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

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

3   [1] University of California, Merced

4   [2] University of Tennessee, Knoxville

5   Author Note

6   School of Natural Sciences, University of California, Merced, Science and Engineering

7 Building 1.

8   Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville,

9 425 Hesler Biology Building, Knoxville, TN 37996, USA.

10   The authors made the following contributions. Luna L. Sánchez Reyes: Data curation,

11 Investigation, Software, Visualization, Validation, Writing - Original Draft Preparation,

12 Writing - Review & Editing; Emily Jane McTavish: Resources, Software, Writing - Review &

13 Editing; Brian O'Meara: Conceptualization, Funding acquisition, Methodology, Resources,

14 Software, Supervision, Writing - Review & Editing.

15   Correspondence concerning this article should be addressed to Luna L. Sánchez Reyes, .

16 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.

⁴³ Summary and newly generated trees are potentially useful to evaluate evolutionary

⁴⁴ hypothesis in different areas of research in biology. How well this chronograms work for this

⁴⁵ purpose still needs to be tested.

⁴⁶ `datelife` will be useful to increase awereness on the existing variation in expert time

⁴⁷ of divergence data, and might foster exploration of the effect of alternative divergence time

⁴⁸ hypothesis on the results of analyses, providing a framework for a more informed

⁴⁹ interpretation of evolutionary results.

⁵⁰ *Keywords:* Tree; Phylogeny; Scaling; Dating; Ages; Divergence times; Open Science;

⁵¹ Congruification; Supertree; Calibrations; Secondary calibrations

⁵² Word count: 2949

₅₃   DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life

₅₄   **Introduction**

₅₅   In the natural sciences, absolute time of evolutionary origin of lineages is represented

₅₆ graphically as a phylogeny with branch lengths proportional to geologic time, also known as

₅₇ a chronogram. Notably, building a chronogram is not an easy task. First, it requires

₅₈ obtaining and curating genetic data to construct an homology hypothesis as a genetic

₅₉ alignment; then, it is important to carefully choose and apply the software that will use the

₆₀ alignment to reconstruct an evolutionary hypothesis in the form of a phylogeny; next,

₆₁ calibrations, i.e., independent age data points from the fossil record or other suitable geologic

₆₂ events have to be obtained from the literature –or by going into the field; fourthly,

₆₃ calibrations have to be carefully placed on the reconstructed phylogeny, demonstrating a

₆₄ biological and geological understanding of their limits and applicability; finally, the

₆₅ appropriate software and model of evolution to estimate ellapsed time since evolutionary

₆₆ events on the phylogeny has to be chosen and applied. In sum, building chronograms require

₆₇ specialized biological training and non-negligible budget, human and time resources. Still,

₆₈ chronograms represent key knowledge for the study of natural processes in many areas of

₆₉ scientific research, from developmental to conservation biology (Felsenstein, 1985; Campbell

₇₀ O. Webb, 2000), from historical biogeography to species diversification (Morlon, 2014;

₇₁ Posadas, Crisci, & Katinas, 2006). Because of their importance for biological research,

₇₂ chronograms are still constantly being built by experts in the field, and have been published

₇₃ in peer-reviewed publications in a steady increase for the last two decades, amounting to at

₇₄ least four thousand individual chronograms (Kumar, Stecher, Suleski, & Hedges, 2017).

₇₅   This state-of-the-art scientific data on time of evolutionary origin that has already been

₇₆ produced could be leveraged for the benefit of biological research, and there has been an urge

₇₇ for promoting, normlizing, and facilitating its reuse in research and education (Stoltzfus et

₇₈ al., 2013; Campbell O. Webb & Donoghue, 2005). We identify that a tool for efficient reuse

⁷⁹ of state-of-the-art scientific data should have an open and fully public database storing data

⁸⁰ in a computer readable format (R. A. Vos et al., 2012), an automatised and programatic way

⁸¹ of accessing the data (Stoltzfus et al., 2013), and straightforward means of comparing and

⁸² summarizing data as needed by the user []. The TreeBASE project started as an open and

⁸³ public database for state-of-the-art chronograms and other evolutionary research data in

⁸⁴ computer readable formats, and it currently supports programatic data accession R. Vos

⁸⁵ (n.d.), unfortunately, it is no longer under active development. The Open Tree of Life project

⁸⁶ (OpenTree, OpenTreeOfLife et al., 2019) has a system for storing chronograms and other

⁸⁷ evolutionary information in an open and public database that is programatically accessible,

⁸⁸ but it does not yet support phylogenetic queries of age data nor chronogram summaries.

⁸⁹   The DateLife project was born as a prototype service aiming to provide tools for easy

⁹⁰ reuse and summary of state-of-the-art time of lineage evolutionary origin, and was developed

⁹¹ over a series of hackathons at the National Evolutionary Synthesis Center, NC, USA

⁹² (Stoltzfus et al., 2013). Here we present the full implementation of the DateLife services,

⁹³ available as an R package `datelife` and an application with a graphical user interface web

⁹⁴ site at www.datelife.org/query/. The current implementation of the `datelife` R package

⁹⁵ features an algorithm for automatic curation and maintenance of an open database of

⁹⁶ chronograms pulled from OpenTree's open repository (McTavish et al., 2015), and methods

⁹⁷ to summarize and compare source chronograms

Description of the R package

⁹⁹   The general `datelife` workflow is shown in figure 1:

¹⁰⁰ 1. It starts with an input consisting of at least two taxon names, which can be provided

¹⁰¹   in two different forms: as a comma separated character string, or as tip labels on a tree.

¹⁰²   If input is a tree, it can be provided as a classic newick character string

¹⁰³   [@archie1986newick], or as a "phylo" R object [@paradis2004]. The input tree is not

104   required to have branch lengths.

105   2. Input taxon names are processed with the Taxonomic Name Resolution Service [TNRS,

106   @Boyle2013] implemented with OpenTree services [@opentreeAPIs]. TNRS detects,

107   corrects and standardizes misspellings and typos, variant spellings and authorities, and

108   nomenclatural synonyms to standardized taxonomic names. This increases the

109   probability of correctly finding the input taxon names in the chronogram database.

110   3. The current version of 'datelife' only accepts scientific taxonomic names as input.

111   Names can belong to any taxonomic group or binomial specific. If an input taxon

112   name belongs to an "inclusive" taxonomic group, i.e., a taxon above the species level,

113   such as genus, family, etc.), 'datelife' has two alternative behaviors defined by the "get

114   species from taxon" flag. If the flag is active, 'datelife' retrieves all species names

115   within the "inclusive" taxonomic group and adds them to the input. If the flag is

116   inactive, 'datelife' will drop the "inclusive" taxon names from input.

117   4. The cleaned input taxon names are saved as a special R object (of a newly defined

118   class `datelifeQuery`) that contains the processed names, the corresponding taxonomic

119   id numbers, and the topology of theinput tree if any was provided. The

120   `datelifeQuery` object is used next to search the chronogram database.

121   5. Chronograms with at least two matching input taxon names are identified and pruned

122   down to preserve only input taxon names as tips. Then, each pruned chronogram is

123   transformed to a patristic distance matrix. This format facilitates and greatly speeds

124   up all downstream analyses and summaries. The matrices are associated to the

125   citation of the original study and stored as an R object of class `datelifeResult`.

126   6. At this point, various summary data can be obtained to inform decisions for the next

127   steps of the analysis workflow. Types of summary information provided are: a) all

128   pruned source chronograms, b) age of the MRCA (most recent common ancestor) of

129 the pruned source chronograms, c) citations of studies where pruned source

130 chronograms were originally published, d) a summary table with all of the above, e) a

131 single summary chronogram of all or a subset of pruned source chronograms, f) a report

132 of successful matches of input taxon names across pruned source chronograms, and g)

133 the single pruned source chronogram with the most matching input taxon names.

7. To construct summary trees we use a fixed topology, either provided by the user, taken

from the literature or obtained from expert phylogenetic information, such as the

OpenTree synthetic tree. 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, @Criscuolo2006] and

any summary statistics such as median, minimum andmaximum ages. Finally, 'datelife'

applies the summarized time distances as node calibrations to date the given topology

with BLADJ [@webb2005phylomatic]. <!–ADD FIGURE: mock example explaining

difference between using summary ages and raw ages to calibrate a topology. Then

show it in the biological example. This will show what happens if a tree (phylogenetic

conflict) or node (age) do not agree.–>

<!–

8. Alternatively, time of lineage divergence obtained from the pruned chronograms can be

used directly as secondary calibration points to date a tree with or without branch

lengths containing some or all input taxon names.

9. 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.

10. 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

154   functions.

Benchmark

156   `datelife`'s code speed was tested on an Apple iMac with one 3.4 GHz Intel Core i5

157   processor. We registered variation in computing time of query processing and search through

158   the database relative to number of queried taxon names. Query processing time increases

159   roughly linearly with number of input taxon names, and increases considerably if TNRS

160   (Boyle et al., 2013) is activated. Up to ten thousand names can be processed and searched in

161   less than 30 minutes with the most time consuming settings. Once names have been

162   processed as described in methods, a name search through the chronogram database can be

163   performed in less than a minute, even with a very large number of taxon names (Fig. 2).

164   `datelife`'s code performance was evaluated with a set of unit tests designed and

165   implemented with the R package testthat (R Core Team, 2018) that were run both locally

166   with the devtools package (R Core Team, 2018), and on a public server –via GitHub, using

167   the continuous integration tool Travis CI (https://travis-ci.org). At present, unit tests cover

168   more than 30% of `datelife`'s code (https://codecov.io/gh/phylotastic/datelife).

Results

170   We illustrate the `datelife` workflow using the family of true finches, Fringillidae as an

171   example.

Case study

173   A college educator wishes to obtain state-of-the-art data on time of evolutionary origin

174   of species belonging to the true finches for their class. They decide to use `datelife` because

175   they are teaching best practices for reproducibility. Students have the option to go to the

176   website at www.datelife.org and perform an interactive run. However, the educator wants

177   the students to practice their R skills. The first step is to run a `datelife` query using the

178   "get species from taxon" flag. This will get all recognised species names within their chosen

179 inclusive taxon. The Fringillidae has 289 species, according to the Open Tree of Life

180 taxonomy. Once with a curated set of species taxon names, the next step is to run a

181 `datelife` search that will find all chronograms that contain at least two species names. The

182 algorithm proceeds to prune the trees to keep matching species names on tips only, and

183 transform the pruned trees to pairwise distance matrices. There are 13 chronograms

184 containing at least two Fringillidae species, published in 9 different studies (Fig. 3). The

185 final step is to summarize the available information using two alternative types of summary

186 chronograms, median and SDM. As explained in the "Description" section, data from source

187 chronograms is first summarised into a single distance matrix and then the available node

188 ages are used as fixed node calibrations over a consensus tree topology, to obtain a fully

189 dated tree with the program BLADJ (Fig. 4). Median summary chronograms are older and

190 have wider variation in maximum ages than chronograms obtained with SDM. With both

191 methods, ages are generally consistent with source ages, but there are some biological

192 examples in which this is not true (see Discussion).

Cross-validation test

194 Data from source chronograms can be also used to date tree topologies with no branch

195 lengths, as well as trees with branch lengths as relative substitution rates (Figs. 5 and 6). As

196 a form of cross validation, we took tree topologies from each study and calibrated them using

197 time of lineage divergence data from all other source chronograms. In the absence of branch

198 lengths, the ages of internal nodes were recovered with a high precision in almost all cases

199 (except for studies 3, and 5; Fig. 5). Maximum tree ages were only recovered in one case

200 (study 2; Fig. 5). We also demonstrate the usage of PATHd8 (Britton, Anderson, Jacquet,

201 Lundqvist, & Bremer, 2007) as an alternative method to BLADJ. For this, we run a

202 `datelife` branch length reconstruction that searches for DNA sequence data from the

203 Barcode of Life Data System [BOLD; ratnasingham2007bold] to generate branch lengths.

204 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 accesible 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. The `datelife` workflow builds off of

functions from several R packages that must be acknowledge (rotl (Michonneau, Brown, &

Winter, 2016), ape (Paradis, Claude, & Strimmer, 2004), geiger (Harmon, Weir, Brock, Glor,

& Challenger, 2008), paleotree (Bapst, 2012), bold (Chamberlain et al., 2019), phytools

(Revell, 2012), taxize (Chamberlain & Szöcs, 2013; Chamberlain et al., 2019), phyloch (Heibl,

2008), and phylocomr (Ooms & Chamberlain, 2018)).

At the time of writing of this manuscript (Mar 03, 2022), `datelife`'s database has 253

chronograms, pulled entirely from OpenTree's database, th ePhylesystem (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 Phylesystem, 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 Phylesystem, at a minimum of once a month and a

231 maximum of every 6 months. Users can also upload new chronograms to OpenTree

232 themselves, trigger and update of the `datelife` database to incorporate the new

233 chronograms, and and have them immediately available for analysis.

234 Incorporation of more chronograms into `datelife`'s database is crucial to improve its

235 services. One option to increase chronogram number in the database is the Dryad data

236 repository. Methods to automatically mine chronograms from Dryad could be designed and

237 implemented. However, Dryad's metadata system has no information to automatically detect

238 branch length units, and those would still need to be determined manually by a curator.

239 The only summary chronogram encompassing all life that is currently openly available

240 from OpenTree was constructed using age data from 2,274 published chronograms (Hedges,

241 Marin, Suleski, Paymer, & Kumar, 2015). However source chronograms are not available in

242 computer readable format for reuse or reanalysis. As this tree is part of datelife's database,

243 the amount of lineages that can be queried using `datelife` (99474 unique terminal taxa) is

244 substantial, yet it can be improved. Consequently, we would like to emphasize on the

245 importance of sharing chronogram data for the benefit of the scientific community as a

246 whole, into repositories that require expert input and manual curation, such as OpenTree's

247 Phylesystem (McTavish et al., 2015).

248 By default, `datelife` currently summarizes all source chronograms that overlap with

249 at least two species names. Users can exclude source chronograms if they have reasons to do

250 so. Strictly speaking, the best chronogram should reflect the real time of lineage divergence

251 accurately and precisely. To our knowledge, there are no good measures to determine if a

252 chronogram is better than another. Some measures that have been proposed are the

253 proportion of lineage sampling and the number of calibrations used Magallón,

254 Gómez-Acevedo, Sánchez-Reyes, & Hernández-Hernández (2015). Scientists usually also

255 favor chronograms constructed using primary calibrations (ages obtained from the fossil or

256 geological record) to ones constructed with secondary calibrations (ages coming from other

257 chronograms)(Schenk, 2016). It has been observed with simulations that divergence times
258 inferred with secondary calibrations are significantly younger than those inferred with
259 primary calibrations in analyses performed with bayesian inference methods when priors are
260 implemented in similar ways in both analyses (Schenk, 2016). However, secondary
261 calibrations can be applied using other dating methods that do not require setting priors,
262 such as penalized likelihood (Sanderson, 2003), or as fixed ages, potentially mitigating the
263 bias reported with bayesian methods. Certainly, further studies are required to fully
264 understand the effect of using secondary calibrations on time estimates and downstream
265 anlyses.

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

277   Summarizing chronograms might also imply summarizing fundamentally distinct
278 evolutionary hypotheses. For example, two different researchers working on the same clade
279 both carefully select and argument their choices of fossil calibrations. Still, if one researcher
280 decides a fossil will calibrate the ingroup of a clade, while another researcher uses the same
281 one to calibrate outside the clade, the resulting age estimates will often differ substantially,
282 as the placement of calibrations as stem or crown group is proved to deeply affect estimated

283 times of lineage divergence (Sauquet, 2013). Trying to summarize the resulting chronograms

284 into a single one using simple summary statistics might erase all types of relevant

285 information from the source chronograms. Accordingly, the prevailing view in our research

286 community is that we should favor time of lineage divergence estimates obtained from a

287 single analysis, using fossil data as primary sources of calibrations, and using fossils that

288 have been widely discussed and curated as calibrations to date other trees, making sure that

289 all data used in the analysis reflect a coherent evolutionary history (Antonelli et al., 2017).

290 However, the exercise of summarizing different chronograms has the potential to help getting

291 a single global evolutionary history for a lineage by putting together evidence from different

292 hypothesis. Choosing the elements of the chronograms that we are going to keep and the

293 ones that we are going to discard is key, since we are potentially loosing important parts of

294 the evolutionary history of a lineage that might only be reflected in source chronograms and

295 not on the summary chronogram (Sauquet et al., 2021).

296   Alternatively, one could try to choose the "best" chronogram from a set of possible

297 evolutionary hypotheses. Several characteristics of the data used for dating analyses as well

298 as from the output chronogram itself, could be used to score quality of source chronograms.

299 Some characteristics that are often cited in published studies as a measure of improved age

300 estimates as compared to previously published estimates are: quality of alignment (missing

301 data, GC content), lineage sampling (strategy and proportion), phylogenetic and dating

302 inference method, number of fossils used as calibrations, support for nodes and ages, and

303 magnitude of confidence intervals. To facilitate subsetting of source chronograms following

304 different criteria by the users, this information should be included as metadata manually

305 entered by curators in the future.

306   In other areas of biological research, such as ecology and conservation biology, it has

307 been shown that at least some data on lineage divergence represents a relevant improvement

308 for testing alternative hypothesis using phylogenetic distance (Campbell O. Webb, Ackerly,

309 & Kembel, 2008). Hence, we integrated into datelife's workflow different ways of creating

310 branch lengths in the absence of starting branch length information for taxa lacking this

311 information (using the BLADJ option). "Making up" branch lengths is an accepted practice

312 in scientific publications: Jetz, Thomas, Joy, Hartmann, and Mooers (2012), created a

313 time-calibrated tree of all 9,993 bird species, where 67% had molecular data and the rest was

314 simulated; Rabosky et al. (2018) created a time-calibrated tree of 31,536 ray-finned fishes, of

315 which only 37% had molecular data; Smith and Brown (2018) constructed a tree of 353,185

316 seed plants where only 23% had molecular data. Taken to the extreme, one could make a

317 fully resolved, calibrated tree of all modern and extinct taxa using a single taxonomy and a

318 single calibration with the polytomy resolution and branch assignation methods. There has

319 yet to be a thorough analysis of what can go wrong when one goes beyond the data in this

320 way, so we urge caution; we also urge readers to follow the example of many of the large tree

321 papers cited above and make sure results are substantially similar between trees fully

322 reconstructed with molecular or other data, and trees that are reconstructed using taxonomy

323 by resolving polytomies at random following a statistical model.

Conclusions

325 Divergence time information is key to many areas of evolutionary studies: trait

326 evolution, diversification, biogeography, macroecology and more. It is also crucial for science

327 communication and education, but generating chronograms *de novo* is difficult, especially for

328 those who want to use phylogenies but who are not systematists, or do not have the time to

329 acquire and develop the necessary knowledge and data curation skills. Moreover, years of

330 primarily public funded research have resulted in vast amounts of chronograms that are

331 already available on scientific publications, but hidden to the public and scientific community

332 for reuse.

333 `datelife` allows easy and fast summarization of publicly available information on time

334 of lineage divergence. This provides a straightforward way to get an informed idea on the

³³⁵ state of knowledge of the time frame of evolution of different regions of the tree of life, and

³³⁶ allows identification of regions that require more research or that have conflicting

³³⁷ information. Both summary and newly generated trees are useful to evaluate evolutionary

³³⁸ hypotheses in different areas of research. `datelife` helps with awareness of the existing

³³⁹ variation in expert time of divergence data, and will foster exploration of the effect of

³⁴⁰ alternative divergence time hypothesis on the results of analyses, nurturing a culture of more

³⁴¹ cautious interpretation of evolutionary results.

Availability

³⁴³ `datelife` is free and open source and it can be used through its current website

³⁴⁴ http://www.datelife.org/query/, through its R package, and through Phylotastic's project

³⁴⁵ web portal http://phylo.cs.nmsu.edu:3000/. `datelife`'s website is maintained using

³⁴⁶ RStudio's shiny server and the shiny package open infrastructure, as well as Docker.

³⁴⁷ `datelife`'s R package stable version is available for installation from the CRAN repository

³⁴⁸ (https://cran.r-project.org/package=datelife) using the command `install.packages(pkgs`

³⁴⁹ `= "datelife")` from within R. Development versions are available from the GitHub

³⁵⁰ repository (https://github.com/phylotastic/datelife) and can be installed using the

³⁵¹ command `devtools::install_github("phylotastic/datelife")`.

Supplementary Material

³⁵³ Code used to generate all versions of this manuscript, the biological examples, as well

³⁵⁴ as the benchmark of functionalities are available at datelifeMS1, datelife_examples, and

³⁵⁵ datelife_benchmark repositories in LLSR's GitHub account.

Funding

Acknowledgements

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. 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.

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. (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., Reyes, L. L. S., Cranston, K. A., Allman, J., Holder,

M. T., & McTavish, E. J. (2019). Open tree of life synthetic tree (Version 12.3).

Zenodo. https://doi.org/10.5281/zenodo.3937742

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and

evolution in R language. *Bioinformatics*, *20*(2), 289–290.

Piel, W. H., Donoghue, M., & Sanderson, M. (2002). TreeBASE : A database of

451 phylogenetic information. In J. Shimura, K. Wilson, & D. Gordon (Eds.), *To the*

452 *interoperable "catalog of life" with partners* (pp. 41–47). Tsukuba, Japan:

453 National Institute for Environmental Studies. Retrieved from

454 http://phylogeny.harvard.edu/treebase.

455 Piel, W., Chan, L., Dominus, M., Ruan, J., Vos, R., & Tannen, V. (2009). Treebase v.

456 2: A database of phylogenetic knowledge. E-biosphere. London.

457 Posadas, P., Crisci, J. V., & Katinas, L. (2006). Historical biogeography: A review of

458 its basic concepts and critical issues. *Journal of Arid Environments*, *66*(3),

459 389–403.

460 Price, T. D., Hooper, D. M., Buchanan, C. D., Johansson, U. S., Tietze, D. T.,

461 Alström, P., . . . others. (2014). Niche filling slows the diversification of himalayan

462 songbirds. *Nature*, *509*(7499), 222.

463 R Core Team. (2018). *R: a language and environment for statistical computing.*

464 Vienna, Austria: R Foundation for Statistical Computing.

465 Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., . . .

466 others. (2018). An inverse latitudinal gradient in speciation rate for marine fishes.

467 *Nature*, *559*(7714), 392.

468 Ramshaw, J., Richardson, D., Meatyard, B., Brown, R., Richardson, M., Thompson,

469 E., & Boulter, D. (1972). The time of origin of the flowering plants determined by

470 using amino acid sequence data of cytochrome c. *New Phytologist*, *71*(5), 773–779.

471 Revell, L. J. (2012). Phytools: An r package for phylogenetic comparative biology

472 (and other things). *Methods in Ecology and Evolution*, *3*, 217–223.

473 Sanderson, M. J. (2003). r8s: Inferring absolute rates of molecular evolution and

474 divergence times in the absence of a molecular clock. *Bioinformatics*, *19*(2),

475 301–302.

476 Sanderson, M. J., & Doyle, J. A. (2001). Sources of error and confidence intervals in

477 estimating the age of angiosperms from rbcL and 18S rDNA data. *American*

478 *Journal of Botany, 88*(8), 1499–1516.

479 Sauquet, H. (2013). A practical guide to molecular dating. *Comptes Rendus Palevol,*
480 *12*(6), 355–367.

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

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

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

487 Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C. M., . . .
488 Jordan, G. (2013). Phylotastic! Making tree-of-life knowledge accessible, reusable
489 and convenient. *BMC Bioinformatics, 14*.
490 https://doi.org/10.1186/1471-2105-14-158

491 Title, P. O., & Rabosky, D. L. (2016). Do Macrophylogenies Yield Stable
492 Macroevolutionary Inferences? An Example from Squamate Reptiles. *Systematic*
493 *Biology*, syw102. https://doi.org/10.1093/sysbio/syw102

494 Vos, R. (n.d.). TreeBASE APIs. Retrieved from
495 %7Bhttps://github.com/TreeBASE/treebase/wiki/API%7D

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

499 Webb, Campbell O. (2000). Exploring the Phylogenetic Structure of Ecological
500 Communities : An Example for Rain Forest Trees. *The American Naturalist,*
501 *156*(2), 145–155.

502 Webb, Campbell O., Ackerly, D. D., & Kembel, S. W. (2008). Phylocom: Software for
503 the analysis of phylogenetic community structure and trait evolution.
504 *Bioinformatics, 24*(18), 2098–2100.

505   https://doi.org/10.1093/bioinformatics/btn358

506   Webb, Campbell O., & Donoghue, M. J. (2005). Phylomatic: Tree assembly for

507   applied phylogenetics. *Molecular Ecology Notes*, *5*(1), 181–183.
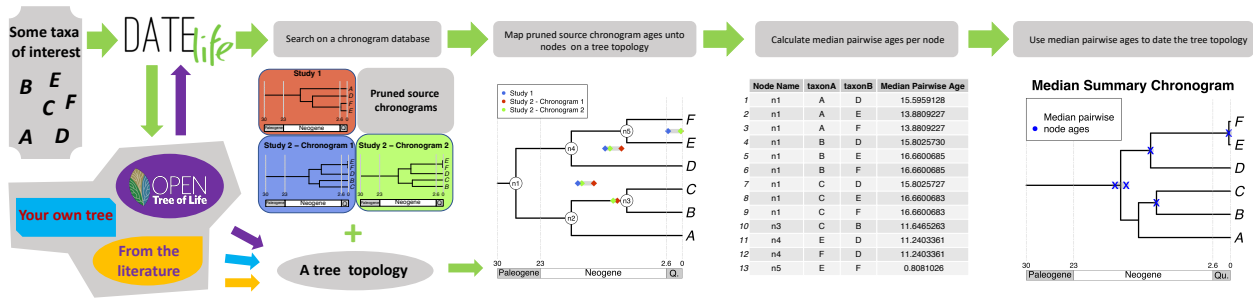
508 FIGURES

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.
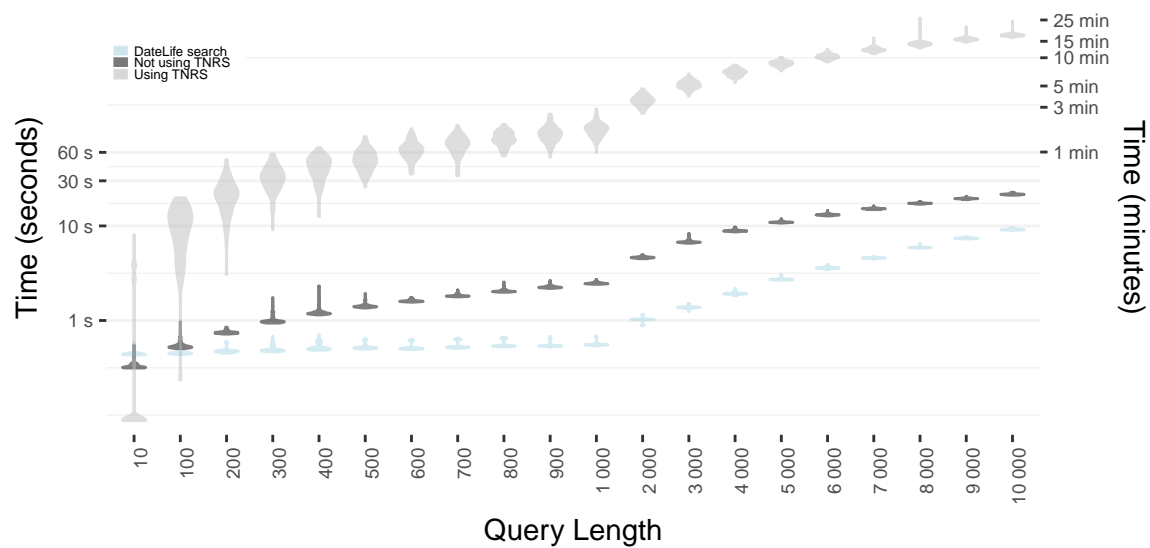
FIGURE 2. Input taxon name processing and chronogram database search computation time increases with number of input taxon names. We sampled N bird species names for each input size class, 100 times, and then performed a `datelife` search using the Taxon Names Resoultion Service (TNRS; dark gray), and without using TNRS (light gray). We also performed a search using the already processed query for comparison (light blue).
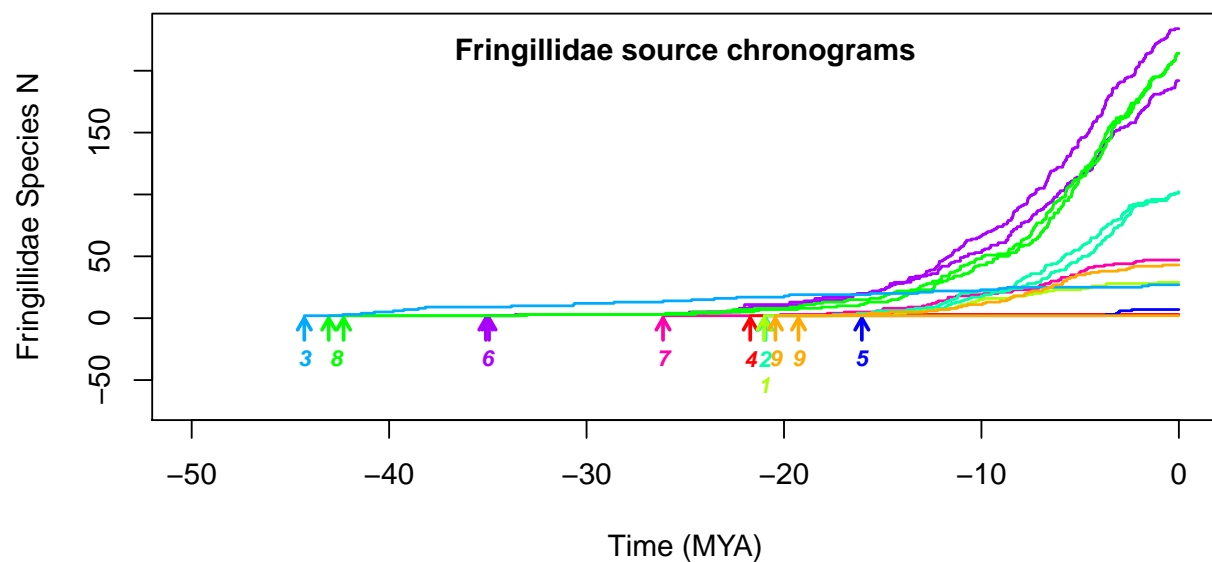
FIGURE 3. Lineage through time (LTT) plots of source chronograms containing all or a subset of species from the bird family Fringillidae of true finches. Arrows indicate maximum age of each chronogram. Numbers reference to chronograms' original publications 1: Barker et al. (2012), 2: Barker et al. (2015), 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).
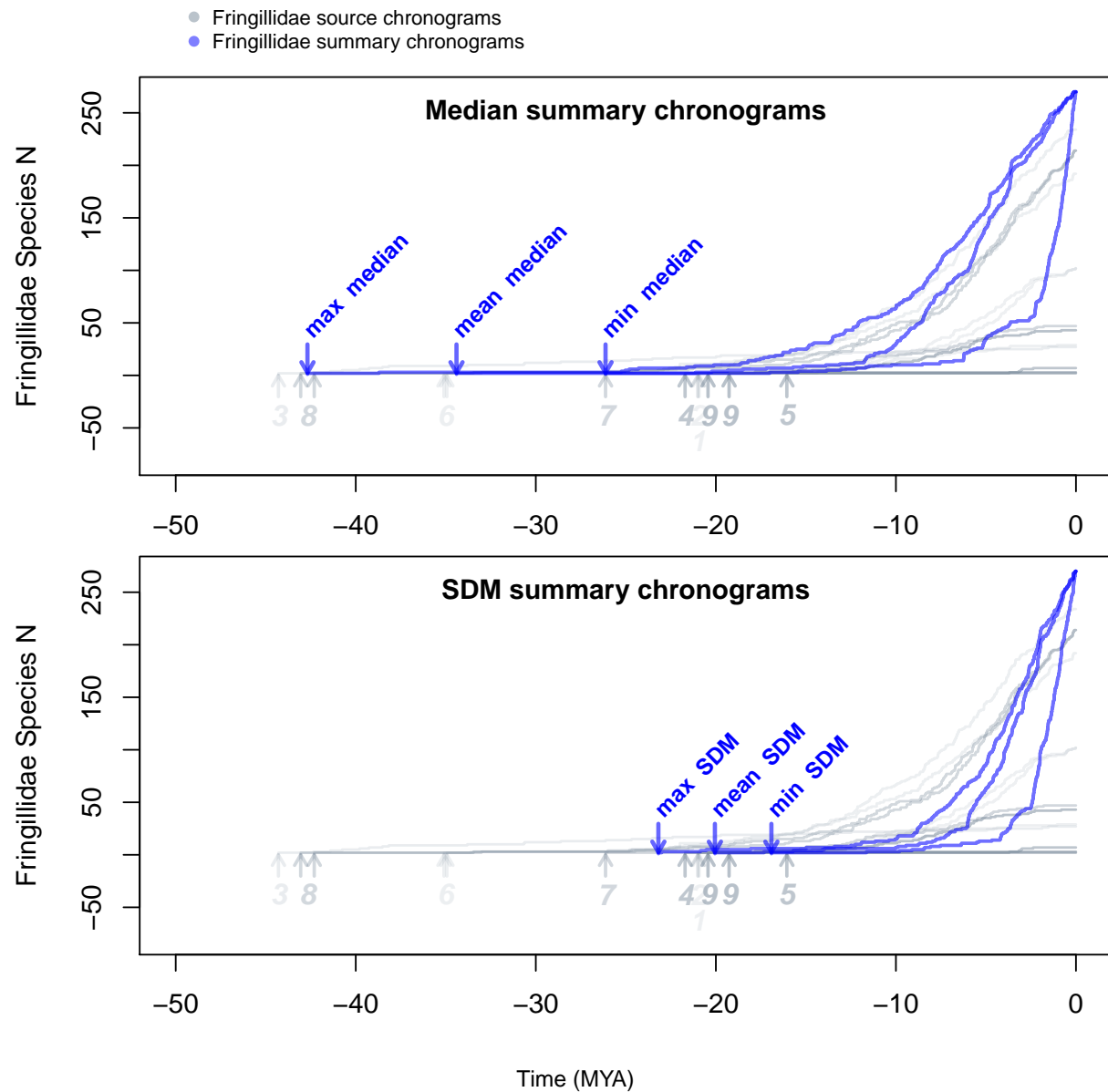
FIGURE 4. LTT plots of median (top) and Supermatrix Distance Method (SDM; bottom) chronograms summarising information from source chronograms found for the Fringillidae. Arrows indicate tree maximum age.
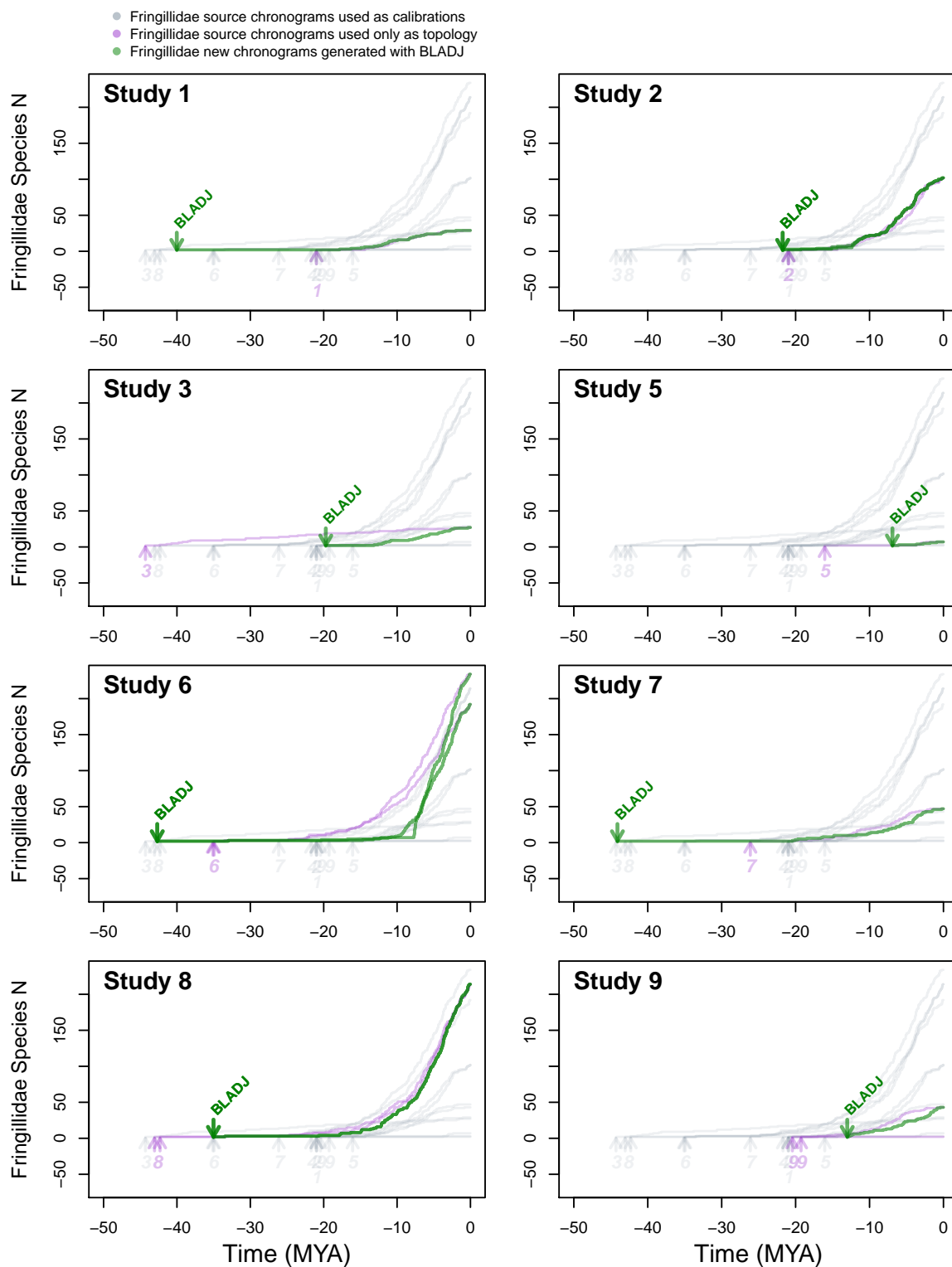
FIGURE 5. LTT plots showing results from the cross-validation analyses of trees without branch lengths dated using BLADJ. The dating analysis can only be performed in trees with more than 2 tips, thus excluding chronogram from study 4; its data was still used as calibration for the other source chronograms.
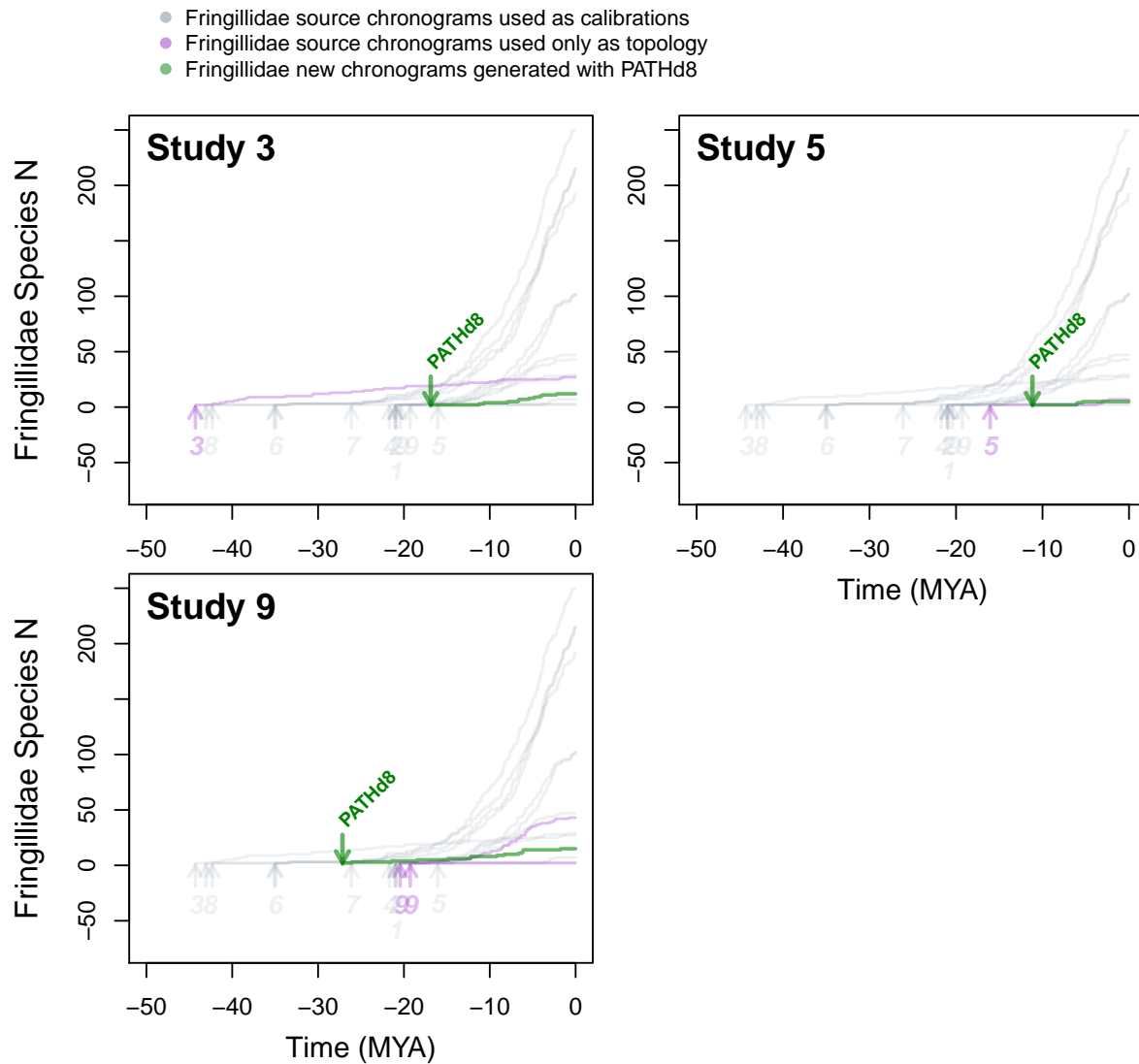
FIGURE 6. LTT plots showing results from the cross-validation analyses of trees with branch length reconstructed with data from the Barcode of Life Database (BOLD) dated using PATHd8. We could construct a tree with branch lengths for all source chronograms. However, dating with PATHd8 was only successful in three source chronograms shown here.