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

email: martha.kandziora@mailbox.org

3. Emily Jane McTavish

School of Natural Sciences, University of California, Merced

email: ejmctavish@gmail.com

Correspondence:

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

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

Running title: Continually updated gene trees with Physcraper

Word count: 4021

Manuscript prepared for submission to Methods in Ecology and Evolution

Article type: Application

1 Abstract

1. Phylogenies are a key part of research in all areas of biology. Tools that automatize some parts of the process of phylogenetic reconstruction (mainly character matrix construction) 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 making use of public DNA sequence data and taxonomic information, providing a framework for comparison of published phylogenies with their updated versions.

3. Physcraper can be used by the nonspecialist, as a tool to generate phylogenetic hypothesis based on already available expert phylogenetic knowledge. Phylogeneticists and group specialists will find it useful as a tool to facilitate comparison of alternative phylogenetic hypotheses (topologies). ***Is physcraper intended for the nonspecialist?? We have two types of nonspecialists: the ones that do not know about phylogenetic methods and the ones that might know about phylogenetic methods but do not know much about a certain biological group.***

4. Physcraper implements node by node/topology comparison of the the original and the updated trees using the conflict API of OToL, and summarizes differences.

5. We hope the physcraper workflow demonstrates the benefits of opening results in phylogenetics and encourages researchers to strive for better data sharing practices.

6. Physcraper can be used with any OS. Detailed instructions for installation and use are available at https://github.com/McTavishLab/physcraper.

Keywords: cross-connectivity, gene tree, open science, open tree of life, phylogeny, public database, python, reproducibility, taxonomy, update alignment

2 Introduction

Phylogenies are important.

Generating phylogenies is not easy and it is largely artisanal. Although many efforts to automatize the process have been done, and the community is using those more and more, automatization of phylogenetic reconstruction is still not a widespread practice and among other benefits, it might be key for adoption of better reproducibility practices in the phylogenetics community. ***paragraph better to end discussion???***

The process of phylogenetic reconstruction implies many steps (that I generalize to the following):

1. Obtention of molecular or morphological character data – get DNA from some organisms and sequence it, or get it from an online nucleotide data repository, such as GenBank (Benson et al. 2000; Wheeler et al. 2000).
2. Assemble a hypothesis of homology – Create a matrix of your character data, by aligning the sequences, in the case of molecular data. Make sure thay are paralogs!
3. Analyse this hypothesis of homology to infer phylogenetic relationships among the organisms you are studying – Use different available programs to infer molecular evolution, trees and times of divergence.
4. Discuss the inferred relationships in the context of previous hypothesis, the biology and biogeography of the organisms, etc. – Answer the question, *is this phylogenetic solution fair/reasonable?*

Each of these steps require different types of specialized training: in the field, in the lab, in front of a computer, discussions with experts in the methods, and/or in the biological group of study. All of these steps also require considerable amounts of time for training and implementation.

In the past decade, various studies have developed solutions to automatize the first and second steps, by creating pipelines that mine already available molecular data from the GenBank repository (Benson et al. 2000; Wheeler et al. 2000), to obtain homologous characters that can be used for phylogenetic reconstruction. These tools have been presented as aid for the nonspecialist to decrease some of the difficulties in the generation of phylogenetic knowledge. However, they are not that often used as so, suggesting that there are

still difficulties for the nonspecialist. The phylogenetic community has some reserves towards these tools, too. Mainly because they sometimes act as a black box. However, automatizing the assembly of the character data set is a crucial step towards reproducibility for a task that was otherwise primarily artisanal and hence largely non-reproducible.

Even if it is hard to obtain phylogenies, we invest copious amounts of time and energy in generating them. Issues such as food security, global warming, global health are crucial to solve and phylogenies might help. There is a lot of phylogenetic knowledge already available in published peer-reviewed studies. In this sense, the non-specialists (and also the specialist) face a new problem: how do I choose the best phylogeny.

Public phylogenies can be updated with the ever increasing amount of genetic data that is available on GenBank (Benson et al. 2000; Wheeler et al. 2000).

We present a way to automatize and standardize the comparison of phylogenetic hypotheses and to allow reproducibility of this last step of the research process.

A key aspect of the standard phylogenetic workflow is comparison with already existing phylogenetic hypotheses and with phylogenies that are considered "best" by experts not only in phylogenetics, but also experts on the focal group of study.

Concerns I think people have about these tools: - Errors in identification of sequences - Little control along the process - Too much of a black box?

Most of these phylogenies are being constructed by people learning about the methods, so they want to know what is going on.

The pipelines are so powerful and they will give you an answer, but there is no way to assess if it is better than previous answers, it just assumes it is better because it used more data.

All these pipelines start tree construction from zero? Yes.

The goal of Physcraper is to build upon previous phylogenetic knowledge, allowing a direct comparison

between existing phylogenies and phylogenies that are constructed using new genetic data retrieved from a public nucleotide database (i.e., GenBank (Benson et al. 2000; Wheeler et al. 2000)).

To achieve this, Physcraper uses the Open Tree of Life phylesystem and connects it to the TreeBase database, to (1) get the original DNA data set matrices (alignments) that produced a phylogeny that was published and then made available in the OToL database, (2) use this DNA alignments as a starting point to get new genetic data belonging to the focal group of study, to (3) finally update the phylogenetic relationships in the group.

A less automated workflow is one in which the alignments that generated the published phylogeny are stored in other public database (such as DRYAD) or elsewhere (the users computer), and are provided by the users.

The original tree is by default used as starting tree for the phylogenetic searches, but it can also be set as a full topological constraint or not used at all, depending on the goals of the user.

Physcraper implements node by node comparison of the the original and the updated trees, using the conflict API of OToL.

3 How does Physcraper work?

3.1 The input: a study tree and an alignment

- The study tree is a published phylogenetic tree stored in the OToL database, phylesystem (McTavish et al. 2015). The main reason for this is that trees in phylesystem have a set of user defined characteristics that are essential for automatizing the phylogeny update process. The most relevant of these being the definition of ingroup and outgroup. Outgroup and ingroup taxa in the original tree are identified and tagged. This allows to automatically set the root for the updated tree on the next steps of the pipeline. A user can choose from the 'r rotl::tol_about()$num_source_trees' published trees supporting the resolved node of the synthetic tree in the OToL website (<>). If the tree you are interested in updating is not in there, you can upload it via OToL's curator tool (<https://tree.opentreeoflife.org/curator).

- The alignment should be a gene alignment that was used to generate the tree. The original alignments are usually stored in a public repository such as TreeBase (Piel et al. 2009; Vos et al. 2012), DRYAD (http://datadryad.org/), or the journal were the tree was originally published. If the alignment is stored in TreeBase, `physcraper` can download it directly, either from the TreeBASE website (https://treebase.org/) or through the TreeBASE GitHub repository (SuperTreeBASE; https://github.com/TreeBASE/supertreebase). If the alignment is on another repository, or provided personally by the owner, a copy of it has to be downloaded by the user, and it's local path has to be provided as an argument.
- A taxon name matching step is performed to verify that all taxon names on the tips of the tree are in the DNA character matrix and vice versa.
- A ".csv" file with the summary of taxon name matching is produced for the user.
- Unmatched taxon names are dropped from both the tree and alignment. Technically, just one matching name is needed to perform the searches. Please, see next section.
- A ".tre" file and a ".fas" file containing only the matched taxa are generated and saved in the `inputs` folder to be used in the following steps.

3.2 DNA sequence search and cleaning

- The next step is to identify the search taxon within the reference taxonomy. The search taxon will be used to constraint the DNA sequence search on the nucleotide database within that taxonomic group. Because we are using the NCBI nucleotide database, by default the reference taxonomy is the NCBI taxonomy. The search taxon can be provided by the user. If none is provided, then the search taxon is identified as the Most Recent Common Ancestor (MRCA) of the matched taxa belonging to the ingroup in the tree, that is also a named clade in the reference taxonomy. This is known as the Most Recent Common Ancestral Taxon (MRCAT; also referred in the literature 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 reference taxonomy. To automatically identify the MRCAT of a group of taxon names, we make use of the OToL taxonomy tool (https:

//github.com/OpenTreeOfLife/germinator/wiki/Taxonomy-API-v3#mrca).

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); altschul1997gapped] is used to identify similarity between DNA sequences within the search taxon in a nucleotide database, and the accepted sequences on the alignment. The blastn function from the BLAST command line tools (Camacho et al. 2009) is used for local-database searches. A modified biopython blast function is used for web-based searches.

- The DNA sequence similarity search can be done on a local database that is easily setup by the user. In this case, the blastn function is used to performs the similarity search (Camacho et al. 2009).

- The search can also be performed remotely, on the NCBI database. In this case, the bioPython BLAST function was modified to accepts is used to perform the similarity search.

- A pairwise alignment-against-all 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 recorded, and matched sequences are saved along with their corresponding identification numbers (accesion numbers in the case of the GenBank database). 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 below an e-value, percentage similarity, and outside a minimum and maximum length threshold are discarded. ***REPORT THE DEFAULT VALUES AND DESCRIBE WHAT THEY MEAN*** This filtering leaves out genomic sequences. All acepted sequences are asigned an

internal identifier, and are further filtered.

- Because the original alignments usually lack database id numbers, a filtering step is needed. Accepted sequences that belong to the same taxon of the query sequence, and that are either identical or shorter than the original sequence are discarded. Only longer sequences belonging to the same taxon as the orignal sequence will be considered further for analysis.

- Among the remaining filtered sequences, there are usually several exemplars per taxon. Although it can be useful to keep some of them to, for example, 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 sequences are identified and translated.

- Users can choose to perform a more "cycles" of sequence similarity search, by blasting the newly found sequences. This can be done iteratively, but by default only sequences in the alignment are blasted. ***Is there an argument to control the number of cycles of blast searches with new sequences?***

- Accepted sequences are downloaded in full, and stored as a local database in a directory that is globally accesible (physcraper/taxonomy), so they are accesible for further runs.

- A fasta file containing all filtered and processed sequences resulting from the BLAST search is generated for the user.

3.3 DNA sequence alignment

- The software MUSCLE (Edgar 2004) is implemented to perform alignments.
- First, all new sequences are aligned using default MUSCLE options.
- Then, a MUSCLE profile alignment is performed, in which the original alignment is used as a template to align new sequences. This ensures that the final alignment follows the homology criteria established

by the original alignment.

- The final alignment is not further processed automatically. We encourage users to check it either by eye and perform manual refinement or using any of the many tools for alignment processing, to eliminate columns with no information.

3.4 Tree reconstruction and comparison

- A gene tree is reconstructed for each alignment provided, using a Maximum Likelihood approach implemented with the software RAxML (Stamatakis 2014) with 100 classic rapid bootstrap (Felsenstein 1985) replicates by default. The number of bootsrap replicates can be modified by the user. Other type of bootstrap that I think is not yet incorporated into physcraper is the Transfer Bootstrap Expectation (TBE) recently proposed in Lemoine et al. (2018).
- Bootstrap results are summarized with Dendropy ADD CITATION
- The final result is an updated phylogenetic hypothesis for each of the genes provided in the alignment.
- Tips on all trees generated by physcraper are defined by a taxon name space, allowing to perform comparisons and conflict analyses.
- Robinson Foulds weighted and unweighted metrics ARE CALCULATED WITH DENDROPY TOO.
- Describe what a conflict analysis is: Node by node comparison of the resulting clades compared to CITE REDELINGS AND HOLDER (**???** and holder)
- For the conflict analysis to be meaningful, the root of the tree ineeds to be accurately defined.
- A SUGGESTED DEFAULT ROOTING BASED ON THE OPEN TREE TAXONOMY is implemented for now. DESCRIBE HOW IT WORKS. SAY THAT IT IS A PROBLEM. Automatic rooting is not that smart yet. The best way right now is for users to define outgroups so trees are better rooted.
- Currently, the root is determined by finding the parent node of the sequences that do not belong to the ingroup/ search taxon. This ensures a correct rooting of the tree even when the search taxon is more inclusive than the ingroup.
- Conflict information can only be generated in the context of the whole Open Tree of Life. Otherwise, it is not really possible to get conflict data. ***- One way to compare two independent phylogenetic***

trees is to compare them both to the synthetic OToL and then measure how well they do against each other

4 Examples

4.1 The hollies

The genus *Ilex* is the only extant clade within the family Aquifoliaceae, order Aquifoliales of flowering plants. It encompasses between 400-600 living species. A review of litterature shows that there are three published phylogenetic trees, showing relationships within the hollies. The first one has been made available both on OToL phylesystem and synth tree, and on treeBASE, it samples 48 species. The second has not been made available anywhere, not even in supplementary data of the journal. ***Contact authors? They seem old school, probably do not wanna share their data.*** The most recent one has been made available in the OToL Phylesystem and DRYAD. It is the best sampled yet, with 200 species. However, it has not been added to the syntehtic 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 (Benson et al. 2000; Wheeler et al. 2000). Updating the oldest tree, we should get something very similar to the newest tree.

4.2 The Ascomycota

Let's be more specific now about our X group and say it is the Ascomycota. The best tree currently available in OToL was published by Schoch et al. (2009). The first step, is to get the Open Tree of Life study id. There are some options to do this: - You can go to the Open Tree of Life website and browse until you find it, or - you can get the study id using R tools: - By using the TreeBase ID of the study (which is not fully exposed on the TreeBase website home page of the study, so you have to really look it up manually):

```
rotl::studies_find_studies(property = "treebaseId", value = "S2137")

##   study_ids n_trees tree_ids candidate study_year title

## 1    pg_238       1  tree109                2009
```

```
##                                    study_doi
## 1 http://dx.doi.org/10.1093/sysbio/syp020
```

- By using the name of the focal clade of study (but this behaved very differently):

```
rotl::studies_find_studies(property="ot:focalCladeOTTTaxonName", value="Ascomycota")
```

Once we have the study id, we can gather the trees published on that study:

```
rotl::get_tree_ids(rotl::get_study_meta("pg_238"))
## [1] "tree109"
rotl::candidate_for_synth(rotl::get_study_meta("pg_238"))
## NULL
my_trees <- rotl::get_study("pg_238")
```

Both trees from this study have NA tips.

Let's check what one of the trees looks like:

1. Download the alignment from TreeBase If you are on the TreeBase home page of the study, you can navigate to the matrix tab, and manually download the alignments that were used to reconstruct the trees reported on the study that were also uploaded to TreeBase and to the Open Tree of Life repository. To make this task easier, you can use a command to download everything into your working folder:

```
physcraper_run.py -s pg_238 -t tree109 -o ../physcraper_example/pg_238
```

In this example, all alignments posted on TreeBase were used to reconstruct both trees.

1. With the study id and the alignment files saved locally, we can do a physcraper run with the command:

```
physcraper_run.py -s pg_238 -t tree109 -a treebase_alns/pg_238tree109.aln -as "nexus" -o pg_238
```

4.3 Testudines example

Phylogeny of the Testudines 6 tips from Crawford et al. (2012) There is just one tree in OToL. There is just one alignment on treebase with all the 1 145 loci.

```
physcraper_run.py -s pg_2573 -t tree5959 -tb -db ~/branchinecta/local_blast_db/ -o pg_2573
```

5 Discussion

Data repositories hold more information than meets the eye. Besides the actual data, they have other types of information that can be used for the advantage of science.

Usually, initial ideas about the data are changed by analyses. We expect that this new ideas on the data can be registered on data bases, exposing new comers to expert understanding about the data.

There are many tools that 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 refinement of 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 been already studied and that have phylogenetic data already.

All these tools are great efforts for advancing towards reproducibility in phylogenetics, a field that has been largely recognised as somewhat artisanal. We propose adding focus to other sources of information available from data repositories. Taking advantage of public DNA data bases have been the main focus. However, phylogenetic knowledge is also accumulating fast in public and open repositories. In this way, the physcraper pipeline can be complemented with other tools that have been developed for other purposes.

We emphasize that physcraper takes advantage of the knowledge and intuition of the expert community to build upon phylogenetic knowledge, using not only data accumulated in DNA repositories, but phylogenetic knowledge accumulated in tree repositories. This might help generate new phylogenetic data. But physcraper does not seek to generate full phylogenies *de novo*.

Describe again statistics to compare phylogenies provided by physcraper via OpenTreeOfLife. Mention statistics provided by other tools: PhyloExplorer (Ranwez et al. 2009). Compare and discuss.

How is physcraper already useful: - to mine targeted sequences, in this way it is similar to baited analyses from PHLAWD and pyPHLAWD. Phylota does not do baited analyses, I think, only clustered analyses. - Finding

How can it be used for the advantage of the field: - rapid phylogenetic placing of newly discovered species, as mentioned in Webb et al. (2010) - obtain trees for ecophylogenetic studies, as mentioned in Helmus and Ives (2012) - one day could be used to sistematize nucleotide databases, such as Genbank (Benson et al. 2000; Wheeler et al. 2000), as mentioned in San Mauro and Agorreta (2010), i.e., curate ncbi taxonomic assignations. - allows to generate custom species trees for downstream analyses, as mentioned in Stoltzfus et al. (2013)

Things that physcraper does not do: - analyse the whole GenBank database (Benson et al. 2000; Wheeler et al. 2000) to find homolog regions suitable to reconstruct phylogenies, as mentioned in Antonelli et al. (2017). There are already some very good tools that do that. - provide basic statistics on data availability to assemble molecular datasets, as mentioned by Ranwez et al. (2009). Phyloexplorer does this? - it is not a tree repo, as phylota is, mentioned in Deepak et al. (2014)

5.1 Tools that automatize any part of the process of phylogenetic reconstruction:

5.1.1 1. Mining DNA databases to generate datasets suitable for phylogenetic reconstruction

Tool	Citation	Cited by	Description	Supermatrix/gene tree/species tree
Phylota	Sanderson et al. (2008)	122 studies	finds sets of DNA homologs on the GenBank database; phylogenetic reconstruction	Supermatrix
AMPHORA	Wu and Eisen (2008)	458 studies	baited search; protein markers on phylogenomic data; personal database of genomes or metagenomic data, manually downloaded either from a public database or from private data; phylogenetic reconstruction	Supermatrix
PHLAWD	Smith et al. (2009)	234 studies	Baited search of DNA markers on the GenBank database; phylogenetic reconstruction	Supermatrix

Tool	Citation	Cited by	Description	Supermatrix/gene tree/species tree
Unnamed ruby pipeline, only available from supplementary data of the journal	Peters et al. (2011)	64 studies	mining public DNA databases, focuses on filtering massive amounts of mined sequences by using established "criteria of compositional homogeneity and defined levels of density and overlap"	Supermatrix
Unnamed	Grant and Katz (2014)	38 studies	predecessor of phylotol; homolog clustering; public and/or personal DNA database; phylogenetic reconstruction; broad taxon analyses; remove contaminant sequences, based on similarity and on phylogenetic position	supermatrix

Tool	Citation	Cited by	Description	Supermatrix/gene tree/species tree
Unnamed	Chesters and Zhu (2014)	10 studies	algorithm that mines GenBank data to delineate species in the insecta. The authors present a nice comparison with the phylota algorithm	Species trees??
PUmPER	Izquierdo-Carrasco et al. (2014)	14 studies	perpetual updating with newly added sequences to GenBank	not sure yet
DarwinTree	Meng et al. (2015a)	6 studies	predecessor is Phylogenetic Analysis of Land Plants Platform (PALPP), takes data from GenBank, EMBL and DDBJ for land plants only	not sure
NCBIminer	Xu et al. (2015)	4 studies	part of darwintree	not sure

Tool	Citation	Cited by	Description	Supermatrix/gene tree/species tree
SUMAC	Freyman (2015)	19 studies	both "baited" analyses and single-linkage clustering methods, as well as a novel means of determining when there are enough overlapping data in the DNA matrix	not sure
STBase	McMahon et al. (2015)	7 studies	pipeline for species tree construction and the public database of one million precomputed species trees	species trees
Unnamed	Papadopoulou et al. (2015)	17 studies	Automated DNA-based plant identification for large-scale biodiversity assessment	not sure

Tool	Citation	Cited by	Description	Supermatrix/gene tree/species tree
BIR	Kumar et al. (2015)	6 studies	blast, align, identify homologs via constructed trees, curate and realign	supermatrix
SUPERSMART	Antonelli et al. (2017)	35 studies	baited analyses up to bayesian divergence time estimation	supermatrix
SOPHI	[Chesters (2017)	17 studies	Searches DNA sequence data from repos other than GenBank, such as transcriptomic and barcoding repos	not sure
phyloSkeleton	Guy (2017)	5 studies	focuses on taxon sampling; baited genomic sequences; public database (NCBI and JGI); marker identification	supermatrix

Tool	Citation	Cited by	Description	Supermatrix/gene tree/species tree
OneTwoTree	Drori et al. (2018)	7 studies	Web-based, user-friendly, online tool for species-tree reconstruction, based on the *supermatrix paradigm* and retrieves all available sequence data from NCBI GenBank	supermatrix
pyPhlawd	Smith and Walker (2019)	6 studies	baited and clustering analyses	Supermatrix or gene tree

Tool	Citation	Cited by	Description	Supermatrix/gene tree/species tree
Phylotol	Cerón-Romero et al. (2019)	5 studies	"phylogenomic pipeline to allow easy incorporation of data from high-throughput sequencing studies, to automate production of both multiple sequence alignments and gene trees, and to identify and remove contaminants. PhyloToL is designed for phylogenomic analyses of diverse lineages across the tree of life", i.e., bacteria and unicellular eukaryotes	supermatrix and gene trees
phylotaR	Bennett et al. (2018)	studies		

According to Cerón-Romero et al. (2019), PhyLoTA and BIR "focus on the identification and collection of homologous and paralog genes from public databases such as GenBank", while both AMPHORA and PHLAWD "focus on the construction and refinement of robust alignments rather than the collection of homologs."

5.1.2 2. Searching phylogenetic tree databases

PhyloFinder (Chen et al. 2008) - cited by 18: a search engine for phylogenetic databases, using trees from TreeBASE - more related to phylotastic's goal than to updating/creating phylogenies

5.1.3 3. Mining phylogenetic tree databases

PhyloExplorer (Ranwez et al. 2009) - cited by 21: a python and MySQL based website to facilitate assessment and management of phylogenetic tree collections. It provides "statistics describing the collection, correcting invalid taxon names, extracting taxonomically relevant parts of the collection using a dedicated query language, and identifying related trees in the TreeBASE database".

5.1.4 4. Pipeline for phylogenetic reconstruction

PhySpeTre (Fang et al. 2019) - no citations yet - no sequence retrieval, just phylogenetic reconstruction pipeline.

5.1.5 5. getting metadata and not sequences from GenBank.

Datataxa Ruiz-Sanchez et al. (2019) - no citations yet - focus on extracting metadata from GenBank sequence information.

5.2 Phylota overview

Phylota was published as a website to summarize and browse the phylogenetic potential of the GenBank database (Sanderson et al. 2008).

Since then, it has been cited 122 times for different reasons.

1. As an example of a tool that mines GenBank data for phylogenetic reconstruction, or that is useful in any way for phylogenetics:
    - original publication of PHLAWD (Smith et al. 2009)
    - an analysis identifying research priorities and data requirements for resolving the red algal tree of

life (Verbruggen et al. 2010)

- Beaulieu et al. (2012a) cites phylota as an example study of very large and comprehensive phylogeny from mined DNA sequence data, (even if no phylogeny was really published there, only the method to do so)
- a review for ecologists about phylogenetic tools (Roquet et al. 2013)
- a study constructing a dated seed plant phylogeny using pyPHLAWD (Smith and Brown 2018)
- a study presenting an "assembly and alignment free" method for phylogenetic reconstruction using genomic data. It aims to be incorporated into a pipeline such as phylota some day (Fan et al. 2015).
- nexml format presentation (Vos et al. 2012) - cites phylota as a tool that uses stored phyloinformatic data that could benefit from adopting nexml, to increase interoperability.
- a study of fruit evolution, analysing a previously published phylogeny of 8911 tips of the Campanulidae, constructed with PHLAWD (Beaulieu and Donoghue 2013)
- a study of Southeast Asia plant biodiversity inventory (Webb et al. 2010) - cites phylota as a tool that would allow rapid phylogenetic placing of newly discovered species, and generation of phylogenetically informed guides for field identification.
- a study of wood density for carbon stock assessments (Flores and Coomes 2011), cites phylota as an initiative to "get supertrees resolved up to species level".
- a study proposing something similar to Open tree but applied only to land plants (Beaulieu et al. 2012b)
- an analysis of the phylogenetic diversity-area curve (Helmus and Ives 2012), cited phylota as a method alternative to phylomatic to "obtain plant phylogenetic trees for ecophylogenetic studies".
- a study generating a phylogeny of 6,098 species of vascular plants from China (Chen et al. 2016) - uses DarwinTree (Meng et al. 2015a) and generates sequence data *de novo* for 781 genera.
- a review of the state of methods and knowledge generated by molecular systematics (San Mauro and Agorreta 2010) cites phylota as a tool "intended to systematize GenBank information for large-scale molecular phylogenetics analysis".

- the first phylotastic paper (Stoltzfus et al. 2013) cites phylota as a "phylogeny related resource that provides ways to generate custom species trees for downstream use".
- Antonelli et al. (2017) cites phylota as a "pipeline that pre-processes entire GenBank releases in pursuit of sufficiently overlapping reciprocal BLAST hits, which are then clustered into candidate data sets". They also use the PHYLOTA database in its own pipeline.
- Deepak et al. (2014) present an algorithm for mining of frequent subtrees (common patterns) in collections of phylogenetic trees, as a way to extract meaningful phylogenetic information from collections of trees when compared to maximum agreement subtrees and majority-rule trees. They cite phylota as one of such tree collections available along with TreeBASE (Piel et al. 2009).
- Ranwez et al. (2009) cites phylota as a "program providing basic statistics on data availability for molecular datasets". They propose a tool to upload and explore user phylogenies to obtain detailed summary statistics on user tree collections.
- Freyman (2015) cites phylota as a tool that "provides a web interface to view all GenBank sequences within taxonomic groups clustered into homologs" but that does not mine for targeted sequences, as opposed to NCBIminer or PHLAWD. They compare the performance of SUMAC to Phylota. This is also presented in their PhD dissertation (Freyman 2017).
- Chesters and Vogler (2013) cites phylota as a data mining tool that compiles metadata from mining of public DNA databases "for construction of large phylogenetic trees and multiple gene sets" and that the authors have recognised that gene annotations in public databases are insufficient and that careful partitioning of orthologous sequences is needed for supermatrix construction. Chesters and Vogler (2013) present a procedure that minimizes the problem of forming multilocus species units in a large phylogenetic data set using algorithms from graph theory.
- Chesters and Zhu (2014) present an algorithm to delineate species form GenBank DNA data, and cites phylota as a tool that partitions "the contents of a database according to homology", by "grouping of database sequences according to internal criteria", searching "from a standardized set of references [...] patterns in sequence similarity and overlap."
- the paper presenting phylotaR, a pipeline that recreates the phylota output but uses the most

updated GenBank release, and is available in R (Bennett et al. 2018), cites phylota as its predecessor and inspiration. The authors mention that phylotaR pipeline mimics phylota's pipeline but with improvements.

- The paper presenging PhyloBase (Jamil 2016), cites phylota as one of its resources to get phylogenies, along with TreeBASE and others.
- The paper presenting STBase, a database of one million precomputed species trees (Deepak 2013; McMahon et al. 2015), cites phylota as a databse of gene trees or mul-trees, "trees having multiple sequences with the same taxon name".
- Drori et al. (2018) present a Web-based, user-friendly, online tool for species-tree reconstruction, based on the *supermatrix paradigm* and retrieves all available sequence data from NCBI GenBank. They cite phylota in the intro as a tool that is "designed to provide users with precomputed sets of clusters that were assembled through a single-linkage clustering approach and additionally provides precomputed gene trees that were reconstructed for each cluster. In particular, the results obtained by PhyLoTa are taxonomically constrained; that is, all sequences of the most recent common ancestor are collected even if one specifies only part of a clade".
- A study developing a tool to link wikipedia data to NCBI taxonomy (Page 2011) cites phylota as a phylogenetic resource that uses the NCBI taxonomy.
- the study that present DarwinTree (Meng et al. 2015a), and all derived studies: the study presenting an approach to screen sequence data for The Platform for Phylogenetic Analysis of Land Plants (PALPP), using the MapReduce paradigm to parallelize BLAST (Yong et al. 2010), as well as Gao et al. (2011), Li et al. (2013), Meng et al. (2014), Meng et al. (2015c), and Meng et al. (2015b), all cite phylota using the exact same introduction and sentence: as one among other "studies based on data mining large numbers of taxa or loci".
- A study presenting a tool to asses gene sequence quality for automatic construction of databases (Meng et al. 2012a), as well as their parallelized version using MapReduce (Meng et al. 2012b), cite phylota (along with Yong et al. (2010)) as a tool that relies on sequence similarity (BLAST) and not taxon name annotations in the database, for mining large numbers of taxa or loci, without

making any control on the quality of the sequencing.

- A review on online plant databases aiming to "provide recommendations for current information managers and developers concerning the user interface and experience; and to provide a picture about the possible directions to take for those in charge of the creation of information at all levels". They cite phylota as a tool allowing researchers "to acces equally and globally, without travel, a [phylogenetic?] model of plants at the kingdom level" (Jones et al. 2014).
- a paper aiming to establish an online information system for the legumes and to outline "best practices for development of a legume portal to enable data sharing and a better understanding of what data are available, missing, or erroneous, and ultimately facilitate cross-analyses and collaboration within the legume-systematics community and with other stakeholders" (Bruneau et al. 2019), cites phylota (along with supersmart and pyphlawd) as a "pipeline for large-scale retrieval of GenBank data of particular taxa or clades". In their Table 1, they also list phylota as a potential data source for developing a legume portal.
- A study on morphological evolution of electric fish skull, that uses phylotaR to retrieve sequences of the family Apteronotidae, order Gymnotiformes (Evans et al. 2019), cites phylota as the inspiration and fundament of phylotaR.
- A phylogenetic revision of the Gymnotidae fish (Teleostei: Gymnotiformes), uses phylotaR to retrieve seqeunces, but cites phylota as "a pipeline that implements BLAST searches to both identify and download sequence clusters for listed taxonomic groups to assemble a robust collection of sequences in a reproducible way based on publicly-available gene sequences while avoiding selection bias on the part of the assembler".
- A master thesis on SearchTree, a "software tool that allows users to query efficiently on an arbitrary user taxon list and returns high scoring matches from approximately one billion phylogenetic trees being constructed from molecular sequence data in GenBank" (Deepak 2010), that seems to be the preliminary work for STBase (McMahon et al. 2015), cites phylota as "a standard strategy, to assemble sets of homologous sequences (clusters) from a database of all-against-all BLAST searches, [in which] clusters are constructed in the context of the NCBI taxonomy tree for

convenience of display, thus child clusters are contained within parent clusters, following the NCBI hierarchy". In opposition, SearchTree uses true agglomerative hierarchical clustering (AHC: Day and Edelsbrunner (1984)) based on the BLAST estimates of sequence dissimilarity rather than the NCBI tree".

- a recent review on the state of large phylogeny (namely insects) generation using tools of the data-driven era (Chesters 2019) cites phylota as a tool for homology inference and retrieval.
- the study presenting phylotol (Cerón-Romero et al. 2019), cites phylota as a tool that "focus on the identification and collection of homologous genes from public databases".
- The iPTOL project cites phylota as a resource of phylogenetic trees.
- Mahmood (2015) PhD dissertation presents a database of avian Raptor sequences (raptorbase), based on the phylota pipeline.
- Ruiz-Sanchez et al. (2019) develops datataxa and cite phylota as "software that has been developed to mine the massive amount of information stored in GenBank", along with its R version (phylotaR; Bennett et al. 2018) and restez https://www.rdocumen-tation.org/packages/restez/versions/1.0.0.
- The phylotastic project (Stoltzfus et al. 2013) cites phylota as a "phylogeny-related resource providing ways to generate custom species trees *de novo* for downstream use" along with CIPRES.

2. When the software was actually used to construct (partially or in full) a DNA data set to be used for phylogenetic reconstruction:

- A 1000 tip phylogeny of the family of the nightshades (Särkinen et al. 2013)
- A 56 tip phylogeny of crustacean zooplancton (Helmus et al. 2010) – ecological study
- A 63 tip phylogeny of the Salmonidae family (Crête-Lafrenière et al. 2012)
- A 321 tip phylogeny of Testudines (Thomson and Shaffer 2010)
- A 69 taxa phylogeny of the family Cyprinodontidae of the pupfish (Martin and Wainwright 2011)
- A 2,957 taxa phylogeny of the class Moniloformopses of living ferns (Lehtonen 2011)
- A 2,573 species phylogeny of the Papilionoidea (Hardy and Otto 2014)
- A 23 taxa phylogeny of the California flora (Anacker et al. 2011)
- Phylogenies of 6 different clades of flowering plants representing an independent evolutionary

origin of extrafloral nectaries: *Byttneria* (Malvaceae), *Pleopeltis* (Polypodiaceae), *Polygoneae* (Polygoneaceae), *Senna* (Fabaceae), *Turnera* (Passifloraceae), and *Viburnum* (Adoxaceae) (Weber and Agrawal 2014).

- To supplement DNA data sets of various pre-existing mammalian phylogenetic trees sampled at different taxonomic levels (Faurby and Svenning 2015)
- A 900 species tree of muroid rodents, Muroidea (Steppan and Schenk 2017), where 300 species were newly added by the study and the rest obtained using phylota.
- A 95 taxa phylogeny of Gymnosperms, focused on Ephedra, Gnetales (Ickert-Bond et al. 2009)
- A 1061 genera phylogeny of the Oscine birds (Selvatti et al. 2015)
- A 268 species phylogeny of sharks, representing all 8 orders and 32 families (Sorenson et al. 2014; Sorenson 2014)
- A 466 species phylogeny of the Proteaceae, focusing on the species found in the Cape Floristic Region (Tucker et al. 2012).
- A series of small phylogenies of unreported exact size, of sister groups of gall-forming insects (Hardy and Cook 2010).
- A 196 species phylogeny of the family Boraginaceae (Nazaire and Hufford 2012). The authors actually found data for 318 Boraginaceae spp using phylota, but decided to reduce their data set to focus on the monophyly of genus *Mertensia*.
- A phylogeny of 401 species of scale insects Coccoidea, Hemiptera (Ross et al. 2013), with some sequences generated *de novo*.
- Two phylogenies sampling all species of two different clades of insectivorous lizards, agamids and diplodactyline geckos, groups considered to be radiating in the Australia's Great Victoria Desert (Rabosky et al. 2011)
- A phylogeny of 91 species of sparid and centracanthid fishes, Sparidae, Percomorpha, plus 2 outgroups, a lethrinid and a nemipterid exemplar (Santini et al. 2014).
- Updating a phylogeny of Arecaceae, constructing relationships in 6 cldes within the group: subfamilies Calamoideae and Coryphoideae, the tribe Ceroxyleae within subfamily Ceroxyloideae

and three groups within subfamily Arecoideae: (1) Iriarteeae,
(2) Cocoseae: Attaleinae except Beccariophoenix and (3) a group containing six tribes; Euterpeae, Leopoldinieae, Pelagodoxeae, Manicarieae, Geonomateae and Areceae (Faurby et al. 2016).

- A phylogeny of 768 Gesneriaceae species and 58 outgroups for a total species sampling of 826 taxa (Roalson and Roberts 2016) some sequence were generated *de novo*.
- A phylogeny of 47 species of scombrid fishes, with 2 outgroups, a gempylid and a trichiurid (Santini and Sorenson 2013).
- to update a dataset underlying a large-scale fern phylogeny (Lehtonen et al. 2017), data set in https://zenodo.org/record/345670#.Xr9QFRPYqqg, also in TreeBASE, but it is one of those studies that is broken.
- A phylogeny of 13 species of billfishes, order Istiophoriformes: Acanthomorpha, and four outgroups (Santini and Sorenson 2013)
- A phylogeny of 765 aphid species, family Aphididae (Hardy et al. 2015)
- A phylogeny of less than 100 taxa of the family Ranunculaceae (Lehtonen et al. 2016), even though they retrieved info from phylota for 194 taxa within the family, they reduced their data set because of low sampling of markers for some taxa.
- A phylogeny of 144 neobatrachian genera, assuming the monophyletic status of genera to increase matrix-filling levels (Frazao et al. 2015).
- A 179 species phylogeny of the bird family Picidae (woodpeckers, piculets, and wrynecks) (Dufort 2015, 2016), augmented with data from an updated GenBank release and newly sequenced data.
- A phylogeny of species of freshwater fish endemic to NorthAmerica (Strecker and Olden 2014), phylota found data for 54 out of 66 spp.
- A phylogeny of 520 species of the order Ericales (Hardy and Cook 2012)
- A phylgeny of 16 fish species of the family Sphyraenidae (Percomorpha), as well as two outgroup species of the Centropomidae (barracudas) (Santini et al. 2015)
- A phylogeny of 34 vole species, Arvicolinae, Rodentia (García-Navas et al. 2016)
- Kolmann et al. (2017) uses phylota to download all 1691 co1 sequences belonging to the order

Carchariniformes, to place phylogenetically DNA samples obtained from fish markets.

- A phylogeny of 329 bird species in the Tyrannidae (77% of the species in the family) (Gómez Bahamón and others 2015; Gómez-Bahamón et al. 2020)
- Retrive 145 sequences registered as *Holothuria* species, but kept 84 as ingroup, plus 4 outgroup sequences from *Stichopus ocellatus*, all belonging to the order Apodida of sea cucumbers (Kamarudin et al. 2016)
- On a master thesis, to get the sequences of the outgroups of Melinidinae, family Poaceae, namely several spp of the subfamily Panicoideae, plus *Gynerium sagittatum*, *Chasmanthium latifolium*, and *Zea mays*, (Salariato 2010). Interestingly, phylota was not used in the published study of the thesis (Salariato et al. 2010). Ingroup sequences were generated *de novo*.
- On a PhD thesis, to construct a phylogeny of Platyrrhini (internal group), Catarrhini (outgroup), and Tarsiiformes Pereira (2013). Have not found a published study.
- The 10k trees project (Arnold et al. 2010) uses phylota to construct a tree of 301 primate species and the outgroup species *Galeopterus variegates*, a tree of 17 extant odd-toed ungulates species and the outgroup species *Bos taurus*, and a tree of 70 different species of carnivorans and *Equus caballus* as outgroup. However, the do not cite it on the paper, but only on their documentation http://www.academia.edu/download/49690788/10kTrees_Documentation.pdf.
- Freyman (2015, also in 2017), use phylota to construct a phylogeny (or maybe only mine genbank???) of the Onagraceae and Lythracea, and compare it to the tool they propose, SUMAC.
- Blackmon (2017) PhD study applies phylota to reconstruct a 822 mite species tree.
- A study of the effect of poliploidy on niche evolution (Baniaga et al. 2018), uses phylota to get a DNA data set for 132 unique taxa of vascular plants from 16 families and 25 genera,and a tree of 33 genera from 20 different families comprising 1706 taxa.

3. When the website was used to identify sequences and markers available in GenBank for a particular group. In this cases, the dataset mining was either performed with other tools, or not performed at all and just used for discussion:
   - A 812 tips phylogeny of the Order Chiroptera (Shi and Rabosky 2015) – dataset constructed with

PHLAWD

- A 1276 tips phylogeny of the Fabaceae (Group et al. 2013) – dataset constructed by hand (I think??)
- A review of dated phylogenies of fire-prone tropical savanna species from Brazil (Simon and Pennington 2012) – just for discussion of the lack of markers available for these species on GenBank
- A review of the phylogeetic sof the Apicomplexa, a parasitic phylum on unicellular protists (Morrison 2009).
- Three data sets from phylota (the suborder Pleurodira of side-necked turtles; the family Cactaceae of cacti; and the Amorpheae, a clade of legumes) were used to demonstrate and exemplify phylogenetic decisiveness (Sanderson et al. 2010)
- Mentioned in a PHD thesis (Gagnon and others 2016), but not on the final publication (Gagnon et al. 2016), phylota was used to state that there are very few sequences available for the Legumes (7,482 out of 19,500 spp) on GenBank's release 194 (Feb2013).

4. Sometimes, it was cited by mistake:
   - In this 630 tip phylogeny of the Caryophyllaceae study (Greenberg and Donoghue 2011) it might have been originally cited as an example of large phylogenies that reflect well supported relationships from previous smaller phylogenies. However, it was removed from the text but not from the final list of references. The DNA data set was constructed by hand most probably.
   - a study reconstructing the insect tree of life with 49,358 species, 13,865 genera, and 760 families within the order Insecta (Chesters 2017), uses its own algorithm (SOPHI) to mine public DNA databases (Chesters and Zhu 2014). It does not cite phylota as it should, but includes it in their references.
5. When phylota was used to extract full trees (not only DNA data sets or markers):
   - Page (2013) uses it to generate phylogenies for the bionames website, a "database linking taxonomic names to their original descriptions, to taxa, and to phylogenies" generated with phylota.
   - Deepak et al. (2013) uses a sample of phylota trees to test their method to remove conflict from MUL-trees (short for multi-labeled trees), that is, phylogenetic trees with two or more leaves

sharing a label, e.g., a species name, which can imply multiple conflicting phylogenetic relationships for the same set of taxa.

- A review by Sanderson et al. (2016), takes 134 595 gene trees from phylota GenBank rel. 176 and estimates its degree of resolutin, calculating that less than half of clades are supported with minilam statistical support (0.53 ± 0.32).

6 Acknowledgements

We acknowledge contributions from

The University of California, Merced cluster, MERCED (Multi-Environment Research Computer for Exploration and Discovery) supported by the National Science Foundation (Grant No. ACI-1429783).

7 Authors' Contributions

8 Data Avilability

9 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.

Anacker B.L., Whittall J.B., Goldberg E.E., Harrison S.P. 2011. Origins and consequences of serpentine endemism in the california flora. Evolution: International Journal of Organic Evolution. 65:365–376.

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.

Arnold C., Matthews L.J., Nunn C.L. 2010. The 10kTrees website: A new online resource for primate phylogeny. Evolutionary Anthropology: Issues, News, and Reviews. 19:114–118.

Baniaga A.E., Marx H.E., Arrigo N., Barker M.S. 2018. Polyploid plants have faster rates of multivariate climatic niche evolution than their diploid relatives. BioRxiv.:406314.

Beaulieu J.M., Donoghue M.J. 2013. Fruit evolution and diversification in campanulid angiosperms. Evolution. 67:3132–3144.

Beaulieu J.M., Jhwueng D.-C., Boettiger C., O'Meara B.C. 2012a. Modeling stabilizing selection: Expanding the ornstein–uhlenbeck model of adaptive evolution. Evolution: International Journal of Organic Evolution. 66:2369–2383.

Beaulieu J.M., Ree R.H., Cavender-Bares J., Weiblen G.D., Donoghue M.J. 2012b. Synthesizing phylogenetic knowledge for ecological research. Ecology. 93:S4–S13.

Bennett D.J., Hettling H., Silvestro D., Zizka A., Bacon C.D., Faurby S., Vos R.A., Antonelli A. 2018. PhylotaR: An automated pipeline for retrieving orthologous dna sequences from genbank in r. Life. 8:20.

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.

Blackmon H. 2017. Synthesis and phylogenetic comparative analyses of the causes and consequences of karyotype evolution in arthropods..

Bruneau A., Borges L.M., Allkin R., Egan A.N., De La Estrella M., Javadi F., Klitgaard B., Miller J.T., Murphy D.J., Sinou C., others. 2019. Towards a new online species-information system for legumes. Australian Systematic Botany. 32:495–518.

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

Cerón-Romero M.A., Maurer-Alcalá X.X., Grattepanche J.-D., Yan Y., Fonseca M.M., Katz L. 2019. PhyloToL: A taxon/gene-rich phylogenomic pipeline to explore genome evolution of diverse eukaryotes. Molecular biology and evolution. 36:1831–1842.

Chen D., Burleigh J.G., Bansal M.S., Fernández-Baca D. 2008. PhyloFinder: An intelligent search engine for phylogenetic tree databases. BMC Evolutionary Biology. 8:90.

Chen Z.-D., Yang T., Lin L., Lu L.-M., Li H.-L., Sun M., Liu B., Chen M., Niu Y.-T., Ye J.-F., others. 2016. Tree of life for the genera of chinese vascular plants. Journal of Systematics and Evolution. 54:277–306.

Chesters D. 2017. Construction of a species-level tree of life for the insects and utility in taxonomic profiling. Systematic biology. 66:426–439.

Chesters D. 2019. The phylogeny of insects in the data-driven era. Systematic Entomology.

Chesters D., Vogler A.P. 2013. Resolving ambiguity of species limits and concatenation in multilocus sequence data for the construction of phylogenetic supermatrices. Systematic Biology. 62:456–466.

Chesters D., Zhu C.-D. 2014. A protocol for species delineation of public dna databases, applied to the insecta. Systematic biology. 63:712–725.

613 Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012. More than
614 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biology letters.
615 8:783–786.

616 Crête-Lafrenière A., Weir L.K., Bernatchez L. 2012. Framing the salmonidae family phylogenetic portrait: A
617 more complete picture from increased taxon sampling. PloS one. 7.

618 Day W.H., Edelsbrunner H. 1984. Efficient algorithms for agglomerative hierarchical clustering methods.
619 Journal of classification. 1:7–24.

620 Deepak A. 2010. SearchTree: Mining robust phylogenetic trees..

621 Deepak A. 2013. Managing and analyzing phylogenetic databases..

622 Deepak A., Fernández-Baca D., McMahon M.M. 2013. Extracting conflict-free information from multi-labeled
623 trees. Algorithms for Molecular Biology. 8:18.

624 Deepak A., Fernández-Baca D., Tirthapura S., Sanderson M.J., McMahon M.M. 2014. EvoMiner: Frequent
625 subtree mining in phylogenetic databases. Knowledge and Information Systems. 41:559–590.

626 Drori M., Rice A., Einhorn M., Chay O., Glick L., Mayrose I. 2018. OneTwoTree: An online tool for phylogeny
627 reconstruction. Molecular ecology resources. 18:1492–1499.

628 Dufort M. 2015. Coexistence, ecomorphology, and diversification in the avian family picidae (woodpeckers
629 and allies)..

630 Dufort M.J. 2016. An augmented supermatrix phylogeny of the avian family picidae reveals uncertainty deep
631 in the family tree. Molecular phylogenetics and evolution. 94:313–326.

632 Edgar R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic
633 acids research. 32:1792–1797.

634 Evans K.M., Vidal-García M., Tagliacollo V.A., Taylor S.J., Fenolio D.B. 2019. Bony patchwork: Mosaic pat-

terns of evolution in the skull of electric fishes (apteronotidae: Gymnotiformes). Integrative and comparative biology. 59:420–431.

Fan H., Ives A.R., Surget-Groba Y., Cannon C.H. 2015. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC genomics. 16:522.

Fang Y., Liu C., Lin J., Li X., Alavian K.N., Yang Y., Niu Y. 2019. PhySpeTree: An automated pipeline for reconstructing phylogenetic species trees. BMC evolutionary biology. 19:1–8.

Faurby S., Eiserhardt W.L., Baker W.J., Svenning J.-C. 2016. An all-evidence species-level supertree for the palms (arecaceae). Molecular Phylogenetics and Evolution. 100:57–69.

Faurby S., Svenning J.-C. 2015. A species-level phylogeny of all extant and late quaternary extinct mammals using a novel heuristic-hierarchical bayesian approach. Molecular phylogenetics and evolution. 84:14–26.

Felsenstein J. 1985. Confidence intervals on phylogenetics: An approach using bootstrap. Evolution. 39:783–791.

Flores O., Coomes D.A. 2011. Estimating the wood density of species for carbon stock assessments. Methods in Ecology and Evolution. 2:214–220.

Frazao A., Silva H.R. da, Moraes Russo C.A. de. 2015. The gondwana breakup and the history of the atlantic and indian oceans unveils two new clades for early neobatrachian diversification. PloS one. 10.

Freyman W.A. 2015. SUMAC: Constructing phylogenetic supermatrices and assessing partially decisive taxon coverage. Evolutionary Bioinformatics. 11:EBO–S35384.

Freyman W.A. 2017. Phylogenetic models linking speciation and extinction to chromosome and mating system evolution..

Gagnon E., Bruneau A., Hughes C.E., Queiroz L.P. de, Lewis G.P. 2016. A new generic system for the pantropical caesalpinia group (leguminosae). PhytoKeys.:1.

Gagnon E., others. 2016. Systématique et biogéographie du groupe caesalpinia (famille leguminosae)..

Gao Y., Meng Z., He X., Liu Y., Zhou Y., Li J. 2011. A solution to integrate data for phylogenetic research. 2011 5th international conference on bioinformatics and biomedical engineering.:1–4.

García-Navas V., Bonnet T., Bonal R., Postma E. 2016. The role of fecundity and sexual selection in the evolution of size and sexual size dimorphism in new world and old world voles (rodentia: Arvicolinae). Oikos. 125:1250–1260.

Gómez-Bahamón V., Márquez R., Jahn A.E., Miyaki C.Y., Tuero D.T., Laverde-R O., Restrepo S., Cadena C.D. 2020. Speciation associated with shifts in migratory behavior in an avian radiation. Current Biology.

Gómez Bahamón V., others. 2015. A behavioral polymorphism as an intermediate stage in the evolution of divergent forms-partial migration in new world flycatchers (aves, tyrannidae)..

Grant J.R., Katz L.A. 2014. Building a phylogenomic pipeline for the eukaryotic tree of life-addressing deep phylogenies with genome-scale data. PLoS currents. 6.

Greenberg A.K., Donoghue M.J. 2011. Molecular systematics and character evolution in caryophyllaceae. Taxon. 60:1637–1652.

Group L.P.W., Bruneau A., Doyle J.J., Herendeen P., Hughes C., Kenicer G., Lewis G., Mackinder B., Pennington R.T., Sanderson M.J., others. 2013. Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species–rich clades. Taxon. 62:217–248.

Guy L. 2017. PhyloSkeleton: Taxon selection, data retrieval and marker identification for phylogenomics. Bioinformatics. 33:1230–1232.

Hardy N.B., Cook L.G. 2010. Gall-induction in insects: Evolutionary dead-end or speciation driver? BMC evolutionary biology. 10:257.

Hardy N.B., Cook L.G. 2012. Testing for ecological limitation of diversification: A case study using parasitic

679 plants. The American Naturalist. 180:438–449.

680 Hardy N.B., Otto S.P. 2014. Specialization and generalization in the diversification of phytophagous insects:
681 Tests of the musical chairs and oscillation hypotheses. Proceedings of the Royal Society B: Biological Sciences.
682 281:20132960.

683 Hardy N.B., Peterson D.A., Dohlen C.D. von. 2015. The evolution of life cycle complexity in aphids:
684 Ecological optimization or historical constraint? Evolution. 69:1423–1432.

685 Helmus M.R., Ives A.R. 2012. Phylogenetic diversity–area curves. Ecology. 93:S31–S43.

686 Helmus M.R., Keller W., Paterson M.J., Yan N.D., Cannon C.H., Rusak J.A. 2010. Communities contain
687 closely related species during ecosystem disturbance. Ecology letters. 13:162–174.

688 Ickert-Bond S.M., Rydin C., Renner S.S. 2009. A fossil-calibrated relaxed clock for ephedra indicates an
689 oligocene age for the divergence of asian and new world clades and miocene dispersal into south america.
690 Journal of Systematics and Evolution. 47:444–456.

691 Izquierdo-Carrasco F., Cazes J., Smith S.A., Stamatakis A. 2014. PUmPER: Phylogenies updated perpetually.
692 Bioinformatics. 30:1476–1477.

693 Jamil H.M. 2016. A visual interface for querying heterogeneous phylogenetic databases. IEEE/ACM
694 transactions on computational biology and bioinformatics. 14:131–144.

695 Jones T.M., Baxter D.G., Hagedorn G., Legler B., Gilbert E., Thiele K., Vargas-Rodriguez Y., Urbatsch L.E.
696 2014. Trends in access of plant biodiversity data revealed by google analytics. Biodiversity data journal.

697 Kamarudin K.R., Rehan A.M., Hashim R., Usup G., Rehan M.M. 2016. Phylogenetic relationships within
698 the genus holothuria inferred from 16S mitochondiral rRNA gene sequences. Sains Malaysiana. 45:1079–1087.

699 Kolmann M.A., Elbassiouny A.A., Liverpool E.A., Lovejoy N.R. 2017. DNA barcoding reveals the diversity
700 of sharks in guyana coastal markets. Neotropical Ichthyology. 15.

Kumar S., Krabberød A.K., Neumann R.S., Michalickova K., Zhao S., Zhang X., Shalchian-Tabrizi K. 2015. BIR pipeline for preparation of phylogenomic data. Evolutionary Bioinformatics. 11:EBO–S10189.

Lehtonen S. 2011. Towards resolving the complete fern tree of life. PLoS One. 6.

Lehtonen S., Christenhusz M.J., Falck D. 2016. Sensitive phylogenetics of clematis and its position in ranunculaceae. Botanical Journal of the Linnean Society. 182:825–867.

Lehtonen S., Silvestro D., Karger D.N., Scotese C., Tuomisto H., Kessler M., Peña C., Wahlberg N., Antonelli A. 2017. Environmentally driven extinction and opportunistic origination explain fern diversification patterns. Scientific Reports. 7:1–12.

Lemoine F., Entfellner J.-B.D., Wilkinson E., Correia D., Felipe M.D., De Oliveira T., Gascuel O. 2018. Renewing felsenstein's phylogenetic bootstrap in the era of big data. Nature. 556:452–456.

Li J., Meng Z., Hou Y., Zhou Y., Gao Y. 2013. PartFastTree: Constructing large phylogenetic trees and estimating their reliability. 2013 ninth international conference on natural computation (icnc).:1052–1056.

Mahmood M.T. 2015. Avian raptor evolution..

Martin C.H., Wainwright P.C. 2011. Trophic novelty is linked to exceptional rates of morphological diversification in two adaptive radiations of cyprinodon pupfish. Evolution: International Journal of Organic Evolution. 65:2197–2212.

McMahon M.M., Deepak A., Fernández-Baca D., Boss D., Sanderson M.J. 2015. STBase: One million species trees for comparative biology. PloS one. 10.

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.

Meng Z., Dong H., Li J., Chen Z., Zhou Y., Wang X., Zhang S. 2015a. Darwintree: A molecular data analysis

723 and application environment for phylogenetic study. Data Science Journal. 14.

724 Meng Z., Li J., Chen Z. 2015b. A solution to phylogeny assembly for ecologists. 2015 12th international
725 conference on fuzzy systems and knowledge discovery (fskd).:1103–1107.

726 Meng Z., Li J., Yang T., Lin L., Chen Z. 2015c. SoTree: An automated phylogeny assembly tool for ecologists
727 from big tree. 2015 ieee international conference on smart city/socialcom/sustaincom (smartcity).:792–797.

728 Meng Z., Li J., Zhou Y., Cao W., Xiao X., Zhao J., Dong H., Zhang S. 2012a. GSQCT: A solution to
729 screening gene sequences for phylogenetics analysis. 2012 9th international conference on fuzzy systems and
730 knowledge discovery.:2929–2933.

731 Meng Z., Shao J., Cao W., Li J., Zhou Y., Wang X. 2014. RapidTree: A solution to rapid reconstruction
732 phylogenetic tree. 2014 11th international conference on fuzzy systems and knowledge discovery (fskd).:513–
733 517.

734 Meng Z., Xiao X., Li J., Zhou Y., Cao W., Shen G. 2012b. Cloud-gsqct: A parallel approach to screen gene
735 sequences for phylogenetics analysis. 2012 international conference on computer science and information
736 processing (csip).:660–663.

737 Morrison D.A. 2009. Evolution of the apicomplexa: Where are we now? Trends in parasitology. 25:375–382.

738 Nazaire M., Hufford L. 2012. A broad phylogenetic analysis of boraginaceae: Implications for the relationships
739 of mertensia. Systematic Botany. 37:758–783.

740 Page R.D. 2011. Linking ncbi to wikipedia: A wiki-based approach. PLoS currents. 3.

741 Page R.D. 2013. BioNames: Linking taxonomy, texts, and trees. PeerJ. 1:e190.

742 Papadopoulou A., Chesters D., Coronado I., De la Cadena G., Cardoso A., Reyes J.C., Maes J.-M., Rueda
743 R.M., Gómez-Zurita J. 2015. Automated dna-based plant identification for large-scale biodiversity assessment.
744 Molecular ecology resources. 15:136–152.

Pereira J.E.S. 2013. Padrões e processos na evolução de primatas neotropicais (platyrrhini, primates)..

Peters R.S., Meyer B., Krogmann L., Borner J., Meusemann K., Schütte K., Niehuis O., Misof B. 2011. The taming of an impossible child: A standardized all-in approach to the phylogeny of hymenoptera using public database sequences. BMC biology. 9:55.

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

Rabosky D.L., Cowan M.A., Talaba A.L., Lovette I.J. 2011. Species interactions mediate phylogenetic community structure in a hyperdiverse lizard assemblage from arid australia. The American Naturalist. 178:579–595.

Ranwez V., Clairon N., Delsuc F., Pourali S., Auberval N., Diser S., Berry V. 2009. PhyloExplorer: A web server to validate, explore and query phylogenetic trees. BMC evolutionary biology. 9:108.

Roalson E.H., Roberts W.R. 2016. Distinct processes drive diversification in different clades of gesneriaceae. Systematic Biology. 65:662–684.

Roquet C., Thuiller W., Lavergne S. 2013. Building megaphylogenies for macroecology: Taking up the challenge. Ecography. 36:13–26.

Ross L., Hardy N.B., Okusu A., Normark B.B. 2013. Large population size predicts the distribution of asexuality in scale insects. Evolution: International Journal of Organic Evolution. 67:196–206.

Ruiz-Sanchez E., Maya-Lastra C.A., Steinmann V.W., Zamudio S., Carranza E., Murillo R.M., Rzedowski J. 2019. Datataxa: A new script to extract metadata sequence information from genbank, the flora of bajío as a case study. Botanical Sciences. 97:754–760.

Salariato D.L. 2010. Filogenia y evolución de la subtribu melinidinae (paniceae: Panicoideae: Poaceae)..

Salariato D.L., Zuloaga F.O., Giussani L.M., Morrone O. 2010. Molecular phylogeny of the subtribe

767 melinidinae (poaceae: Panicoideae: Paniceae) and evolutionary trends in the homogenization of inflorescences.
768 Molecular Phylogenetics and Evolution. 56:355–369.

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

771 Sanderson M.J., McMahon M.M., Steel M. 2010. Phylogenomics with incomplete taxon coverage: The limits
772 to inference. BMC Evolutionary Biology. 10:155.

773 Sanderson M.J., Olson P., Hughes J., Cotton J. 2016. Perspective: Challenges in assembling the 'next
774 generation'Tree of life. Olson PD, Hughes J and Cotton JA.:13–27.

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

777 Santini F., Carnevale G., Sorenson L. 2014. First multi-locus timetree of seabreams and porgies (percomorpha:
778 Sparidae). Italian Journal of Zoology. 81:55–71.

779 Santini F., Carnevale G., Sorenson L. 2015. First timetree of sphyraenidae (percomorpha) reveals a middle
780 eocene crown age and an oligo–miocene radiation of barracudas. Italian Journal of Zoology. 82:133–142.

781 Santini F., Sorenson L. 2013. First molecular timetree of billfishes (istiophoriformes: Acanthomorpha) shows
782 a late miocene radiation of marlins and allies. Italian journal of zoology. 80:481–489.

783 Särkinen T., Bohs L., Olmstead R.G., Knapp S. 2013. A phylogenetic framework for evolutionary study of
784 the nightshades (solanaceae): A dated 1000-tip tree. BMC evolutionary biology. 13:214.

785 Schoch C.L., Sung G.-H., López-Giráldez F., Townsend J.P., Miadlikowska J., Hofstetter V., Robbertse B.,
786 Matheny P.B., Kauff F., Wang Z., others. 2009. The ascomycota tree of life: A phylum-wide phylogeny
787 clarifies the origin and evolution of fundamental reproductive and ecological traits. Systematic biology.
788 58:224–239.

789 Selvatti A.P., Gonzaga L.P., Moraes Russo C.A. de. 2015. A paleogene origin for crown passerines and the
790 diversification of the oscines in the new world. Molecular phylogenetics and evolution. 88:1–15.

791 Shi J.J., Rabosky D.L. 2015. Speciation dynamics during the global radiation of extant bats. Evolution.
792 69:1528–1545.

793 Simon M.F., Pennington T. 2012. Evidence for adaptation to fire regimes in the tropical savannas of the
794 brazilian cerrado. International Journal of Plant Sciences. 173:711–723.

795 Smith S.A., Beaulieu J.M., Donoghue M.J. 2009. Mega-phylogeny approach for comparative biology: An
796 alternative to supertree and supermatrix approaches. BMC evolutionary biology. 9:37.

797 Smith S.A., Brown J.W. 2018. Constructing a broadly inclusive seed plant phylogeny. American Journal of
798 Botany. 105:302–314.

799 Smith S.A., Walker J.F. 2019. PyPHLAWD: A python tool for phylogenetic dataset construction. Methods
800 in Ecology and Evolution. 10:104–108.

801 Sorenson L. 2014. Evolution of marine fish biodiversity: Phylogenomics and ecological processes shaping
802 diversification..

803 Sorenson L., Santini F., Alfaro M. 2014. The effect of habitat on modern shark diversification. Journal of
804 Evolutionary Biology. 27:1536–1548.

805 Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies.
806 Bioinformatics. 30:1312–1313.

807 Steppan S.J., Schenk J.J. 2017. Muroid rodent phylogenetics: 900-species tree reveals increasing diversification
808 rates. PLoS One. 12.

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

bioinformatics. 14:158.

Strecker A.L., Olden J.D. 2014. Fish species introductions provide novel insights into the patterns and drivers of phylogenetic structure in freshwaters. Proceedings of the Royal Society B: Biological Sciences. 281:20133003.

Thomson R.C., Shaffer H.B. 2010. Sparse supermatrices for phylogenetic inference: Taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Systematic biology. 59:42–58.

Tucker C.M., Cadotte M.W., Davies T.J., Rebelo T.G. 2012. Incorporating geographical and evolutionary rarity into conservation prioritization. Conservation Biology. 26:593–601.

Verbruggen H., Maggs C.A., Saunders G.W., Le Gall L., Yoon H.S., De Clerck O. 2010. Data mining approach identifies research priorities and data requirements for resolving the red algal tree of life. BMC evolutionary biology. 10:16.

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.

Weber M.G., Agrawal A.A. 2014. Defense mutualisms enhance plant diversification. Proceedings of the National Academy of Sciences. 111:16442–16447.

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.

Wu M., Eisen J.A. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome biology. 9:R151.

834 Xu X., Dimitrov D., Rahbek C., Wang Z. 2015. NCBIminer: Sequences harvest from genbank. Ecography.
835 38:426–430.

836 Yong L., Zhen M., Qi L., Yanping G., Yuanchun Z., Jianhui L. 2010. Screening data for phylogenetic
837 analysis of land plants: A parallel approach. 2010 first international conference on networking and distributed
838 computing.:305–308.