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

School of Natural Sciences, University of California, Merced

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

email: kandziom@natur.cuni.cz

3. Emily Jane McTavish

School of Natural Sciences, University of California, Merced

email: ejmctavish@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: 3948

Manuscript prepared for submission to Methods in Ecology and Evolution

Article type: Application

Abstract

1. Phylogenies are a key part of research in many areas of biology. Tools that automate some parts of the process of phylogenetic reconstruction, mainly molecular character matrix assembly, have been developed for the advantage of both specialists in the field of phylogenetics and nonspecialists. However, interpretation of results, comparison with previously available phylogenetic hypotheses, and choice of one phylogeny for downstream analyses and discussion still impose difficulties to one that is not a specialist either on phylogenetic methods or on a particular group of study.

2. Physcraper is a command-line Python program that automates the update of published phylogenies by adding public DNA sequences to underlying alignments of previously published phylogenies. It also provides a framework for straightforward comparison of published phylogenies with their updated versions, by leveraging upon tools from the Open Tree of Life project to link taxonomic information across databases.

3. Physcraper can be used by the nonspecialist, as a tool to generate phylogenetic hypotheses based on publicly available expert phylogenetic knowledge. Phylogeneticists and taxonomic group specialists will find it useful as a tool to facilitate molecular dataset gathering and comparison of alternative phylogenetic hypotheses (topologies).

4. The Physcraper workflow demonstrates the benefits of doing open science for phylogenetics, encouraging 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, and other metadata in a reproducible and continuous manner is possible by automating connections between biological databases. Here, we introduce Physcraper, a tool that uses existing phylogenetic research from public biological databases to update a starting tree and single locus alignments. Physcraper lets researchers build upon molecular data that taxon specialists have assessed and deemed appropriate for a specific phylogenetic scope.

The prevalence of taxonomic idiosyncrasies across databases represent a key challenge to automatically connecting data from disparate biological databases in a phylogenetic context. To standardize taxonomic names, a unified system is needed. The main aim of the Open Tree of Life project (OpenTree) 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/]. This provideslinks taxonomic information across databases 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 multiple single locus data sets taking into account the multispecies coalescent model are seen as the gold standard for inferring

species relationships (Song *et al.* 2012). Genomics has, and will continue to, revolutionize phylogenetic inference. Yet, different research questions call for different genomic sequencing approaches, from whole genomes, to transcriptomes, restriction-site associated DNA sequencing, single nucleotide polymorphisms, microsatellites, and ultra-conserved elements, which has lead to largely non-overlapping genomic data sets across taxa, creating difficulty in wide scale phylogenetic reconstructions. While phylogenomics ameliorates the problem of non-overlapping genomic data sets by focusing on targeted capture of informative characters from independent and single-copy genetic markers (Jones & Good 2016; Andermann *et al.* 2020), decades of single locus sequencing have already generated homologous DNA data sets 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 DNA sequences. These sequences were gauged to have the potential to resolve phylogenetic relationships of most (98.05%) of the almost 241, 000 distinct taxa in the NCBI taxonomy at the time (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 approach. Computational pipelines that make DNA sequence search faster and more efficient, as well as more reproducible, have been applied to study evolutionary relationships among a variety of organisms (e.g., Smith *et al.* 2009; Izquierdo-Carrasco *et al.* 2014; Antonelli *et al.* 2017). However, even in phylogenomic reconstructions, thoughtfully curated markers and alignments can improve 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 accessible alignments, providing information on evolutionary relationships of around 100, 000 distinct taxa (see TreeBASE's website about), representing a source of valuable expert knowledge. Linking published alignments with public molecular data that has not yet been included in any public phylogenetic estimate, has the potential to accelerate the enrichment and updating of phylogenetic relationships in many regions of the tree of life. The phylogenies associated with TreeBASE alignments have been integrated to the OpenTree's

datastore, the Phylesystem (McTavish *et al.* 2015), and metadata linking them to their corresponding alignment repository is available, providing a non-automated way of linking trees with the alignment that generated them.

Physcraper relies on programmatic access protocols (API's) available to automatically link molecular data from GenBank to alignments from TreeBASE and phylogenies from OpenTree's Phylesystem, to continually 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 comparison of existing phylogenetic hypotheses with newly generated ones. Physcraper is coded as a Python pipeline that focuses on data interoperability, by integrating taxonomic name matching across biological databases. This integration also allows users to rapidly place new data from a diverse range of biological databases in an evolutionary context, making a variety of downstream analyses straightforward.

2 The Physcraper framework

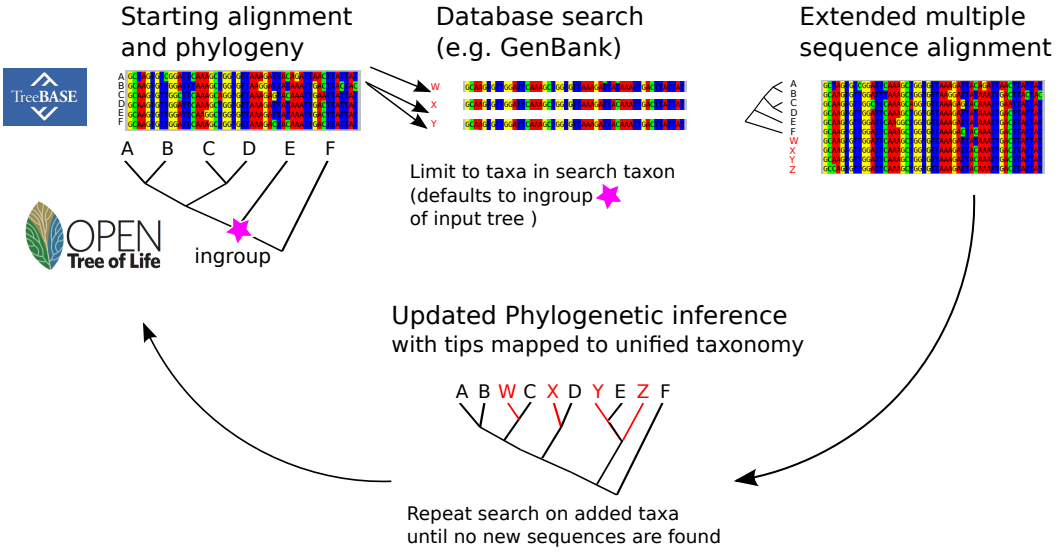

Figure 1: The Physcraper framework consists of 4 steps (see text). The software is fully described on its documentation website at physcraper.readthedocs.io, along with installation instructions, function usage descriptions, examples and tutorials.

The general Physcraper framework is depicted in Figure 1. Briefly, it consists of 4 steps: 1) identifying and processing a phylogenetic tree to update and its underlying alignment; 2) performing a constrained BLAST

search of sequences from the original alignment on the GenBank DNA database, and filtering of newly found sequences; 3) profile-aligning new sequences that passed the filtering to the original alignment; 4) performing a phylogenetic analysis and comparing the updated tree to previous phylogenetic estimates in the focus group. Next, we will describe technical details for each step.

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, or submitted via OpenTree's curator application (McTavish *et al.* 2015). Currently, only trees connected to a published study can be stored in the Phylesystem. Users can choose from among the 2, 950 studies in OpenTree's Phylesystem that have alignments on TreeBASE. If the user is not ready to make the input tree public, tree tip labels can be standardized to the unified OpenTree taxonomy using OpenTree's bulk Taxonomic Name Resolution Service TNRS tool. This step is referred to as taxonomic name mapping and Phylesystem stored trees are processed in this way upon submission. Physcraper saves a summary "csv" file with results from the taxon name standardization for advantage of the user, including the mappings to unique identifiers in the OpenTree and NCBI taxonomies. If taxon names can't be mapped, their taxonomic information is not used in comparison analysis. These taxa will still be used in the sequence search and phylogenetic reconstruction steps. 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 standardized 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.

The input alignment should be a single locus alignment that was used in part or in whole, 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 a data repository associated with the journal where the tree was originally published. If the alignment is stored in TreeBase, Physcraper can download 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.

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. 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 API (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. 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 OTT taxonomy or the NCBI 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 would 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 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 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 Example: The hollies

To illustrate the utility of Physcraper we propose a scenario in which a user is interested in phylogenetic relationships within the genus *Ilex*. Commonly known as "hollies", the genus encompasses between 400-700 living species, and is the only extant clade within the family Aquifoliaceae, order Aquifoliales of flowering plants.

An online literature review in June 2020 (google scholar search for "ilex phylogeny") reveals that there are several published 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 original tree and alignment data available in TreeBASE. The "Gottlieb2005" tree sampling 41 species was added to the OpenTree Phylesystem and its information has been integrated into 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. The "Yao2020" tree, is the best sampled phylogenetic tree yet available for the hollies, with 175 tips. This makes it a great case to highlight the functionalities of Physcraper and assess its limitations. A tutorial as well as illustrated examples of functions implemented on each step of the analysis are available in Physcraper's documentation website. Figure 2 shows results from a Physcraper analysis applied to an alignment of internal transcribed spacer DNA regions (ITS) from Gottlieb *et al.* (2005). Physcraper ran on a local BLAST database, on a laptop Linux computer for 19hrs 45min to perform BLAST and RAxML analyses, with bootstrap analyses taking an additional 13hrs. **MTH: you probably need some details 252 about the hardware, given the fact that you are discussing running times** The updated Gottlieb2005 tree contains all 41 distinct taxa from the original study plus 231 new tips, contributing phylogenetic data to 84 additional taxa within *Ilex*. The best RaxML tree is 99% resolved, while the consensus tree is 75% resolved, meaning that 25% of nodes have a bootstrap < 0.1, while 48% nodes have bootstrap support > 0.75. A large portion of internal branches are negligibly small, with 30 branches < 0.00001 substitution rate units, from which only 9 have a bootstrap support > 0.75 (Fig. 2). For comparison, the Yao2020 tree also contains all 41 distinct taxa from the original Gottlieb2005 study, and contributes phylogenetic data to 134 additional taxa within *Ilex*, from which 67 are also in the updated Gottlieb2005 tree. Yao *et al.* (2020) also used ITS as a marker, but their data in GenBank is not public yet, so Physcraper was unable to incorporate 68 additional taxa into the analysis. When those sequences are available online, repeating the physcraper run will easily incorporate them into the expanded analysis. The Yao2020 tree also lacks 18 taxa that were added by Physcraper. This might be caused by the method they used to download existing ITS *Ilex* sequences from GenBank, which is not fully explained in the publication.

Other examples of Physcraper runs on empirical data at a range of taxonomic scales are available in the Physcraper documentation online.

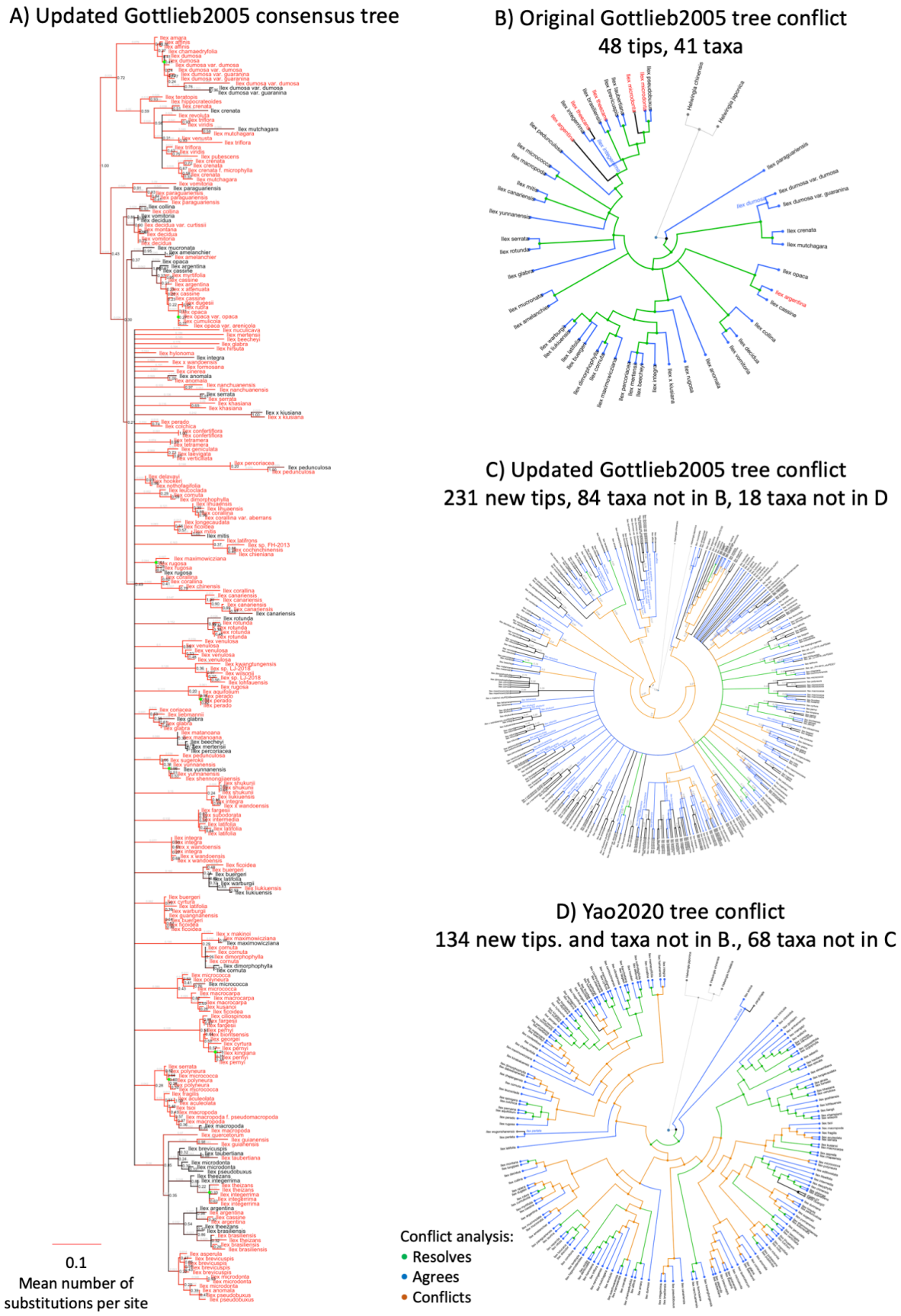

Figure 2: A) Phylogenetic tree obtained by updating the Gottlieb et al. 2005 tree in B) using Physcraper.

Figure 2 caption continued: Tips in original alignment and new tips added with Physcraper are depicted in black and red, respectively. Physcraper obtained sequences from the GenBank database via local BLAST of all sequences in the original alignment that generated tree in B), filtered them following criteria specified 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 show results of the conflict analysis performed with OpenTree tools.

4 Discussion

Data repositories designed to preserve and democratize access to biological data constitute an essential resource for evolutionary estimation. While data keep accumulating, our ability to reanalyze and incorporate new data with available knowledge is being challenged.

Various phylogenetic pipelines have been designed to make evolutionary sense of the vast amount of public molecular data available in different ways. Pipelines like Phylota (Sanderson *et al.* 2008), PyPHLAWD (Smith *et al.* 2009), and SUPERSMART (Antonelli *et al.* 2017) focus mainly on generating full trees *de novo* (i.e., inferring phylogenetic relationships from a newly generated homology hypothesis, as opposed to e.g., supertrees, that are generated by assembling previous phylogenetic estimates) for regions of the Tree of Life that have no phylogenetic assessment yet in published studies. While Physcraper does not generate phylogenies *de novo* in a traditional sense, it successfully generates new phylogenetic knowledge, revealing the potential of pre-existing phylogenetic knowledge available in tree repositories to facilitate phylogenetic placement of public molecular data. The PUMPER pipeline (Izquierdo-Carrasco *et al.* 2014) also uses the concept of updating pre-existing alignments to incorporate public molecular data into phylogenies. Unfortunately, installation of the tool was unsuccesful following instructions from the author, and we were unable to conduct a comparison analysis to Physcraper. **MTH: add some more details about what aspect of the PUMPER install failed** Physcraper on it's own generates individual gene trees, which do not capture the full complexity of species' evolutionary history (Song *et al.* 2012). However, using Physcraper it is straightforward to gather alignments and gene trees for multiple loci for a taxonomic group of interest, which can be used as inputs

for downstream tools to perform coalescent analyses and assess species tree relationships, (e.g. ASTRAL (Mirarab *et al.* 2014), BEAST2 (Bouckaert *et al.* 2019), SVD Quartets (Chifman & Kubatko 2014)).

As far as we know, Physcraper is the only phylogenetic pipeline that links molecular and tree data repositories. It can also link to various biological databases, such as GBIF, the Paleobiology Database, and others, by leveraging the integrated taxonomy from OpenTree (Rees & Cranston 2017), a resource that is contributing to the Encyclopedia of Life main goal of "exploring and analysing biodiversity at an accelerated pace, and returning systematics into the mainstream of science" (Wilson 2003).

As such, Physcraper can be used to rapidly (in a matter of hours) solve challenges overarching both fields of ecology and evolution, such as placing newly discovered species phylogenetically (Webb *et al.* 2010), obtaining trees for ecophylogenetic studies (Helmus & Ives 2012), systematizing molecular (and other) databases, i.e., curating taxonomic assignations (San Mauro & Agorreta 2010), and generating custom species trees for ecological and evolutionary downstream analyses (Stoltzfus *et al.* 2013).

Data repositories hold more information than meets the eye. Besides 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. Usually, initial ideas about the data are changed by new analyses. Ideally, new understanding of the data can be continually registered on databases to consistently expose newcomers to the most up to date knowledge about the data.

5 Acknowledgements

Research was supported by the grant "Sustaining the Open Tree of Life", National Science Foundation ABI No. 1759838, and ABI No. 1759846. Computer time was provided by the Multi-Environment Research Computer for Exploration and Discovery (MERCED) cluster from the University of California, Merced (UCM), supported by the NSF Grant No. ACI-1429783.

We thank the members of the OpenTree development team and the "short bar" Science and Engineering Building 1, UCM, joint lab paper discussion group for valuable comments on this manuscript.

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

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

8 References

References

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

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

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. (2020). A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. *Frontiers in Genetics*, **10**. Retrieved July 28, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7047930/

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

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

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. du, Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M.A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T. & Drummond, A.J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, **15**, e1006650. Retrieved August 18, 2020, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006650

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

Castresana, J. (2002). GBLOCKS: Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Version 0.91 b. Copyrighted by J. Castresana, EMBL.*

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, **17**, 540–552.

Chifman, J. & Kubatko, L. (2014). Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, **30**, 3317–3324. Retrieved August 18, 2020, from https://academic.oup.com/bioinformatics/article/30/23/3317/206559

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

Cuénoud, P., Martinez, M.A. del P., Loizeay, P.-A., Spichiger, R., Andrews, S. & Manen, J.-F. (2000). Molecular phylogeny and biogeography of the genus ilex l.(Aquifoliaceae). *Annals of Botany*, **85**, 111–122.

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.

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

Jones, M.R. & Good, J.M. (2016). TARGETED CAPTURE IN EVOLUTIONARY AND ECOLOGICAL GENOMICS. *Molecular ecology*, **25**, 185–202. Retrieved July 28, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4823023/

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., 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.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S. & Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548. Retrieved July 25, 2015, from http://bioinformatics.oxfordjournals.org/content/30/17/i541

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.

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.

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

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.

Wilson, E.O. (2003). The encyclopedia of life. *Trends in Ecology & Evolution*, **18**, 77–80.

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*.