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

Luna L. Sánchez Reyes[1,2], Emily Jane McTavish[1], & Brian O'Meara[2]

[1] University of California, Merced

[2] University of Tennessee, Knoxville

Author Note

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

Time of evolutionary origin is fundamental for research in the natural sciences, as well as for education, science communication and policy. Despite an increased availability of fossil and molecular data, and time-efficient analytical techniques, 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 for a majority of interested parties. Yet, the amount of published chronograms has increased significantly in the past two decades, and a non-negligeable proportion of these data have been steadily accumulating in public, open databases such as TreeBASE and Open Tree of Life, exposing a wealth of expertly-curated and peer-reviewed data on time of evolutionary origin in a programatic and reusable way, for a large quantity and diversity of organisms. This trend results from intensive and localized efforts for improving data sharing practices, as well as incentivizing open science in biology. Despite these trends, accessibility to state-of-the-art knowledge on time of evolutionary origin is still reduced.

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 workflow of `datelife` is to construct a chronogram for any given combination of taxon names, by searching a local chronogram database constructed and curated from the Open Tree of Life (OpenTree), which incorporates phylogenetic data from the TreeBASE database as well. We implement and test methods for summarizing time data from multiple source chronograms using supertree and congruification algorithms. Additionally, time data extracted from source chronograms can be usedas secondary calibration points to add branch lengths proportional to absolute time to a tree topology using alternative dating methods.

Summary and newly generated trees are potentially useful to evaluate evolutionary hypothesis in different areas of research in biology. How well this chronograms work for this purpose still needs to be tested.

`datelife` will be useful to increase awereness on the existing variation in expert time of divergence data, and might 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; Congruification; Supertree; Calibrations; Secondary calibrations

Word count: 2949

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

Introduction

Time of evolutionary origin represents a fundamental piece of information for understanding biological processes in many areas of research, from developmental to conservation biology (Felsenstein, 1985; Webb, 2000), from historical biogeography to species diversification studies (Morlon, 2014; Posadas, Crisci, & Katinas, 2006). The number of studies publishing phylogenies with branch lengths proportional to absolute time (hereafter chronograms) have constantly increased in number for the last two decades (Kumar, Stecher, Suleski, & Hedges, 2017). Still, generating a chronogram is not an easy task unless you have specialized training, and a wealth of time and resources: it requires inferring a phylogenetic tree using genetic markers, obtaining independent time data from the fossil record, and understanding the placement of fossils on the tree as well as their limits for analysis. That is why there has been an urge for promoting and facilitating the reuse of the vast amount of phylogenetic and evolutionary time data that has been made available in publications, for the advantage of research relying on this information (Stoltzfus et al., 2013; Webb & Donoghue, 2005).

A tool for efficient reuse of expert, published data on time of evolutionary origin should have an open and fully public chronogram database storing data in a format suitable for scientific reuse, an automatised way of accessing the information, and straightforward means of comparing and summarizing chronogram information as needed by the user. A prototype service aiming to meet this criteria was developed over a series of hackathons at the National Evolutionary Synthesis Center (Stoltzfus et al., 2013). Here we present the full implementation of the `datelife` service, constituted by an R package and a web site available at www.datelife.org/query/. The current implementation of `datelife` performs the tasks described above. It features an algorithm for automatic curation and maintenance of an open database of chronograms pulled from the OpenTree public repository, methods to summarize and compare source chronograms, and new functions to visualize and graphically compare source and summary chronograms.

Implementation/Description/Workflow

The `datelife` workflow builds off of functions from several R packages (rotl (Michonneau, Brown, & Winter, 2016), ape (Paradis, Claude, & Strimmer, 2004), geiger (Harmon, Weir, Brock, Glor, & Challenger, 2008), paleotree (Bapst, 2012), bold (Chamberlain, 2018), phytools (Revell, 2012), taxize (Chamberlain, 2018; Chamberlain & Szöcs, 2013), phyloch (Heibl, 2008), and phylocomr (Ooms & Chamberlain, 2018)).

The basic `datelife` workflow is shown in figure 1, largely:

1. It starts with an input consisting of at least two taxon names, which can be provided as a comma separated character string, or as tip labels on a tree. The tree can be provided in newick format, also as a character string, or as a "phylo" R object, and can have any type of branch lengths or none.

2. The input taxon names are cleaned with TNRS and saved as a ‘datelifeQuery‘ object. If taxon names are taxonomic groups above the species level, ‘datelife‘ has two alternative behaviors. If the "get species from taxon" flag is active, ‘datelife‘ will retrieve all species within a higher taxon name and add the species names to the input. If the flag is inactive, ‘datelife‘ will drop the higher taxon names from the input. The cleaned input taxon names are searched across the source chronogram database. Source chronograms with at least two matching input taxon names are singled out and pruned down to preserve only input taxon names in the tips of the chronograms. Then, each pruned source 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 a ‘datelifeResult‘ object.

3. At this point, various summary data can be obtained to inform decisions for the next steps of the analysis workflow. Types of summary information provided are: a) all pruned source chronograms, b) age of the MRCA (most recent common ancestor) of the pruned source chronograms, c) citations of studies where pruned source chronograms were originally published, d) a summary table with all of the above, e) a single summary chronogram of all or a subset of pruned source chronograms, f) a report of successful matches of input taxon names across pruned source chronograms, and g) the single pruned source chronogram with the most matching input taxon names.

4. Finally, time of lineage divcergence obtained from the pruned source chronograms can be used as secondary calibration points to date a tree with or without branch lengths containing some or all input taxon names.

5. 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 described below.

6. Users can easily save all source and summary chronograms in formats that permit easy 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.

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 the taxonomic name resolution service (TNRS; Boyle et al., 2013) 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/phylotastic/datelife).

Results

Case study

We illustrate the `datelife` workflow using the family of true finches, Fringillidae as an example. To contextualize, a college educator wishes to know the state-of-the-art on time of evolutionary origin of species belonging to the true finches using `datelife`. One option is 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 higher-taxon-name `datelife` query. This will get taxon names for all recognised species within any higher taxon. The Fringillidae has 289 species, according to the Open Tree of Life taxonomy. Once with a curated set of query taxon names, the next step is to run a `datelife` search. This will find all chronograms that contain at least two queried taxon names, and will save the information on time of lineage divergence as (an R "data frame") table. 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 the 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 (using the median and the SDM method respectively) and then the available node ages are used as fixed ages 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; ratnasingham2007bold] 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 expert information on time of lineage divergence easily accesible for comparison, reuse, and reanalysis, to researchers in all areas of science and with all levels of expertise in the matter. It is a very fast tool that fulfills the quality of openness and 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. However, it has many flaws. Some of them can be overcome, some of them might represent limitations.

Up to the time this manuscript was written, `datelife`'s chronogram database had 231 chronograms, pulled entirely from OpenTree's tree repository, the only public tree repository from where `datelife` can currently get chronograms to construct its database. This represents 5.79% of the largest existing chronogram database, TimeTree, which has a collection of 3,998 chronograms as of November 02 2021. Unfortunately, TimeTree's database is not open for scientific reuse nor automatised data mining (Kumar et al., 2017). In 2015, a synthetic chronogram was constructed from 2,274 chronograms available at the time on the TimeTree database (Hedges, Marin, Suleski, Paymer, & Kumar, 2015). This is the only synthetic TimeTree chronogram that has been made publicly available and deposited on the OpenTree repository, and is part of datelife's database now. Hence, the amount of lineages represented in datelife's database is at least as substantial as TimeTree's, ensuring that some information will be available for any given taxon or lineage. Regrettably, this does not ensure that the full state of knowledge of time of divergence of the taxon/lineage will be available. Incorporation of more published chronograms into `datelife`'s database is crucial to improve its services. One option to increase our 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 on a second step, by a curator. 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 tree repository (McTavish et al., 2015).

Another potential concern comes from summary chronograms. We currently summarize by default all source chronograms that overlap with at least two taxa. Users can subset source data if they have reasons to choose some source chronograms over others. Strictly speaking, a good chronogram should reflect the real time of lineage divergence accurately and precisely. To our knowledge, there is no objective way to determine if an expert chronogram is better than another. Some criteria that have been put forward are the level of lineage sampling and the number of calibrations used. 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). 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). Yet, there are different ways to use secondary calibrations and that same bias might not be encountered with dating methods that do not require setting priors, i.e., Maximum Likelihood methods such as r8s (Sanderson, 2003). Certainly, further studies are required to fully understand the effect of using secondary calibrations on time estimates and downstream anlyses.

Furthermore, even chronograms obtained with primary fossil data can show substantial variation in time estimates between clades, 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, Gómez-Acevedo, Sánchez-Reyes, & Hernández-Hernández, 2015; Ramshaw et al., 1972; Sanderson & Doyle, 2001). 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 here 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 teh same one to calibrate outside the clade, the resulting age estimates will probably differ substantially (the placement of calibrations is proved to deeply affect estimated times of lineage divergence). 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 ha sthe 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 tp keep and the ones that we are going to discard is key, since we are potentially loosing important parts of the evolutionary history of a lineage that might only be reflected in source chronograms and not on the summary chronogram.

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 (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 (BLADJ option). Making up branch lengths in this or other ways is accepted 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 imputation 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 *de novo* 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.

`datelife` 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. `datelife` 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 will be 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

We thank colleagues from the O'Meara Lab at the University of Tennesse 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

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. (2018). *bold: Interface to Bold Systems API*. Retrieved from https://CRAN.R-project.org/package=bold

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

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.

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

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

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

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 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.

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.

R Core Team. (2018). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Revell, L. J. (2012). Phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*, 217–223.

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.

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 Macroevolutionary Inferences? An Example from Squamate Reptiles. *Systematic Biology*, syw102. https://doi.org/10.1093/sysbio/syw102

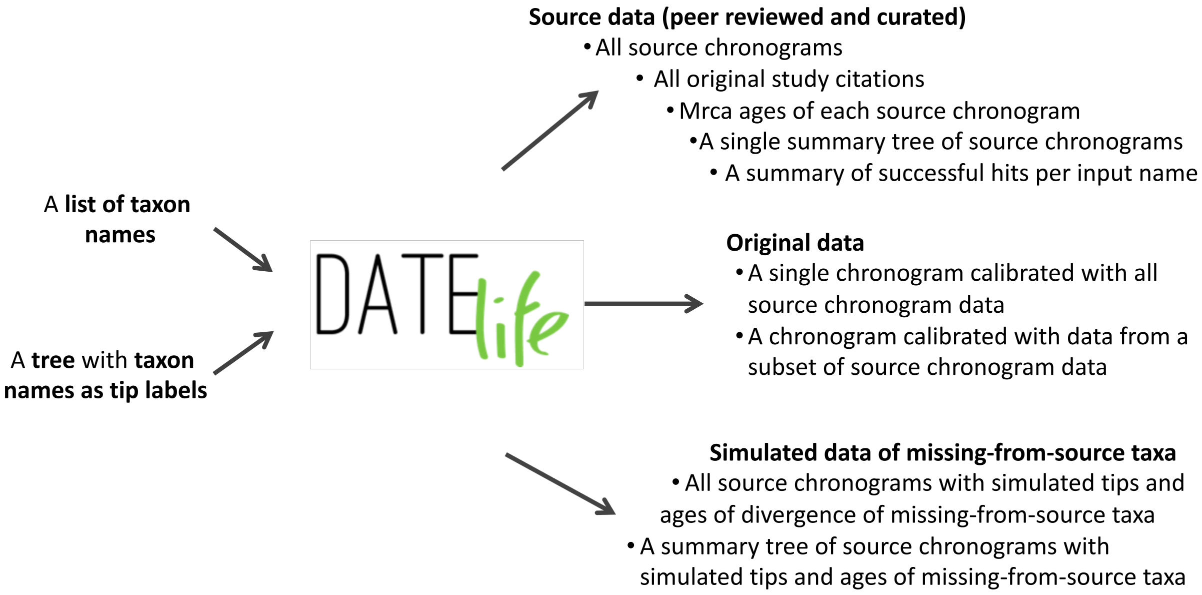
Webb, C. O. (2000). Exploring the Phylogenetic Structure of Ecological Communities : An Example for Rain Forest Trees. *The American Naturalist*, *156*(2), 145–155.

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

to be formatted in the same way as the general text (double spaced and linenumbered)

FIGURES



A **list of taxon names**

A **tree** with **taxon names as tip labels**

Source data (peer reviewed and curated)

- All source chronograms
- All original study citations
- Mrca ages of each source chronogram
- A single summary tree of source chronograms
- A summary of successful hits per input name

Original data

- A single chronogram calibrated with all source chronogram data
- A chronogram calibrated with data from a subset of source chronogram data

Simulated data of missing-from-source taxa

- All source chronograms with simulated tips and ages of divergence of missing-from-source taxa
- A summary tree of source chronograms with simulated tips and ages of missing-from-source taxa

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.

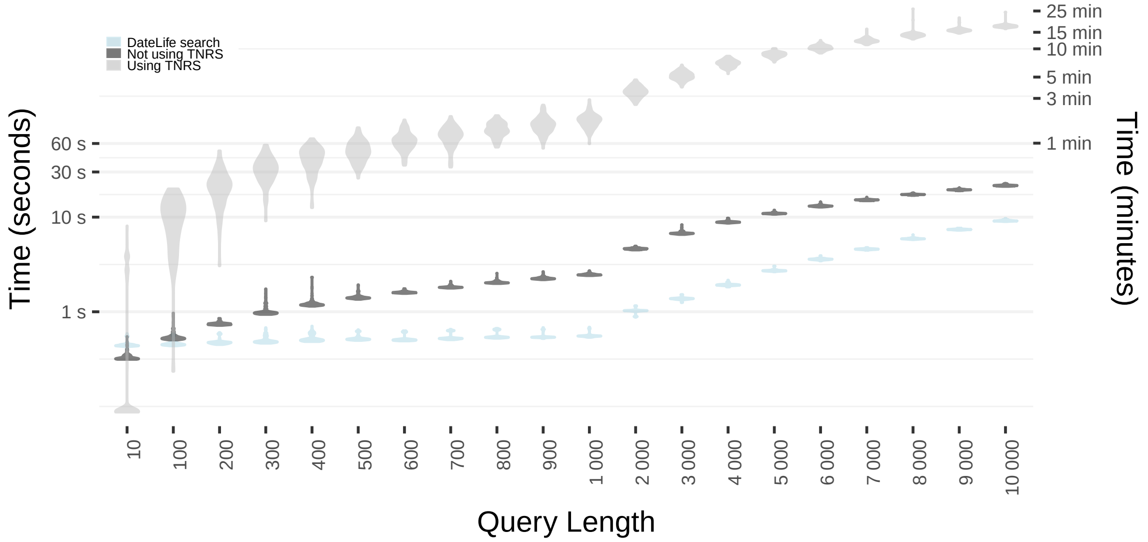


Figure 2. Computation time of query processing and search across `datelife`'s chronogram database relative to number of input taxon names. We sampled N names from the class Aves for each cohort 100 times and then performed a search with query processing not using the Taxon Names Resoultion Service (TNRS; dark gray), and using TNRS (light gray). We also performed a search using the already processed query for comparison (light blue).

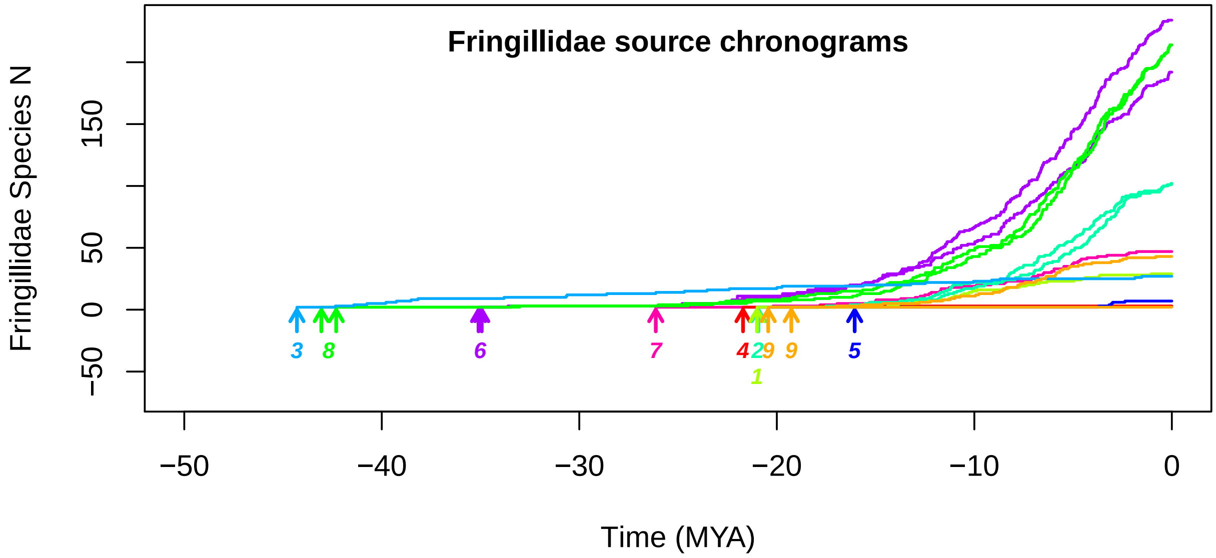


Figure 3. Lineage through time (LTT) plots of source chronograms containing all or a subset of species from the bird family Fringillidae of true finches. Arrows indicate maximum age of each chronogram. Numbers reference to chronograms' original publications 1: Barker et al. (2012), 2: Barker et al. (2015), 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).

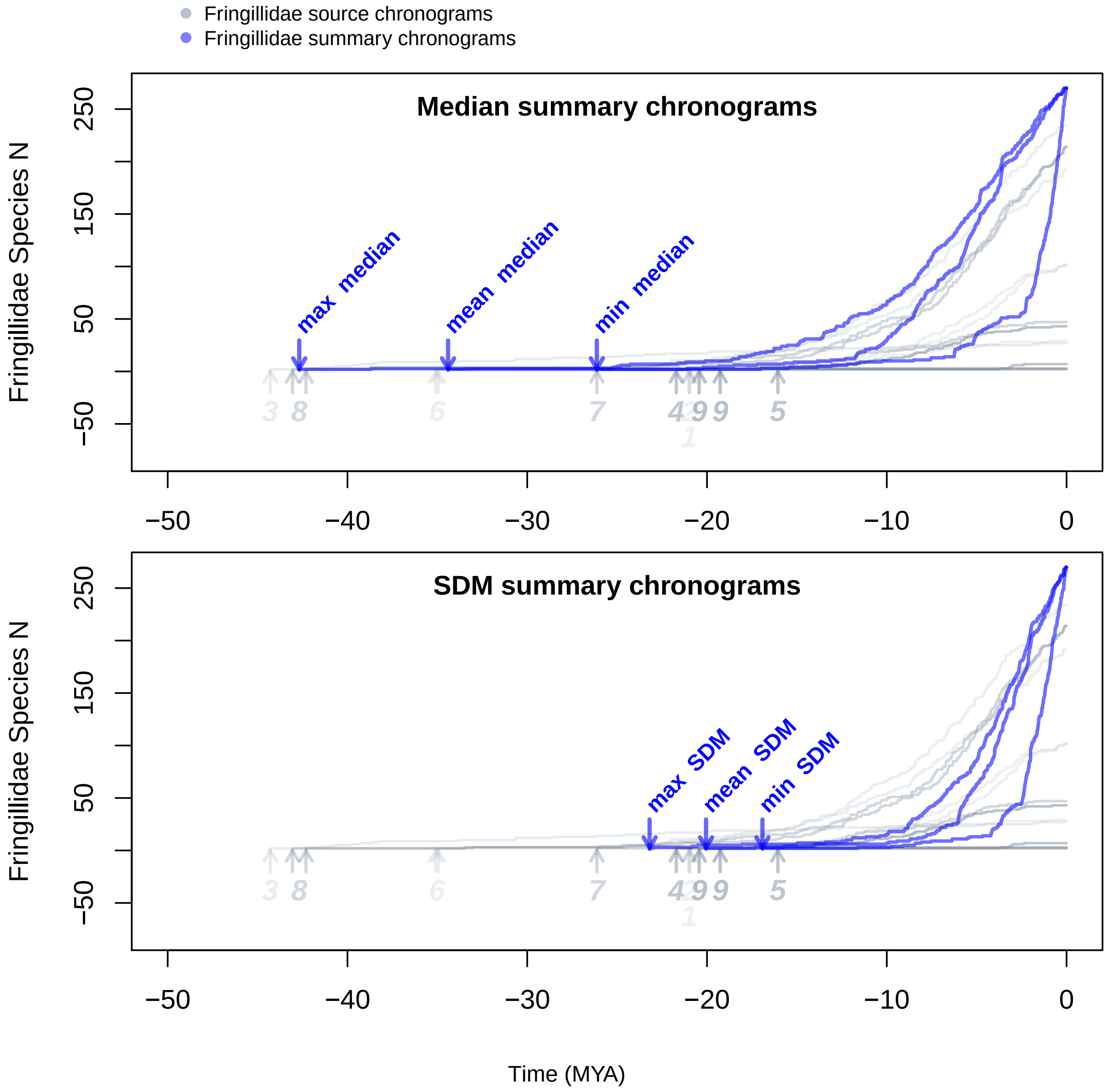


Figure 4. LTT plots of median (top) and Supermatrix Distance Method (SDM; bottom) chronograms summarising information from source chronograms found for the Fringillidae. Arrows indicate tree maximum age.

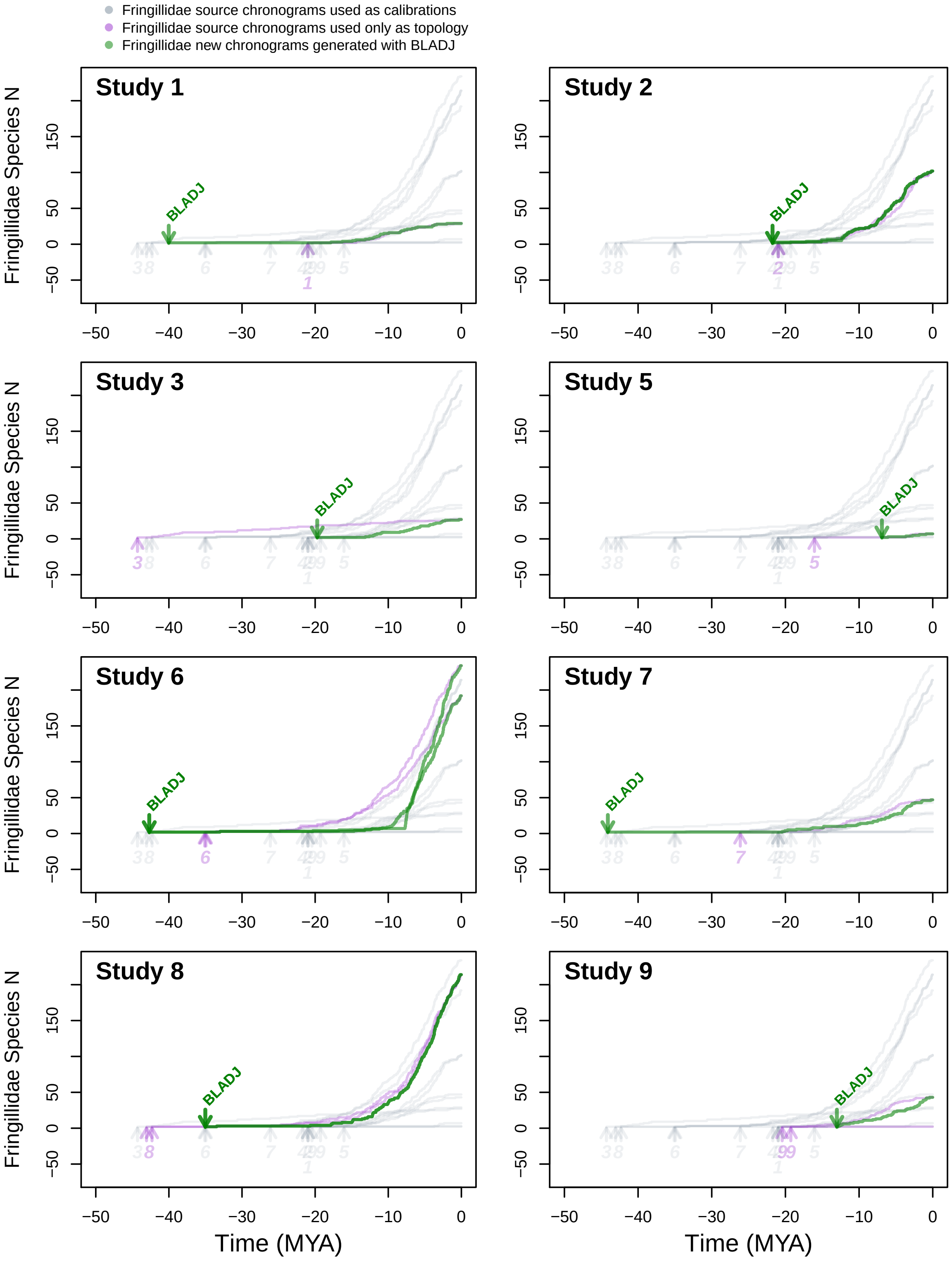


Figure 5. LTT plots showing results from the cross-validation analyses of trees without branch lengths dated using BLADJ. The dating analysis can only be performed in trees with more than 2 tips, thus excluding chronogram from study 4; its data was still used as calibration for the other source chronograms.

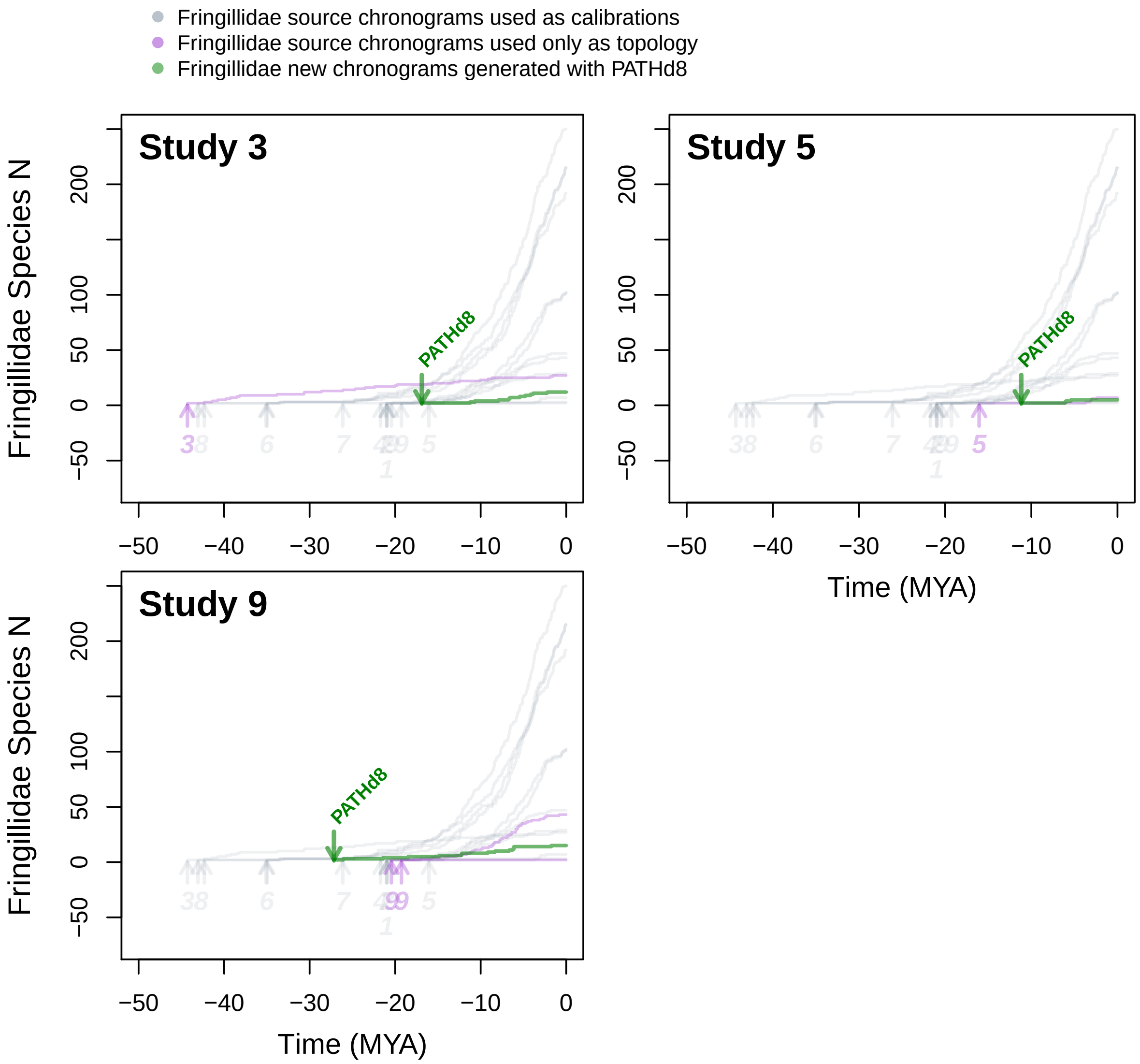


Figure 6. LTT plots showing results from the cross-validation analyses of trees with branch length reconstructed with data from the Barcode of Life Database (BOLD) dated using PATHd8. We could construct a tree with branch lengths for all source chronograms. However, dating with PATHd8 was only successful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