than half of the taxonomic diversity with phylogenetically informative sequence data available in GenBank more than a decade ago.

The discrepancy between molecular data availability and phylogenetic estimates can be partially explained by the many phylogenies that are generated and published and not shared publicly in an accesible way (Drew *et al.* 2013; Magee *et al.* 2014; McTavish *et al.* 2017). However, there is also a lag between the amount of new DNA data generated and the analysis of these data in a phylogenetic context.

We address this gap by extending existing phylogenetic estimates with publicly available sequence data. By using a starting tree and single locus alignment, Physcraper, takes advantage of existing research, and extends trees using loci that taxon specialists have assessed and deemed appropriate for the phylogenetic scope. The sequences added in the search are limited to a user specified taxon or monophyletic group, or within the taxonomic scope of the in-group of the starting tree. These automated trees can provide a quick inference of potential relationships, of problems in the taxonomic assignments of sequences, and flag areas of potential systematic interest.

Physcraper leverages public phylogenetic data stored in Open Tree of Life and in TreeBase. The Open Tree of Life (OpenTree from now on https://opentreeoflife.github.io/) is a project that unites phylogenetic inferences

and taxonomy to provide a synthetic estimate of species relationships across the entire tree of life. OpenTree aims to construct a comprehensive, dynamic and digitally-available tree of life by synthesizing published phylogenetic trees along with taxonomic data. This "synthetic" tree comprises 2.3 million tips, of which around 90,000 of those taxa are represented by phylogenetic estimates - the rest are placed in the tree based on their taxonomic names.

The Open Tree of Life data store, the Phylesystem, contains more than 4,500 phylogenetic trees from published studies. The tips in these trees are mapped to a unified taxonomy, which makes these data searchable in a phylogenetically explicit way. This provides a resource for finding existing estimates of phylogenetic relationships, and assessing which regions of the tree of life are lacking available phylogenetic estimates.

By linking molecular data, available from databases such as the GenBank (Benson *et al.* 2000; Wheeler *et al.* 2000), to alignments and phylogenies, available in the TreeBASE repository (Piel *et al.* 2009) and OpenTree's Phylesystem, we can place new biological data in an evolutionary context.

ARGUMENT - GENES ARE STILL USEFUL IN THE GENOMICS ERA, AND HERE'S WHY: GENOMIC MARKERS SUCH AS RADSEQ, SNP, MISCROSATS AND UCES - ARE HOMOLOGY HYPOTHESIS THE SAME ON THESE MARKERS THAN FOR PROTEIN CODING AND NON CODING LOCI? WHAT ARE THE PROS AND CONS OF USING GENE ALIGNMENTS ONLY AND NOT GENOMIC MARKERS? Martha: Genomic data is not available for a large number of taxa. While the focus on single locus and gene sequence alignments could appear backwards-looking, in the age of genomics, single locus data has a lot to offer phylogenetics. One major challenge of inferring phylogenies from genome scale data is inference of homology, and acquiring homologous data across divergent species. Different research questions call for on different approaches to genomic sequencing, from whole genomes, to transcriptomes, to RadSeq, SNP, mirosats and UCE's. This variety of approaches results in non-overlapping data sets across taxa. Even when the same sequencing approach is applied, such as RadSeq, phylogenetic distance can cause allelic dropout at deeper divergences (Eaton *et al.* 2016) In contrast, single locus sequencing generates homologous data across large phylogenetic scales.

Indeed, some systematics support a classic phylogenetics approach (few markers thoughtfully curated) over the genomics approach (a massive amount of DNA markers that will overcome potential errors in the alignment coming from a lack of human curation). Species tree reconstructions from multi-gene data sets taking into account the multispecies coalescent model are considered the gold standard for inferring species relationships [Song *et al.* (2012); ROJAS ET AL. bats paper, take citations from there]. It has also been suggested that manual curation of locus alignments produces better phylogenetic reconstructions and this has been demonstrated for genomic alignments (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). This expert-curated alignments can be continuously enriched and updated by incorporating newly released data from public molecular databases.

In leveraging existing homology statements in the form of alignments, this approach differs from existing approaches that automatize the assembly of DNA alignments from the GenBank database for phylogenetic reconstruction ("phylogenetic pipelines") such as PHYLOTA (Sanderson *et al.* 2008), PHLAWD (Smith *et al.* 2009), and SUPERSMART (Antonelli *et al.* 2017). Physcraper shares a similar conceptual framework to Pumper (Izquierdo-Carrasco *et al.* 2014), but that software is not currently supported or developed (*or runnable at all honestly...*)

Data input availability: As of April 2014, the TreeBASE repository hosted about 8 200 curated alignments, providing information on evolutionary relationships of around 100 000 distinct taxa (see TreeBASE's website about). This database provides an untapped source of valuable expert knowledge with the potential to update phylogenetic relationships in several different regions of the tree of life.

The Phylesystem (OpenTree's datastore) (McTavish *et al.* 2015) automatically incorporates phylogenies from TreeBASE, and saves metadata linking the original tree to its corresponding alignment repository in TreeBASE. If there are multiple alignments, TreeBASE does not always indicate how they were used to generate the tree. This provides a loose means of linking the tree with the exact alignment that generated it.

Often, linking data in an original alignment with its corresponding phylogeny has to be done by a human curator. Moreover, different data repositories follow different systems for taxon and study identification, posing a real challenge to automatically link data from across databases that belong to the same taxon and study. OpenTree's metadata system incorporates taxon identifiers from a variety of taxonomies and repositories, including the NCBI taxonomy, GBIF, etc., MORE EXAMPLES OF DATABASES providing a way to automatically link data from different databases.

Physcraper is a Python encoded pipeline designed to update previously known phylogenetic relationships in a continuous manner, by connecting phylogenies stored in the OpenTree Phylesystem with alignments from TreeBASE and newly released DNA data from GenBank, by using the OpenTree metadata system to connect independent databases through their unique taxon identifiers, automatizing taxonomic name matching across them. By design, this approach focuses on data interoperability. By automating taxonomic name matching across NCBI, OpenTree, GBIF and virtually any biological database, users can perform downstream analyses straightforwardly. For example, it automatizes and standardizes comparison of phylogenetic hypotheses with currently known relationships from the synthetic Open Tree of Life, the Open Tree of Life taxonomy tree and any phylogeny that is stored in the Phylesystem.

We propose Physcraper as a tool to make data connections across biological databases in a phylogenetic context for the advantage of phylogenetics and comparative biology, as well as an effort towards establishing fully reproducible workflows in phylogenetics.

2 The Physcraper framework

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 reccommended that the input tree is either stored in the OpenTree [Phylesystem] (https://github.com/opentreeoflife/phylesystem), or submitted

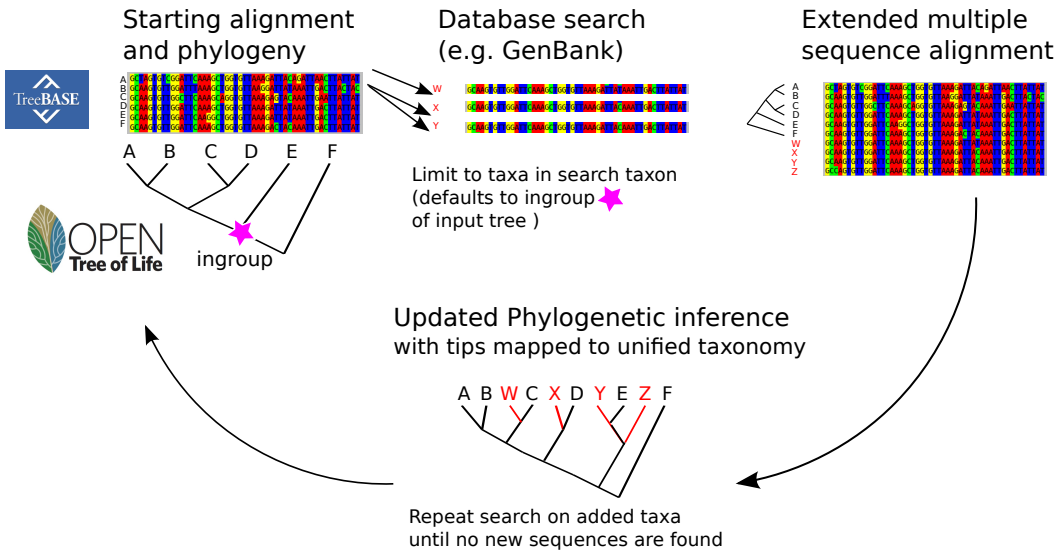

Figure 1: The Physcraper software is fully described on its documentation website at https://physcraper.readthedocs.io/en/latest/, along with installation instructions, tutorials, examples and function usage documentation.

via OpenTree's curator application. 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. Ix taxonomic names can't be standardized, they will be excluded for further analysis. Adding the tree to OpenTree's Phylesystem is recommended because it saves a set of user defined characteristics that are essential for automatizing the phylogeny updating process. The most relevant of these is the standardization of taxonomic names 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 Phylesystem. Users can choose from among the 1216 published trees supporting the resolved nodes of the synthetic tree in the OpenTree website (See OpenTree's website about). *MK suggests that WE SHOULD ACTUALLY SAY HERE HOW MANY OPENTREE TREES HAVE ALIGNMENT DATA FROM TREEBASE, I think it's a good idea*

- 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), or 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)

or through the TreeBASE GitHub repository (SuperTreeBASE; 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.

- When dealing with single locus alignments, it is common for alignments to have less taxa than the tree, simply because a single molecular marker usually does not cover all the taxa included in the phylogenetic analysis. Hence, a pruning step to reconcile taxon presence in the tree and alignment is performed. This 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 coresonding sequence in the alignment) is needed to continue the algorithm. See next section. *MK mentioned that nothing shoul be dropped here bc the name standardization should have matched everything, so I explained a bit more what this step is. It is not an unmatched taxa pruning but more of an absent taxa pruning. Also, taxa that are not matched should be dropped before right?*
- A "csv" file with the summary of taxon name standardization and pruning is produced for the user.
- A "newick" file and a "fasta" file containing the tree and alignment respectively with matched taxa only are generated and saved 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. Using the latter speeds up the process. Detailed instructions to setup a local database are provided on the 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 as the Most Recent Common Ancestor (MRCA) of the matched taxa belonging to the ingroup 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)

I looked for a citation but it seems that it is a concept used on the open tree of life API wikis only. The MRCAT can be different from the phylogenetic MRCA when the latter is an unnamed clade in the reference taxonomy. 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 wait times, which slow down the searches markedly. *MK suggest explaining here what the ncbi waiting times are, but I don;t think it's needed, what do you think?*

- A pairwise BLAST search is performed. This means that each sequence in the alignment is BLASTed against DNA sequences in a nucleotide database constrained to the search taxon. Results from each one of these BLAST runs are written down, and matched sequences are saved along with their corresponding identification numbers, i.e., their GenBank accession numbers. This information will be used later to store the whole sequences in a dedicated library within the "physcraper" folder, allowing for secondary analyses to run significantly faster.

- Matched sequences will be discarded if they fall below a default e-value of 0.00001, and outside a default

minimum and maximum length of 80% and 120%, respectively, of the average length (gaps dropped) of sequences in the checked alignment . These parameters can be configured for each run. This filtering guarantees the exclusion of whole genome sequences. EXPLAIN WHY THIS IS IMPORTANT. All accepted sequences are assigned an internal identifier, and are further filtered.

- New sequences that are identical to existing sequences, or to subsets of an existing sequence are discarded, unless they represent a different taxon than the existing sequence *MK mentioned it was unclear to her if this was an NCBI or OTT taxon. I think it is the unified taxonomy, hence an OTT taxon? I think it doesn;t harm to be super explicit about it.* Longer sequences belonging to the same taxon as the original sequence will be considered further for analysis.

- 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.

- Reverse, complement, and reverse-complement BLAST result sequences are identified and translated using BioPython internal functions.

- 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 processed original alignment are blasted.

- To speed up future runs, accepted sequences are downloaded in full, and stored in a local directory (default to "physcraper/taxonomy" folder) that is globally accessible to users.

- 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

- 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 alignment processing EXAMPLES OF SUCH TOOLS.
- 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 siftware and the manually curated aligment will be ideal. Versioning alignments and adding them to a global database is a next step for us!
- Users may also use Physcraper to gather matched sequences only, and 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 bootsrap 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 each of the genes provided in the input alignment.
- Tips on all trees generated by Physcraper are defined by a taxon "name space", which 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. *It would be a nice addiion to have users give the output of the input tree as an argument at some point, or maybe we could add a super outgroup at random based on the search taxon*

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

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

3.2 The Malvaceae

A postdoc started working with a new reserach group. They are interested in solving relationships among lineages of the Malvaceae, a family of flowering plants with almost 6 000 known species, containing the relatives of cacao, cotton, durian and okra. A review of the literature shows them that there are many phylogenetic trees encompassing some of the linegaes in the group. However, the head of the research group wants to use a particular marker they believe to be the best one to be able to solve the relationships in the

group. They have been working on the alignment for a long time and they want to incorporate new data into the hypothesis of homology that they have been curating and that they trust.

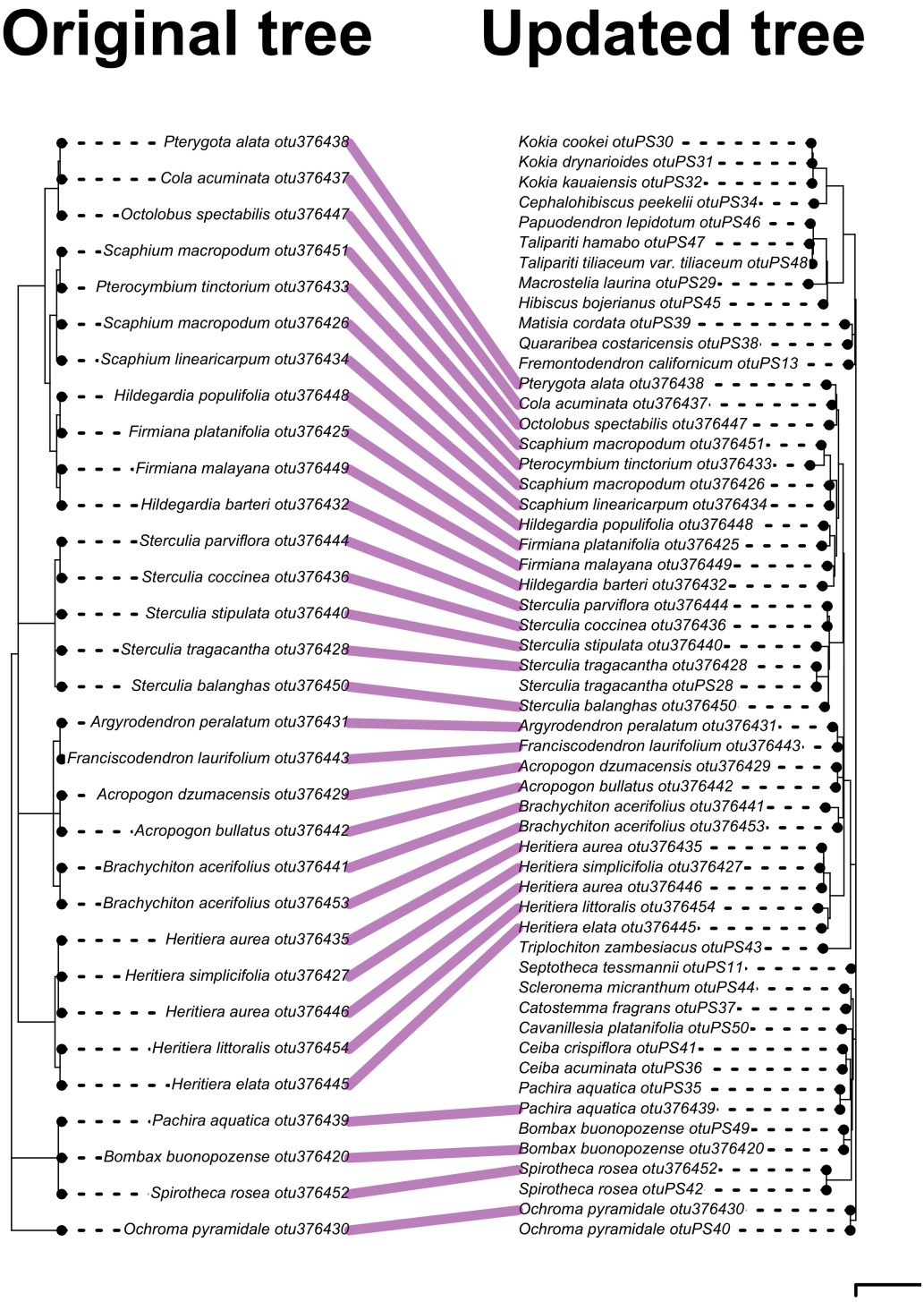

Figure 2: Comparison of original tree and tree updated with Physcraper, family Malvaceae.

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 assignations (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, UCM, 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 Avilability

8 References

354 ## References

355 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool.
356 *Journal of molecular biology*, **215**, 403–410.

357 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped
358 blast and psi-blast: A new generation of protein database search programs. *Nucleic acids research*, **25**,
359 3389–3402.

360 Antonelli, A., Hettling, H., Condamine, F.L., Vos, K., Nilsson, R.H., Sanderson, M.J., Sauquet, H., Scharn, R.,
361 Silvestro, D., Töpel, M. & others. (2017). Toward a self-updating platform for estimating rates of speciation
362 and migration, ages, and relationships of taxa. *Systematic Biology*, **66**, 152–166.

363 Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. & Wheeler, D.L. (2000). GenBank.
364 *Nucleic acids research*, **28**, 15–18.

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

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

370 CUÉNOUD, P., MARTINEZ, M.A.D.P., LOIZEAU, P.-A., SPICHIGER, R., ANDREWS, S. & MANEN,
371 J.-F. (2000). Molecular phylogeny and biogeography of the genus ilex l.(Aquifoliaceae). *Annals of Botany*,
372 **85**, 111–122.

373 Drew, B.T., Gazis, R., Cabezas, P., Swithers, K.S., Deng, J., Rodriguez, R., Katz, L.A., Crandall, K.A.,
374 Hibbett, D.S. & Soltis, D.E. (2013). Lost branches on the tree of life. *PLoS biology*, **11**.

375 Eaton, D.A.R., Spriggs, E.L., Park, B. & Donoghue, M.J. (2016). Misconceptions on Missing Data in RAD-seq
376 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 aflp 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.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R. & others. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, **112**, 12764–12769.

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

Magee, A.F., May, M.R. & Moore, B.R. (2014). The dawn of open access to phylogenetic data. *PLoS One*, **9**.

Manen, J.-F., Barriera, 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., Drew, B.T., Redelings, B. & Cranston, K.A. (2017). How and Why to Build a Unified Tree of Life. *BioEssays*, **39**. Retrieved April 10, 2018, from https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201700114

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). Phylesystem: 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.

OpenTreeOfLife, Redelings, B., Reyes, L.L.S., Cranston, K.A., Allman, J., Holder, M.T. & McTavish, E.J. (2019). Open tree of life synthetic tree. Retrieved from https://doi.org/10.5281/zenodo.3937742

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

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

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

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

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

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

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

Smith, S.A., Beaulieu, J.M. & Donoghue, M.J. (2009). Mega-phylogeny approach for comparative biology:

An alternative to supertree and supermatrix approaches. *BMC evolutionary biology*, **9**, 37.

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

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

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

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

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. & others. (2012). NeXML: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic biology*, **61**, 675–689.

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

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

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