The Journal of Open Source Software

Physcraper: A Python package for continually updated gene trees

Luna L. Sanchez Reyes[1], **Martha Kandziora**[1], **and Emily Jane McTavish**[1]

1 University of California, Merced

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

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 phylogenies.
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 comparison of the the original and the updated trees using the conflict API of OToL.
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.

Introduction

Phylogenies are important. Generating phylogenies is not easy. The process of phylogenetic reconstruction implies many steps that can be generalized 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 repository, such as GenBank. 1. Assemble a hypothesis of homology – Create a matrix of your character data, by aligning the sequences, in the case of molecular data. 1. 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. 1. 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, 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. They are crucial to solve problems such as food security, global warming, global health. 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.

A way to automatize the comparison of phylogenetic hypotheses and to allow reproducibility of the last step of the 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.

It is well known that GenBank holds enormous amounts of genetic data, and it continues to grow. A lot of this genetic data has the potential to be used to reconstruct the phylogenetic history of various organisms (Sanderson, Boss, Chen, Cranston, & Wehe, 2008). Pipelines that harness this potential have been available for over a decade now, such as the Phylota browser, and PHLAWD. New ones keep on being developed, such as SUPERSMART and the upgraded version of PHLAWD, PyPHLAWD. Notably large phylogenies have been constructed using some of these tools, Some other have not been used that much. So, how well accepted is this approach in the community?

Concerns with 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?

The goal of Physcraper is to build upon previous phylogenetic knowledge, allowing a direct comparison of existing phylogenies to phylogenies that are constructed using new genetic data from GenBank

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.

How does Physcraper work?

The input: a study tree and an alignment

- The phylogenetic tree has to be in the Open Tree of Life store (McTavish et al., 2015). You can choose from a variety of published trees supporting any node of the Tree of Life. If the tree you are interested in is not in Open Tree of Life, you can easily upload it via the curator tool.
- The alignment should be a gene alignment that was used to generate the tree. The alignments are usually stored in a public repository such as TreeBase (Piel et al., 2009; Vos et al., 2012), DRYAD (http://datadryad.org/), or th ejournal 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 TreeBASE GitHub repository, SuperTreeBASE (https://github.com/TreeBASE/supertreebase). If the alignment is on another repository, 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 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 the tree and alignment. Technically, just one matching name is needed to perform the searches. See below.
- A ".tre" and ".aln" files are generated and saved for a `physcraper` run.

DNA sequence search and cleaning

- The next step is to identify the search taxon. This must be a taxon (a named clade) from the NCBI taxonomy. It will be used to constraint the DNA sequence search on the GenBank database within that taxonomic group. By default, the search taxon is the most recent common ancestor (MRCA) of the matched taxa that is also a named clade in the NCBI taxonomy. This is refered to as the most recent common ancestral taxon (MRCAT) or the least inclusive common ancestral taxon (LICA). It can be different from the phylogenetic MRCA when the latter is an unnamed clade. This is done using the Open Tree API taxonomy/mrca. A search taxon can also be given by the user. It can be a more inclusive clade, if the user wants to perform a wider search, outside the MRCAT of the matched taxa, e.g., including all taxa within the family or the order. It can also be a less inclusive clade, if the user only wants to focus on enriching a particular clade/region within the tree. When only one taxon is matched in both the tree and alignment, an MRCAT can be found for that single taxon, and thus a DNA sequence search can be performed even with only one sequence in the alignment.

- The BLAST algorithm is used to identify similarity among DNA sequences in the GenBank nucleotide database within the search taxon and the remaining sequences on the alignment.
- The DNA sequence similarity search can be done on a local database that is easily setup by the user. In this case it uses the BLASTn algorithm.
- The search can also be performed remotely, using the bioPython BLASt algorithm.
- A BLAST run is performed for each sequence in the alignment. Results of each BLAST are recorded. All matched sequences are saved with their corresponding GenBank accesion numbers that will be used to download the whole sequences into a local library.
- Matched sequences below an e-value, percentage similarity, and outside a minimum and maximum length threshold are discarded. This will leave out genomic sequences. All acepted sequences are asigned an internal identifier, and are further filtered.
- Because we usually do not have the accession number from sequences in the original alignment, a filtering process is needed. Accepted sequences that belong to the same taxon in the query sequence and that are either identical or shorter than the original sequence are also discarded. Only longer sequences belonging to the same taxon as the orignal sequence will be considered for further analyses.
- 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 kept for further analyses, by default 5 sequences per taxon are chosen at random. This number can be controlled by the user.
- Reverse complement sequences are identified and translated.
- This cycle of sequence search is performed two times. ***Is there an arhument to control the number of cycles of blast searches with new sequences***
- A fasta file containing all sequences resulting from the BLAST searches is generated for the user.

DNA sequence alignment

- The software MUSCLE (Edgar, 2004) is implemented for profile alignment, in which the original alignment is used as a template to align all new sequences.

Tree reconstruction

- A gene tree is reconstructed for each alignment provided, using RAxML with bootstrap replicates.
- The final result is a gene tree coupled to the conlict info.

Tree comparison

- 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***

Use case/ example

Imagine you are starting to work on a new biological group X. You have not much of an idea about its phylogenetic relationships, you are a newly established researcher, and the group is not anything any of your collaborators have worked on before. A good idea is to start an intensive literature review on the phylogenetics of the group. Rapidly, you find

out there are 5 different phylogenies, that used different markers, and that the papers, published at different times, do not discuss which phylogeny is the one accepted by the expert community on X. You might need to go to the annual conference of X, and even then, you might only find different and contrasting opinions. Somewhere along these months or even years doing this task, you looked into the the OToL database. You found in there some or all the published trees of X, along with a tree that has been deemed the best tree by curators and ideally experts on X?

Ascomycota Example

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       2 tree109, tree110                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" "tree110"
rotl::candidate_for_synth(rotl::get_study_meta("pg_238"))
## NULL
my_trees <- rotl::get_study("pg_238")
```

Both trees from this study have 434 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 "nex
```

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
```

Tools on a similar track

Tools that do similar things at different levels:

Phylota (Sanderson et al., 2008) - cited by 122 studies.

PHLAWD (Smith, Beaulieu, & Donoghue, 2009) and pyPhlawd (Smith & Walker, 2019) - baited analyses

DarwinTree (Meng et al., 2015) predecessor is Phylogenetic Analysis of Land Plants Platform (PALPP) - takes data from GenBank, EMBL and DDBJ for land plants only.

NCBIminer (Xu, Dimitrov, Rahbek, & Wang, 2015)

A ruby pipeline, only available from the supplementary data of the journal (Peters et al., 2011)

PUmPER (Izquierdo-Carrasco, Cazes, Smith, & Stamatakis, 2014) - perpetual updating with newly added sequences to GenBank

SUMAC (Freyman, 2015) - 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

SUPERSMART (Antonelli et al., 2017) - baited analyses up to bayesian divergente time estimation

SOPHI - (Chesters, 2017) - Searches DNA sequence data from repos other than GenBank, such as transcriptomic and barcoding repos.

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

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, Thuiller, & Lavergne, 2013)
   - a study constructing a dated seed plant phylogeny using pyPHLAWD (Smith & Brown, 2018)
   - a study presenting an assembly and alignment free method for phylogenetic reconstruction using genomic data, that aims to be incorporated in a tool as phylota some day (Fan, Ives, Surget-Groba, & Cannon, 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 & Donoghue, 2013)
- a study of Southeast Asia plant biodiversity inventory (Webb, Slik, & Triono, 2010) - cites phylota as a tool that would allow rapid phylogentic placing of newly discovered species, and generation of phylogenetically informed guides for field identification.
- a study of wood density for carbon stock assessments (Flores & 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 & 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., 2015) and generates sequence data *de novo* for 781 genera.
- a review of the state of methods and knowledge generated by molecular systematics (San Mauro & 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". I also uses the PHYLOTA database in its own pipeline.
- Deepak, Fernández-Baca, Tirthapura, Sanderson, & McMahon (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 ta xonomic 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

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, Bohs, Olmstead, & Knapp, 2013)
    - A 56 tip phylogeny of crustacean zooplancton (Helmus et al., 2010) – ecological study
    - A 63 tip phylogeny of the Salmonidae family (Crête-Lafrenière, Weir, & Bernatchez, 2012)
    - A 321 tip phylogeny of Testudines (Thomson & Shaffer, 2010)
    - A 69 taxa phylogeny of the family Cyprinodontidae of the pupfish (Martin & 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 & Otto, 2014)
- A 23 taxa phylogeny of the California flora (Anacker, Whittall, Goldberg, & Harrison, 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 & Agrawal, 2014).
- To supplement DNA data sets of various pre-existing mammalian phylogenetic trees sampled at different taxonomic levels (Faurby & Svenning, 2015)
- A 900 species tree of muroid rodents, Muroidea (Steppan & 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, Rydin, & Renner, 2009)
- A 1061 genera phylogeny of the Oscine birds (Selvatti, Gonzaga, & Moraes Russo, 2015)
- A 268 species phylogeny of sharks, representing all orders and 32 families (Sorenson, Santini, & Alfaro, 2014)
- A 466 species phylogeny of the Proteaceae, focusing on the species found in the Cape Floristic Region (Tucker, Cadotte, Davies, & Rebelo, 2012).
- A series of small phylogenies of unreported exact size, of sister groups of gall-forming insects (Hardy & Cook, 2010).
- A 196 species phylogeny of the family Boraginaceae (Nazaire & 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, Hardy, Okusu, & Normark, 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, Cowan, Talaba, & Lovette, 2011)
- A phylogeny of 91 species of sparid and centracanthid fishes, Sparidae, Percomorpha, plus 2 outgroups, a lethrinid and a nemipterid exemplar (Santini, Carnevale, & Sorenson, 2014).
- Updating a phylogeny of Arecaceae, constructing relationships in 6 clades 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, Eiserhardt, Baker, & Svenning, 2016).

- A phylogeny of 768 Gesneriaceae species and 58 outgroups for a total species sampling of 826 taxa (Roalson & 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 & 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 & Sorenson, 2013)
- A phylogeny of 765 aphid species, family Aphididae (Hardy, Peterson, & Dohlen, 2015)
- A phylogeny of less than 100 taxa of the family Ranunculaceae (Lehtonen,

Christenhusz, & Falck, 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, Silva, & Moraes Russo, 2015).
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 & 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 & 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, McMahon, & Steel, 2010)
4. Sometimes, it was cited by mistake:
    - In this 630 tip phylogeny of the Caryophyllaceae study (Greenberg & 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.
5. Other miscellaneous uses of phylota:
    - 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.

Acknowledgements

We acknowledge contributions from

References

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*(2), 365–376. doi:10.1111/j.1558-5646.2010.01114.x

Antonelli, A., Hettling, H., Condamine, F. L., Vos, K., Nilsson, R. H., Sanderson, M. J., Sauquet, H., et al. (2017). Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of taxa. *Systematic Biology*, *66*(2), 152–166. doi:10.1093/sysbio/syw066

Beaulieu, J. M., & Donoghue, M. J. (2013). Fruit evolution and diversification in campanulid angiosperms. *Evolution*, *67*(11), 3132–3144. doi:10.1111/evo.12180

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*(8), 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*(sp8), S4–S13. doi:10.1890/11-0638.1

Chen, Z.-D., Yang, T., Lin, L., Lu, L.-M., Li, H.-L., Sun, M., Liu, B., et al. (2016). Tree of life for the genera of chinese vascular plants. *Journal of Systematics and Evolution*, *54*(4), 277–306. doi:10.1111/jse.12219

Chesters, D. (2017). Construction of a species-level tree of life for the insects and utility in taxonomic profiling. *Systematic biology*, *66*(3), 426–439. doi:10.1093/sysbio/syw099

Crawford, N. G., Faircloth, B. C., McCormack, J. E., Brumfield, R. T., Winker, K., & Glenn, T. C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology letters*, *8*(5), 783–786. doi:10.1098/rsbl.2012.0331

Crête-Lafrenière, A., Weir, L. K., & Bernatchez, L. (2012). Framing the salmonidae family phylogenetic portrait: A more complete picture from increased taxon sampling. *PloS one*, *7*(10).

Deepak, A., Fernández-Baca, D., Tirthapura, S., Sanderson, M. J., & McMahon, M. M. (2014). EvoMiner: Frequent subtree mining in phylogenetic databases. *Knowledge and Information Systems*, *41*(3), 559–590. doi:10.1007/s10115-013-0676-0

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792–1797. doi:10.1093/nar/gkh340

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*(1), 522. doi:10.1186/s12864-015-1647-5

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), 1–8. doi:10.1186/s12862-019-1541-x

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. doi:10.1016/j.ympev.2016.03.002

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. doi:10.1016/j.ympev.2014.11.001

Flores, O., & Coomes, D. A. (2011). Estimating the wood density of species for carbon stock assessments. *Methods in Ecology and Evolution*, *2*(2), 214–220. doi:10.1111/j.2041-210X.2010.00068.x

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*(11). doi:10.1371/journal.pone.0143926

Freyman, W. A. (2015). SUMAC: Constructing phylogenetic supermatrices and assessing partially decisive taxon coverage. *Evolutionary Bioinformatics*, *11*, EBO–S35384. doi:10.4137/EBO.S35384

Greenberg, A. K., & Donoghue, M. J. (2011). Molecular systematics and character evolution in caryophyllaceae. *Taxon*, *60*(6), 1637–1652. doi:10.1002/tax.606009

Group, L. P. W., Bruneau, A., Doyle, J. J., Herendeen, P., Hughes, C., Kenicer, G., Lewis, G., et al. (2013). Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species–rich clades. *Taxon*, *62*(2), 217–248. doi:10.12705/622.8

Hardy, N. B., & Cook, L. G. (2010). Gall-induction in insects: Evolutionary dead-end or speciation driver? *BMC evolutionary biology*, *10*(1), 257. doi:10.1186/1471-2148-10-257

Hardy, N. B., & Otto, S. P. (2014). Specialization and generalization in the diversification of phytophagous insects: Tests of the musical chairs and oscillation hypotheses. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1795), 20132960. doi:10.1098/rspb.2013.2960

Hardy, N. B., Peterson, D. A., & Dohlen, C. D. von. (2015). The evolution of life cycle complexity in aphids: Ecological optimization or historical constraint? *Evolution*, *69*(6), 1423–1432. doi:10.1111/evo.12643

Helmus, M. R., & Ives, A. R. (2012). Phylogenetic diversity–area curves. *Ecology*, *93*(sp8), S31–S43. doi:10.1890/11-0435.1

Helmus, M. R., Keller, W., Paterson, M. J., Yan, N. D., Cannon, C. H., & Rusak, J. A. (2010). Communities contain closely related species during ecosystem disturbance. *Ecology letters*, *13*(2), 162–174. doi:10.1111/j.1461-0248.2009.01411.x

Ickert-Bond, S. M., Rydin, C., & Renner, S. S. (2009). A fossil-calibrated relaxed clock for ephedra indicates an oligocene age for the divergence of asian and new world clades and miocene dispersal into south america. *Journal of Systematics and Evolution*, *47*(5), 444–456. doi:10.1111/j.1759-6831.2009.00053.x

Izquierdo-Carrasco, F., Cazes, J., Smith, S. A., & Stamatakis, A. (2014). PUmPER: Phylogenies updated perpetually. *Bioinformatics*, *30*(10), 1476–1477. doi:10.1093/bioinformatics/btu05

Lehtonen, S. (2011). Towards resolving the complete fern tree of life. *PLoS One*, *6*(10). doi:10.1371/journal.pone.0024851

Lehtonen, S., Christenhusz, M. J., & Falck, D. (2016). Sensitive phylogenetics of clematis and its position in ranunculaceae. *Botanical Journal of the Linnean Society*, *182*(4), 825–867. doi:10.1111/boj.12477

Lehtonen, S., Silvestro, D., Karger, D. N., Scotese, C., Tuomisto, H., Kessler, M., Peña, C., et al. (2017). Environmentally driven extinction and opportunistic origination explain fern diversification patterns. *Scientific Reports*, *7*(1), 1–12. doi:10.1038/s41598-017-05263-7

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*(8), 2197–2212. doi:10.1111/j.1558-5646.2011.01294.x

McTavish, E. J., Hinchliff, C. E., Allman, J. F., Brown, J. W., Cranston, K. A., Holder, M. T., Rees, J. A., et al. (2015). Phylesystem: A git-based data store for community-curated phylogenetic estimates. *Bioinformatics*, *31*(17), 2794–2800. doi:10.1093/bioinformatics/btv276

Meng, Z., Dong, H., Li, J., Chen, Z., Zhou, Y., Wang, X., & Zhang, S. (2015). Darwintree: A molecular data analysis and application environment for phylogenetic study. *Data Science Journal*, *14*. doi:10.5334/dsj-2015-010

Morrison, D. A. (2009). Evolution of the apicomplexa: Where are we now? *Trends in parasitology*, *25*(8), 375–382. doi:10.1016/j.pt.2009.05.010

Nazaire, M., & Hufford, L. (2012). A broad phylogenetic analysis of boraginaceae: Implications for the relationships of mertensia. *Systematic Botany*, *37*(3), 758–783. doi:10.1600/036364412X648715

Page, R. D. (2013). BioNames: Linking taxonomy, texts, and trees. *PeerJ*, *1*, e190. doi:10.7717/peerj.190

Peters, R. S., Meyer, B., Krogmann, L., Borner, J., Meusemann, K., Schütte, K., Niehuis, O., et al. (2011). The taming of an impossible child: A standardized all-in approach to the phylogeny of hymenoptera using public database sequences. *BMC biology*, *9*(1), 55. doi:10.1186/1741-7007-9-55

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

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*(5), 579–595. doi:10.1086/662162

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*(1), 108. doi:10.1186/1471-2148-9-108

Roalson, E. H., & Roberts, W. R. (2016). Distinct processes drive diversification in different clades of gesneriaceae. *Systematic Biology*, *65*(4), 662–684. doi:10.1093/sysbio/syw012

Roquet, C., Thuiller, W., & Lavergne, S. (2013). Building megaphylogenies for macroecology: Taking up the challenge. *Ecography*, *36*(1), 13–26. doi:10.1111/j.1600-0587.2012.07773.x

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*(1), 196–206. doi:10.1111/j.1558-5646.2012.01784.x

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*(3), 335–346. doi:10.1080/10635150802158688

Sanderson, M. J., McMahon, M. M., & Steel, M. (2010). Phylogenomics with incomplete taxon coverage: The limits to inference. *BMC Evolutionary Biology*, *10*(1), 155. doi:10.1186/1471-2148-10-155

San Mauro, D., & Agorreta, A. (2010). Molecular systematics: A synthesis of the common methods and the state of knowledge. *Cellular & Molecular Biology Letters*, *15*(2), 311. doi:10.2478/s11658-010-0010-8

Santini, F., Carnevale, G., & Sorenson, L. (2014). First multi-locus timetree of seabreams and porgies (percomorpha: Sparidae). *Italian Journal of Zoology*, *81*(1), 55–71. doi:10.1080/11250003.2013.878960

Santini, F., & Sorenson, L. (2013). First molecular timetree of billfishes (istiophoriformes: Acanthomorpha) shows a late miocene radiation of marlins and allies. *Italian journal of zoology*, *80*(4), 481–489. doi:10.1080/11250003.2013.848945

Särkinen, T., Bohs, L., Olmstead, R. G., & Knapp, S. (2013). A phylogenetic framework for evolutionary study of the nightshades (solanaceae): A dated 1000-tip tree. *BMC evolutionary biology*, *13*(1), 214.

Schoch, C. L., Sung, G.-H., López-Giráldez, F., Townsend, J. P., Miadlikowska, J., Hofstetter, V., Robbertse, B., et al. (2009). The ascomycota tree of life: A phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic biology*, *58*(2), 224–239.

Selvatti, A. P., Gonzaga, L. P., & Moraes Russo, C. A. de. (2015). A paleogene origin for crown passerines and the diversification of the oscines in the new world. *Molecular phylogenetics and evolution*, *88*, 1–15. doi:10.1016/j.ympev.2015.03.018

Shi, J. J., & Rabosky, D. L. (2015). Speciation dynamics during the global radiation of extant bats. *Evolution*, *69*(6), 1528–1545. doi:10.1111/evo.12681

Simon, M. F., & Pennington, T. (2012). Evidence for adaptation to fire regimes in the tropical savannas of the brazilian cerrado. *International Journal of Plant Sciences*, *173*(6), 711–723. doi:10.1086/665973

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*(1), 37.

Smith, S. A., & Brown, J. W. (2018). Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany*, *105*(3), 302–314. doi:10.1002/ajb2.1019

Smith, S. A., & Walker, J. F. (2019). PyPHLAWD: A python tool for phylogenetic dataset construction. *Methods in Ecology and Evolution*, *10*(1), 104–108.

Sorenson, L., Santini, F., & Alfaro, M. (2014). The effect of habitat on modern shark diversification. *Journal of Evolutionary Biology*, *27*(8), 1536–1548. doi:10.1111/jeb.12405

Steppan, S. J., & Schenk, J. J. (2017). Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates. *PLoS One*, *12*(8). doi:10.1371/journal.pone.0183070

Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C. M., Vaidya, G., et al. (2013). Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC bioinformatics*, *14*(1), 158. doi:10.1186/1471-2105-14-158

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*(1), 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*(4), 593–601. doi:10.1111/j.1523-1739.2012.01845.x

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*(1), 16. doi:10.1186/1471-2148-10-16

Vos, R. A., Balhoff, J. P., Caravas, J. A., Holder, M. T., Lapp, H., Maddison, W. P., Midford, P. E., et al. (2012). NeXML: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic biology*, *61*(4), 675–689. doi:10.1093/sysbio/sys025

Webb, C. O., Slik, J. F., & Triono, T. (2010). Biodiversity inventory and informatics in southeast asia. *Biodiversity and Conservation*, *19*(4), 955–972. doi:10.1007/s10531-010-9817-x

Weber, M. G., & Agrawal, A. A. (2014). Defense mutualisms enhance plant diversification. *Proceedings of the National Academy of Sciences*, *111*(46), 16442–16447. doi:10.1073/pnas.1413253111

Xu, X., Dimitrov, D., Rahbek, C., & Wang, Z. (2015). NCBIminer: Sequences harvest from genbank. *Ecography*, *38*(4), 426–430. doi:10.1111/ecog.01055