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.

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

16 E-mail: sanchez.reyes.luna@gmail.com

17   Abstract

18   Time of evolutionary origin is fundamental for research in the natural sciences, as well as for

19   education, science communication and policy. Despite an increased availability of fossil and

20   molecular data, and time-efficient analytical techniques, achieving a high-quality

21   reconstruction of time of evolutionary origin as a phylogenetic tree with branch lengths

22   proportional to absolute time (chronogram), is still a difficult and time-consuming task for a

23   majority of interested parties. Yet, the amount of published chronograms has increased

24   significantly in the past two decades, and a non-negligeable proportion of these data have

25   been steadily accumulating in public, open databases such as TreeBASE and Open Tree of

26   Life, exposing a wealth of expertly-curated and peer-reviewed data on time of evolutionary

27   origin in a programatic and reusable way, for a large quantity and diversity of organisms.

28   This trend results from intensive and localized efforts for improving data sharing practices,

29   as well as incentivizing open science in biology. Despite these trends, accessibility to

30   state-of-the-art knowledge on time of evolutionary origin is still reduced.

31   Here we present `datelife`, a service implemented as an R package and an Rshiny

32   website application available at www.datelife.org/query/, that provides functionalities for

33   efficient and easy finding, summary, reuse, and reanalysis of expert, peer-reviewed, public

34   data on time of evolutionary origin.

35   The main workflow of `datelife` is to construct a chronogram for any given

36   combination of taxon names, by searching a local chronogram database constructed and

37   curated from the Open Tree of Life (OpenTree), which incorporates phylogenetic data from

38   the TreeBASE database as well. We implement and test methods for summarizing time data

39   from multiple source chronograms using supertree and congruification algorithms.

40   Additionally, time data extracted from source chronograms can be usedas secondary

41   calibration points to add branch lengths proportional to absolute time to a tree topology

42   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: 3670

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

Introduction

In the natural sciences, absolute time of evolutionary origin of lineages is represented graphically as a phylogeny with branch lengths proportional to geologic time, also known as a chronogram. Notably, building a chronogram is not an easy task. First, it requires obtaining and curating genetic data to construct an homology hypothesis as a genetic alignment; then, it is important to carefully choose and apply the software that will use the alignment to reconstruct an evolutionary hypothesis in the form of a phylogeny; thirdly, calibrations, i.e., independent age data points from the fossil record or other suitable geologic events have to be obtained from the literature –or by going into the field; next, calibrations have to be carefully placed on the reconstructed phylogeny, demonstrating a biological and geological understanding of their limits and applicability; finally, the appropriate software and model of evolution to estimate ellapsed time since evolutionary events on the phylogeny has to be chosen and applied. In sum, building chronograms require specialized biological training and a non-negligible amount of human, time and monetary resources.

Because of their importance for biological research, chronograms are constantly being built by experts in the field, and have been published in a steady increase for the last two decades, amounting to around four thousand individual chronograms published in peer-reviewed journals (Kumar, Stecher, Suleski, & Hedges, 2017). Chronograms represent key knowledge for the study of natural processes in many areas of scientific research, from developmental to conservation biology (Felsenstein, 1985; Campbell O. Webb, 2000), from historical biogeography to species diversification (Morlon, 2014; Posadas, Crisci, & Katinas, 2006). As such, the state-of-the-art scientific data on time of evolutionary origin that has already been produced and curated could be leveraged with beneficial results, and there has been an urge for promoting, normalizing, and facilitating reuse of scientific data for research and education (Stoltzfus et al., 2013; Campbell O. Webb & Donoghue, 2005). We identify

79 that any tool for efficient reuse of state-of-the-art scientific data should have a versioned,

80 open and fully public database storing information in a computer readable format (Vos et al.,

81 2012), an automatised and programatic way of accessing the data (Stoltzfus et al., 2013),

82 and straightforward means of comparing and summarizing data as needed by the user [@].

83 The DateLife project was born as a prototype tool aiming to provide these services, and was

84 developed over a series of hackathons at the National Evolutionary Synthesis Center, NC,

85 USA (Stoltzfus et al., 2013). Here we present the full implementation of DateLife, available

86 as an R package `datelife` and as a website application with a graphical user interface at

87 www.datelife.org/query/. The current implementation of the `datelife` R package is

88 available on CRAN [@] and features an algorithm for automatic curation and maintenance of

89 an open database of chronograms pulled from OpenTree's public repository (McTavish et al.,

90 2015), and methods to summarize and compare source chronograms

Description of the R package

92 The general `datelife` workflow is shown in figure 1, briefly:

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

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

95   If input is a tree, it can be provided as a classic newick character string (Archie et al.,

96   1986), or as a "phylo" R object (Paradis, Claude, & Strimmer, 2004). The input tree is

97   not required to have branch lengths.

98 2. Input taxon names are processed to detect, correct and standardize name misspellings

99   and typos, variant spellings and authorities, and nomenclatural synonyms to a single

100   taxonomic standard. To do this, `datelife` implements the Taxonomic Name

101   Resolution Service [TNRS; Boyle et al. (2013)] provided by OpenTree

102   (OpenTreeOfLife et al., n.d.). TNRS increases the probability of correctly finding the

103   input taxon names in the chronogram database.

104 3. The current version of `datelife` only accepts scientific taxonomic names as input.

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

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

107 such as genus, family, etc.), `datelife` has two alternative behaviors defined by the

108 "get species from taxon" flag. If the flag is active, `datelife` retrieves all species names

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

110 inactive, `datelife` will drop the "inclusive" taxon names from input.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

130 OpenTree synthetic tree. Then we summarize source chronograms into a single patristic

131 distance matrix using a method chosen by the user. Summarizing method options

132   available include Super Distance Matrix method (SDM, Criscuolo, Berry, Douzery, &

133   Gascuel, 2006) and any summary statistics such as median, minimum andmaximum

134   ages. Finally, `datelife` applies the summarized time distances as node calibrations to

135   date the given topology with BLADJ (Campbell O. Webb & Donoghue, 2005).

136 8. If there is no information available for any input taxon name, users can also create

137   both age and phylogenetic data for the missing branches with a variety of algorithms.

138 9. Users can save all source and summary chronograms in formats that permit reuse and

139   reanalyses (newick and R "phylo" format), as well as view and compare results

140   graphically, or construct their own graphs using `datelife`'s graphic generation

141   functions.

142   The `datelife` workflow relies on functions from several R packages that we

143 acknowledge: rotl (Michonneau, Brown, & Winter, 2016), ape (Paradis et al., 2004), geiger

144 (Harmon, Weir, Brock, Glor, & Challenger, 2008), paleotree (Bapst, 2012), bold

145 (Chamberlain et al., 2019), phytools (Revell, 2012), taxize (Chamberlain & Szöcs, 2013;

146 Chamberlain et al., 2019), phyloch (Heibl, 2008), and phylocomr (Ooms & Chamberlain,

147 2018).

Benchmark

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

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

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

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

153 activated. Up to ten thousand names can be processed and searched in less than 30 minutes

154 with the most time consuming settings. Once names have been processed as described in

155 methods, a name search through the chronogram database can be performed in less than a

156 minute, even with a very large number of taxon names (Fig. 2). `datelife`'s code

157 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. 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

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

Cross-validation test

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

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

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

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

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

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

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

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

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

195 Barcode of Life Data System [BOLD; Ratnasingham and Hebert (2007)] to generate branch

196 lengths. We were able to successfully generate a tree with BOLD branch lengths for all of

197 the Fringillidae source chronograms. However, dating with PATHd8 using congruified

198 calibrations, was only successful in three cases (studies 3, 5, and 9, shown in Fig. 6). From

199 these, two trees have a different sampling than the original source chronogram, mainly

200 because DNA BOLD data for some species is absent from the database. Maximum ages are

201 quite different from source chronograms, but this might be explained also by the differences

202 in sampling between source chronograms and BOLD trees. More examples and code used to

203 generate these trees were developed on an open repository that is available for consultation

204 and reuse at https://github.com/LunaSare/datelife_examples.

Discussion

206   The main goal of `datelife` is to make state-of-the-art information on time of lineage

207 divergence easily accesible for comparison, reuse, and reanalysis, to researchers in all areas of

208 science and with all levels of expertise in the matter. It is an open service that does not

209 require any expert biological knowledge from users –besides the names of the organisms they

210 want to work with, for any of its functionalities.

211   At the time of writing of this manuscript (Mar 04, 2022), `datelife`'s database has 253

212 chronograms, pulled entirely from OpenTree's database, the Phylesystem (McTavish et al.,

213 2015). A unique feature of OpenTree's Phylesystem is that the community can add new

214 state-of-the-art chronograms any time. As chronograms are added to Phylesystem, they are

215 incorporated into an updated `datelife`'s database that is assigned a new version number,

216 followed by a package release on CRAN [@]. `datelife`'s chronogram database is updated as

217 new chronogram data is added to Phylesystem, at a minimum of once a month and a

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

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

220 chronograms, to have them immediately available for analysis.

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

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

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

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

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

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

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

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

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

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

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

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

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

234 Phylesystem (McTavish et al., 2015).

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

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

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

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

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

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

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

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

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

244 chronograms)(Schenk, 2016). It has been observed with simulations that divergence times

245 inferred with secondary calibrations are significantly younger than those inferred with

246 primary calibrations in analyses performed with bayesian inference methods when priors are

247 implemented in similar ways in both analyses (Schenk, 2016). However, secondary

248 calibrations can be applied using other dating methods that do not require setting priors,

249 such as penalized likelihood (Sanderson, 2003), or as fixed ages, potentially mitigating the

250 bias reported with bayesian methods. Certainly, further studies are required to fully

251 understand the effect of using secondary calibrations on time estimates and downstream

252 anlyses.

253   Furthermore, even chronograms obtained with primary fossil data can vary

254 substantially in time estimates between lineages, as observed from the comparison of source

255 chronograms in the Fringillidae example. This observation is often encountered in the

256 literature (see, for example, the ongoing debate about crown group age of angiosperms

257 (Barba-Montoya, Reis, Schneider, Donoghue, & Yang, 2018; Magallón et al., 2015; Ramshaw

258 et al., 1972; Sanderson & Doyle, 2001; Sauquet, Ramírez-Barahona, & Magallón, 2021). For

259 some studies, especially ones based on branch lengths (e.g., studies of species diversification,

260 timing of evolutionary events, phenotypic trait evolution), using a different chronogram may

261 return different results (Title & Rabosky, 2016). Stitching together these chronograms can

262 create a larger tree that uses information from multiple studies, but the effect of

263 uncertainties and errors at this level on downstream analyses is still largely unknown.

264   Summarizing chronograms might also imply summarizing fundamentally distinct

265 evolutionary hypotheses. For example, two different researchers working on the same clade

266 both carefully select and argument their choices of fossil calibrations. Still, if one researcher

267 decides a fossil will calibrate the ingroup of a clade, while another researcher uses the same

268 one to calibrate outside the clade, the resulting age estimates will often differ substantially,

269 as the placement of calibrations as stem or crown group is proved to deeply affect estimated

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

292 entered by curators in the future.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

310 by resolving polytomies at random following a statistical model.

Conclusions

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

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

³¹⁴ communication and education, but generating chronograms is difficult, especially for those

³¹⁵ who want to use phylogenies but who are not systematists, or do not have the time to

³¹⁶ acquire and develop the necessary knowledge and data curation skills. Moreover, years of

³¹⁷ primarily public funded research have resulted in vast amounts of chronograms that are

³¹⁸ already available on scientific publications, but hidden to the public and scientific community

³¹⁹ for reuse.

³²⁰ The `datelife` R package allows easy and fast summarization of publicly available

³²¹ information on time of lineage divergence. This provides a straightforward way to get an

³²² informed idea on the state of knowledge of the time frame of evolution of different regions of

³²³ the tree of life, and allows identification of regions that require more research or that have

³²⁴ conflicting information. Both summary and newly generated trees are useful to evaluate

³²⁵ evolutionary hypotheses in different areas of research. The DateLife project helps with

³²⁶ awareness of the existing variation in expert time of divergence data, and will foster

³²⁷ exploration of the effect of alternative divergence time hypothesis on the results of analyses,

³²⁸ nurturing a culture of more cautious interpretation of evolutionary results.

Availability

³²⁹

³³⁰ `datelife` is free and open source and it can be used through its current website

³³¹ http://www.datelife.org/query/, through its R package, and through Phylotastic's project

³³² web portal http://phylo.cs.nmsu.edu:3000/. `datelife`'s website is maintained using

³³³ RStudio's shiny server and the shiny package open infrastructure, as well as Docker.

³³⁴ `datelife`'s R package stable version is available for installation from the CRAN repository

³³⁵ (https://cran.r-project.org/package=datelife) using the command `install.packages(pkgs`

³³⁶ `= "datelife")` from within R. Development versions are available from the GitHub

³³⁷ repository (https://github.com/phylotastic/datelife) and can be installed using the

³³⁸ command `devtools::install_github("phylotastic/datelife")`.

Supplementary Material

₃₃₉

₃₄₀ Code used to generate all versions of this manuscript, the biological examples, as well

₃₄₁ as the benchmark of functionalities are available at datelifeMS1, datelife_examples, and

₃₄₂ datelife_benchmark repositories in LLSR's GitHub account.

Funding

₃₄₃

Acknowledgements

₃₄₇

357   **References**

358   Antonelli, A., Hettling, H., Condamine, F. L., Vos, K., Nilsson, R. H., Sanderson, M.

359   J., . . . Vos, R. A. (2017). Toward a self-updating platform for estimating rates of

360   speciation and migration, ages, and relationships of Taxa. *Systematic Biology*,

361   *66*(2), 153–166. https://doi.org/10.1093/sysbio/syw066

362   Archie, J., Day, W. H., Felsenstein, J., Maddison, W., Meacham, C., Rohlf, F. J., &

363   Swofford, D. (1986). The Newick tree format. Retrieved from

364   %7Bhttps://evolution.genetics.washington.edu/phylip/newicktree.html%7D

365   Bapst, D. W. (2012). Paleotree: An R package for paleontological and phylogenetic

366   analyses of evolution. *Methods in Ecology and Evolution*, *3*(5), 803–807.

367   https://doi.org/10.1111/j.2041-210X.2012.00223.x

368   Barba-Montoya, J., Reis, M. dos, Schneider, H., Donoghue, P. C., & Yang, Z. (2018).

369   Constraining uncertainty in the timescale of angiosperm evolution and the

370   veracity of a cretaceous terrestrial revolution. *New Phytologist*, *218*(2), 819–834.

371   Barker, F. K., Burns, K. J., Klicka, J., Lanyon, S. M., & Lovette, I. J. (2012). Going

372   to extremes: Contrasting rates of diversification in a recent radiation of new world

373   passerine birds. *Systematic Biology*, *62*(2), 298–320.

374   Barker, F. K., Burns, K. J., Klicka, J., Lanyon, S. M., & Lovette, I. J. (2015). New

375   insights into new world biogeography: An integrated view from the phylogeny of

376   blackbirds, cardinals, sparrows, tanagers, warblers, and allies. *The Auk:*

377   *Ornithological Advances*, *132*(2), 333–348.

378   Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., . . .

379   Enquist, B. J. (2013). The taxonomic name resolution service: An online tool for

380   automated standardization of plant names. *BMC Bioinformatics*, *14*(1).

381   https://doi.org/10.1186/1471-2105-14-16

382   Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S., & Bremer, K. (2007).

383   Estimating Divergence Times in Large Phylogenetic Trees. *Systematic Biology*,

384   *56*(788777878), 741–752. https://doi.org/10.1080/10635150701613783

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

386   Lovette, I. J. (2014). Phylogenetics and diversification of tanagers (passeriformes:

387   Thraupidae), the largest radiation of neotropical songbirds. *Molecular*

388   *Phylogenetics and Evolution*, *75*, 41–77.

389   Chamberlain, S. A., & Szöcs, E. (2013). taxize : taxonomic search and retrieval in R

390   [version 2; referees: 3 approved]. *F1000Research*, *2*(191), 1–29.

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

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

393   Li, G. (2019). *taxize: Taxonomic information from around the web*. Retrieved

394   from https://github.com/ropensci/taxize

395   Claramunt, S., & Cracraft, J. (2015). A new time tree reveals earth history's imprint

396   on the evolution of modern birds. *Science Advances*, *1*(11), e1501005.

397   Criscuolo, A., Berry, V., Douzery, E. J. P., & Gascuel, O. (2006). SDM: A fast

398   distance-based approach for (super)tree building in phylogenomics. *Systematic*

399   *Biology*, *55*(5), 740–755. https://doi.org/10.1080/10635150600969872

400   Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American*

401   *Naturalist*, *125*(1), 1–15. Retrieved from http://www.jstor.org/stable/2461605

402   Gibb, G. C., England, R., Hartig, G., McLenachan, P. A., Taylor Smith, B. L.,

403   McComish, B. J., . . . Penny, D. (2015). New zealand passerines help clarify the

404   diversification of major songbird lineages during the oligocene. *Genome Biology*

405   *and Evolution*, *7*(11), 2983–2995.

406   Harmon, L., Weir, J., Brock, C., Glor, R., & Challenger, W. (2008). GEIGER:

407   investigating evolutionary radiations. *Bioinformatics*, *24*, 129–131.

408   Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life

409   reveals clock-like speciation and diversification. *Molecular Biology and Evolution*,

410   *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., Cranston, K. A., Allman, J., Holder, M. T., &

McTavish, E. J. (n.d.). Open Tree of Life APIs v. 3.0. Retrieved from

%7Bhttps://github.com/OpenTreeOfLife/germinator/wiki/Open-Tree-of-Life-

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). *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.

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 Macroevolutionary 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.

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

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

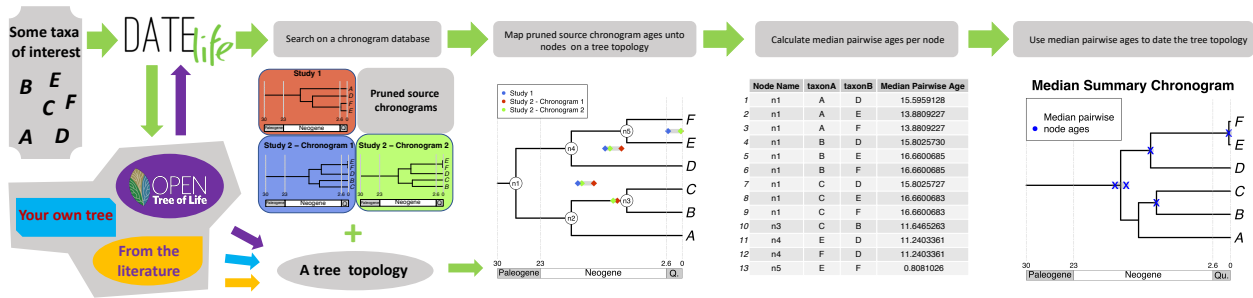494   applied phylogenetics. *Molecular Ecology Notes*, *5*(1), 181–183.

495 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.
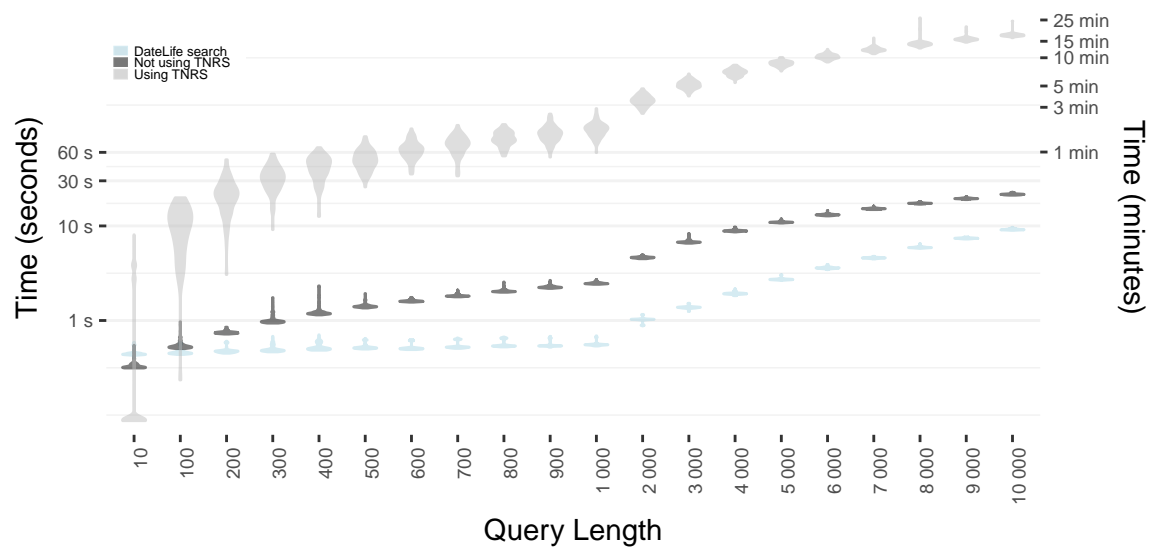
FIGURE 2. Input taxon name processing and chronogram database search computation time increases with number of input taxon names. We sampled N bird species names for each input size class, 100 times, and then performed a `datelife` search using the Taxon Names Resoultion Service (TNRS; dark gray), and without using TNRS (light gray). We also performed a search using the already processed query for comparison (light blue).
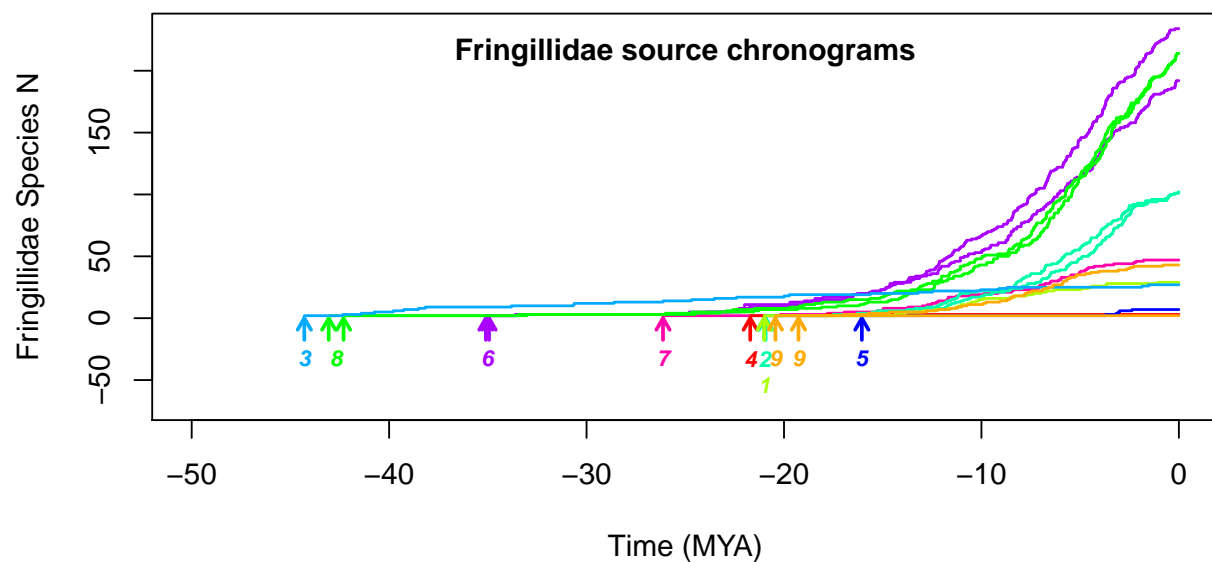
FIGURE 3. Lineage through time (LTT) plots of source chronograms containing all or a subset of species from the bird family Fringillidae of true finches. Arrows indicate maximum age of each chronogram. Numbers reference to chronograms' original publications 1: Barker et al. (2012), 2: Barker et al. (2015), 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).
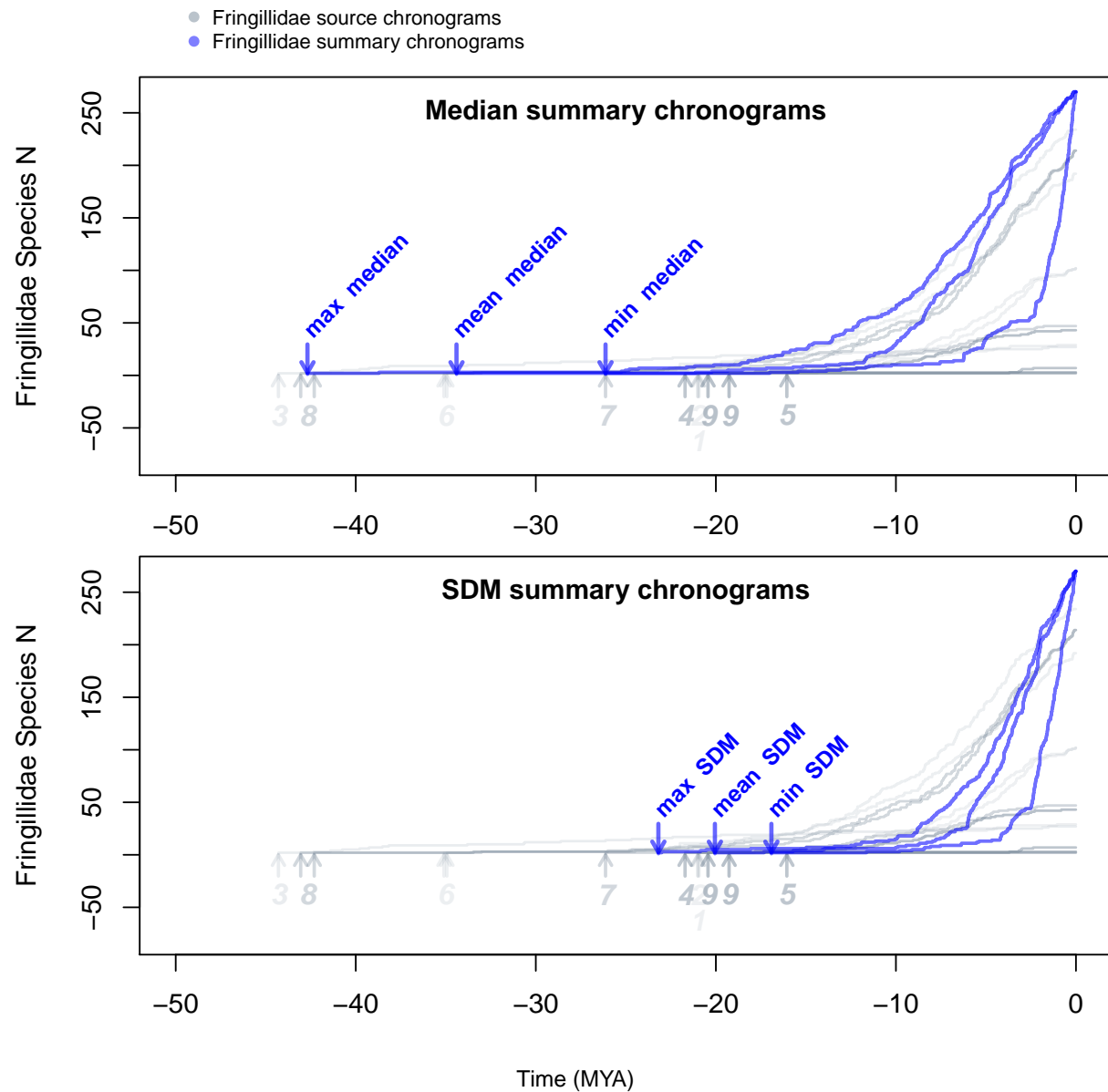
FIGURE 4. LTT plots of median (top) and Supermatrix Distance Method (SDM; bottom) chronograms summarising information from source chronograms found for the Fringillidae. Arrows indicate tree maximum age.
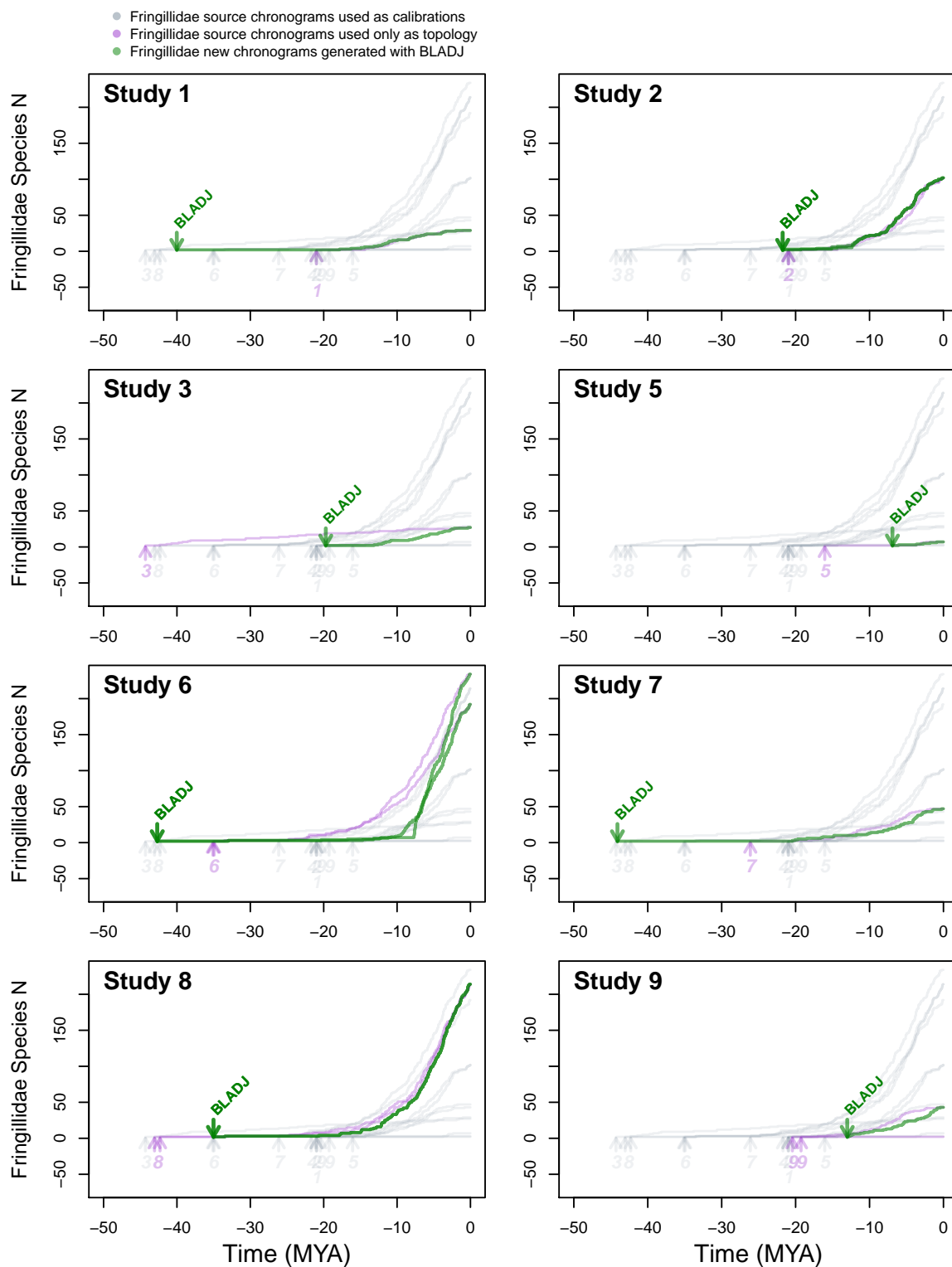
FIGURE 5. LTT plots showing results from the cross-validation analyses of trees without branch lengths dated using BLADJ. The dating analysis can only be performed in trees with more than 2 tips, thus excluding chronogram from study 4; its data was still used as calibration for the other source chronograms.
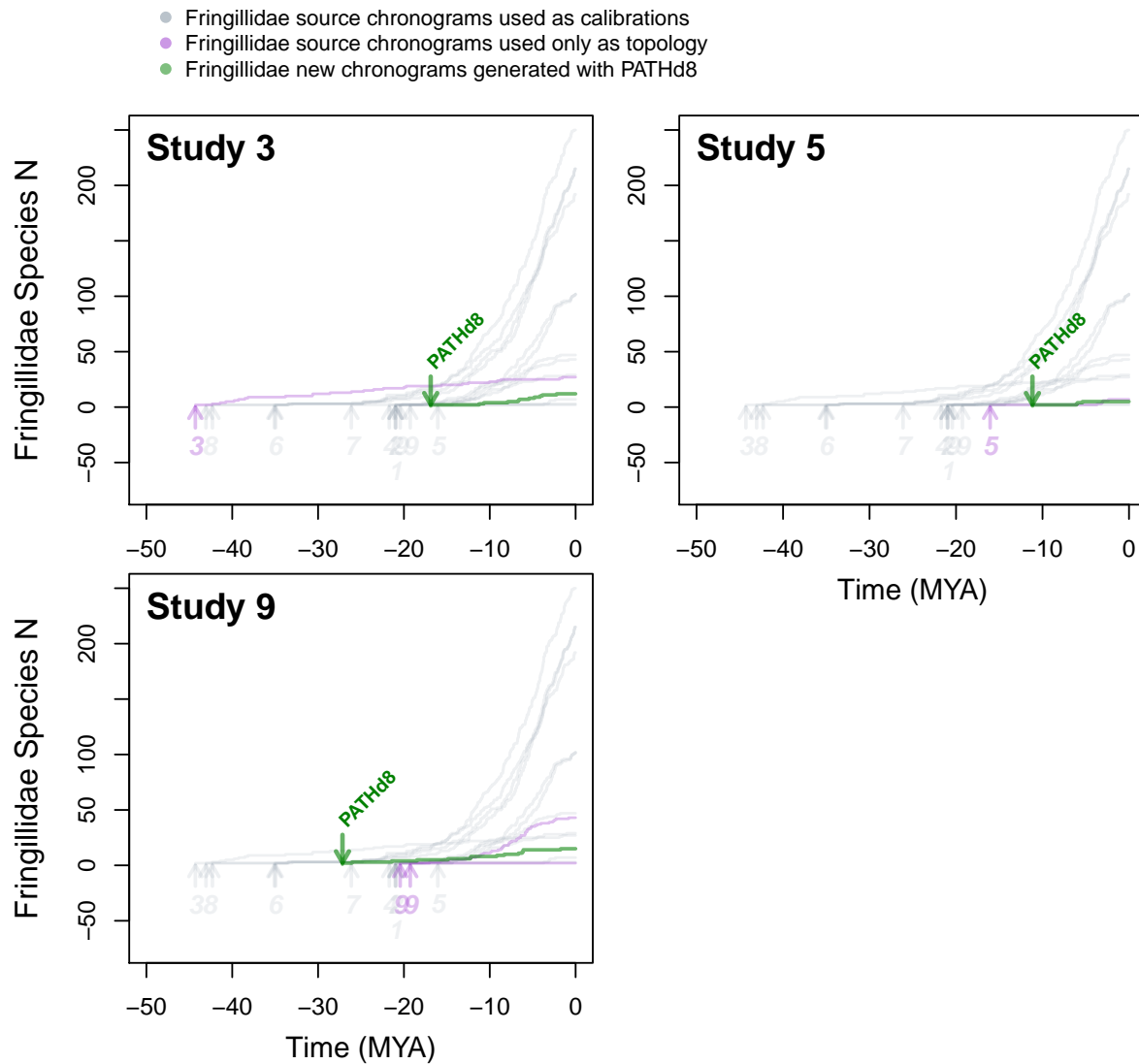
FIGURE 6. LTT plots showing results from the cross-validation analyses of trees with branch length reconstructed with data from the Barcode of Life Database (BOLD) dated using PATHd8. We could construct a tree with branch lengths for all source chronograms. However, dating with PATHd8 was only successful in three source chronograms shown here.