

1 DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life

2 Luna L. Sánchez Reyes^{1,2}, Emily Jane McTavish¹, & Brian O'Meara²

3 ¹ University of California, Merced

4 ² University of Tennessee, Knoxville

School of Natural Sciences, University of California, Merced, Science and Engineering
Building 1.

Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville,
425 Hesler Biology Building, Knoxville, TN 37996, USA.

The authors made the following contributions. Luna L. Sánchez Reyes: Data curation,
Investigation, Software, Visualization, Validation, Writing - Original Draft Preparation,
Writing - Review & Editing; Emily Jane McTavish: Resources, Software, Writing - Review &
Editing; Brian O'Meara: Conceptualization, Funding acquisition, Methodology, Resources,
Software, Supervision, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Luna L. Sánchez Reyes, .
E-mail: sanchez.reyes.luna@gmail.com

Abstract

Date estimates for times of evolutionary divergences are key data for research in the natural sciences. These estimates also provide valuable information for for education, science communication and policy decisions. Although achieving a high-quality reconstruction of time of evolutionary origin as a phylogenetic tree with branch lengths proportional to absolute time (chronogram), is still a difficult and time-consuming task, the increased availability of fossil and molecular data, and time-efficient analytical techniques has resulted in many recent publications of large chronograms for a large number and wide diversity of organisms. When these estimates are shared in public, open databases this wealth of expertly-curated and peer-reviewed data on time of evolutionary origin is exposed in a programatic and reusable way. Intensive and localized efforts have improved data sharing practices, as well as incentivized open science in biology. Here we present DateLife, a service implemented as an R package and an Rshiny website application available at www.datelife.org/query/, that provides functionalities for efficient and easy finding, summary, reuse, and reanalysis of expert, peer-reviewed, public data on time of evolutionary origin. The main DateLife workflow constructs a chronogram for any given combination of taxon names, by searching a local chronogram database constructed and curated from the Open Tree of Life Phylsystem phylogenetic database, which incorporates phylogenetic data from TreeBASE database as well. We implement and test methods for summarizing time data from multiple source chronograms using supertree and congruification algorithms, and using age data extracted from source chronograms as secondary calibration points to add branch lengths proportional to absolute time to a tree topology. DateLife will be useful to increase awareness on the existing variation in expert time of divergence data, and can foster exploration of the effect of alternative divergence time hypothesis on the results of analyses, providing a framework for a more informed interpretation of evolutionary results.

Keywords: Tree; Phylogeny; Scaling; Dating; Ages; Divergence times; Open Science;

⁴³ Congruification; Supertree; Calibrations; Secondary calibrations

⁴⁴ Word count: 3542

DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life

Introduction

Chronograms –phylogenies with branch lengths proportional to time, provide key data for the study of natural processes in many areas of biological research, such as developmental biology (cite) to conservation biology (Felsenstein, 1985; Campbell O. Webb, 2000), historical biogeography, a species diversification (Morlon, 2014; Posadas, Crisci, & Katinas, 2006).

Building a chronogram is not an easy task. It requires obtaining and curating data to construct a phylogeny; selecting and placing appropriate calibrations on the phylogeny using independent age data points from the fossil record or other dated events, and inferring the full dated tree. Estimating accurate chronograms generally requires specialized biological training, taxonomic domain knowledge, and a non-negligible amount of research time, computational resources and funding.

Here we present the DateLife software application, available as an R package and as an online Rshiny interactive website at www.datelife.org/query/, which captures data from published chronograms, and make these data readily accessible to users. DateLife features a versioned, open and fully public chronogram database (McTavish et al., 2015) storing age information in a computer readable format (Vos et al., 2012), an automated and programmatic way of accessing the data (Stoltzfus et al., 2013) and methods to summarize and compare age data.

The DateLife algorithm

Providing an input.– DateLife starts with an input query consisting of at least two taxon names, which can be provided as a comma separated character string, or as tip labels on a tree. If the input is a tree, it can be provided as a classic newick character string (Archie et al., 1986), or as a “phylo” R object (Paradis, Claude, & Strimmer, 2004). The input tree is not required to have branch lengths, and its topology is used on the summary

steps described below.

Currently, DateLife only accepts scientific names as input, and they can belong to any inclusive taxonomic group (e.g., genus, family, tribe, etc.) or binomial specific. Subspecies and variants are ignored. If an input taxon name belongs to an “inclusive” taxonomic group the algorithm has two alternative behaviors defined by the “get species from taxon” flag. If the flag is active, the DateLife algorithm retrieves all species names within the “inclusive” taxonomic group and adds them to the input. If the flag is inactive, DateLife ignores the “inclusive” taxon names from the input.

Input scientific names are processed using the Taxonomic Name Resolution Service (TNRS; Boyle et al. (2013)) provided by OpenTree (OpenTreeOfLife et al., n.d.). TNRS increases the probability of correctly finding the query organisms in the chronogram database, as it detects, corrects and standardizes name misspellings and typos, variant spellings and authorities, and nomenclatural synonyms to a single taxonomic standard.

The processed input taxon names are saved as a special R object (of a newly defined class `datelifeQuery`) that contains the processed names, the corresponding taxonomic id numbers, and the topology of the input tree if any was provided. The `datelifeQuery` object is used next to search the chronogram database.

Searching the database.—Chronograms in the DateLife database with at least two matching processed taxon names are identified and pruned down to preserve only the matched names as tips. Then, each pruned chronogram is transformed to a patristic distance matrix. This format facilitates and greatly speeds up all downstream analyses and summaries. The matrices are associated to the citation of the original study and stored as an R object of class `datelifeResult`.

Summarizing search results.—At this point, summary data is collected to inform decisions for ulterior steps in the user workflow. Types of summary information provided are:

1. The pruned source chronograms are used to get pairwise node ages of your query taxa (Figure 1 subfigure X).
2. Age of the root of pruned source chronograms. This can correspond to the mrca age of your group if the pruned source chronogram has all your query taxa. If not, the root corresponds to the mrca of subgroup withing your query taxa.
3. Citations of studies where pruned source chronograms were originally published.
4. A summary table with all of the above.
5. A single chronogram summarizing all pruned source chronograms, or a subset defined by the user. To construct summary trees, **datelife** requires a fixed topology. Users can provide one from the literature, or one of their own making. If no topology is provided, **datelife** chooses one automatically retrieved from an expert phylogenetic information database such as the OpenTree synthetic tree.

????? More meat on the methods needed: e.g. how the distances are generated and summarized, how conflict is handled. . .

Then we summarize source chronograms into a single patristic distance matrix using a method chosen by the user. Summarizing method options available include Super Distance Matrix method (SDM, Criscuolo, Berry, Douzery, & Gascuel, 2006) and any summary statistics such as median, minimum and maximum ages. Finally, **datelife** applies the summarized time distances as node calibrations to date the given topology with BLADJ (Campbell O. Webb & Donoghue, 2005).

6. A report of successful matches of input taxon names across pruned source chronograms.
7. A single pruned source chronogram with the most matching input taxon names.

If there is no information available for any input taxon name, users can also create both age and phylogenetic data for the missing branches with a variety of algorithms, ??? like how?

Visualizing summaries.— Users can save all source and summary chronograms in formats that permit reuse and reanalyses (newick and R “phylo” format), as well as view and compare results graphically, or construct their own graphs using `datelife`’s graphic generation functions.

The DateLife algorithm illustrated in figure 1 is fully implemented using the R language. The latest stable version of the R package `datelife` is available from the CRAN repository (v0.6.2; Sanchez-Reyes et al. (2022)), and relies on functionalities from various biological R packages: `ape` (Paradis et al., 2004), `bold` (Chamberlain et al., 2019), `geiger` (Harmon, Weir, Brock, Glor, & Challenger, 2008), `paleotree` (Bapst, 2012), `phyloch` (Heibl, 2008), `phylocomr` (Ooms & Chamberlain, 2018), `phytools` (Revell, 2012), `rotl` (Michonneau, Brown, & Winter, 2016), and `taxize` (Chamberlain & Szöcs, 2013; Chamberlain et al., 2019).

Benchmark

`datelife`’s code speed was tested on an Apple iMac with one 3.4 GHz Intel Core i5 processor. We registered variation in computing time of query processing and search through the database relative to number of queried taxon names. Query processing time increases roughly linearly with number of input taxon names, and increases considerably if TNRS is activated. Up to ten thousand names can be processed and searched in less than 30 minutes with the most time consuming settings. Once names have been processed as described in methods, a name search through the chronogram database can be performed in less than a minute, even with a very large number of taxon names (Fig. 2). `datelife`’s code performance was evaluated with a set of unit tests designed and implemented with the R package `testthat` (R Core Team, 2018) that were run both locally with the `devtools` package (R Core Team, 2018), and on a public server –via GitHub, using the continuous integration tool Travis CI (<https://travis-ci.org>). At present, unit tests cover more than 30% of `datelife`’s code (<https://codecov.io/gh/phylostatic/datelife>).

Results

Case study

We illustrate the **datelife** algorithm using the family of true finches, Fringillidae as an example. A college educator wishes to obtain state-of-the-art data on time of evolutionary origin of species belonging to the true finches for their class. They decide to use **datelife** because they are teaching best practices for reproducibility. Students have the option to go to the website at www.datelife.org and perform an interactive run. However, the educator wants the students to practice their R skills. The first step is to run a **datelife** query using the “get species from taxon” flag. This will get all recognised species names within their chosen inclusive taxon. The Fringillidae has 289 species, according to the Open Tree of Life taxonomy. Once with a curated set of species taxon names, the next step is to run a **datelife** search that will find all chronograms that contain at least two species names. The algorithm proceeds to prune the trees to keep matching species names on tips only, and transform the pruned trees to pairwise distance matrices. There are 13 chronograms containing at least two Fringillidae species, published in 9 different studies (Fig. 3). The final step is to summarize the available information using two alternative types of summary chronograms, median and SDM. As explained in the “Description” section, data from source chronograms is first summarised into a single distance matrix and then the available node ages are used as fixed node calibrations over a consensus tree topology, to obtain a fully dated tree with the program BLADJ (Fig. 4). Median summary chronograms are older and have wider variation in maximum ages than chronograms obtained with SDM. With both methods, ages are generally consistent with source ages, but there are some biological examples in which this is not true (see Discussion).

Cross-validation test

Data from source chronograms can be also used to date tree topologies with no branch lengths, as well as trees with branch lengths as relative substitution rates (Figs. 5 and 6). As

a form of cross validation, we took tree topologies from each study and calibrated them using time of lineage divergence data from all other source chronograms. In the absence of branch lengths, the ages of internal nodes were recovered with a high precision in almost all cases (except for studies 3, and 5; Fig. 5). Maximum tree ages were only recovered in one case (study 2; Fig. 5). We also demonstrate the usage of PATHd8 (Britton, Anderson, Jacquet, Lundqvist, & Bremer, 2007) as an alternative method to BLADJ. For this, we run a **datelife** branch length reconstruction that searches for DNA sequence data from the Barcode of Life Data System [BOLD; Ratnasingham and Hebert (2007)] to generate branch lengths. We were able to successfully generate a tree with BOLD branch lengths for all of the Fringillidae source chronograms. However, dating with PATHd8 using congruified calibrations, was only successful in three cases (studies 3, 5, and 9, shown in Fig. 6). From these, two trees have a different sampling than the original source chronogram, mainly because DNA BOLD data for some species is absent from the database. Maximum ages are quite different from source chronograms, but this might be explained also by the differences in sampling between source chronograms and BOLD trees. More examples and code used to generate these trees were developed on an open repository that is available for consultation and reuse at https://github.com/LunaSare/datelife_examples.

Discussion

The main goal of **datelife** is to make state-of-the-art information on time of lineage divergence easily accessible for comparison, reuse, and reanalysis, to researchers in all areas of science and with all levels of expertise in the matter. It is an open service that does not require any expert biological knowledge from users –besides the names of the organisms they want to work with, for any of its functionalities.

At the time of writing of this manuscript (Mar 07, 2022), **datelife**'s database has 253 chronograms, pulled entirely from OpenTree's database, the Phylesystem (McTavish et al., 2015). A unique feature of OpenTree's Phylesystem is that the community can add new

state-of-the-art chronograms any time. As chronograms are added to Phylsystem, they are incorporated into an updated **datelife**'s database that is assigned a new version number, followed by a package release on CRAN. **datelife**'s chronogram database is updated as new chronogram data is added to Phylsystem, at a minimum of once a month and a maximum of every 6 months. Users can also upload new chronograms to OpenTree themselves, and trigger an update of the **datelife** database to incorporate the new chronograms, to have them immediately available for analysis.

Incorporation of more chronograms into **datelife**'s database is crucial to improve its services. One option to increase chronogram number in the database is the Dryad data repository. Methods to automatically mine chronograms from Dryad could be designed and implemented. However, Dryad's metadata system has no information to automatically detect branch length units, and those would still need to be determined manually by a curator.

The only summary chronogram encompassing all life that is currently openly available from OpenTree was constructed using age data from 2,274 published chronograms (Hedges, Marin, Suleski, Paymer, & Kumar, 2015). However source chronograms are not available in computer readable format for reuse or reanalysis. As this tree is part of **datelife**'s database, the amount of lineages that can be queried using **datelife** (99474 unique terminal taxa) is substantial, yet it can be improved. Consequently, we would like to emphasize on the importance of sharing chronogram data for the benefit of the scientific community as a whole, into repositories that require expert input and manual curation, such as OpenTree's Phylsystem (McTavish et al., 2015).

By default, **datelife** currently summarizes all source chronograms that overlap with at least two species names. Users can exclude source chronograms if they have reasons to do so. Strictly speaking, the best chronogram should reflect the real time of lineage divergence accurately and precisely. To our knowledge, there are no good measures to determine if a chronogram is better than another. Some measures that have been proposed are the

proportion of lineage sampling and the number of calibrations used Magallón, Gómez-Acevedo, Sánchez-Reyes, & Hernández-Hernández (2015). Scientists usually also favor chronograms constructed using primary calibrations (ages obtained from the fossil or geological record) to ones constructed with secondary calibrations (ages coming from other chronograms)(Schenk, 2016). It has been observed with simulations that divergence times inferred with secondary calibrations are significantly younger than those inferred with primary calibrations in analyses performed with bayesian inference methods when priors are implemented in similar ways in both analyses (Schenk, 2016). However, secondary calibrations can be applied using other dating methods that do not require setting priors, such as penalized likelihood (Sanderson, 2003), or as fixed ages, potentially mitigating the bias reported with bayesian methods. Certainly, further studies are required to fully understand the effect of using secondary calibrations on time estimates and downstream analyses.

Furthermore, even chronograms obtained with primary fossil data can vary substantially in time estimates between lineages, as observed from the comparison of source chronograms in the Fringillidae example. This observation is often encountered in the literature (see, for example, the ongoing debate about crown group age of angiosperms (Barba-Montoya, Reis, Schneider, Donoghue, & Yang, 2018; Magallón et al., 2015; Ramshaw et al., 1972; Sanderson & Doyle, 2001; Sauquet, Ramírez-Barahona, & Magallón, 2021). For some studies, especially ones based on branch lengths (e.g., studies of species diversification, timing of evolutionary events, phenotypic trait evolution), using a different chronogram may return different results (Title & Rabosky, 2016). Stitching together these chronograms can create a larger tree that uses information from multiple studies, but the effect of uncertainties and errors at this level on downstream analyses is still largely unknown.

Summarizing chronograms might also imply summarizing fundamentally distinct evolutionary hypotheses. For example, two different researchers working on the same clade

both carefully select and argument their choices of fossil calibrations. Still, if one researcher decides a fossil will calibrate the ingroup of a clade, while another researcher uses the same one to calibrate outside the clade, the resulting age estimates will often differ substantially, as the placement of calibrations as stem or crown group is proved to deeply affect estimated times of lineage divergence (Sauquet, 2013). Trying to summarize the resulting chronograms into a single one using simple summary statistics might erase all types of relevant information from the source chronograms. Accordingly, the prevailing view in our research community is that we should favor time of lineage divergence estimates obtained from a single analysis, using fossil data as primary sources of calibrations, and using fossils that have been widely discussed and curated as calibrations to date other trees, making sure that all data used in the analysis reflect a coherent evolutionary history (Antonelli et al., 2017). However, the exercise of summarizing different chronograms has the potential to help getting a single global evolutionary history for a lineage by putting together evidence from different hypothesis. Choosing the elements of the chronograms that we are going to keep and the ones that we are going to discard is key, since we are potentially losing important parts of the evolutionary history of a lineage that might only be reflected in source chronograms and not on the summary chronogram (Sauquet et al., 2021).

Alternatively, one could try to choose the “best” chronogram from a set of possible evolutionary hypotheses. Several characteristics of the data used for dating analyses as well as from the output chronogram itself, could be used to score quality of source chronograms. Some characteristics that are often cited in published studies as a measure of improved age estimates as compared to previously published estimates are: quality of alignment (missing data, GC content), lineage sampling (strategy and proportion), phylogenetic and dating inference method, number of fossils used as calibrations, support for nodes and ages, and magnitude of confidence intervals. To facilitate subsetting of source chronograms following different criteria by the users, this information should be included as metadata manually entered by curators in the future.

In other areas of biological research, such as ecology and conservation biology, it has been shown that at least some data on lineage divergence represents a relevant improvement for testing alternative hypothesis using phylogenetic distance (Campbell O. Webb, Ackerly, & Kembel, 2008). Hence, we integrated into datelife’s workflow different ways of creating branch lengths in the absence of starting branch length information for taxa lacking this information (using the BLADJ option). “Making up” branch lengths is an accepted practice in scientific publications: Jetz, Thomas, Joy, Hartmann, and Mooers (2012), created a time-calibrated tree of all 9,993 bird species, where 67% had molecular data and the rest was simulated; Rabosky et al. (2018) created a time-calibrated tree of 31,536 ray-finned fishes, of which only 37% had molecular data; Smith and Brown (2018) constructed a tree of 353,185 seed plants where only 23% had molecular data. Taken to the extreme, one could make a fully resolved, calibrated tree of all modern and extinct taxa using a single taxonomy and a single calibration with the polytomy resolution and branch assignation methods. There has yet to be a thorough analysis of what can go wrong when one goes beyond the data in this way, so we urge caution; we also urge readers to follow the example of many of the large tree papers cited above and make sure results are substantially similar between trees fully reconstructed with molecular or other data, and trees that are reconstructed using taxonomy by resolving polytomies at random following a statistical model.

Conclusions

Divergence time information is key to many areas of evolutionary studies: trait evolution, diversification, biogeography, macroecology and more. It is also crucial for science communication and education, but generating chronograms is difficult, especially for those who want to use phylogenies but who are not systematists, or do not have the time to acquire and develop the necessary knowledge and data curation skills. Moreover, years of primarily public funded research have resulted in vast amounts of chronograms that are already available on scientific publications, but hidden to the public and scientific community

for reuse.

The **datelife** R package allows easy and fast summarization of publicly available information on time of lineage divergence. This provides a straightforward way to get an informed idea on the state of knowledge of the time frame of evolution of different regions of the tree of life, and allows identification of regions that require more research or that have conflicting information. Both summary and newly generated trees are useful to evaluate evolutionary hypotheses in different areas of research. The DateLife project helps with awareness of the existing variation in expert time of divergence data, and will foster exploration of the effect of alternative divergence time hypothesis on the results of analyses, nurturing a culture of more cautious interpretation of evolutionary results.

Availability

datelife is free and open source and it can be used through its current website <http://www.datelife.org/query/>, through its R package, and through Phylotastic's project web portal <http://phylo.cs.nmsu.edu:3000/>. **datelife**'s website is maintained using RStudio's shiny server and the shiny package open infrastructure, as well as Docker. **datelife**'s R package stable version is available for installation from the CRAN repository (<https://cran.r-project.org/package=datelife>) using the command `install.packages(pkgs = "datelife")` from within R. Development versions are available from the GitHub repository (<https://github.com/phylotastic/datelife>) and can be installed using the command `devtools::install_github("phylotastic/datelife")`.

Supplementary Material

Code used to generate all versions of this manuscript, the biological examples, as well as the benchmark of functionalities are available at `datelifeMS1`, `datelife_examples`, and `datelife_benchmark` repositories in LLSR's GitHub account.

Funding

Funding was provided by the US National Science Foundation (NSF) grants ABI-1458603 to Datelife project and DBI-0905606 to the National Evolutionary Synthesis Center (NESCent), and the Phylotastic project Grant ABI-1458572.

Acknowledgements

The DateLife project was born as a prototype tool aiming to provide these services, and was developed over a series of hackathons at the National Evolutionary Synthesis Center, NC, USA (Stoltzfus et al., 2013). We thank colleagues from the O’Meara Lab at the University of Tennessee Knoxville for suggestions, discussions and software testing. The late National Evolutionary Synthesis Center (NESCent), which sponsored hackathons that led to initial work on this project. The team that assembled **datelife**’s first proof of concept: Tracy Heath, Jonathan Eastman, Peter Midford, Joseph Brown, Matt Pennell, Mike Alfaro, and Luke Harmon. The Open Tree of Life project that provides the open, metadata rich repository of trees used for **datelife**. The many scientists who publish their chronograms in an open, reusable form, and the scientists who curate them for deposition in the Open Tree of Life repository. The NSF for funding nearly all the above, in addition to the ABI grant that funded this project itself.

References

- Antonelli, A., Hettling, H., Condamine, F. L., Vos, K., Nilsson, R. H., Sanderson, M. J., ... Vos, R. A. (2017). Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of Taxa. *Systematic Biology*, 66(2), 153–166. <https://doi.org/10.1093/sysbio/syw066>
- Archie, J., Day, W. H., Felsenstein, J., Maddison, W., Meacham, C., Rohlf, F. J., & Swofford, D. (1986). The Newick tree format. Retrieved from <https://evolution.genetics.washington.edu/phylip/newicktree.html>
- Bapst, D. W. (2012). Paleotree: An R package for paleontological and phylogenetic analyses of evolution. *Methods in Ecology and Evolution*, 3(5), 803–807. <https://doi.org/10.1111/j.2041-210X.2012.00223.x>
- Barba-Montoya, J., Reis, M. dos, Schneider, H., Donoghue, P. C., & Yang, Z. (2018). Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a cretaceous terrestrial revolution. *New Phytologist*, 218(2), 819–834.
- Barker, F. K., Burns, K. J., Klicka, J., Lanyon, S. M., & Lovette, I. J. (2012). Going to extremes: Contrasting rates of diversification in a recent radiation of new world passerine birds. *Systematic Biology*, 62(2), 298–320.
- Barker, F. K., Burns, K. J., Klicka, J., Lanyon, S. M., & Lovette, I. J. (2015). New insights into new world biogeography: An integrated view from the phylogeny of blackbirds, cardinals, sparrows, tanagers, warblers, and allies. *The Auk: Ornithological Advances*, 132(2), 333–348.
- Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., ... Enquist, B. J. (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics*, 14(1). <https://doi.org/10.1186/1471-2105-14-16>
- Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S., & Bremer, K. (2007). Estimating Divergence Times in Large Phylogenetic Trees. *Systematic Biology*,

56(788777878), 741–752. <https://doi.org/10.1080/10635150701613783>

Burns, K. J., Shultz, A. J., Title, P. O., Mason, N. A., Barker, F. K., Klicka, J., ...

Lovette, I. J. (2014). Phylogenetics and diversification of tanagers (passeriformes: Thraupidae), the largest radiation of neotropical songbirds. *Molecular Phylogenetics and Evolution*, 75, 41–77.

Chamberlain, S. A., & Szöcs, E. (2013). taxize : taxonomic search and retrieval in R [version 2; referees: 3 approved]. *F1000Research*, 2(191), 1–29.

<https://doi.org/10.12688/f1000research.2-191.v2>

Chamberlain, S. A., Szöcs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., ...

Li, G. (2019). *taxize: Taxonomic information from around the web*. Retrieved from <https://github.com/ropensci/taxize>

Claramunt, S., & Cracraft, J. (2015). A new time tree reveals earth history’s imprint on the evolution of modern birds. *Science Advances*, 1(11), e1501005.

Criscuolo, A., Berry, V., Douzery, E. J. P., & Gascuel, O. (2006). SDM: A fast distance-based approach for (super)tree building in phylogenomics. *Systematic Biology*, 55(5), 740–755. <https://doi.org/10.1080/10635150600969872>

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1), 1–15. Retrieved from <http://www.jstor.org/stable/2461605>

Gibb, G. C., England, R., Hartig, G., McLenachan, P. A., Taylor Smith, B. L., McComish, B. J., ... Penny, D. (2015). New zealand passerines help clarify the diversification of major songbird lineages during the oligocene. *Genome Biology and Evolution*, 7(11), 2983–2995.

Harmon, L., Weir, J., Brock, C., Glor, R., & Challenger, W. (2008). GEIGER: investigating evolutionary radiations. *Bioinformatics*, 24, 129–131.

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, 32(4), 835–845. <https://doi.org/10.1093/molbev/msv037>

- Heibl, C. (2008). *PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages*. Retrieved from <http://www.christophheibl.de/Rpackages.html>
- Hooper, D. M., & Price, T. D. (2017). Chromosomal inversion differences correlate with range overlap in passerine birds. *Nature Ecology & Evolution*, 1(10), 1526.
- Jetz, W., Thomas, G., Joy, J. J. B., Hartmann, K., & Mooers, A. (2012). The global diversity of birds in space and time. *Nature*, 491(7424), 444–448. <https://doi.org/10.1038/nature11631>
- Magallón, S. (2010). Using fossils to break long branches in molecular dating: A comparison of relaxed clocks applied to the origin of angiosperms. *Systematic Biology*, 59(4), 384–399.
- Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L., & Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist*, 207(2), 437–453.
- McTavish, E. J., Hinchliff, C. E., Allman, J. F., Brown, J. W., Cranston, K. A., Holder, M. T., . . . Smith, S. A. (2015). Phylesystem: A git-based data store for community-curated phylogenetic estimates. *Bioinformatics*, 31(17), 2794–2800.
- Michonneau, F., Brown, J. W., & Winter, D. J. (2016). rotl: an R package to interact with the Open Tree of Life data. *Methods in Ecology and Evolution*, 7(12), 1476–1481. <https://doi.org/10.1111/2041-210X.12593>
- Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, 17(4), 508–525. <https://doi.org/10.1111/ele.12251>
- Ooms, J., & Chamberlain, S. (2018). *Phylocomr: Interface to 'phylocom'*. Retrieved from <https://CRAN.R-project.org/package=phylocomr>
- OpenTreeOfLife, Redelings, B., Cranston, K. A., Allman, J., Holder, M. T., & McTavish, E. J. (n.d.). Open Tree of Life APIs v. 3.0. Retrieved from [https://github.com/OpenTreeOfLife/germinator/wiki/Open-Tree-of-Life-](https://github.com/OpenTreeOfLife/germinator/wiki/Open-Tree-of-Life-%7B)

Web-APIs%7D

- Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2), 289–290.
- Posadas, P., Crisci, J. V., & Katinas, L. (2006). Historical biogeography: A review of its basic concepts and critical issues. *Journal of Arid Environments*, 66(3), 389–403.
- Price, T. D., Hooper, D. M., Buchanan, C. D., Johansson, U. S., Tietze, D. T., Alström, P., . . . others. (2014). Niche filling slows the diversification of himalayan songbirds. *Nature*, 509(7499), 222.
- R Core Team. (2018). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., . . . others. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, 559(7714), 392.
- Ramshaw, J., Richardson, D., Meatyard, B., Brown, R., Richardson, M., Thompson, E., & Boulter, D. (1972). The time of origin of the flowering plants determined by using amino acid sequence data of cytochrome c. *New Phytologist*, 71(5), 773–779.
- Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The barcode of life data system ([http://www. Barcodinglife. org](http://www.Barcodinglife.org)). *Molecular Ecology Notes*, 7(3), 355–364.
- Revell, L. J. (2012). Phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3, 217–223.
- Sanchez-Reyes, L. L., O’Meara, B., Eastman, J., Heath, T., Wright, A., Schliep, K., . . . Alfaro, M. (2022). datelife: Scientific Data on Time of Lineage Divergence for Your Taxa. *R Package Version 0.6.2*. Retrieved from <https://doi.org/10.5281/zenodo.593938>
- Sanderson, M. J. (2003). r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2),

301–302.

Sanderson, M. J., & Doyle, J. A. (2001). Sources of error and confidence intervals in estimating the age of angiosperms from rbcL and 18S rDNA data. *American Journal of Botany*, 88(8), 1499–1516.

Sauquet, H. (2013). A practical guide to molecular dating. *Comptes Rendus Palevol*, 12(6), 355–367.

Sauquet, H., Ramírez-Barahona, S., & Magallón, S. (2021). The age of flowering plants is unknown.

Schenk, J. J. (2016). Consequences of secondary calibrations on divergence time estimates. *PLoS ONE*, 11(1). <https://doi.org/10.1371/journal.pone.0148228>

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

Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C. M., ...

Jordan, G. (2013). Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC Bioinformatics*, 14.

<https://doi.org/10.1186/1471-2105-14-158>

Title, P. O., & Rabosky, D. L. (2016). Do Macrophylogenies Yield Stable Macroevo-
lutionary Inferences? An Example from Squamate Reptiles. *Systematic Biology*, syw102. <https://doi.org/10.1093/sysbio/syw102>

Vos, R. A., Balhoff, J. P., Caravas, J. A., Holder, M. T., Lapp, H., Maddison, W. P., ... others. (2012). NeXML: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic Biology*, 61(4), 675–689.

Webb, Campbell O. (2000). Exploring the Phylogenetic Structure of Ecological Communities : An Example for Rain Forest Trees. *The American Naturalist*, 156(2), 145–155.

Webb, Campbell O., Ackerly, D. D., & Kembel, S. W. (2008). Phylocom: Software for the analysis of phylogenetic community structure and trait evolution.

Bioinformatics, 24(18), 2098–2100.

<https://doi.org/10.1093/bioinformatics/btn358>

Webb, Campbell O., & Donoghue, M. J. (2005). Phylomatic: Tree assembly for applied phylogenetics. *Molecular Ecology Notes*, 5(1), 181–183.

FIGURES

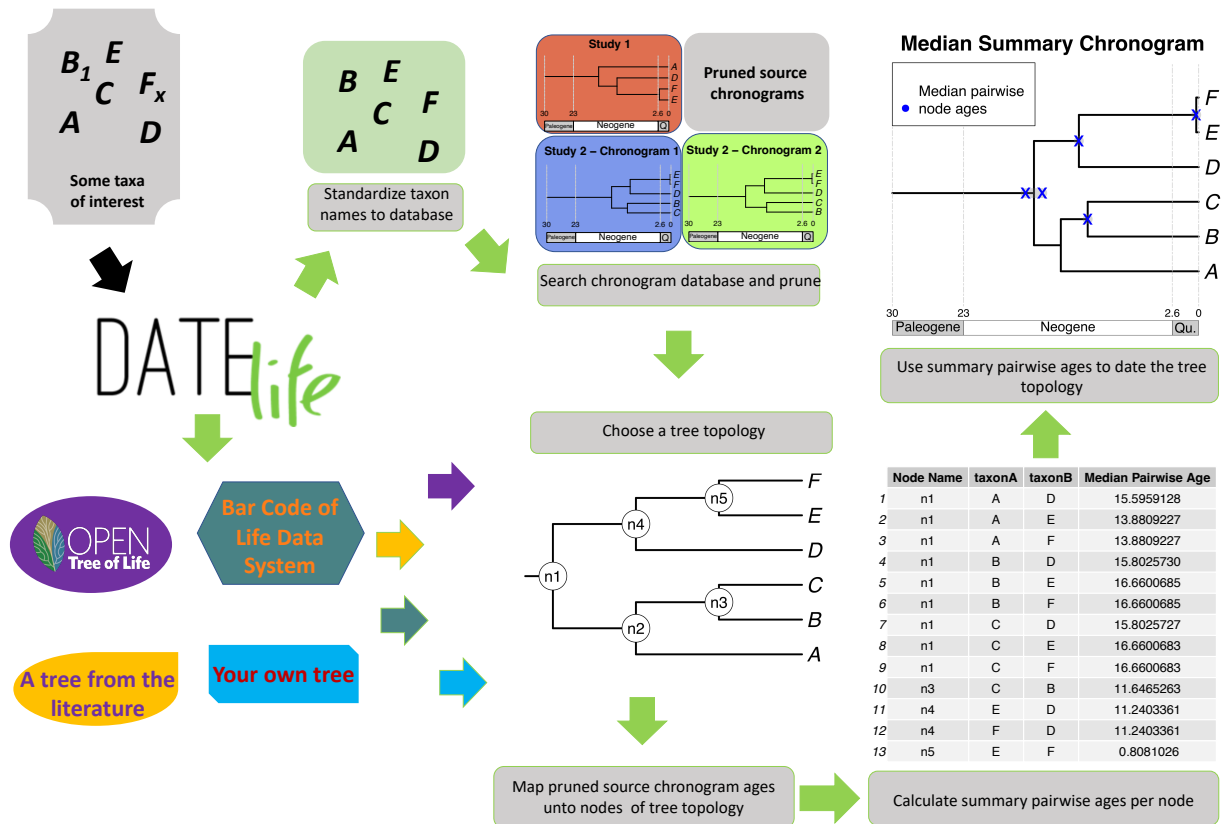


FIGURE 1. Stylized DateLife workflow. This shows the general workflows and analyses that can be performed with `datelife`, via the R package or through the website at www.datelife.org/query/. Details on the functions involved on each workflow are shown in `datelife`'s R package vignette.

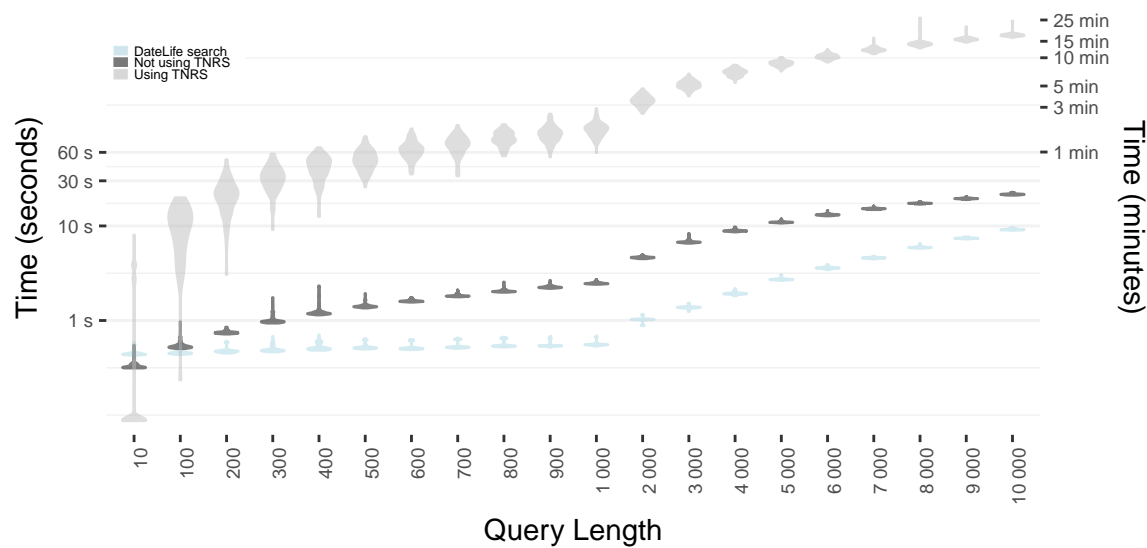


FIGURE 2. Input taxon name processing and chronogram database search computation time increases with number of input taxon names. We sampled N bird species names for each input size class, 100 times, and then performed a **datelife** search using the Taxon Names Resolution Service (TNRS; dark gray), and without using TNRS (light gray). We also performed a search using the already processed query for comparison (light blue).

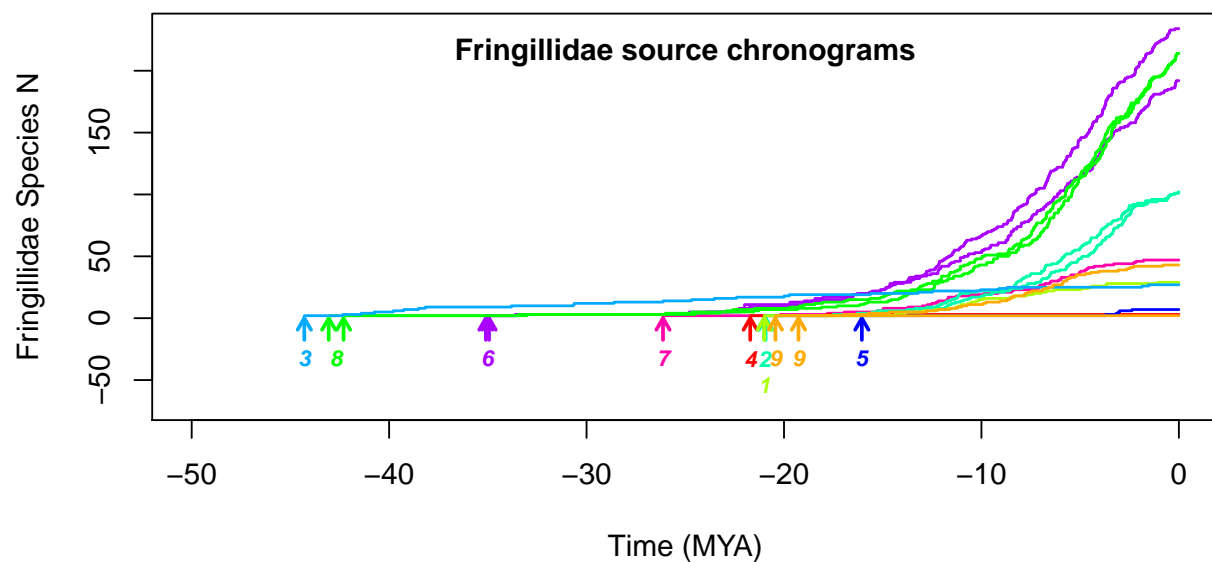


FIGURE 3. Lineage through time (LTT) plots of source chronograms containing all or a subset of species from the bird family Fringillidae of true finches. Arrows indicate maximum age of each chronogram. Numbers reference to chronograms' original publications 1: Barker et al. (2012), 2: Barker et al. (2015), 3: Burns et al. (2014), 4: Claramunt and Cracraft (2015), 5: Gibb et al. (2015), 6: Hedges et al. (2015), 7: Hooper and Price (2017), 8: Jetz et al. (2012), 9: Price et al. (2014).

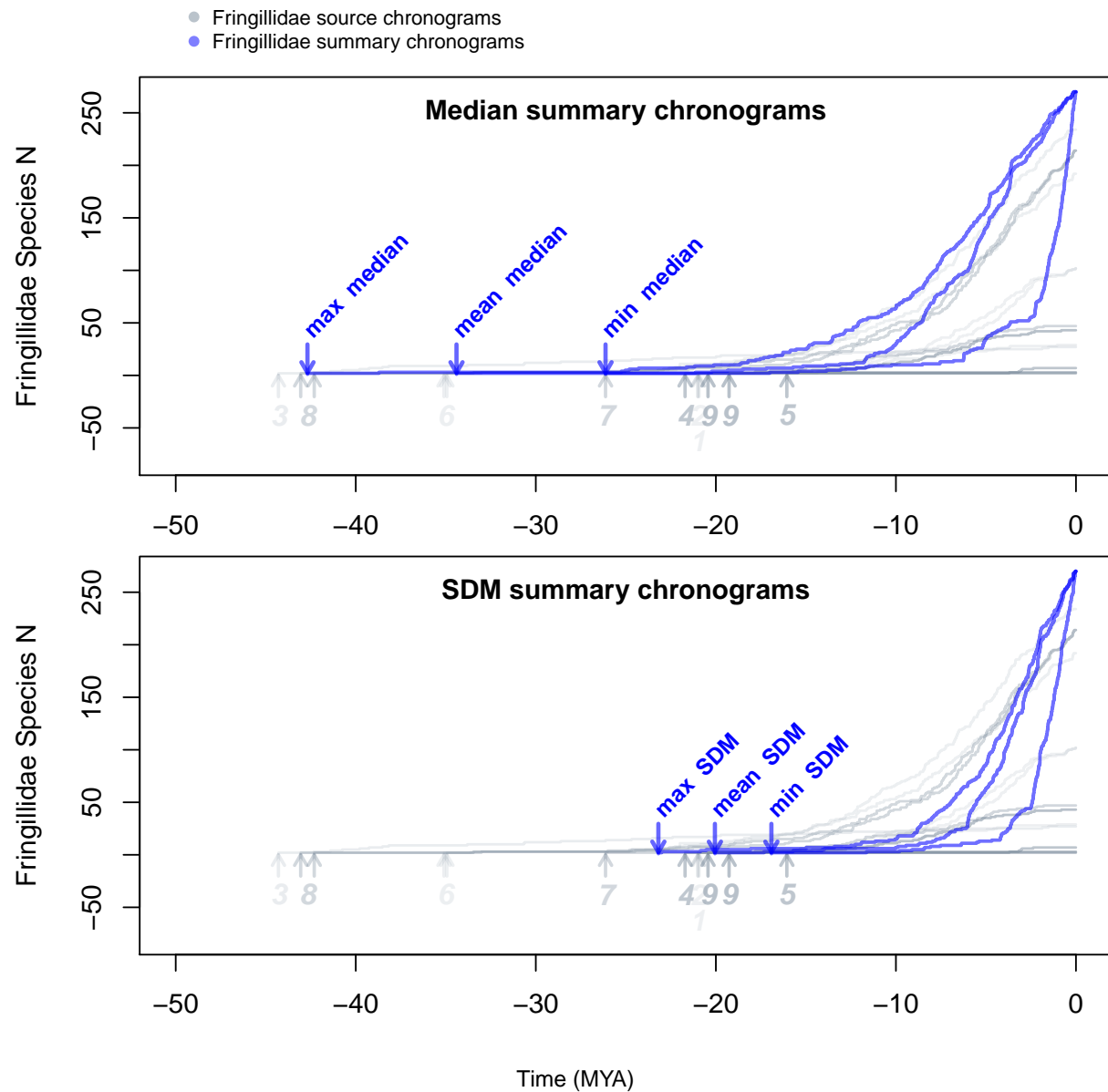


FIGURE 4. LTT plots of median (top) and Supermatrix Distance Method (SDM; bottom) chronograms summarising information from source chronograms found for the Fringillidae. Arrows indicate tree maximum age.

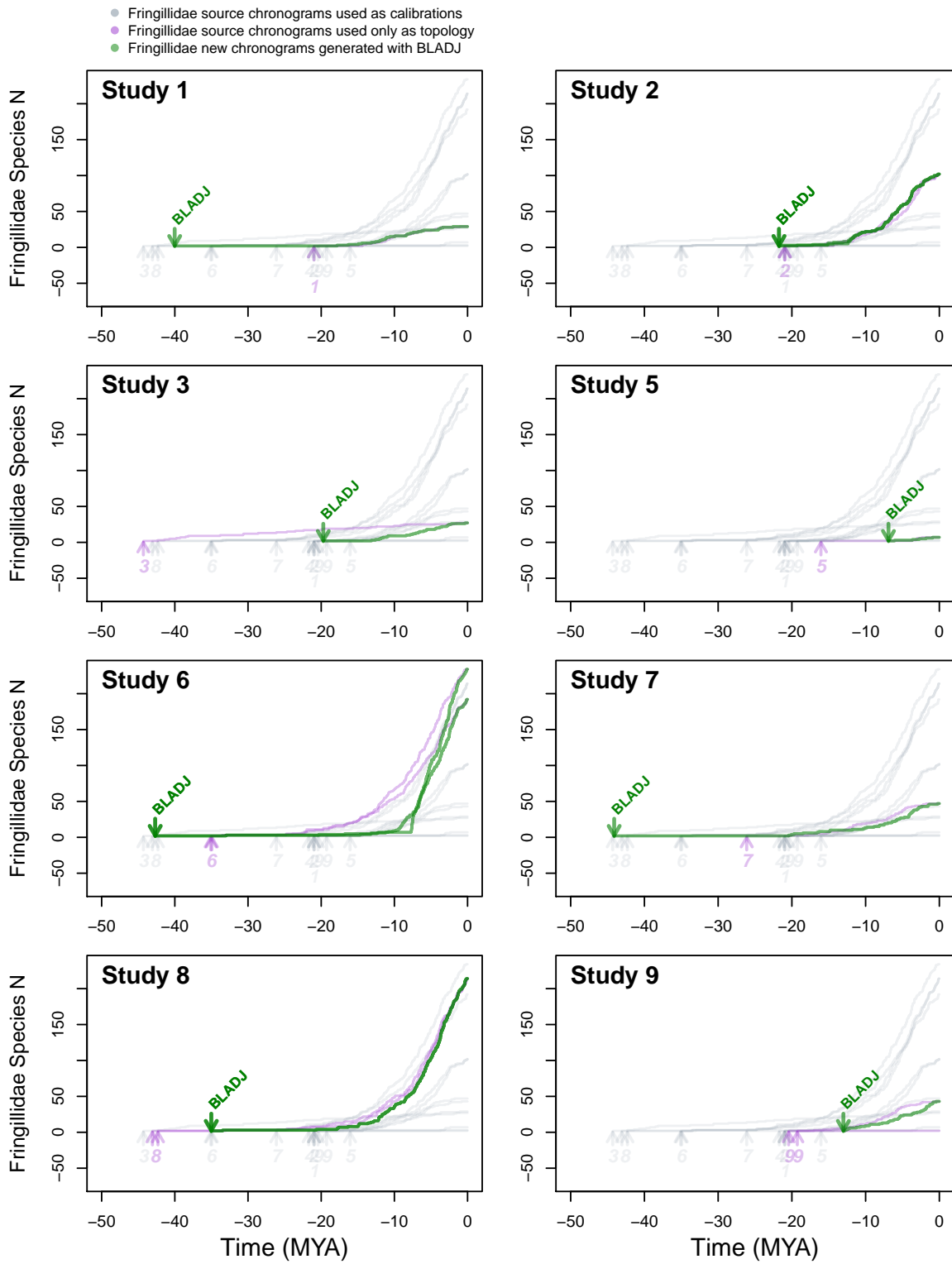


FIGURE 5. LTT plots showing results from the cross-validation analyses of trees without branch lengths dated using BLADJ. The dating analysis can only be performed in trees with more than 2 tips, thus excluding chronogram from study 4; its data was still used as calibration for the other source chronograms.

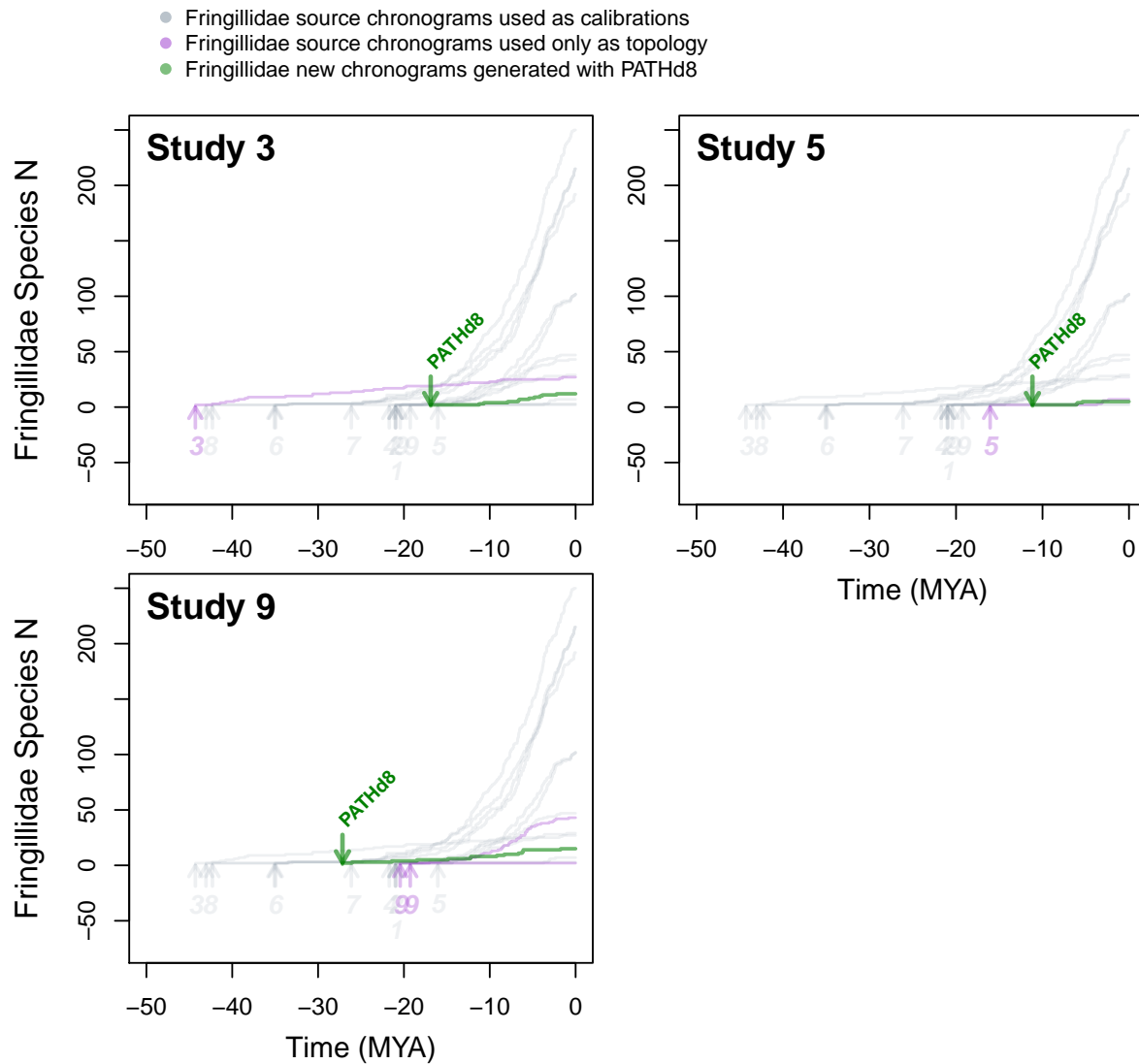


FIGURE 6. LTT plots showing results from the cross-validation analyses of trees with branch length reconstructed with data from the Barcode of Life Database (BOLD) dated using PATHd8. We could construct a tree with branch lengths for all source chronograms. However, dating with PATHd8 was only successful in three source chronograms shown here.