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

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

17   Abstract

18   Date estimates for times of evolutionary divergences are key data for research in the

19   natural sciences. These estimates also provide valuable information for for education,

20   science communication and policy decisions. Although achieving a high-quality

21   reconstruction of a phylogenetic tree with branch lengths proportional to absolute time

22   (chronogram), is a difficult and time-consuming task, the increased availability of fossil and

23   molecular data, and time-efficient analytical techniques has resulted in many recent

24   publications of large chronograms for a large number and wide diversity of organisms.

25   When these estimates are shared in public, open databases this wealth of expertly-curated

26   and peer-reviewed data on time of evolutionary origin is exposed in a programatic and

27   reusable way. Intensive and localized efforts have improved data sharing practices, as well

28   as incentivizited open science in biology. Here we present DateLife, a service implemented

29   as an R package and an Rshiny website application available at www.datelife.org/query/,

30   that provides functionalities for efficient and easy finding, summary, reuse, and reanalysis

31   of expert, peer-reviewed, public data on time of evolutionary origin. The main DateLife

32   workflow constructs a chronogram for any given combination of taxon names, by searching

33   a local chronogram database constructed and curated from the Open Tree of Life

34   Phylesystem phylogenetic database, which incorporates phylogenetic data from TreeBASE

35   database as well. We implement and test methods for summarizing time data from

36   multiple source chronograms using supertree and congruification algorithms, and using age

37   data extracted from source chronograms as secondary calibration points to add branch

38   lengths proportional to absolute time to a tree topology. DateLife will be useful to increase

39   awereness on the existing variation in expert time of divergence data, and can foster

40   exploration of the effect of alternative divergence time hypothesis on the results of

41   analyses, providing a framework for a more informed interpretation of evolutionary results.

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

43   Congruification; Supertree; Calibrations; Secondary calibrations

44   Word count: 3868

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

Introduction

47   Chronograms –phylogenies with branch lengths proportional to time– provide key

48 data for the study of natural processes in many areas of biological research, such as

49 developmental biology (Delsuc et al., 2018; Laubichler & Maienschein, 2009), conservation

50 biology (Felsenstein, 1985; Campbell O. Webb, 2000), historical biogeography (Posadas,

51 Crisci, & Katinas, 2006), and species diversification (Magallon & Sanderson, 2001; Morlon,

52 2014).

53   Building a chronogram is not an easy task. It requires obtaining and curating data to

54 construct a phylogeny; selecting and placing appropriate calibrations on the phylogeny

55 using independent age data points from the fossil record or other dated events, and

56 inferring the full dated tree. Estimating accurate chronograms generally requires

57 specialized biological training, taxonomic domain knowledge, and a non-negligible amount

58 of research time, computational resources and funding.

59   Here we present the DateLife software application, available as an R package and as

60 an online Rshiny interactive website at www.datelife.org/query/, which captures data from

61 published chronograms, and make these data readily accessible to users. DateLife features

62 a versioned, open and fully public chronogram database (McTavish et al., 2015) storing age

63 information in a computer readable format (Vos et al., 2012), an automated and

64 programmatic way of accessing the data (Stoltzfus et al., 2013) and methods to summarize

65 and compare age data.

Description

67   The DateLife algorithm is fully implemented using the R language. The latest stable

68 version of the R package `datelife` is available from the CRAN repository (v0.6.1;

⁶⁹ Sanchez-Reyes et al. (2022)), and relies on functionalities from various biological R

⁷⁰ packages: ape (Paradis, Claude, & Strimmer, 2004), bold (Chamberlain et al., 2019), geiger

⁷¹ (Harmon, Weir, Brock, Glor, & Challenger, 2008), paleotree (Bapst, 2012), phyloch (Heibl,

⁷² 2008), phylocomr (Ooms & Chamberlain, 2018), phytools (Revell, 2012), rotl (Michonneau,

⁷³ Brown, & Winter, 2016), and taxize (Chamberlain & Szöcs, 2013; Chamberlain et al.,

⁷⁴ 2019). Figure 1 provides a graphical summary of the three main steps of the DateLife

⁷⁵ algorithm: providing an input, searching a chronogram database, and summarizing results

⁷⁶ from the search.

Providing an input

⁷⁸ DateLife starts with an input query consisting of at least two taxon names, which can

⁷⁹ be provided as a comma separated character string, or as tip labels on a tree. If the input

⁸⁰ is a tree, it can be provided as a classic newick character string (Archie et al., 1986), or as

⁸¹ a "phylo" R object (Paradis et al., 2004). The input tree is not required to have branch

⁸² lengths, and its topology is used on the summary steps described below.

⁸³ Currently, DateLife only accepts scientific names as input, and they can belong to

⁸⁴ any inclusive taxonomic group (e.g., genus, family, tribe, etc.) or binomial specific.

⁸⁵ Subspecies and variants are ignored. If an input taxon name belongs to an inclusive

⁸⁶ taxonomic group the algorithm has two alternative behaviors defined by the "get species

⁸⁷ from taxon" flag. If the flag is active, the DateLife algorithm retrieves all species names

⁸⁸ within the inclusive taxonomic group and adds them to the input. If the flag is inactive,

⁸⁹ DateLife ignores the inclusive taxon names from the input.

⁹⁰ Input scientific names are processed using the Taxonomic Name Resolution Service

⁹¹ (TNRS; Boyle et al. (2013)), which increases the probability of correctly finding the

⁹² queried taxon names in the chronogram database. TNRS detects, corrects and standardizes

⁹³ name misspellings and typos, variant spellings and authorities, and nomenclatural

94 synonyms to a single taxonomic standard. DateLife implements TNRS using OpenTree's

95 taxonomic as standard (Open Tree Of Life et al., 2016; Rees & Cranston, 2017).

96   The processed input taxon names are saved as an R object of a newly defined class

97 `datelifeQuery` that is used in the following steps. It contains the processed names, the

98 corresponding OpenTree taxonomic id numbers, and the topology of the input tree if any

99 was provided.

Searching the database

101   A DateLife search consists of matching processed taxon names to tip labels in a

102 chronogram database. Chronograms with at least two matching tip labels are identified

103 and pruned down to preserve only the matched tips.

104   Matching pruned chronograms are stored as individual patristic distance matrices

105 (Figure 1 subfigure X). This format speeds up extraction of pairwise taxon ages of the

106 queried taxa, as opposed to searching the ancestor node of a pair of taxa in a "phylo"

107 object or newick string. The patristic matrices are also associated to the study citation

108 where the original chronogram was published, and stored as an R object of the newly

109 defined class `datelifeResult`.

110   DateLife's chronogram database latest version consist of 231 chronograms published

111 in 171 different studies. It is constructed from OpenTree's phylogenetic database, the

112 Phylesystem, which constitutes an open source of expert phylogenetic knowledge with rich

113 metadata (McTavish et al., 2015) that allows automatic and reproducible construction of a

114 chronogram database.

Summarizing search results

116   At this point, summary information is extracted from the `datelifeResult` object to

117 inform decisions for ulterior steps in the user workflow. Age data from matching pruned

chronograms is summarized and used to generate a single summary chronogram. Other
basic summary information available is:

1. The matching pruned chronograms as newick strings or "phylo" objects.
2. The ages of the root of all matching pruned chronograms. This can correspond to the
   age of the most recent common ancestor (mrca) of your group of interest if the
   pruned chronograms have all taxa belonging to the group. If not, the root
   corresponds to the mrca ofa subgroup withing your group of interest.
3. Study citations where original chronograms were published.
4. A report of input taxon names matches across pruned chronograms.
5. The single matching pruned chronogram with the most input taxon names.

Identifying groves.– To generate a single summary chronogram, the DateLife
algorithm starts by identifying the matching pruned chronograms that form a grove,
roughly, a sufficiently overlapping set of taxa between trees, by implementing definition 2.8
for n-overlap from Ané et al. (2009). In rare cases, a group of trees can have multiple
groves. By default, DateLife chooses the grove with the most taxa, however, the "criterion
= trees" flag allows the user to choose the grove with the most trees instead.

Choosing a topology.– Due to sampling, DateLife requires a tree topology to
summarize age data upon it. Users can provide one as input from the literature, or one of
their own making. If no topology is provided, DateLife automatically subsets one from the
OpenTree synthetic tree (Open Tree Of Life et al., 2019).

DateLife can also reconstruct branch lengths proportional to substitution rates on the
fixed tree topology from available genetic data from BOLD.

Congruifying nodes.– DateLife then implements the congruification method
(Eastman, Harmon, & Tank, 2013) to find nodes belonging to the same clade across
matching pruned chronograms. Congruified node ages stored as a

143 `congruifiedCalibrations` object are then matched to nodes in the chosen tree topology

144 and stored as a `matchedCalibrations` object.

145 ***Summarizing node ages.*** – DateLife summarizes matched calibrations into a single

146 patristic distance matrix using different methods. Summarizing options implemented

147 include Super Distance Matrix method (SDM, Criscuolo, Berry, Douzery, & Gascuel, 2006)

148 and summary statistics such as median, minimum and maximum ages.

149 ***Dating the tree topology.*** – Summarized calibrations can be applied as secondary

150 calibrations with different dating methods currently supported within DateLife: MrBayes

151 (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003), PATHd8 (Britton,

152 Anderson, Jacquet, Lundqvist, & Bremer, 2007), BLADJ (Campbell O. Webb, Ackerly, &

153 Kembel, 2008; Campbell O. Webb & Donoghue, 2005), and treePL (Stephen A. Smith &

154 O'Meara, 2012). By default, DateLife implements the Branch Length Adjuster (BLADJ)

155 algorithm that assigns ages to nodes with no data evenly between nodes with age data,

156 which minimizes age variance in the resulting chronogram (Campbell O. Webb et al.,

157 2008). When there is conflict in ages across node with age data, the algorithm ignores ages

158 that are older than parent nodes and/or younger than descendant nodes.

159 If there is no information on the age of the root in the chronogram database, users

160 can provide an estimate from the literature. If none is provided, DateLife assigns an

161 arbitrary age to the root as 10% older than the oldest age available within the group.

162 ***Visualizing results.*** – Finally, users can save all source and summary chronograms

163 in formats that permit reuse and reanalyses (newick and R "phylo" format), as well as view

164 and compare results graphically, or construct their own graphs using `datelife`'s

165 chronogram plot generation functions.

166   **Benchmark**

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

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

169   through the database relative to number of queried taxon names. Query processing time

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

171   TNRS is activated. Up to ten thousand names can be processed and searched in less than

172   30 minutes with the most time consuming settings. Once names have been processed as

173   described in methods, a name search through the chronogram database can be performed

174   in less than a minute, even with a very large number of taxon names (Fig. 2). `datelife`'s

175   code performance was evaluated with a set of unit tests designed and implemented with

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

177   package (R Core Team, 2018), and on a public server –via GitHub, using the continuous

178   integration tool Travis CI (https://travis-ci.org). At present, unit tests cover more than

179   30% of `datelife`'s code (https://codecov.io/gh/phylotastic/datelife).

180   **Case study**

181   We illustrate the `datelife` algorithm using the family of true finches, Fringillidae as

182   an example. A college educator wishes to obtain state-of-the-art data on time of

183   evolutionary origin of species belonging to the true finches for their class. They decide to

184   use `datelife` because they are teaching best practices for reproducibility. Students have

185   the option to go to the website at www.datelife.org and perform an interactive run.

186   However, the educator wants the students to practice their R skills. The first step is to run

187   a `datelife` query using the "get species from taxon" flag. This will get all recognised

188   species names within their chosen inclusive taxon. The Fringillidae has 289 species,

189   according to the Open Tree of Life taxonomy. Once with a curated set of species taxon

190   names, the next step is to run a `datelife` search that will find all chronograms that

191 contain at least two species names. The algorithm proceeds to prune the trees to keep

192 matching species names on tips only, and transform the pruned trees to pairwise distance

193 matrices. There are 13 chronograms containing at least two Fringillidae species, published

194 in 9 different studies (Fig. 3). The final step is to summarize the available information

195 using two alternative types of summary chronograms, median and SDM. As explained in

196 the "Description" section, data from source chronograms is first summarised into a single

197 distance matrix and then the available node ages are used as fixed node calibrations over a

198 consensus tree topology, to obtain a fully dated tree with the program BLADJ (Fig. 4).

199 Median summary chronograms are older and have wider variation in maximum ages than

200 chronograms obtained with SDM. With both methods, ages are generally consistent with

201 source ages, but there are some biological examples in which this is not true (see

202 Discussion).

Cross-validation test

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

205 branch lengths, as well as trees with branch lengths as relative substitution rates (Figs. 5

206 and 6). As a form of cross validation, we took tree topologies from each study and

207 calibrated them using time of lineage divergence data from all other source chronograms.

208 In the absence of branch lengths, the ages of internal nodes were recovered with a high

209 precision in almost all cases (except for studies 3, and 5; Fig. 5). Maximum tree ages were

210 only recovered in one case (study 2; Fig. 5). We also demonstrate the usage of PATHd8

211 (Britton et al., 2007) as an alternative method to BLADJ. For this, we run a `datelife`

212 branch length reconstruction that searches for DNA sequence data from the Barcode of

213 Life Data System [BOLD; Ratnasingham and Hebert (2007)] to generate branch lengths.

214 We were able to successfully generate a tree with BOLD branch lengths for all of the

215 Fringillidae source chronograms. However, dating with PATHd8 using congruified

216 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.

At the time of writing of this manuscript (Mar 10, 2022), `datelife`'s database has

231 chronograms, pulled entirely from OpenTree's database, the Phylesystem (McTavish et

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

new state-of-the-art chronograms any time. As chronograms are added to 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 maximum of every 6 months. Users can also upload new chronograms to OpenTree

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

chronograms, to have them immediately available for analysis.

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

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

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

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

243 detect branch length units, and those would still need to be determined manually by a

244 curator.

245   The only summary chronogram encompassing all life that is currently openly

246 available from OpenTree was constructed using age data from 2,274 published chronograms

247 (Hedges, Marin, Suleski, Paymer, & Kumar, 2015). However source chronograms are not

248 available in computer readable format for reuse or reanalysis. As this tree is part of

249 datelife's database, the amount of lineages that can be queried using `datelife` (86879

250 unique terminal taxa) is substantial, yet it can be improved. Consequently, we would like

251 to emphasize on the importance of sharing chronogram data for the benefit of the scientific

252 community as a whole, into repositories that require expert input and manual curation,

253 such as OpenTree's Phylesystem (McTavish et al., 2015).

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

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

256 do so. Strictly speaking, the best chronogram should reflect the real time of lineage

257 divergence accurately and precisely. To our knowledge, there are no good measures to

258 determine if a chronogram is better than another. Some measures that have been proposed

259 are the proportion of lineage sampling and the number of calibrations used Magallón,

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

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

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

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

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

265 primary calibrations in analyses performed with bayesian inference methods when priors

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

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

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

269 bias reported with bayesian methods. Certainly, further studies are required to fully
270 understand the effect of using secondary calibrations on time estimates and downstream
271 anlyses.

272   Furthermore, even chronograms obtained with primary fossil data can vary
273 substantially in time estimates between lineages, as observed from the comparison of source
274 chronograms in the Fringillidae example. This observation is often encountered in the
275 literature (see, for example, the ongoing debate about crown group age of angiosperms
276 (Barba-Montoya, Reis, Schneider, Donoghue, & Yang, 2018; Magallón et al., 2015;
277 Ramshaw et al., 1972; Sanderson & Doyle, 2001; Sauquet, Ramírez-Barahona, & Magallón,
278 2021). For some studies, especially ones based on branch lengths (e.g., studies of species
279 diversification, timing of evolutionary events, phenotypic trait evolution), using a different
280 chronogram may return different results (Title & Rabosky, 2016). Stitching together these
281 chronograms can create a larger tree that uses information from multiple studies, but the
282 effect of uncertainties and errors at this level on downstream analyses is still largely
283 unknown.

284   Summarizing chronograms might also imply summarizing fundamentally distinct
285 evolutionary hypotheses. For example, two different researchers working on the same clade
286 both carefully select and argument their choices of fossil calibrations. Still, if one researcher
287 decides a fossil will calibrate the ingroup of a clade, while another researcher uses the same
288 one to calibrate outside the clade, the resulting age estimates will often differ substantially,
289 as the placement of calibrations as stem or crown group is proved to deeply affect
290 estimated times of lineage divergence (Sauquet, 2013). Trying to summarize the resulting
291 chronograms into a single one using simple summary statistics might erase all types of
292 relevant information from the source chronograms. Accordingly, the prevailing view in our
293 research community is that we should favor time of lineage divergence estimates obtained
294 from a single analysis, using fossil data as primary sources of calibrations, and using fossils
295 that have been widely discussed and curated as calibrations to date other trees, making

296 sure that all data used in the analysis reflect a coherent evolutionary history (Antonelli et

297 al., 2017). However, the exercise of summarizing different chronograms has the potential to

298 help getting a single global evolutionary history for a lineage by putting together evidence

299 from different hypothesis. Choosing the elements of the chronograms that we are going to

300 keep and the ones that we are going to discard is key, since we are potentially loosing

301 important parts of the evolutionary history of a lineage that might only be reflected in

302 source chronograms and not on the summary chronogram (Sauquet et al., 2021).

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

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

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

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

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

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

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

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

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

312 entered by curators in the future.

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

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

315 improvement for testing alternative hypothesis using phylogenetic distance (Campbell O.

316 Webb et al., 2008). Hence, we integrated into datelife's workflow different ways of creating

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

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

319 practice in scientific publications: Jetz, Thomas, Joy, Hartmann, and Mooers (2012),

320 created a time-calibrated tree of all 9,993 bird species, where 67% had molecular data and

321 the rest was simulated; Rabosky et al. (2018) created a time-calibrated tree of 31,536

322 ray-finned fishes, of which only 37% had molecular data; Stephen A. Smith and Brown

₃₂₃ (2018) constructed a tree of 353,185 seed plants where only 23% had molecular data.

₃₂₄ Taken to the extreme, one could make a fully resolved, calibrated tree of all modern and

₃₂₅ extinct taxa using a single taxonomy and a single calibration with the polytomy resolution

₃₂₆ and branch assignation methods. There has yet to be a thorough analysis of what can go

₃₂₇ wrong when one goes beyond the data in this way, so we urge caution; we also urge readers

₃₂₈ to follow the example of many of the large tree papers cited above and make sure results

₃₂₉ are substantially similar between trees fully reconstructed with molecular or other data,

₃₃₀ and trees that are reconstructed using taxonomy by resolving polytomies at random

₃₃₁ following a statistical model.

Conclusions

₃₃₃ Divergence time information is key to many areas of evolutionary studies: trait

₃₃₄ evolution, diversification, biogeography, macroecology and more. It is also crucial for

₃₃₅ science communication and education, but generating chronograms is difficult, especially

₃₃₆ for those who want to use phylogenies but who are not systematists, or do not have the

₃₃₇ time to acquire and develop the necessary knowledge and data curation skills. Moreover,

₃₃₈ years of primarily public funded research have resulted in vast amounts of chronograms

₃₃₉ that are already available on scientific publications, but hidden to the public and scientific

₃₄₀ community for reuse.

₃₄₁ The `datelife` R package allows easy and fast summarization of publicly available

₃₄₂ information on time of lineage divergence. This provides a straightforward way to get an

₃₄₃ informed idea on the state of knowledge of the time frame of evolution of different regions

₃₄₄ of the tree of life, and allows identification of regions that require more research or that

₃₄₅ have conflicting information. Both summary and newly generated trees are useful to

₃₄₆ evaluate evolutionary hypotheses in different areas of research. The DateLife project helps

₃₄₇ with awareness of the existing variation in expert time of divergence data, and will foster

₃₄₈ exploration of the effect of alternative divergence time hypothesis on the results of

349 analyses, nurturing a culture of more cautious interpretation of evolutionary results.

Availability

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

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

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

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

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

356 (https://cran.r-project.org/package=datelife) using the command

357 `install.packages(pkgs = "datelife")` from within R. Development versions are

358 available from the GitHub repository (https://github.com/phylotastic/datelife) and can be

359 installed using the command `devtools::install_github("phylotastic/datelife")`.

Supplementary Material

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

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

363 datelife_benchmark repositories in LLSR's GitHub account.

Funding

Acknowledgements

References

Ané, C., Eulenstein, O., Piaggio-Talice, R., & Sanderson, M. J. (2009). Groves of phylogenetic trees. *Annals of Combinatorics*, *13*(2), 139–167.

Antonelli, A., Hettling, H., Condamine, F. L., Vos, K., Nilsson, R. H., Sanderson, M. J., . . . Vos, R. A. (2017). Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of Taxa. *Systematic Biology*, *66*(2), 153–166. https://doi.org/10.1093/sysbio/syw066

Archie, J., Day, W. H., Felsenstein, J., Maddison, W., Meacham, C., Rohlf, F. J., & Swofford, D. (1986). The Newick tree format. Retrieved from %7Bhttps://evolution.genetics.washington.edu/phylip/newicktree.html%7D

Bapst, D. W. (2012). Paleotree: An R package for paleontological and phylogenetic analyses of evolution. *Methods in Ecology and Evolution*, *3*(5), 803–807. https://doi.org/10.1111/j.2041-210X.2012.00223.x

Barba-Montoya, J., Reis, M. dos, Schneider, H., Donoghue, P. C., & Yang, Z. (2018). Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a cretaceous terrestrial revolution. *New Phytologist*, *218*(2), 819–834.

Barker, F. K., Burns, K. J., Klicka, J., Lanyon, S. M., & Lovette, I. J. (2012). Going to extremes: Contrasting rates of diversification in a recent radiation of new world passerine birds. *Systematic Biology*, *62*(2), 298–320.

Barker, F. K., Burns, K. J., Klicka, J., Lanyon, S. M., & Lovette, I. J. (2015). New insights into new world biogeography: An integrated view from the phylogeny of blackbirds, cardinals, sparrows, tanagers, warblers, and allies. *The Auk: Ornithological Advances*, *132*(2), 333–348.

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., . . . Enquist, B. J. (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-16

Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S., & Bremer, K. (2007). Estimating Divergence Times in Large Phylogenetic Trees. *Systematic Biology*, *56*(788777878), 741–752. https://doi.org/10.1080/10635150701613783

Burns, K. J., Shultz, A. J., Title, P. O., Mason, N. A., Barker, F. K., Klicka, J., . . . Lovette, I. J. (2014). Phylogenetics and diversification of tanagers (passeriformes: Thraupidae), the largest radiation of neotropical songbirds. *Molecular Phylogenetics and Evolution*, *75*, 41–77.

Chamberlain, S. A., & Szöcs, E. (2013). taxize : taxonomic search and retrieval in R [version 2; referees: 3 approved]. *F1000Research*, *2*(191), 1–29. https://doi.org/10.12688/f1000research.2-191.v2

Chamberlain, S. A., Szöcs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., . . . Li, G. (2019). *taxize: Taxonomic information from around the web*. Retrieved from https://github.com/ropensci/taxize

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

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

Delsuc, F., Philippe, H., Tsagkogeorga, G., Simion, P., Tilak, M.-K., Turon, X., . . . Douzery, E. J. (2018). A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biology*, *16*(1), 1–14.

Eastman, J. M., Harmon, L. J., & Tank, D. C. (2013). Congruification: Support for time scaling large phylogenetic trees. *Methods in Ecology and Evolution*, *4*(7), 688–691. https://doi.org/10.1111/2041-210X.12051

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.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

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

Laubichler, M. D., & Maienschein, J. (2009). *Form and function in developmental evolution.* Cambridge University Press.

Magallon, S., & Sanderson, M. J. (2001). Absolute diversification rates in angiosperm clades. *Evolution*, *55*(9), 1762–1780.

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

Open Tree Of Life, Redelings, B., Cranston, K. A., Allman, J., Holder, M. T., & McTavish, E. J. (2016). Open Tree of Life APIs v3.0. *Open Tree of Life Project*, (Online Resources). Retrieved from %7Bhttps://github.com/OpenTreeOfLife/germinator/wiki/Open-Tree-of-Life-Web-APIs%7D

Open Tree Of Life, Redelings, B., Sánchez Reyes, L. L., Cranston, K. A., Allman, J., Holder, M. T., & McTavish, E. J. (2019). Open tree of life synthetic tree v12.3. *Zenodo*. Retrieved from 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.

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.

Rees, J. A., & Cranston, K. (2017). Automated assembly of a reference taxonomy
for phylogenetic data synthesis. *Biodiversity Data Journal*, (5).

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

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic
inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574.
https://doi.org/10.1093/bioinformatics/btg180

Sanchez-Reyes, L. L., O'Meara, B., Eastman, J., Heath, T., Wright, A., Schliep, K.,
. . . Alfaro, M. (2022). datelife: Scientific Data on Time of Lineage Divergence
for Your Taxa. *R Package Version 0.6.2.* Retrieved from
https://doi.org/10.5281/zenodo.593938

Sanderson, M. J. (2003). r8s: Inferring absolute rates of molecular evolution and
divergence times in the absence of a molecular clock. *Bioinformatics*, *19*(2),
301–302.

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

estimating the age of angiosperms from rbcL and 18S rDNA data. *American Journal of Botany*, *88*(8), 1499–1516.

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

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

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

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

Smith, Stephen A., & O'Meara, B. C. (2012). TreePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*, *28*(20), 2689–2690. https://doi.org/10.1093/bioinformatics/bts492

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

544   for the analysis of phylogenetic community structure and trait evolution.

545   *Bioinformatics*, *24*(18), 2098–2100.

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

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

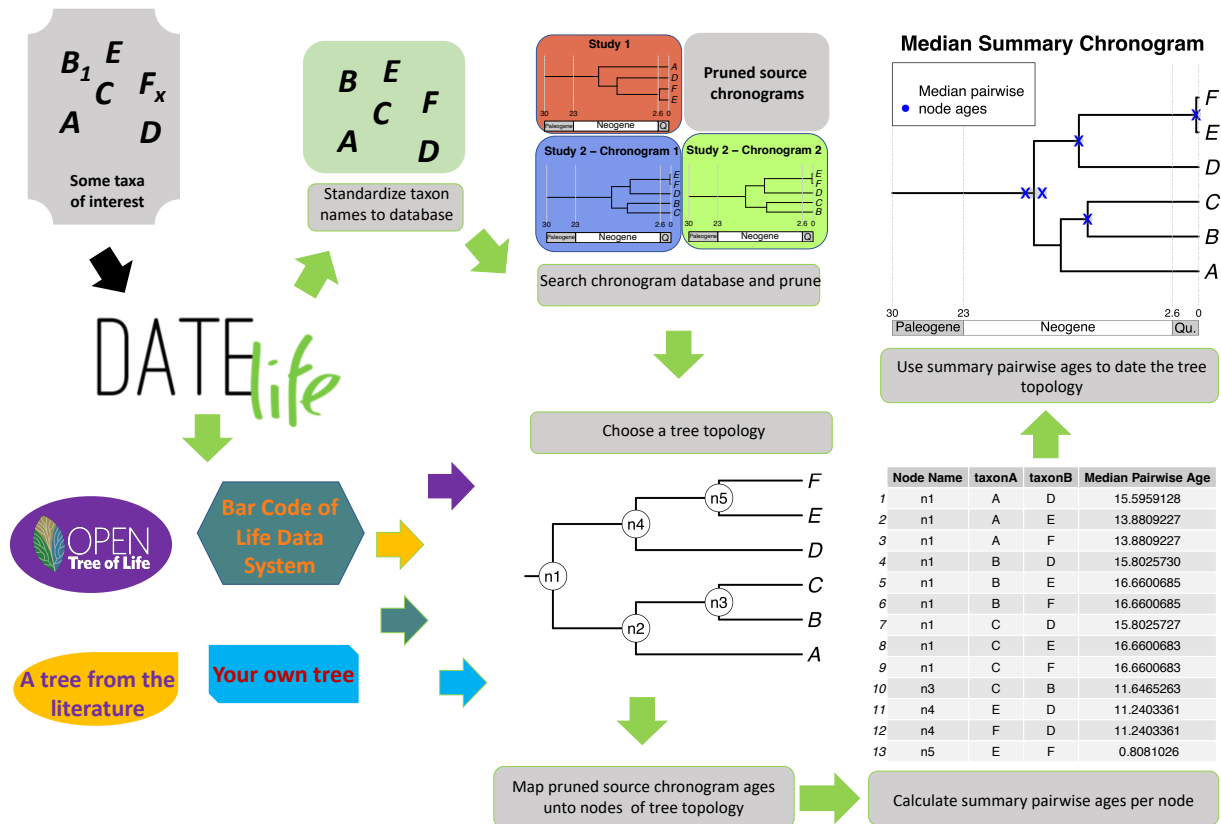548   applied phylogenetics. *Molecular Ecology Notes*, *5*(1), 181–183.

549   FIGURES

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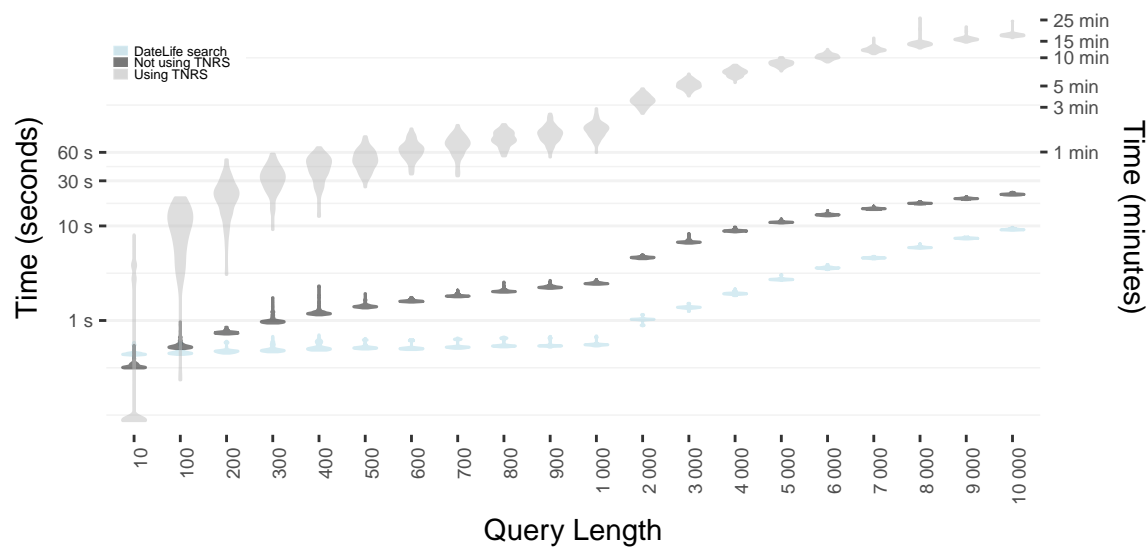
FIGURE 2. Input taxon name processing and chronogram database search computation time increases with number of input taxon names. We sampled N bird species names for each input size class, 100 times, and then performed a `datelife` search using the Taxon Names Resoultion Service (TNRS; light gray), and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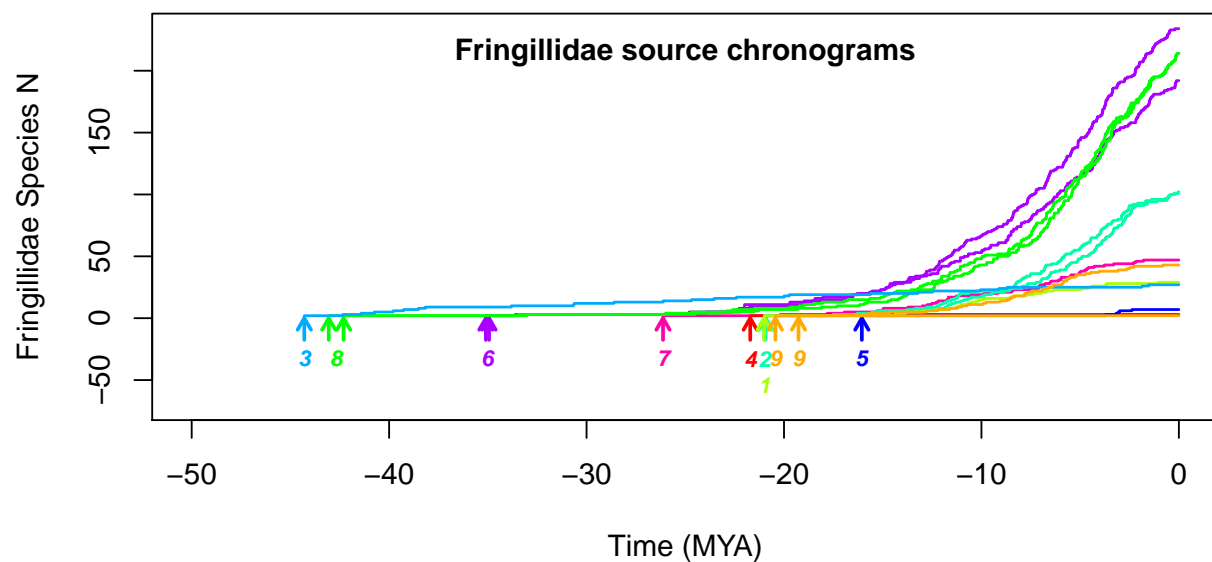
FIGURE 3. Lineage through time (LTT) plots of source chronograms containing all or a subset of species from the bird family Fringillidae of true finches. Arrows indicate maximum age of each chronogram. Numbers reference to chronograms' original publications 1: Barker et al. (2012), 2: Barker et al. (2015),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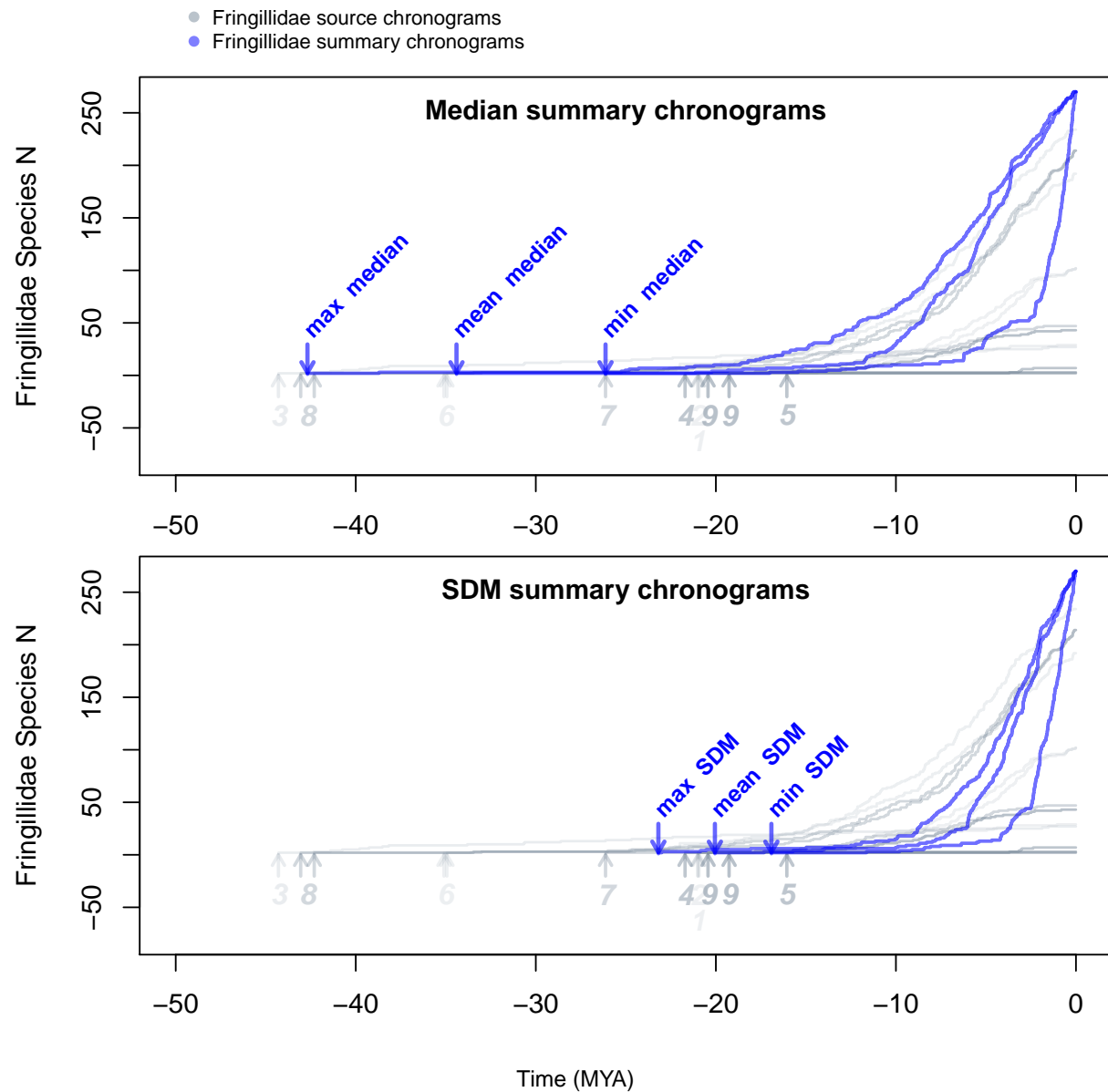
FIGURE 4. LTT plots of median (top) and Supermatrix Distance Method (SDM; bottom) chronograms summarising information from source chronograms found for the Fringillidae. Arrows indicate tree maximum age.
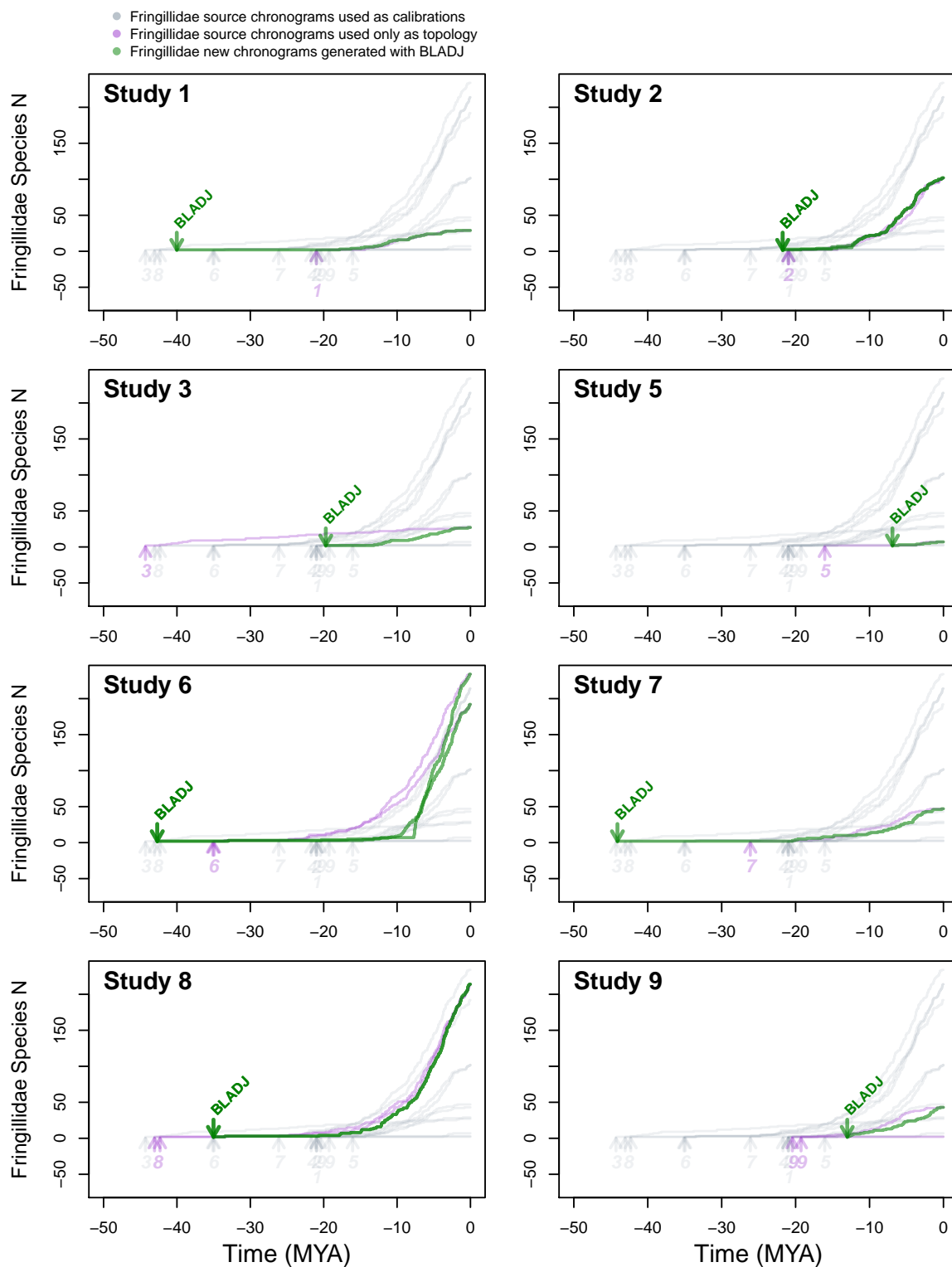
FIGURE 5. LTT plots showing results from the cross-validation analyses of trees without branch lengths dated using BLADJ. The dating analysis can only be performed in trees with more than 2 tips, thus excluding chronogram from study 4; its data was still used as calibration for the other source chronograms.
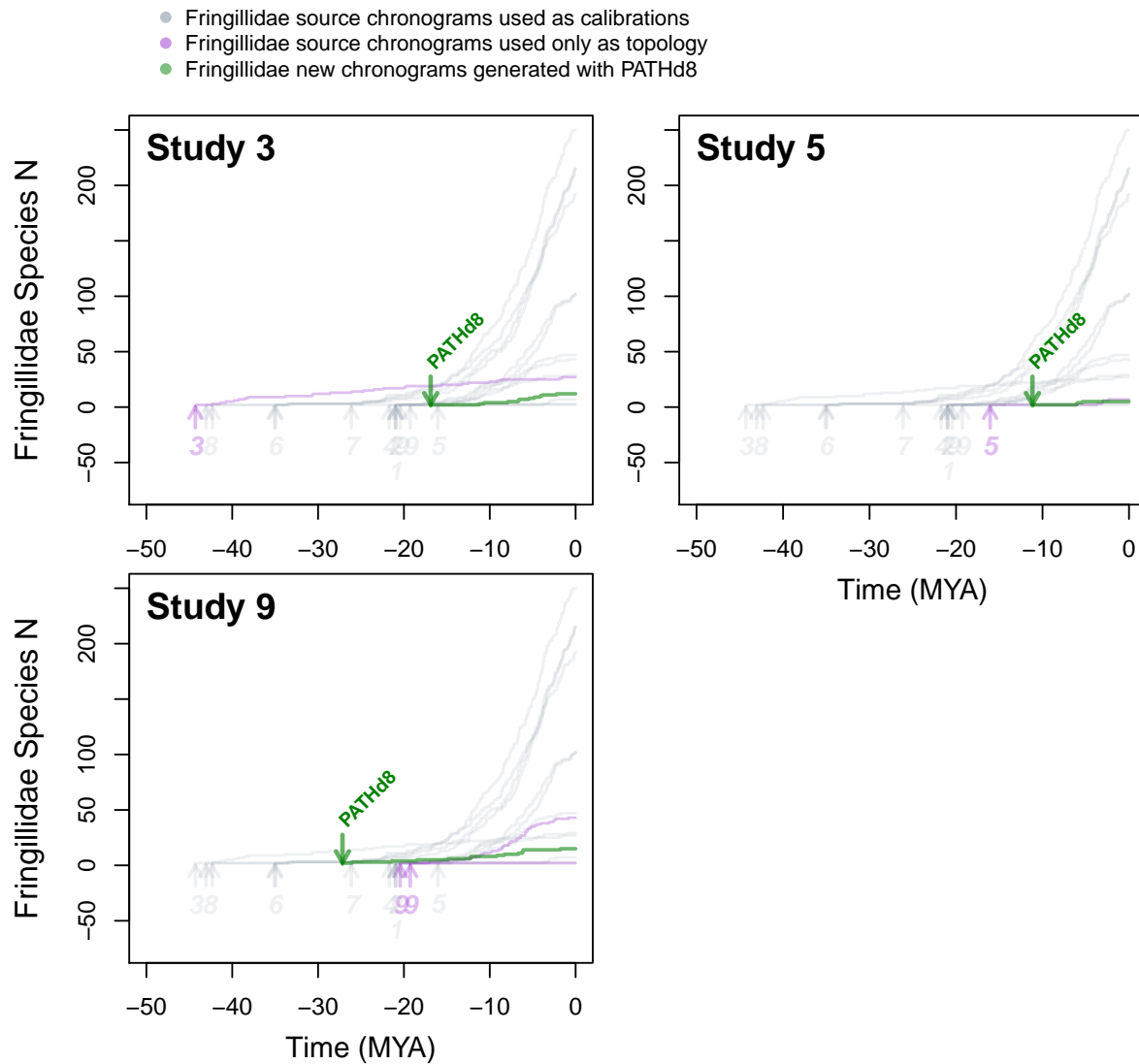
FIGURE 6.   LTT plots showing results from the cross-validation analyses of trees with branch length reconstructed with data from the Barcode of Life Database (BOLD) dated using PATHd8. We could construct a tree with branch lengths for all source chronograms. However, dating with PATHd8 was only successful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