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, USA

4   [2] University of Tennessee, Knoxville, USA

5   Author Note

⁶   Department of Life and Environmental Sciences, University of California, Merced, CA

⁷ 95343, USA.

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

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

18   Abstract

19   Chronograms –phylogenies with branch lengths proportional to time– represent key

20 data on timing of evolutionary events for the study of natural processes in many areas of

21 biological research. Chronograms also provide valuable information that can be used for

22 education, science communication, and conservation policy decisions. Yet, achieving a

23 high-quality reconstruction of a chronogram is a difficult and resource-consuming task. Here

24 we present DateLife, a phylogenetic software implemented as an R package and an R Shiny

25 web application available at www.datelife.org, that provides services for efficient and easy

26 discovery, summary, reuse, and reanalysis of node age data mined from a curated database of

27 expert, peer-reviewed, and openly available chronograms. The main DateLife workflow starts

28 with one or more scientific taxon names provided by a user. Names are processed and

29 standardized to a unified taxonomy, allowing DateLife to run a name match across its local

30 chronogram database that is curated from Open Tree of Life's phylogenetic repository, and

31 extract all chronograms that contain at least two queried taxon names, along with their

32 metadata. Finally, node ages from matching chronograms are mapped using the

33 congruification algorithm to corresponding nodes on a tree topology, either extracted from

34 Open Tree of Life's synthetic phylogeny or one provided by the user. Congruified node ages

35 are used as secondary calibrations to date the chosen topology, with or without initial

36 branch lengths, using different phylogenetic dating methods such as BLADJ, treePL,

37 PATHd8 and MrBayes. We performed a cross-validation test to compare node ages resulting

38 from a DateLife analysis (i.e, phylogenetic dating using secondary calibrations) to those from

39 the original chronograms (i.e, obtained with primary calibrations), and found that DateLife's

40 node age estimates are consistent with the age estimates from the original chronograms, with

41 the largest variation in ages occurring around topologically deeper nodes. Because the

42 results from any software for scientific analysis can only be as good as the data used as input,

we highlight the importance of considering the results of a DateLife analysis in the context of
the input chronograms. DateLife can help to increase awareness of the existing disparities
among alternative hypotheses of dates for the same diversification events, and to support
exploration of the effect of alternative chronogram hypotheses on downstream analyses,
providing a framework for a more informed interpretation of evolutionary results.

Keywords: Tree; Phylogeny; Scaling; Dating; Ages; Divergence times; Open Science;
Congruification; Supertree; Calibrations; Secondary calibrations.

Word count: 7478

51 Chronograms –phylogenies with branch lengths proportional to time– provide key data

52 on evolutionary time frame for the study of natural processes in many areas of biological

53 research, such as comparative analysis (Freckleton, Harvey, & Pagel, 2002; Harvey, Pagel, &

54 others, 1991), developmental biology (Delsuc et al., 2018; Laubichler & Maienschein, 2009),

55 conservation biology and ecology (Felsenstein, 1985; Webb, 2000), historical biogeography

56 (Posadas, Crisci, & Katinas, 2006), and species diversification (Magallon & Sanderson, 2001;

57 Morlon, 2014).

58 Building a chronogram is not an easy task. It requires obtaining and curating a

59 homology hypothesis to construct a phylogeny, selecting and placing appropriate calibrations

60 on the phylogeny using independent age data points from the fossil record or other dated

61 events, and inferring a full dated tree. All of this entails specialized biological training,

62 taxonomic domain knowledge, and a significant amount of research time, computational

63 resources and funding.

64 Here we present the DateLife project which has the main goal of extracting and

65 exposing age data from published chronograms, making age data readily accessible to a

66 wider community for reuse and reanalysis in research, teaching, science communication and

67 conservation policy. DateLife's core software application is available as an R package

68 (Sanchez-Reyes et al., 2022), and as an online Rshiny interactive website at www.datelife.org.

69 It features key elements for scientific reproducibility, such as a curated, versioned, open and

70 fully public chronogram database (McTavish et al., 2015) that stores data in a

71 computer-readable format (Vos et al., 2012); automated and programmatic ways of accessing

72 and downloading the data, also in a computer-readable format (Stoltzfus et al., 2013); and

73 methods to summarize and compare the data.

Description

75 DateLife's core software applications are implemented in the R package `datelife`, and

76 relies on functionalities from other biological R packages: ape (Paradis, Claude, & Strimmer,

77 2004), bold (Chamberlain, 2018), geiger (Pennell et al., 2014), msa (Bodenhofer, Bonatesta,

78 Horejš-Kainrath, & Hochreiter, 2015), paleotree (Bapst, 2012), phyloch (Heibl, 2008),

79 phylocomr (Ooms & Chamberlain, 2018), phytools (Revell, 2012), rotl (Michonneau, Brown,

80 & Winter, 2016), and taxize (Chamberlain, 2018; Chamberlain & Szöcs, 2013). There are

81 three main steps to the DateLife workflow: 1) creating a search query, 2) searching a

82 database, and 3) summarizing results from the search.

Creating a Search Query

84 DateLife starts by processing an input consisting of the scientific name of at least one

85 taxon. Multiple input names can be provided as a comma separated character string or as

86 tip labels on a tree. If the input is a tree, it can be provided as a classic newick character

87 string (Archie et al., 1986), or as a "phylo" R object (Paradis et al., 2004). The input tree is

88 not required to have branch lengths, and its topology is used in the summary steps described

89 in the next section.

90 DateLife processes input scientific names using a Taxonomic Name Resolution Service

91 (TNRS), which increases the probability of correctly finding the queried taxon names in the

92 chronogram database. TNRS detects, corrects and standardizes name misspellings and typos,

93 variant spellings and authorities, and nomenclatural synonyms to a single taxonomic

94 standard (Boyle et al., 2013). TNRS also allows to correctly choose between homonyms, by

95 considering other taxa provided as input to infer the taxonomic context of the homonym.

96 DateLife implements TNRS using the Open Tree of Life (OpenTree) unified Taxonomy

97 (OTT, Open Tree Of Life et al., 2016; Rees & Cranston, 2017) as standard, storing

98 taxonomic identification numbers (OTT ids) for further processing and analysis. Other

99 taxonomies currently supported by DateLife are the National Center of Biotechnology

100 Information (NCBI) taxonomic database (Schoch et al., 2020), the Global Biodiversity

101 Information Facility (GBIF) taxonomic backbone (GBIF Secretariat, 2022), and the Interim

102 Register of Marine and Non-marine Genera (IRMNG) database (Rees et al., 2017).

103   Besides binomial species names, DateLife accepts scientific names from any inclusive

104 taxonomic group (e.g., genus, family, tribe), as well as subspecific taxonomic variants (e.g.,

105 subspecies, variants, strains). If a taxon name belongs to an inclusive taxonomic group,

106 DateLife has two alternative behaviors defined by the "get species from taxon" flag. If the

107 flag is active, DateLife retrieves all species names within a taxonomic group provided, from a

108 standard taxonomy of choice, and adds them to the search query. In this case, subspecific

109 variants are excluded. If the flag is inactive, DateLife excludes inclusive taxon names from

110 the search query, and species and subspecific variant names are processed as provided by the

111 user. The processed taxon names are saved as an R object of a newly defined class,

112 `datelifeQuery`, that is used in the following steps. This object contains the input names

113 standardized to a taxonomy of choice (OTT by default), the corresponding OTT id numbers,

114 and the topology of an input tree, if one was provided.

115   *Searching a Chronogram Database*

116   At the time of writing of this manuscript (Mar 08, 2024), DateLife's chronogram

117 database latest version consist of 253 chronograms published in 187 studies, encompassing 99

118 474 species. It is curated from OpenTree's phylogenetic database, the Phylesystem, an open

119 database of expert and peer-reviewed phylogenetic knowledge with rich metadata and a wide

120 taxonomic scope (McTavish et al., 2015). We expect DateLife's database to largely overlap

121 with OpenTree's phylogenetic database taxonomic coverage, where Chordata and

122 Embryophyta are nearly fully sampled. In contrast, Bacteria, Fungi, Nematoda, and Insecta,

123 currently present a large gap between the number of named species and what has

124 phylogenetic information in OpenTree's synthetic tree. It is likely that users working with

125 the former groups will get results from a DateLife analysis. If none of the user's species are

126 found, the software will indicate the lack of age data for the queried taxa in the database.

¹²⁷ A unique feature of the Phylesystem is that any user can add new published,

¹²⁸ state-of-the-art chronograms any time, through OpenTree's curator application

¹²⁹ (https://tree.opentreeoflife.org/curator). Relying on an open source database permits an

¹³⁰ automatic and reproducible assembly of DateLife's chronogram database, which is stored and

¹³¹ navigable as an R data object within the `datelife` R package. As chronograms are added to

¹³² Phylesystem, they can be incorporated into the chronogram database within the `datelife`

¹³³ R package, by manually triggering an update. The updated `datelife` database is assigned a

¹³⁴ new version number, followed by a package release on CRAN. We encourage users to submit

¹³⁵ published chronograms to OpenTree's phylogenetic database, so that their taxon of interest

¹³⁶ can be included in future DateLife searches. Users can directly run `datelife` functions to

¹³⁷ trigger an update of their local chronogram database, to incorporate any new chronograms

¹³⁸ to their DateLife analysis before a `datelife` database update is released on CRAN.

¹³⁹ A DateLife search is implemented by matching processed taxon names provided by the

¹⁴⁰ user to tip labels in the chronogram database. Chronograms with at least two matching

¹⁴¹ taxon names on their tip labels are identified and pruned down to preserve only the matched

¹⁴² taxa. These matching pruned chronograms are referred to as source chronograms. Total

¹⁴³ distance in units of million years (myr) between taxon pairs within each source chronogram

¹⁴⁴ are stored as a patristic distance matrix. The matrix format speeds up extraction of pairwise

¹⁴⁵ taxon ages of any queried taxa, as opposed to searching the ancestor node of a pair of taxa

¹⁴⁶ in a "phylo" object or newick string. Finally, the patristic matrices are associated to the

¹⁴⁷ study citation where the original chronogram was published, and stored as an R object of

¹⁴⁸ the newly defined class `datelifeResult`.

¹⁴⁹ *Summarizing Search Results*

¹⁵⁰ Summary information is extracted from the `datelifeResult` object to inform

¹⁵¹ decisions for subsequent steps in the analysis workflow. Basic summary information available

¹⁵² to the user includes:

1. The matching pruned chronograms as newick strings or "phylo" objects.

2. The ages of the root of all source chronograms. These ages can correspond to the age of the most recent common ancestor (mrca) of the user's group of interest if the source chronograms have all taxa belonging to the group. If not, the root corresponds to the mrca of a subgroup within the group of interest.

3. Study citations where original chronograms were published.

4. A report of input taxon names matches across source chronograms.

5. The source chronogram(s) with the most input taxon names.

6. Various single summary chronograms resulting from summarizing age data, generated using the methodology described next.

Choosing a Topology

DateLife requires a tree topology to summarize age data upon. We recommend that users provide as input a tree topology from the literature, or one of their own making. If no topology is provided, DateLife automatically extracts one from the OpenTree synthetic tree, a phylogeny currently encompassing 2.3 million taxa across all life, assembled from 1,239 published phylogenetic trees and OpenTree's unified Taxonomy, OTT (Open Tree Of Life et al., 2019). Alternatively, DateLife can combine topologies from source chronograms using a supertree approach (Criscuolo, Berry, Douzery, & Gascuel, 2006). To do this, DateLife first identifies the source 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). If the source chronograms do not form a grove, the supertree reconstruction will fail. 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. The result is a single summary (i.e., supertree) topology, that combines topologies from source chronograms in a grove.

Applying Secondary Calibrations

179   Once a topology is chosen, DateLife applies the congruification method (Eastman,

180 Harmon, & Tank, 2013) that find nodes belonging to the same clade across source

181 chronograms, and then extracts the corresponding node ages from patristic distance matrices

182 stored as a `datelifeResult` object. Note that by definition, these matrices store total

183 distance (time from tip to tip), assuming that the terminal taxa are coeval and occur at the

184 present. Hence, node ages correspond to half the values stored in the `datelifeResult`

185 matrices. A table of congruified node ages that can be used as calibrations for a dating

186 analysis is stored as a `congruifiedCalibrations` object.

187   For each congruent node, the pairwise distances that traverse that node are summarized

188 into a single summary matrix using classic summary statistics (i.e., mean, median, minimum

189 and maximum ages), and the Supermatrix Distance Method (SDM; Criscuolo et al., 2006),

190 which deforms patristic distance matrices by minimizing variance and then averaging them.

191 These single summary taxon pair age matrices are stored as summarized calibrations that

192 can be used as secondary calibrations to date a tree topology - with or without initial branch

193 lengths, using phylogenetic dating methods currently supported within DateLife: BLADJ

194 (Webb, Ackerly, & Kembel, 2008; Webb & Donoghue, 2005), MrBayes (Huelsenbeck &

195 Ronquist, 2001; Ronquist & Huelsenbeck, 2003), PATHd8 (Britton, Anderson, Jacquet,

196 Lundqvist, & Bremer, 2007), and treePL (Smith & O'Meara, 2012).

Dating a Tree Topology

197

Dating a tree without branch lengths.– To date a tree topology when initial

198

199 branch lengths are unavailable, DateLife implements the Branch Length Adjuster (BLADJ)

200 algorithm (Webb et al., 2008; Webb & Donoghue, 2005), which only requires a tree topology

201 with no branch lengths and at least two node ages to use as calibrations, one for the tree root

202 and one for any internal node of the topology. The BLADJ algorithm fixes ages for nodes

203 with calibration data upon the given tree topology. Then, it assigns ages to nodes with no

204 available age information by distributing time evenly between calibrated nodes, minimizing

205 age variance in the resulting chronogram. This approach has proven useful for ecological

206 analyses that require a phylogenetic time context (Webb et al., 2008). When there is conflict

207 between ages of calibrated nodes, BLADJ ignores node ages that are older than the age of a

208 parent node. The BLADJ algorithm requires a root age to run. Users can provide an

209 appropriate root age estimate of their own or one obtained from the literature. If a root age

210 is not provided and there is no information on the age of the root in the chronogram

211 database, DateLife chooses a random age for the root, so that a dated tree topology can be

212 generated with BLADJ. In this case, DateLife will provide a conspicuous warning message,

213 so that users are aware that the root of the chronogram was chosen at random because there

214 was no information available for it in the chronogram database, along with suggestions on

215 how the user can find and provide an appropriate age for the root of the initial topology.

216   An alternative to BLADJ to date tree topologies in the absence of initial branch

217 lengths that is common practice in the literature is to use a birth-death model to draw

218 branch lengths (Jetz, Thomas, Joy, Hartmann, & Mooers, 2012; Rabosky et al., 2018; Smith

219 & Brown, 2018). In addition to the initial tree topology and nodes with age data, these

220 methods require initial values of speciation and extinction rate parameters provided by the

221 user. DateLife implements this approach with MrBayes (Huelsenbeck & Ronquist, 2001;

222 Ronquist & Huelsenbeck, 2003), using nodes with published age data as calibration priors on

223 nodes of a tree topology with no branch lengths, a simple birth-death model with speciation

224 and extinction rate parameters that are provided by the user, and no genetic data. However,

225 BLADJ is the default option in DateLife, as it does not require any information on

226 diversification rates for the phylogenetic sample to draw from a branch length distribution.

227   *Dating a tree with branch lengths.*– Relative branch lengths can provide key

228 information for phylogenetic dating, specifically for nodes without any calibration data

229 available. While using initial branch length data is the golden standard for phylogenetic

230 dating analyses, obtaining such information from scratch is not an easy task: it requires

²³¹ obtaining primary data, assembling and curating a homology (orthology) hypothesis, and

²³² choosing and implementing a method for phylogenetic inference. DateLife implements a

²³³ workflow to streamline this process by applying open data from the Barcode of Life Data

²³⁴ System, BOLD (Ratnasingham & Hebert, 2007) to obtain genetic markers for input taxa.

²³⁵ By default, BOLD genetic sequences are aligned with MUSCLE (Edgar, 2004) using

²³⁶ functions from the msa R package (Bodenhofer et al., 2015). Alternatively, sequences can be

²³⁷ aligned with MAFFT (Katoh, Asimenos, & Toh, 2009), using functions from the ape R

²³⁸ package (Paradis et al., 2004). The BOLD sequence alignment is then used to obtain initial

²³⁹ branch lengths with the accelerated transformation (ACCTRAN) parsimony algorithm,

²⁴⁰ which resolves ambiguous character optimization by assigning changes along branches of the

²⁴¹ tree as close to the root as possible (Agnarsson & Miller, 2008), resulting in older internal

²⁴² nodes as compared to other parsimony algorithms (Forest et al., 2005). The parsimony

²⁴³ branch lengths are then optimized using Maximum Likelihood, given the alignment, the

²⁴⁴ topology and a simple Jukes-Cantor model, producing a BOLD tree with branch lengths

²⁴⁵ proportional to expected number of substitutions per site. Both parsimony and ML

²⁴⁶ optimizations are done with functions from the `phangorn` package (Schliep, 2011). Due to

²⁴⁷ the computing load it requires, the BOLD workflow is currently only supported through

²⁴⁸ DateLife's R package. It is not yet available through the web application.

²⁴⁹ Phylogenetic dating methods supported in DateLife that incorporate branch length

²⁵⁰ information from the input topology in combination with the secondary calibrations include:

²⁵¹ PATHd8, a non-clock, rate-smoothing method to date trees (Britton et al., 2007); treePL

²⁵² (Smith & O'Meara, 2012), a semi-parametric, rate-smoothing, penalized likelihood dating

²⁵³ method (Sanderson, 2002); and MrBayes (Huelsenbeck & Ronquist, 2001; Ronquist &

²⁵⁴ Huelsenbeck, 2003), a Bayesian inference program implementing Markov chain Monte Carlo

²⁵⁵ (MCMC) methods to estimate a posterior distribution of model parameters.

²⁵⁶ *Visualizing Results*

257 Finally, users can save all source and summary chronograms in formats allowing for

258 reuse and reanalysis, such as newick and the R "phylo" format. Input and summary

259 chronograms can be visualized and compared graphically, and users can construct their own

260 graphs using DateLife's chronogram plot generation functions available from the R package

261 `datelifeplot` (Sanchez-Reyes & O'Meara, 2022).

Benchmark

263 R package `datelife` code speed was tested on an Apple iMac with one 3.4 GHz Intel

264 Core i5 processor. We registered variation in computing time of query processing and search

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

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

267 Taxonomic Name Resolution Service (TNRS) is activated. Up to ten thousand names can be

268 processed and searched in less than 30 minutes with the most time consuming settings. Once

269 names have been processed as described in methods, a name search through the chronogram

270 database can be performed in less than a minute, even with a very large number of taxon

271 names (Fig. 1).

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

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

274 with the devtools package (R Core Team, 2018), and on a public server using the continuous

275 integration tool of GitHub actions (https://docs.github.com/en/actions). At present, unit

276 tests cover more than 40% of `datelife`'s code (https://codecov.io/gh/phylotastic/datelife).

277 Unit testing helps identify potential issues as code is updated or, more critically, as services

278 code relies upon may change.

Case Studies

280 We illustrate the DateLife workflow using a family within the passeriform birds

281 encompassing the true finches, Fringillidae, as case study. On a small example, we analysed 6

²⁸² bird species, and results from each step of the workflow are shown in Figure 2. As a second

²⁸³ example, we analysed 289 bird species in the family Fringillidae that are included in the

²⁸⁴ NCBI taxonomy. One clade from the full summary chronogram result from the DateLife

²⁸⁵ analysis is shown Figure 3. The full chronogram for all 289 species and the results from

²⁸⁶ previous steps of the workflow are available as Supplementary Figures.

²⁸⁷ *A Small Example*

²⁸⁸ ***Creating a search query.***– We chose 6 bird species within the Passeriformes. The

²⁸⁹ sample includes two species of cardinals: the black-thighed grosbeak – *Pheucticus tibialis*

²⁹⁰ and the crimson-collared grosbeak – *Rhodothraupis celaeno*; three species of buntings: the

²⁹¹ yellowhammer – *Emberiza citrinella*, the pine bunting – *Emberiza leucocephalos* and the

²⁹² yellow-throated bunting – *Emberiza elegans*; and one species of tanager, the vegetarian finch –

²⁹³ *Platyspiza crassirostris*. Processing of input names found that *Emberiza elegans* is synonym

²⁹⁴ for *Schoeniclus elegans* in the default reference taxonomy (OTT v3.3, June 1, 2021). For a

²⁹⁵ detailed discussion on the state of the synonym, refer to Avibase (Avibase, 2022; Lepage,

²⁹⁶ 2004; Lepage, Vaidya, & Guralnick, 2014). Discovering this synonym allowed assigning five

²⁹⁷ age data points for the parent node of *Emberiza elegans*, shown as *Schoeniclus elegans* in

²⁹⁸ Figure 2, which would not have had any data otherwise.

²⁹⁹ ***Searching the database.***– DateLife used the processed input names to search the

³⁰⁰ local chronogram database and found 9 matching chronograms from 6 different studies (Fig.

³⁰¹ 2c). Three studies matched five input names (Barker, Burns, Klicka, Lanyon, & Lovette,

³⁰² 2015; Hedges, Marin, Suleski, Paymer, & Kumar, 2015; Jetz et al., 2012), one study matched

³⁰³ four input names (Hooper & Price, 2017) and two studies matched two input names (Barker,

³⁰⁴ Burns, Klicka, Lanyon, & Lovette, 2013; Burns et al., 2014). No studies matched all input

³⁰⁵ names. Together, source chronograms provide 28 unique age data points, covering all nodes

³⁰⁶ on our chosen tree topology to date (Table 1).

307   ***Summarizing search results.*** – DateLife obtained OpenTree's synthetic tree

308 topology for these taxa (Fig. 2d), and congruified and mapped age data to nodes in this

309 chosen topology, shown in Table 1. The name processing step allowed including five data

310 points for node "n4" (parent of *Schoeniclus elegans*) that would not have had any data

311 otherwise due to name mismatch. Age summary statistics per node were calculated (Table 2)

312 and used as calibrations to date the tree topology using the BLADJ algorithm. As expected,

313 more inclusive nodes (e.g., node "n1") have more variance in age data than less inclusive

314 nodes (e.g., node "n5"). Median summary age data for node "n2" was excluded as final

315 calibration because it is older than the median age of a more inclusive node, "n1" (Fig. 2g).

An Example with the Family of True Finches

317   ***Creating a query.*** – To obtain ages for all species within the family of true finches,

318 Fringillidae, we ran a DateLife query using the "get species from taxon" flag, which gets all

319 recognized species names within a named group from a taxonomy of choice. Following the

320 NCBI taxonomy, our DateLife query has 289 Fringillidae species names. This

321 taxon-constrained approach implies that the full DateLife analysis will be performed using a

322 tree topology and ages available for species names from a given taxonomic group, which do

323 not necessarily correspond to a monophyletic group. Users can change this behavior by

324 providing all species names corresponding to a monophyletic group as input for a DateLife

325 search, or a monophyletic tree to construct a DateLife summary.

326   ***Searching the database.*** – Next, we used the processed species names in our

327 DateLife query to identify chronograms with at least two Fringillidae species as tip taxa.

328 The DateLife search identified 19 chronograms matching this criteria, published in 13

329 different studies (Barker et al., 2013, 2015; Burns et al., 2014; Claramunt & Cracraft, 2015;

330 Gibb et al., 2015; Hedges et al., 2015; Hooper & Price, 2017; Jetz et al., 2012; Kimball et al.,

331 2019; Oliveros et al., 2019; Price et al., 2014; Roquet, Lavergne, & Thuiller, 2014; Uyeda,

332 Pennell, Miller, Maia, & McClain, 2017). Once identified, DateLife pruned these matching

333 chronograms to remove tips that do not belong to the queried taxon names, and transformed

334 these pruned chronograms to pairwise distance matrices, revealing 1,206 different age data

335 points available for species within the Fringillidae (Supplementary Table S1).

336   ***Summarizing search results.*** – The final step entailed congruifying and

337 summarizing the age data available for the Fringillidae species into two single summary

338 chronograms, using two different types of summary ages, median and SDM. As explained in

339 the "Description" section, a tree topology to summarize age data upon is required. By

340 default, DateLife uses the topology from OpenTree's synthetic tree that contains all taxa

341 from the search query. According to OpenTree's synthetic tree, species belonging to the

342 family Fringillidae do not form a monophyletic group (Supplementary Fig. S1). Hence, a

343 topology containing only the 289 species from the original query was extracted from Open

344 Tree of Life's synthetic tree v12.3 (Supplementary Fig. S2; Open Tree Of Life et al., 2019).

345   All 19 source chronograms (Supplementary Figs. S5-S23) were congruified to

346 OpenTree's topology shown in Supplementary Figure S2, reducing the original 1,206 node

347 age data set to 818 different data points (Supplementary Table S2) that could be used as

348 calibrations for that chosen topology. The congruent node age data points were summarized

349 for each node, resulting in 194 summary node ages. From these 21 were excluded as

350 secondary calibrations because they were older than the ancestral node. The remaining 173

351 summary node ages were used as secondary calibrations to obtain a fully dated (and

352 resolved) phylogeny with the program BLADJ (Supplementary Figure S3). Results for a

353 subgroup are shown in Figure 3.

Cross-Validation Test

355   We performed a cross validation test of a DateLife analysis using the Fringillidae

356 source chronograms obtained above (Supplementary Figs. S5-S23). As inputs for a DateLife

357 analysis, we used all individual tree topologies from each of the 19 source chronograms from

358 13 studies, treating their node ages as unknown. We congruified node ages extracted from

359 chronograms from all other studies upon the individual topologies, effectively excluding

360 original ages from each topology. Finally, average node ages per node were applied as

361 secondary calibrations and smoothed with the BLADJ algorithm. We found that node ages

362 from the original studies, and ages estimated using all other age data available are generally

363 correlated (Fig. 4). For five studies, DateLife tended to underestimate ages for topologically

364 deeper nodes (those with many descendant taxa, aka "closer to the root") relative to the

365 original estimate, and overestimate ages for nodes closer to the tips. Accordingly, root ages

366 are generally older in the original study than estimated using cross-validated ages

367 (Supplementary Fig. S4). In general, topologically deeper nodes display the largest age

368 variation between node ages from the original chronograms and ages summarized with

369 DateLife.

                                    DISCUSSION

371   DateLife's goal is to improve availability, accessibility, and reusability of

372 state-of-the-art data on evolutionary time frame of organisms, to allow users from all areas of

373 science and with all levels of expertise to compare, use and reanalyse expert age data for

374 their own applications. As such, it is designed as an open service that does not require any

375 expert biological knowledge –besides the scientific names of the species or group that users

376 want to work with– to use any of its functionalities.

377   A total of 99,474 unique terminal taxa are represented in DateLife's database.

378 Incorporation of more chronograms into the database will continue to improve DateLife's

379 services. One option to increase the number of chronograms in the DateLife database is the

380 Dryad data repository. Methods to automatically mine chronograms from Dryad could be

381 designed and implemented. However, Dryad's metadata system has no information to

382 automatically detect branch length units, and those would still need to be determined

383 manually by a human curator. We would like to emphasize on the importance of sharing

384 chronogram data, including systematically curated metadata, into open repositories, such as

385 OpenTree's Phylesystem (McTavish et al., 2015) for the benefit of research and the scientific

386 community as a whole. Another important source of expert data on time of lineage

387 divergence is TimeTree's database (Hedges, Dudley, & Kumar, 2006), which holds

388 chronograms from more than 4k published studies, and is fully browsable using its graphical

389 user interface (timetree.org). TimeTree's chronogram database was not accessible in

390 computer readable format until very recently (Kumar et al., 2022), when its terms of use and

391 website application were updated, now allowing some kinds of reuse, but not redistribution.

392   The inaccessibility of TimeTree's database was an inspiration for the DateLife project,

393 which was born as a prototype tool initially developed over a series of hackathons at the

National Evolutionary Synthesis Center, NC, USA (Stoltzfus et al., 2013), as the need to make scientific information that is funded by the public practically available to the public was acknowledged and prioritized.

As we envision that DateLife will have many interesting applications in research and beyond, we emphasize that DateLife's results –as well as any insights gleaned from them, largely depend on the quality of the source chronograms: low quality chronograms will produce low quality results. The "garbage in, garbage out" problem has long been recognised in supertree methods for summarizing phylogenetic trees (Bininda-Emonds et al., 2004). We note that this is a surfacing issue of any automated tool for biological data analysis. For example, DNA riddled with sequencing errors will produce generally poor alignments that will return biased evolutionary hypothesis, independently of the quality of the analysis software used. Again, we urge readers and DateLife users to explore all input chronograms before using a summary chronogram resulting from a DateLife workflow.

Finally, uncertainty and variability of chronogram node age estimates might pose larger issues in some research areas than others. For example, in ecological and conservation biology studies, it has been shown that incorporating some chronogram data provides better results than when not using any age data at all, even if the node ages are not the best quality (Webb et al., 2008). In the following sections we discuss the particularities of divergence times from DateLife's summary chronograms and their impact on certain evolutionary analyses, for consideration of the readers and users in different research areas.

Age Variation in Source Chronograms

Conflict in estimated ages among alternative studies is common in the literature. See, for example, the robust ongoing debate about crown group age of angiosperms (Barba-Montoya, Reis, Schneider, Donoghue, & Yang, 2018; Magallón, Gómez-Acevedo, Sánchez-Reyes, & Hernández-Hernández, 2015; Ramshaw et al., 1972; Sanderson & Doyle,

2001; Sauquet, Ramírez-Barahona, & Magallón, 2021). Alternative source chronograms available for the same taxa have potentially been estimated implementing different types of calibrations, which affects the resulting node age estimates. For example, in the DateLife analysis of the Fringillidae shown above, the chronograms from one study (Burns et al., 2014) were inferred using molecular substitution rate estimates across birds (Weir & Schluter, 2008), and have much older age estimates for the same nodes than chronograms that were inferred using fossil calibrations (Figs. 3, 4c, Supplementary Figs. S4c, S10). Another source of conflict in estimated node ages can arise from different placements for the same calibration, which would imply fundamentally distinct evolutionary hypotheses (Antonelli et al., 2017). For example, two independent researchers working on the same clade should both carefully select and justify their choices of fossil calibration placement. Yet, if one researcher concludes that a fossil should calibrate the ingroup of a clade, while another researcher concludes that the same fossil should calibrate the outgroup of the clade, the resulting age estimates will differ, as the placement of calibrations as stem or crown group is known to significantly affect estimates of time of lineage divergence (Sauquet, 2013). Finally, placement of calibrations also affects uncertainty of node age estimates. For example, nodes that are sandwiched between a calibrated node and a calibrated root have less freedom of movement and hence narrower confidence intervals (Vos & Mooers, 2004), which inflates precision for nodes without calibrations but does not necessarily improve accuracy of the estimated ages.

DateLife's summary chronograms are intended to represent all variation in estimated node ages from source chronograms. Node age distribution ranges allow to visually explore ages from source chronograms individually and contextualize and compare them against other chronograms. Researchers that wish to use summary chronograms in downstream evolutionary analysis may select multiple trees sampled from the summary distribution of node ages, to account for variation in source chronograms.

Primary vs Secondary Calibrations

⁴⁴⁵ DateLife constructs summary chronograms using node ages extracted from existing

⁴⁴⁶ chronograms, i.e. secondary calibrations. In general, the scientific community has more

⁴⁴⁷ confidence in chronograms using primary calibrations, where the dated tree is generated from

⁴⁴⁸ a single analysis where carefully chosen fossil calibrations are the source of absolute time

⁴⁴⁹ information, than in analyses dated using secondary calibrations (Antonelli et al., 2017;

⁴⁵⁰ Garzón-Orduña, Silva-Brandão, Willmott, Freitas, & Brower, 2015; Graur & Martin, 2004;

⁴⁵¹ Sauquet, 2013; Sauquet et al., 2012; Schenk, 2016; Shaul & Graur, 2002). However,

⁴⁵² implementation of primary calibrations is difficult: it requires specialized expertise and

⁴⁵³ training to discover, place and apply calibrations appropriately (Hipsley & Müller, 2014;

⁴⁵⁴ Ksepka et al., 2011). One approach is to use fossils that have been widely discussed and

⁴⁵⁵ previously curated as calibrations to date other trees (Ksepka et al., 2011; Sauquet, 2013),

⁴⁵⁶ and making sure that all data reflect a coherent evolutionary history (Sauquet, 2013), as for

⁴⁵⁷ example done by Antonelli et al. (2017). The Fossil Calibration Database provides data for

⁴⁵⁸ 220 primary calibration points encompassing flowering plants and metazoans, that have been

⁴⁵⁹ curated by experts and used for dating analysis in peer-reviewed publications (Ksepka et al.,

⁴⁶⁰ 2015). This database facilitates the use of expert primary fossil calibrations in new

⁴⁶¹ phylogenetic dating analyses. Yet, users still require the expertise to locate and calibrate

⁴⁶² appropriate nodes in their phylogenies which correspond with fossils available in the

⁴⁶³ database.

⁴⁶⁴ Recently, Powell, Waskin, and Battistuzzi (2020) showed in a simulation study that

⁴⁶⁵ secondary calibrations using node ages based on previous molecular clock analyses can be as

⁴⁶⁶ good as primary calibrations. Using several secondary calibrations (as opposed to just one)

⁴⁶⁷ can provide sufficient information to alleviate or even neutralize potential biases (Graur &

⁴⁶⁸ Martin, 2004; Sauquet, 2013; Shaul & Graur, 2002). Our cross validation analysis also

⁴⁶⁹ provides insight into the application of secondary calibrations. Node ages summarized with

⁴⁷⁰ DateLife and those from the original studies are well correlated (Supplementary Figs.

⁴⁷¹ S5-S23). We also note that DateLife estimates for nodes closer to the root tend to be slightly

younger than ages from the original studies. In contrast, nodes closer to the tips tend to be slightly older when estimated using our secondary calibrations than ages from the original studies. The only exception to this trend was observed in Burns et al. (2014) chronogram, which generally displays much younger node ages when estimated using secondary calibrations than the original study (Supplementary Figs. S4c, S10), supporting previous observations (Sauquet et al., 2012; Schenk, 2016). However, these younger dates are more likely an example of how multiple secondary calibrations can correct erroneous estimates, as dates on the Burns et al. (2014) tree were obtained using a single secondary calibration based on a previously estimated molecular evolution rate across birds from Weir and Schluter (2008), and appear as major outliers compared to alternate estimates for the same nodes based on primary fossil calibrations (Fig. 3, Supplementary Fig. S3).

Sumarizing Chronograms

By default, DateLife currently summarizes all source chronograms that overlap with at least two species names. Users can exclude source chronograms if they have reasons to do so. Strictly speaking, a good chronogram should reflect the real time of lineage divergence accurately and precisely. To our knowledge, there are no tested measures to determine independently when a chronogram is better than another. Yet, several characteristics of the data used for dating analyses, as well as from the output chronogram itself, could be used to score the quality of source chronograms.

Some measures that have been proposed are the proportion of lineage sampling and the number of calibrations used (Magallón, 2010; Magallón et al., 2015). Some characteristics that are often cited in published studies as a measure of improved age estimates as compared to previously published estimates are: quality of alignment (missing data, GC content), lineage sampling (strategy and proportion), phylogenetic and dating inference method, number of fossils used as calibrations, support for nodes and ages, and magnitude of confidence intervals.

498 DateLife provides an opportunity to capture concordance and conflict among date

499 estimates, which can also be used as a metric for chronogram reliability. Its open database of

500 chronograms allows other researchers to do such analyses themselves reproducibly, and

501 without needing permission. Though, of course, they should follow proper citation practices,

502 especially for the source chronogram studies.

503 The exercise of summarizing age data from across multiple studies is a common

504 resource in research, as it provides the opportunity to work with a chronogram that reflects a

505 unified evolutionary history for a lineage, by putting together evidence from different

506 hypotheses. For example, the largest, and taxonomically broadest chronogram currently

507 available from OpenTree was constructed summarizing age data from 2,274 published

508 chronograms using NCBI's taxonomic tree as backbone (Hedges et al., 2015), which has been

509 widely reused for research. Finally, we note that summarizing chronograms should be done

510 with caution, as it may amplify the effect of uncertainty and errors in source data, and blur

511 parts of the evolutionary history of a lineage that might only be reflected in source

512 chronograms and lost on the summary chronogram (Sauquet et al., 2021).

513 *Effects of Taxon Sampling on Downstream Analyses*

514 Analysis of species diversification of simulated and empirical phylogenies suggest that

515 using a more completely sampled phylogeny provides estimates that are closer to the true

516 diversification history than when analysing incompletely sampled phylogenies (Chang,

517 Rabosky, & Alfaro, 2020; Cusimano, Stadler, & Renner, 2012; Sun et al., 2020). Ideally,

518 phylogenies should be completed using genetic data, but this is a time-consuming and

519 difficult task to achieve for many biological groups. Hence, DateLife's workflow features

520 different ways of assigning divergence times to taxa with missing the absence of branch

521 length data and calibrations and branch lengths for certain taxa.

522 Completing a phylogeny using a stochastic birth-death polytomy resolver and a

backbone taxonomy is a common practice in scientific publications: Jetz et al. (2012),
created a chronogram of all 9,993 bird species, where 67% had molecular data and the rest
was simulated; Rabosky et al. (2018) created a chronogram of 31,536 ray-finned fishes, of
which only 37% had molecular data; Smith and Brown (2018) constructed a chronogram of
353,185 seed plants where only 23% had molecular data. Stochastically resolved chronograms
can return diversification rates estimates that appear less biased than those estimated from
their incompletely sampled counterparts, even with methods that account for missing
lineages by using sampling fractions (Chang et al., 2020; Cusimano et al., 2012), but can also
introduce spurious patterns of early bursts of diversification (Cusimano & Renner, 2010; Sun
et al., 2020).

Taxonomy-based stochastic polytomy resolvers also introduce topological differences in
phylogenetic trees. The study of macroevolutionary processes largely depends on an
understanding of the timing of species diversification events, and different phylogenetic and
chronogram hypothesis can provide very different overviews of the macroevolutionary history
of a biological group. For example, alternative topologies in chronograms from the same
biological group can infer very different species diversification patterns (Rabosky, 2015; Title
& Rabosky, 2016). Similarly, there are worries that patterns of morphological evolution
cannot be accurately inferred with phylogenies that have been resolved stochastically over a
taxonomic backbone, as any patterns would be erased by randomization (Rabosky, 2015).
We note that the same applies for geography- and morphology-dependent diversification
analysis. Hence, we suggest that phylogenies that have been processed with taxonomy-based
stochastic polytomy resolvers, including certain summary chronograms from a DateLife
analysis, can be useful as null or neutral models, representing the case of a diversification
process that is independent of traits and geographical scenario.

Taxonomy-based stochastic polytomy resolvers have been used to advance research in
evolution, still, risks come with this practice. Taken to the extreme, one could generate a

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

550 single calibration, and assigning branch lengths following a birth-death diversification model.

551 Clearly, this can lead to a misrepresentation of the true evolutionary history. We urge

552 DateLife users to follow the example of the large tree papers cited above, by carefully

553 considering the statistical assumptions being made, potential biases, and assessing the

554 consistency of DateLife's results with prior work.

Conclusions

556 Knowledge of the evolutionary time frame of organisms is key to many research areas:

557 trait evolution, species diversification, biogeography, macroecology and more. It is also

558 crucial for education, science communication and policy, but generating chronograms is

559 difficult, especially for those who want to use phylogenies but who are not systematists, or

560 do not have the time to acquire and develop the necessary knowledge and skills to construct

561 them on their own. Importantly, years of primarily publicly funded research have resulted in

562 vast amounts of chronograms that are already available in scientific publications, but

563 functionally hidden from the public and scientific community for reuse.

564 The DateLife project allows for easy and fast summarization of public and

565 state-of-the-art data on time of lineage divergence. It is available as an R package, and as a

566 web-based R shiny application at www.datelife.org. DateLife provides a straightforward way

567 to get an informed picture of the state of knowledge for the time frame of evolution of

568 different regions of the tree of life, and allows identifying regions that require more research,

569 or that have conflicting information. Additionally, both summary and newly generated trees

570 using the DateLife workflow are useful to evaluate evolutionary hypotheses in different areas

571 of research. We hope that the DateLife project will increase awareness of the existing

572 variation in expert estimations of time of divergence, and foster exploration of the effect of

573 alternative divergence time hypotheses on the results of analyses, nurturing a culture of more

574 cautious interpretation of evolutionary results.

₅₇₅   AVAILABILITY

₅₇₆   The DateLife software is free and open source. It can be used online through its R

₅₇₇ shiny web application hosted at http://www.datelife.org, and locally through the `datelife`

₅₇₈ R package, available from Zenodo (https://doi.org/10.5281/zenodo.593938 and the CRAN

₅₇₉ repository (Sanchez-Reyes et al., 2022). DateLife's web application is maintained using

₅₈₀ RStudio's shiny server and the shiny package open infrastructure, as well as Docker and

₅₈₁ OpenTree's infrastructure (datelife.opentreeoflife.org). `datelife`'s stable version can be

₅₈₂ installed from the CRAN repository using the command `install.packages(pkgs =`

₅₈₃ `"datelife")` from within R. Development versions are available from DateLife's GitHub

₅₈₄ repository (https://github.com/phylotastic/datelife) and can be installed using the

₅₈₅ command `devtools::install_github("phylotastic/datelife")`.

₅₈₆   SUPPLEMENTARY MATERIAL

₅₈₇   Supplementary material, including figures, tables, code, biological examples,

₅₈₈ benchmark results, data files and online-only appendices, can be viewed and downloaded

₅₈₉ from the Dryad data repository (https://doi.org/10.5061/dryad.cnp5hqc6w), as well as from

₅₉₀ the Zenodo stable repositories that host the reproducible manuscript

₅₉₁ (https://doi.org/10.5281/zenodo.7435094), the biological examples

₅₉₂ (https://doi.org/10.5281/zenodo.7435101), the software benchmark

₅₉₃ (https://doi.org/10.5281/zenodo.7435106), and the figures

₅₉₄ (https://doi.org/10.5281/zenodo.6683667). Dryad's data publication fee is covered by the

₅₉₅ Society of Systematic Biologists. Development versions corresponding to all of the above are

₅₉₆ hosted on GitHub, accessible at https://github.com/LunaSare/datelifeMS1,

₅₉₇ https://github.com/LunaSare/datelife_examples, and

₅₉₈ https://github.com/LunaSare/datelife_benchmark.

₅₉₉   FUNDING

References

Agnarsson, I., & Miller, J. A. (2008). Is ACCTRAN better than DELTRAN? *Cladistics*, *24*(6), 1032–1038.

Alström, P., Hooper, D. M., Liu, Y., Olsson, U., Mohan, D., Gelang, M., . . . Price, T. D. (2014). Discovery of a relict lineage and monotypic family of passerine birds. *Biology Letters*, *10*(3), 20131067.

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 {https://evolution.genetics.washington.edu/phylip/newicktree.html}

Avibase. (2022). Yellow-throated Bunting. *Avibase - the World Bird Database*, (Online Resource). Retrieved from {https://avibase.bsc-eoc.org/species.jsp?lang=EN&avibaseid=82D1EE0049D8D927}

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. (2014). Mitogenomic data resolve basal relationships among passeriform and passeridan birds. *Molecular Phylogenetics and Evolution, 79*, 313–324.

Barker, F. K., Burns, K. J., Klicka, J., Lanyon, S. M., & Lovette, I. J. (2013). 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.

Barker, F. K., Cibois, A., Schikler, P., Feinstein, J., & Cracraft, J. (2004). Phylogeny and diversification of the largest avian radiation. *Proceedings of the National Academy of Sciences, 101*(30), 11040–11045.

Beresford, P., Barker, F., Ryan, P., & Crowe, T. (2005). African endemics span the tree of songbirds (passeri): Molecular systematics of several evolutionary "enigmas". *Proceedings of the Royal Society B: Biological Sciences, 272*(1565), 849–858.

Bininda-Emonds, O. R., Jones, K. E., Price, S. A., Cardillo, M., Grenyer, R., & Purvis, A. (2004). Garbage in, garbage out: Data issues in supertree construction. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, 267–280.

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., & Hochreiter, S. (2015). Msa: An r package for multiple sequence alignment. *Bioinformatics, 31*(24), 3997–3999.

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

Bryson Jr, R. W., Chaves, J., Smith, B. T., Miller, M. J., Winker, K., Pérez-Emán, J. L., &

Klicka, J. (2014). Diversification across the new world within the 'blue'cardinalids

(aves: Cardinalidae). *Journal of Biogeography*, *41*(3), 587–599.

Burleigh, J. G., Kimball, R. T., & Braun, E. L. (2015). Building the avian tree of life using a

large-scale, sparse supermatrix. *Molecular Phylogenetics and Evolution*, *84*, 53–63.

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

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

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

and Evolution*, *75*, 41–77.

Chamberlain, S. (2018). *bold: Interface to Bold Systems API*. Retrieved from

https://CRAN.R-project.org/package=bold

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

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

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

Chang, J., Rabosky, D. L., & Alfaro, M. E. (2020). Estimating diversification rates on

incompletely sampled phylogenies: Theoretical concerns and practical solutions.

Systematic Biology, *69*(3), 602–611.

Chaves, J. A., Hidalgo, J. R., & Klicka, J. (2013). Biogeography and evolutionary history of

the n eotropical genus s altator (a ves: T hraupini). *Journal of Biogeography*, *40*(11), 2180–2190.

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., & 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

Cusimano, N., & Renner, S. S. (2010). Slowdowns in diversification rates from real phylogenies may not be real. *Systematic Biology*, *59*(4), 458–464.

Cusimano, N., Stadler, T., & Renner, S. S. (2012). A new method for handling missing species in diversification analysis applicable to randomly or nonrandomly sampled phylogenies. *Systematic Biology*, *61*(5), 785–792.

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

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

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

Forest, F., Savolainen, V., Chase, M. W., Lupia, R., Bruneau, A., & Crane, P. R. (2005).

708   Teasing apart molecular-versus fossil-based error estimates when dating phylogenetic

709   trees: A case study in the birch family (betulaceae). *Systematic Botany*, *30*(1),

710   118–133.

711 Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative

712   data: A test and review of evidence. *The American Naturalist*.

713 Garzón-Orduña, I. J., Silva-Brandão, K. L., Willmott, K. R., Freitas, A. V., & Brower, A. V.

714   (2015). Incompatible ages for clearwing butterflies based on alternative secondary

715   calibrations. *Systematic Biology*, *64*(5), 752–767.

716 GBIF Secretariat. (2022). GBIF Backbone Taxonomy. *Checklist dataset*, (Online Resource

717   accessed via GBIF.org). Retrieved from {https://doi.org/10.15468/39omei }

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

719   B. J., . . . Penny, D. (2015). New zealand passerines help clarify the diversification of

720   major songbird lineages during the oligocene. *Genome Biology and Evolution*, *7*(11),

721   2983–2995.

722 Graur, D., & Martin, W. (2004). Reading the entrails of chickens: Molecular timescales of

723   evolution and the illusion of precision. *TRENDS in Genetics*, *20*(2), 80–86.

724 Hackett, S. J., Kimball, R. T., Reddy, S., Bowie, R. C., Braun, E. L., Braun, M. J., . . .

725   others. (2008). A phylogenomic study of birds reveals their evolutionary history.

726   *Science*, *320*(5884), 1763–1768.

727 Harvey, P. H., Pagel, M. D., & others. (1991). *The comparative method in evolutionary

728   biology* (Vol. 239). Oxford university press Oxford.

729 Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: A public knowledge-base of

730   divergence times among organisms. *Bioinformatics*, *22*(23), 2971–2972.

731   https://doi.org/10.1093/bioinformatics/btl505

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

733   clock-like speciation and diversification. *Molecular Biology and Evolution*, *32*(4),

734   835–845. https://doi.org/10.1093/molbev/msv037

735   Heibl, C. (2008). *PHYLOCH: R language tree plotting tools and interfaces to diverse*

736   *phylogenetic software packages.* Retrieved from

737   http://www.christophheibl.de/Rpackages.html

738   Hipsley, C. A., & Müller, J. (2014). Beyond fossil calibrations: Realities of molecular clock

739   practices in evolutionary biology. *Frontiers in Genetics*, *5*, 138.

740   Hooper, D. M., & Price, T. D. (2017). Chromosomal inversion differences correlate with

741   range overlap in passerine birds. *Nature Ecology & Evolution*, *1*(10), 1526.

742   Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic

743   trees. *Bioinformatics*, *17*(8), 754–755.

744   https://doi.org/10.1093/bioinformatics/17.8.754

745   Jetz, W., Thomas, G., Joy, J. J., Hartmann, K., & Mooers, A. (2012). The global diversity

746   of birds in space and time. *Nature*, *491*(7424), 444–448.

747   https://doi.org/10.1038/nature11631

748   Johansson, U. S., Fjeldså, J., & Bowie, R. C. (2008). Phylogenetic relationships within

749   passerida (aves: Passeriformes): A review and a new molecular phylogeny based on

750   three nuclear intron markers. *Molecular Phylogenetics and Evolution*, *48*(3), 858–876.

751   Katoh, K., Asimenos, G., & Toh, H. (2009). Multiple alignment of dna sequences with mafft.

752   In *Bioinformatics for dna sequence analysis* (pp. 39–64). Springer.

753   Kimball, R. T., Oliveros, C. H., Wang, N., White, N. D., Barker, F. K., Field, D. J., . . .

others. (2019). A phylogenomic supertree of birds. *Diversity*, *11*(7), 109.

Klicka, J., Barker, F. K., Burns, K. J., Lanyon, S. M., Lovette, I. J., Chaves, J. A., & Bryson
Jr, R. W. (2014). A comprehensive multilocus assessment of sparrow (aves:
Passerellidae) relationships. *Molecular Phylogenetics and Evolution*, *77*, 177–182.

Ksepka, D. T., Benton, M. J., Carrano, M. T., Gandolfo, M. A., Head, J. J., Hermsen, E. J.,
... others. (2011). *Synthesizing and databasing fossil calibrations: Divergence dating
and beyond*. The Royal Society.

Ksepka, D. T., Parham, J. F., Allman, J. F., Benton, M. J., Carrano, M. T., Cranston, K.
A., ... others. (2015). The fossil calibration database—a new resource for divergence
dating. *Systematic Biology*, *64*(5), 853–859.

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., ... Hedges,
S. B. (2022). TimeTree 5: An expanded resource for species divergence times.
Molecular Biology and Evolution, *39*(8), msac174.

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio,
A., ... others. (2015). Evolution of darwin's finches and their beaks revealed by
genome sequencing. *Nature*, *518*(7539), 371–375.

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

Lepage, D. (2004). *Avibase: The world bird database*. Bird Studies Canada.

Lepage, D., Vaidya, G., & Guralnick, R. (2014). Avibase–a database system for managing
and organizing taxonomic concepts. *ZooKeys*, (420), 117.

Lerner, H. R., Meyer, M., James, H. F., Hofreiter, M., & Fleischer, R. C. (2011). Multilocus
resolution of phylogeny and timescale in the extant adaptive radiation of hawaiian

777   honeycreepers. *Current Biology*, *21*(21), 1838–1844.

778 Lovette, I. J., Pérez-Emán, J. L., Sullivan, J. P., Banks, R. C., Fiorentino, I.,

779   Córdoba-Córdoba, S., . . . others. (2010). A comprehensive multilocus phylogeny for

780   the wood-warblers and a revised classification of the parulidae (aves). *Molecular*

781   *Phylogenetics and Evolution*, *57*(2), 753–770.

782 Magallon, S., & Sanderson, M. (2001). Absolute diversification rates in angiosperm clades.

783   *Evolution*, *55*(9), 1762–1780.

784 Magallón, S. (2010). Using fossils to break long branches in molecular dating: A comparison

785   of relaxed clocks applied to the origin of angiosperms. *Systematic Biology*, *59*(4),

786   384–399.

787 Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L., & Hernández-Hernández, T. (2015).

788   A metacalibrated time-tree documents the early rise of flowering plant phylogenetic

789   diversity. *New Phytologist*, *207*(2), 437–453.

790 McTavish, E. J., Hinchliff, C. E., Allman, J. F., Brown, J. W., Cranston, K. A., Holder, M.

791   T., . . . Smith, S. (2015). Phylesystem: A git-based data store for community-curated

792   phylogenetic estimates. *Bioinformatics*, *31*(17), 2794–2800.

793 Michonneau, F., Brown, J. W., & Winter, D. J. (2016). rotl: an R package to interact with

794   the Open Tree of Life data. *Methods in Ecology and Evolution*, *7*(12), 1476–1481.

795   https://doi.org/10.1111/2041-210X.12593

796 Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*,

797   *17*(4), 508–525. https://doi.org/10.1111/ele.12251

798 Moyle, R. G., Oliveros, C. H., Andersen, M. J., Hosner, P. A., Benz, B. W., Manthey, J. D.,

799   . . . Faircloth, B. C. (2016). Tectonic collision and uplift of wallacea triggered the

global songbird radiation. *Nature Communications*, *7*(1), 1–7.

Oliveros, C. H., Field, D. J., Ksepka, D. T., Barker, F. K., Aleixo, A., Andersen, M. J., . . . others. (2019). Earth history and the passerine superradiation. *Proceedings of the National Academy of Sciences*, *116*(16), 7916–7925.

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 {https://github.com/OpenTreeOfLife/germinator/wiki/Open-Tree-of-Life-Web-APIs}

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

Ödeen, A., Håstad, O., & Alström, P. (2011). Evolution of ultraviolet vision in the largest avian radiation-the passerines. *BMC Evolutionary Biology*, *11*(1), 1–8.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289–290.

Parchman, T. L., Benkman, C. W., & Mezquida, E. T. (2007). Coevolution between hispaniolan crossbills and pine: Does more time allow for greater phenotypic escalation at lower latitude? *Evolution*, *61*(9), 2142–2153.

Päckert, M., Martens, J., Sun, Y.-H., Severinghaus, L. L., Nazarenko, A. A., Ting, J., . . . Tietze, D. T. (2012). Horizontal and elevational phylogeographic patterns of

himalayan and southeast asian forest passerines (aves: Passeriformes). *Journal of Biogeography*, *39*(3), 556–573.

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., . . . Harmon, L. J. (2014). Geiger v2. 0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*(15), 2216–2218.

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.

Powell, A. F., Barker, F. K., Lanyon, S. M., Burns, K. J., Klicka, J., & Lovette, I. J. (2014). A comprehensive species-level molecular phylogeny of the new world blackbirds (icteridae). *Molecular Phylogenetics and Evolution*, *71*, 94–112.

Powell, C. L. E., Waskin, S., & Battistuzzi, F. U. (2020). Quantifying the error of secondary vs. Distant primary calibrations in a simulated environment. *Frontiers in Genetics*, *11*, 252.

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.

Pulgarín-R, P. C., Smith, B. T., Bryson Jr, R. W., Spellman, G. M., & Klicka, J. (2013). Multilocus phylogeny and biogeography of the new world pheucticus grosbeaks (aves: Cardinalidae). *Molecular Phylogenetics and Evolution*, *69*(3), 1222–1227.

Rabosky, D. L. (2015). No substitute for real data: A cautionary note on the use of phylogenies from birth–death polytomy resolvers for downstream comparative analyses. *Evolution*, *69*(12), 3207–3216.

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.

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

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

Rees, Vandepitte, L., Decock, W., & Vanhoorne, B. (2017). IRMNG 2006–2016: 10 Years of a Global Taxonomic Database. *Biodiversity Informatics*, *12*.

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

Roquet, C., Lavergne, S., & Thuiller, W. (2014). One tree to link them all: A phylogenetic dataset for the european tetrapoda. *PLoS Currents*, *6*.

Sanchez-Reyes, L. L., & O'Meara, B. (2022). `datelifeplot`: Methods to plot chronograms and outputs of the `datelife` package. *R Package Release V0.2.2.* Retrieved from https://zenodo.org/badge/latestdoi/381501451

869 Sanchez-Reyes, L. L., O'Meara, B., Eastman, J., Heath, T., Wright, A., Schliep, K., . . .

870   Alfaro, M. (2022). `datelife`: Scientific Data on Time of Lineage Divergence for Your

871   Taxa. In *R package version 0.6.6.* Retrieved from

872   https://CRAN.R-project.org/package=datelife and

873   https://doi.org/10.5281/zenodo.593938

874 Sanderson, M. (2002). Estimating Absolute Rates of Molecular Evolution and Divergence

875   Times: A Penalized Likelihood Approach. *Molecular Biology and Evolution*, *19*(1),

876   101–109. https://doi.org/10.1093/oxfordjournals.molbev.a003974

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

878   the age of angiosperms from rbcL and 18S rDNA data. *American Journal of Botany*,

879   *88*(8), 1499–1516.

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

881   355–367.

882 Sauquet, H., Ho, S. Y. W., Gandolfo, M. a, Jordan, G. J., Wilf, P., Cantrill, D. J., . . .

883   Udovicic, F. (2012). Testing the impact of calibration on molecular divergence times

884   using a fossil-rich group: the case of Nothofagus (Fagales). *Systematic Biology*, *61*(2),

885   289–313. https://doi.org/10.1093/sysbio/syr116

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

887   *unknown.*

888 Schenk, J. J. (2016). Consequences of secondary calibrations on divergence time estimates.

889   *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0148228

890 Schliep, K. P. (2011). Phangorn: Phylogenetic analysis in r. *Bioinformatics*, *27*(4), 592–593.

891 Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., . . .

others. (2020). NCBI Taxonomy: a Comprehensive Update on Curation, Resources
and Tools. *Database*, *2020*.

Selvatti, A. P., Gonzaga, L. P., & Moraes Russo, C. A. de. (2015). A paleogene origin for
crown passerines and the diversification of the oscines in the new world. *Molecular
Phylogenetics and Evolution*, *88*, 1–15.

Shaul, S., & Graur, D. (2002). Playing chicken (gallus gallus): Methodological
inconsistencies of molecular divergence date estimates due to secondary calibration
points. *Gene*, *300*(1-2), 59–61.

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

Smith, S., & O'Meara, B. (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

Sun, M., Folk, R. A., Gitzendanner, M. A., Soltis, P. S., Chen, Z., Soltis, D. E., & Guralnick,
R. P. (2020). Estimating rates and patterns of diversification with incomplete
sampling: A case study in the rosids. *American Journal of Botany*, *107*(6), 895–909.

Tietze, D. T., Päckert, M., Martens, J., Lehmann, H., & Sun, Y.-H. (2013). Complete
phylogeny and historical biogeography of true rosefinches (aves: Carpodacus).
Zoological Journal of the Linnean Society, *169*(1), 215–234.

Title, P. O., & Rabosky, D. L. (2016). Do Macrophylogenies Yield Stable Macroevolutionary

915   Inferences? An Example from Squamate Reptiles. *Systematic Biology*, syw102.

916   https://doi.org/10.1093/sysbio/syw102

917   Treplin, S., Siegert, R., Bleidorn, C., Thompson, H. S., Fotso, R., & Tiedemann, R. (2008).

918   Molecular phylogeny of songbirds (aves: Passeriformes) and the relative utility of

919   common nuclear marker loci. *Cladistics*, *24*(3), 328–349.

920   Uyeda, J. C., Pennell, M. W., Miller, E. T., Maia, R., & McClain, C. R. (2017). The

921   evolution of energetic scaling across the vertebrate tree of life. *The American*

922   *Naturalist*, *190*(2), 185–199.

923   Vos, R. A., Balhoff, J. P., Caravas, J. A., Holder, M. T., Lapp, H., Maddison, W. P., . . .

924   others. (2012). NeXML: Rich, extensible, and verifiable representation of

925   comparative data and metadata. *Systematic Biology*, *61*(4), 675–689.

926   https://doi.org/10.1093/sysbio/sys025

927   Vos, R. A., & Mooers, A. Ø. (2004). Reconstructing divergence times for supertrees: A

928   molecular approach. *Phylogenetic Supertrees: Combining Information to Reveal the*

929   *Tree of Life*, 281–299.

930   Webb, C. (2000). Exploring the Phylogenetic Structure of Ecological Communities : An

931   Example for Rain Forest Trees. *The American Naturalist*, *156*(2), 145–155.

932   Webb, C., Ackerly, D., & Kembel, S. (2008). Phylocom: Software for the analysis of

933   phylogenetic community structure and trait evolution. *Bioinformatics*, *24*(18),

934   2098–2100. https://doi.org/10.1093/bioinformatics/btn358

935   Webb, C., & Donoghue, M. (2005). Phylomatic: Tree assembly for applied phylogenetics.

936   *Molecular Ecology Notes*, *5*(1), 181–183.

937   Weir, J., & Schluter, D. (2008). Calibrating the avian molecular clock. *Molecular Ecology*,

938   *17*(10), 2321–2328.

939   Zuccon, D., Prŷs-Jones, R., Rasmussen, P. C., & Ericson, P. G. (2012). The phylogenetic

940   relationships and generic limits of finches (fringillidae). *Molecular Phylogenetics and*

941   *Evolution*, *62*(2), 581–596.

942   **Figure Captions**

943   Figure 1. DateLife's benchmarking results showing computation time used for taxon

944   name processing and search across `datelife`'s chronogram database, as a function of

945   number of input taxon names (N). For each N = {10, 100, 200, . . . , 1 000, . . . , 9 000, 10

946   000}, we randomly sampled N species names from the class Aves, a hundred times, and then

947   performed a `datelife` search processing the input names using the Taxon Names Resolution

948   Service (TNRS; light gray), and without processing input names (dark gray). For comparison,

949   we performed a chronogram search using names that have been pre-processed with TNRS.

950   Figure 2. DateLife results of an analysis of a small sample of 6 bird species within the

951   Passeriformes (a, b). Processed species names were found across 9 chronograms within 6

952   independent studies (c; Barker et al. (2012), Barker et al. (2015), Burns et al. (2014),

953   Hedges et al. (2015), Hooper and Price (2017), Jetz et al. (2012). This revealed 28 source

954   age data points for the queried species names (e; Table 1). Summarized age data (f; Table 2)

955   was used as secondary calibrations to date a tree topology obtained from OpenTree's

956   synthetic tree v13.4 (d), resulting in the chronogram of summary source ages shown in (g).

957   Figure 3. Median summary chronogram resulting from a DateLife analysis of bird

958   species within the family Fringillidae. For visualization purposes, we are showing a portion of

959   the final median summary chronogram encompassing 57 species out of the 289 total included

960   in the analysis. The complete final chronogram is available as Supplementary Figure S3. The

961   starting tree topology (Supplementary Fig. S2) has 289 tip species and 253 nodes; DateLife

962   revealed age data for 194 of these nodes from at least one published chronogram. In total, 19

different chronograms from 13 different studies contributed 818 age data points, which were
summarized to obtain a single value for each one of the 194 nodes with age data. From the
194 summary ages available, 21 were discarded and not used as calibrations (asterisk, *),
because they were older than a parent node or younger than a descendant node; the
remaining 173 summary ages were used as secondary calibrations (forward slash, /) with the
Branch Length Adjuster (BLADJ) software from Webb et al., (2008).

Figure 4. Cross validation of results from a DateLife analysis of the family Fringillidae,
shown in Figure 3 and Supplementary Figure S3. Each plot compares the original node age
estimates from an input source study chronogram (x axis) with the corresponding node age
resulting from a dating analysis using the DateLife workflow, excluding data from that study
(y axis).