bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Authors: Luna L. Sánchez-Reyes[1], Brian C. O'Meara[1]

Correspondence address:

1. *Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, 425 Hesler Biology Building, Knoxville, TN 37996, USA*

Corresponding authors: sanchez.reyes.luna@gmail.com, bomeara@utk.edu

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

abstract.- The combination of new analytical techniques, availability of more fossil and molecular data, and better practices in data sharing has resulted in a steady accumulation of chronograms in public and open databases such as TreeBASE, Dryad, and Open Tree of Life for a large quantity and diversity of organisms in the last few decades. However, getting a tree with branch lengths proportional to time remains difficult for many biologists and the non-academic community, despite its importance in many areas of research, education, and science communication. `datelife` is a service implemented via an R package and a web site (http://www.datelife.org/) for efficient reuse, summary and reanalysis of published data on lineage divergence times. The main workflow starts with at least two taxon names as input, either as tip labels on a tree, or as a simple comma separated character string. A name search is then performed across the chronogram database and positively identified source trees are pruned to maintain queried taxa only and stored as a named list of patristic distance matrices. Source chronogram data can be summarised using branch length summary statistics or variance minimizing approaches to generate a single summary chronogram. Source chronogram data can also be used as calibration points to date a tree containing some or all names from the initial query. If there is no information available for any queried taxa, data can be simulated. All source and summary chronograms can be saved in formats that permit easy reuse and reanalysis. Summary and newly generated trees are potentially useful to evaluate evolutionary hypothesis in different areas of research in biology. How well this trees work for this purpose still needs to be tested. `datelife` will be useful to increase awereness on the existing variation in expert time of divergence data, and might foster exploration of the effect of alternative divergence time hypothesis on the results of analyses, nurturing a culture of more cautious interpretation of evolutionary results.

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

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Clade ages represent a fundamental piece of information for evolutionary understanding in many areas of research, from developmental to conservation biology (Felsenstein 1985; Webb 2000), from historical biogeography to species diversification studies (Posadas et al. 2006; Morlon 2014). The primary information needed for these time estimates comes from the fossil record. Coupled with phylogenies with branch lengths based on molecular and/or morphological data, the time of divergence of extant and extinct lineages can be reconstructed with molecular dating methods. The number of studies publishing phylogenies with branch lengths proportional to geological time (hereafter chronograms) have constantly increased in number for the last two decades (Kumar et al. 2017). Still, generating a chronogram is not an easy task unless you have specialized training: it requires inferring a tree, understanding what fossil data are available and their limits, and where fossils go on the tree. That is why there has been an urge for promoting and facilitating reuse of the vast amount of phylogenetic and time of lineage divergence data that has been generated and made available in publications, for the advantage of research relying on this information (Webb and Donoghue 2005; Stoltzfus et al. 2013).

Wide interest from the scientific community to make information from phylogenies in general and chronograms in particular available for consultation and reuse has spurred the creation of public platforms with various goals and characteristics. TreeBASE (Morell 1996; Piel et al. 2002), the Dryad repository (http://datadryad.org/), and the Open Tree of Life (OToL; Hinchliff et al. 2015) are platforms that store and make available published phylogenies and chronograms for easy scientific reuse. Tree in all of them can be queried using automatised web procedures, which permit personalized, large scale searches that are also very fast. OToL stores trees with branch length information from a wide range of living organisms, implementing a metadata structure that stores the branch length units (i.e., time or relative susbtitution rates). Treebase and Dryad repositories also contain trees from all groups of life, but the former did not store branch length information until recently (and lacks consistent metadata on what any branch lengths stored mean) and Dryad stores many other types of biological data using metadata that does not allow automatic distinction of types of trees and branch length units, impairing the automatised access to time of lineage divergence information.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Besides keeping a repository to easily store and share expert phylogenetic and chronogram knowledge, OToL also has the primary goal of synthesising all trees in their repository to expose to the community a single tree of all life depicting the phylogenetic relationships among known lineages. All or parts of this synthetic tree can be reused for any purpose. However, it currently only focus on synthesizing tree topology, meaning that it does not expose branch length data of any type. The Timetree of Life project focuses on the synthesis of a single chronogram of life (Hedges et al. 2006) and presents a very accessible, attractive interface. However, the thousands of chronograms this NSF-funded project have compiled for synthesis are only publicly available for visual examination in their website or for download as images, but large scale download remains prohibited by their site. The latest version of their synthetic chronogram (Kumar et al. 2017) can be queried only through their website in a non-automatised fashion, and only subsets of it can be reused for analyses with the permission of the authors. Other platforms such as SuperSmart (Antonelli et al. 2017) and phylogenerator (Pearse and Purvis 2013) are focused in automatised *de novo* chronogram inference, by reusing DNA sequence data to reconstruct phylogenetic trees. However, expert fossil information necessary for subsequent molecular dating analyses still needs to be compiled and curated by the user, rendering them a challenging tool to obtain data on time of lineage divergence for the non-specialist. Moreover, these tools do not provide information from already created expert chronograms.

A tool for efficient reuse of expert, published data on time of lineage divergence should have an open and fully public chronogram database storing data in a format suitable for scientific reuse, an automatised way of accessing the information, and straightforward means of comparing and summarizing chronogram information as needed by the user. A prototype service aiming to meet this characteristics was developed over a series of hackathons at the National Evolutionary Synthesis Center (Stoltzfus et al. 2013). In here we present the formal description and implementation of the `datelife` service, constituted by an R package and a web site (http://www.datelife.org/). There is still much room for improvement, and flaws and limitations are addressed below. We strived for the current implementation of `datelife` to perform the basic tasks described above, featuring a system for maintenance of an open database of chronograms pulled from public repositories, methods to summarize and compare source chronograms, and new functions to visualize and graphically compare source and summary chronograms.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Description

The basic `datelife` workflow is shown in figure 1 and consists of:

1. A user providing at least two taxon names as input, either as tip labels on a tree, or as a simple comma separated character string. The tree can be in newick or phylo format, and can be with or without branch lengths.

2. A name search is then performed across the chronogram database; source trees with at least two matching input names are identified; all other taxa that do not match the original query are then dropped from the positively identified source trees. These pruned chronograms are hereafter referred as source chronograms. Finally, each source chronogram is transformed to a patristic matrix named by the citation of the original study. This format facilitates and greatly speeds up all further analyses and summarizing algorithms.

3. The user can obtain different summary information including: a) all source chronograms, b) maximum ages of source chronograms, c) citations of studies where source chronograms were originally published, d) a summary table with all of the above, e) a single summary tree of all or a subset of source chronograms, f) a report of succesful matches of input taxon names across source chronograms, and g) the single source chronogram with the greatest number of taxa. Summary information can be used to make decisions on the next steps of the workflow.

4. Then, source chronogram data can be used as calibration points to date a tree with or without branch lengths containing some or all names from the initial query.

5. If there is no information available for any queried taxa, users can also create both age and phylogenetic data for this missing taxa with a variety of algorithms described below.

6. Finally, users can easily save all source and summary chronograms in formats that permit easy reuse and reanalyses (newick and R "phylo" format), as well as view and compare results graphically, or construct their own graphs using inbuilt `datelife` graphic generation functions.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To gather, process, and present information, `datelife` builds up from functions available in several R packages including rotl (Michonneau et al. 2016), ape (Paradis et al. 2004), geiger (Harmon et al. 2008), paleotree (Bapst 2012), bold (Chamberlain 2018), phytools (Revell 2012), taxize (Chamberlain and Szöcs 2013; Chamberlain 2018), phyloch (Heibl 2008), phylocomr (Ooms and Chamberlain 2018) and rphylotastic (O'Meara et al. 2019).

A `datelife` search currently accepts scientific names only. It can be any named clade or binomial specific. The search is performed at the species level, so when input names correspond to named clades, `datelife` pulls all accepted species names within the clade from OToL's reference taxonomy and uses all of those. Searches at the infraspecies level are not currently allowed, so input names belonging to subspecies or any other infraspecific category are collapsed to the species level. `datelife` processes input names with the taxon name resolution service (TNRS; Boyle et al. 2013), which corrects potentially misspelled names and typos, and standardizes spelling variations and synonyms , increasing the probability of correctly finding the queried taxa in `datelife`'s chronogram database.

Then, a name matching procedure is performed across `datelife`'s chronogram database which is assembled from OToL's tree repository. Compared to other existing open tree repositories OToL's metadata rich tree store is the only one that supports search, identification, and handling of chronograms in an automatised fashion. Also, the tip names in their stored trees correspond to scientific names at or below the species level. Finally, all their chronograms come from peer-reviewed published studies generated by specialists in the targeted lineages, arguably representing expert knowledge on time of lineage divergence.

Information from source chronograms can be summarised with a summary statistic of tree branch lengths, such as median or mean. A much slower, but possibly more accurate Super Distance Matrix (SDM) approach for supertree reconstruction with branch lengths (Criscuolo et al. 2006) is also implemented via the ape package (Paradis et al. 2004). The resulting summary patristic distance matrix could be clustered with classic algorithms to return a tree. However, the resulting trees are often non-ultrameric and do not reflect the source chronogram data (see datelife_examples package). Instead, we obtained a distribution of age data from the summary matrix available for nodes on a consensus tree. The Branch Length Adjuster (BLADJ)

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

algorithm (Webb et al. 2008) was then used to fix known node ages and then distribute nodes with unknown ages evenly over the consensus tree, minimizing age variance in the resulting chronogram.

For tree dating, the congruification algorithm described by Eastman et al. (2013) is implemented to find shared nodes between trees (congruent nodes). The ages of these nodes are then used as calibrations to date any given tree. Currently implemented methods for tree dating are BLADJ, MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) and PATHd8 (Britton et al. 2007), a non-clock, rate-smoothing dating method.

Benchmark

`datelife`'s code speed was tested on an Apple iMac with one 3.4 GHz Intel Core i5 processor. We registered variation in computing time of query processing and search through the database relative to number of queried taxon names. Query processing increases roughly linearly with number of input taxon names, and increases considerably if TNRS service is activated. Up to ten thousand names can be processed and searched in less than 30 minutes. A name search through the chronogram database with an already processed query can be performed in less than a minute, even with a very large number of taxon names (Fig. 2). `datelife`'s code performance was evaluated with a set of unit tests designed and implemented with the R package testthat (R Core Team 2018) that were run locally –using the devtools package (R Core Team 2018), and on a public server –via GitHub, using the continuous integration tool Travis CI (https://travis-ci.org). At present, unit tests cover more than 50% of `datelife`'s code (https://codecov.io/gh/phylotastic/datelife).

Example

In this section we demonstrate the types of outputs that can be obtained with `datelife`, using as an example the bird family Fringillidae of true finches. We performed a higher-taxon search to obtain all data on lineage divergence available in `datelife`'s database for all recognised species within the Fringillidae (475 spp. according to the Open Tree of Life taxonomy). There are 13 chronograms containing at least two Fringillidae species, published in 9 different studies (Fig. 3). Data from these source chronograms was used to generate two types of summary chronograms, median and SDM. As explained in the "Description" section, data from

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

source chronograms was first summarised into a single distance matrix (using the median and the SDM method respectively) and then the available node ages were used as fixed ages over a consensus tree topology, to obtain a fully dated tree with the program BLADJ (Fig. 4). Median summary chronograms are older and have wider variation in maximum ages than chronograms obtained with SDM. In both cases, ages are generally consistent with source ages, but see next section.

Data from source chronograms was also used to date tree topologies with no branch length information and trees with branch lengths in relative substitution rates (Figs. 5 and 6). As a form of cross validation, we used tree topologies from each study and calibrated them using information from all other source chronograms. In the absence of branch length data, the ages of internal nodes were approximately recovered in almost all cases (except for studies 3, and 5; Fig. 5). Maximum tree ages were only approximately recovered in one case (study 2; Fig. 5). To exemplify dating a tree with branch lengths provided by the user, we obtained DNA sequence data from the Barcode of Life Database (BOLD) and branch lengths were successfully generated for all source chronograms. However, dating with PATHd8 (using congruified calibrations) was only successful in three cases (studies 3, 5, and 9, shown in Fig. 6). From these, two trees have a different sampling than the original source chronogram, mainly because DNA data for some species is absent from the BOLD. 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 can be consulted in https://github.com/LunaSare/datelife_examples.

Flaws, Limitations and Prospects

The main goal of `datelife` is to make expert information on time of lineage divergence easily accesible for comparison, reuse, and reanalysis, to researchers in all areas of science and with all levels of expertise in the matter. It is a very fast tool that fulfills the quality of openness and does not require any expert biological knowledge from users –besides the names of the organisms they want to work with– for any of its functionalities. However, it has many flaws. Some of them can be overcome, some of them might represent limitations.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

At the moment, `datelife`'s chronogram database is not very large, storing 231 chronograms up to the time the manuscript was written. This represents 5.79% of the largest existing chronogram database, which is not open for scientific reuse nor automatised data mining (Kumar et al. 2017). OToL is the only public tree repository from where `datelife` can currently pull chronograms to construct its database. A previous version of TimeTree's synthetic chronogram (Hedges et al. 2015) was made available in the OToL repository, hence the amount of lineages represented in datelife's database is at least as substantial as TimeTree's. This ensures that some information will be available for any given query, but it does not ensure that the full state of knowledge of time of divergence data will be available for any given lineage. Thus, incorporation of more published chronograms deposited in OpenTree, or perhaps directly pulled from the Dryad repository, to `datelife`'s database is crucial to improve its services. Methods to automatically mine chronogram data from Dryad could be designed and implemented. However, the unit of branch lengths would still need to be determined by hand. Consequently, we would like to emphasize on the importance of sharing chronogram data for the scientific community, in repositories that require expert input and manual curation, such as OToL's tree repository.

Another potential concern comes from summary chronograms. We currently summarize by default all source chronograms that overlap with at least two taxa. Users can subset source data if they have reasons to favor some source chronograms over others. Strictly speaking, a good chronogram should reflect the real time of lineage divergence accurately and precisely. To our knowledge, there is no objective way to determine if an expert chronogram is better than other. Some criteria that have been put forward are the level of lineage sampling and the number of calibrations used. Scientists usually also favor chronograms coming from studies with primary calibrations to ones from secondary calibrations. It has been observed with simulations that divergence times inferred with secondary calibrations are significantly younger than those inferred with primary calibrations in analyses performed with bayesian inference methods when priors are implemented in similar ways in both analyses (Schenk 2016). Yet, there are different ways to use secondary calibrations and the bias might not be encountered with other dating methods that do not require setting priors (such as ML methods). In sum, further studies are required to fully understand the effect of using secondary calibrations on time estimates and downstream anlyses.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Furthermore, even chronograms obtained with primary fossil data can show substantial variation in time estimates between clades, as observed from the comparison of source chronograms in the Fringillidae example. This observation is often encountered in the literature (see, for example, the ongoing debate about crown group age of angiosperms (Ramshaw et al. 1972; Sanderson and Doyle 2001; Magallón et al. 2015; Barba-Montoya et al. 2018). For some studies, especially ones based on branch lengths (e.g., studies of species diversification, timing of evolutionary events, phenotypic trait evolution), using a different chronogram may return different results (Title and Rabosky 2016). Stitching together these chronograms can create a larger tree that uses information from multiple studies, but the effect of uncertainties and errors here on downstream analyses is still largely unknown.

Summarizing high-quality chronograms might also imply summarizing evolutionary hypothesis. This could be enlightening from certain point of view, since it could help getting a single global evolutionary history for a lineage. It could also be really misleading, since we are potentially loosing important parts of the evolutionary history of a lineage that might only be reflected in some chronograms and not from the summary chronogram. Ideally, we should still rely on time of lineage divergence data obtained from a single analysis using fossil data as primary sources of calibrations, and using fossils that have already been curated as calibrations to date other trees, which should reflect a more homogeneous evolutionary history (Antonelli et al. 2017). This will be implemented in future `datelife` versions.

Alternatively, one could try to choose the "best" chronogram. Several characteristics of the data used for dating analyses as well as from the output chronogram itself, could be used to score quality of source chronograms. Some characteristics that are often cited as a measure of improved quality in dating studies addressing long-studied organisms are: quality of alignment (missing data, GC content), lineage sampling (strategy and proportion), phylogenetic and dating inference method, number of fossils used as calibrations, support for nodes and ages, and magnitude of confidence intervals. To facilitate subsetting of source chronograms following different criteria by the users, this information should be included as metadata manually entered by curators in the near future.

In other areas of biological research, such as ecology and conservation biology, it has been shown that at least

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

some data on lineage divergence represents a relevant improvement for testing alternative hypothesis using phylogenetic distance. Hence, we allow accepted ways of creating branch lengths in the absence of starting branch length information (such as BLADJ (Webb et al. 2008)) for several taxa lacking this information. Making up branch lengths in this or other ways is accepted in scientific publications: Rabosky et al. (2018) created a time-calibrated tree of 31,536 ray-finned fishes, of which only 37% had molecular data; Jetz et al. (2012), created a time-calibrated tree of all 9,993 bird species, where 67% had molecular data; Smith and Brown (2018) constructed a tree of 353,185 seed plants where only 23% had molecular data. Taken to the extreme, one could make a fully resolved, calibrated tree of all modern and extinct taxa using a single taxonomy and a single calibration with the polytomy resolution and branch imputation methods. There has yet to be a thorough analysis of what can go wrong when one goes beyond the data in this way, so we urge caution; we also urge readers to follow the example of many of the large tree papers cited above and make sure results are substantially similar between trees fully reconstructed with molecular or other data and trees that are reconstructed using taxonomy.

Conclusions

Divergence time information is key to many areas of evolutionary studies: trait evolution, diversification, biogeography, macroecology and more. Generating this information is difficult, especially for those who want to use phylogenies but who are not systematists, or do not have the time to acquire and develop the necessary knowledge and data curation skills to produce chronograms *de novo*. Knowledge on clade ages is also crucial for science communication and education.

`datelife` allows an easy and fast obtention, as well as comparison of publicly available information on time of lineage divergence, providing a straightforward way to get an informed idea on the state of knowledge of the time frame of evolution of different regions of the tree of life, allowing identification of regions that require more research or that have conflicting information. Both summary and newly generated trees are potentially useful to evaluate evolutionary hypothesis in different areas of research. `datelife` helps with awereness on the existing variation in expert time of divergence data, and might foster exploration of the effect of alternative divergence time hypothesis on the results of analyses, nurturing a culture of more cautious

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interpretation of evolutionary results.

Availability

`datelife` is free and open source and it can be used through its current website http://www.datelife.org/query/, through its R package, and through Phylotastic's project web portal http://phylo.cs.nmsu.edu:3000/. `datelife`'s website is maintained using RStudio's shiny server and the shiny package open infrastructure, as well as Docker. `datelife`'s R package stable version will be available for installation from the CRAN repository (https://cran.r-project.org/package=datelife) using the command `install.packages(pkgs = "datelife")` from within R. Development versions are available from the GitHub repository (https://github.com/phylotastic/datelife) and can be installed using the command `devtools::install_github("phylotastic/datelife")`.

Supplementary Material

Code used to generate all versions of this manuscript, the biological examples, as well as the benchmark of functionalities are respectively in the datelife_paper1, datelife_examples, and datelife_benchmark repositories in LLSR GitHub account.

Funding

Funding was provided by the US National Science Foundation (NSF) grants ABI-1458603 to Datelife project and DBI-0905606 to the National Evolutionary Synthesis Center (NESCent), and the Phylotastic project Grant ABI-1458572.

Acknowledgements

We thank colleagues from the O'Meara Lab at the University of Tennesse Knoxville for suggestions, discussions and software testing. The late National Evolutionary Synthesis Center (NESCent), which sponsored hackathons that led to initial work on this project. The team that assembled `datelife`'s first proof of concept: Tracy Heath, Jonathan Eastman, Peter Midford, Joseph Brown, Matt Pennell, Mike Alfaro, and

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Luke Harmon. The Open Tree of Life project that provides the open, metadata rich repository of trees used for `datelife`. The many scientists who publish their chronograms in an open, reusable form, and the scientists who curate them for deposition in the Open Tree of Life repository. The NSF for funding nearly all the above, in addition to the ABI grant that funded this project itself.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Antonelli A., Hettling H., Condamine F.L., Vos K., Nilsson R.H., Sanderson M.J., Sauquet H., Scharn R., Silvestro D., Töpel M., Bacon C.D., Oxelman B., Vos R.A. 2017. Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of Taxa. Systematic Biology. 66:153–166.

Bapst D.W. 2012. Paleotree: An R package for paleontological and phylogenetic analyses of evolution. Methods in Ecology and Evolution. 3:803–807.

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. 218:819–834.

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. 62:298–320.

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. 132:333–348.

Boyle B., Hopkins N., Lu Z., Raygoza Garay J.A., Mozzherin D., Rees T., Matasci N., Narro M.L., Piel W.H., Mckay S.J., Lowry S., Freeland C., Peet R.K., Enquist B.J. 2013. The taxonomic name resolution service: An online tool for automated standardization of plant names. BMC Bioinformatics. 14.

Britton T., Anderson C.L., Jacquet D., Lundqvist S., Bremer K. 2007. Estimating Divergence Times in Large Phylogenetic Trees. Systematic Biology. 56:741–752.

Burns K.J., Shultz A.J., Title P.O., Mason N.A., Barker F.K., Klicka J., Lanyon S.M., 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..

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Chamberlain S.A., Szöcs E. 2013. taxize : taxonomic search and retrieval in R [version 2; referees: 3 approved]. F1000Research. 2:1–29.

Claramunt S., Cracraft J. 2015. A new time tree reveals earth history's imprint on the evolution of modern birds. Science advances. 1: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. 55:740–755.

Eastman J.M., Harmon L.J., Tank D.C. 2013. Congruification: Support for time scaling large phylogenetic trees. Methods in Ecology and Evolution. 4:688–691.

Felsenstein J. 1985. Phylogenies and the Comparative Method. The American Naturalist. 125:1–15.

Gibb G.C., England R., Hartig G., McLenachan P.A., Taylor Smith B.L., McComish B.J., Cooper A., Penny D. 2015. New zealand passerines help clarify the diversification of major songbird lineages during the oligocene. Genome biology and evolution. 7:2983–2995.

Harmon L., Weir J., Brock C., Glor R., Challenger W. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics. 24:129–131.

Hedges S.B., Dudley J., Kumar S. 2006. TimeTree: A public knowledge-base of divergence times among organisms. Bioinformatics. 22:2971–2972.

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. 32:835–845.

Heibl C. 2008. PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages..

Hinchliff C.E., Smith S.A., Allman J.F., Burleigh J.G., Chaudhary R., Coghill L.M., Crandall K.A., Deng J., Drew B.T., Gazis R., Gude K., Hibbett D.S., Katz L.A., Laughinghouse H.D., McTavish E.J., Midford P.E.,

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Owen C.L., Ree R.H., Rees J.A., Soltis D.E., Williams T., Cranston K.A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proceedings of the National Academy of Sciences. 112:12764–12769.

Hooper D.M., Price T.D. 2017. Chromosomal inversion differences correlate with range overlap in passerine birds. Nature ecology & evolution. 1:1526.

Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Jetz W., Thomas G., Joy J.J., Hartmann K., Mooers A. 2012. The global diversity of birds in space and time. Nature. 491:444–448.

Kumar S., Stecher G., Suleski M., Hedges S.B. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular biology and evolution. 34:1812–1819.

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. 207:437–453.

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. 7:1476–1481.

Morell V. 1996. The roots of phylogeny. Science. 273:569.

Morlon H. 2014. Phylogenetic approaches for studying diversification. Ecology Letters. 17:508–525.

O'Meara B., Md Tayeen A.S., Sanchez Reyes L.L. 2019. Rphylotastic: An r interface to 'phylotastic' web services..

Ooms J., Chamberlain S. 2018. Phylocomr: Interface to 'phylocom'..

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

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pearse W.D., Purvis A. 2013. PhyloGenerator: An automated phylogeny generation tool for ecologists. Methods in Ecology and Evolution. 4:692–698.

Piel W.H., Donoghue M., Sanderson M. 2002. TreeBASE : A database of phylogenetic information. In: Shimura J., Wilson K., Gordon D., editors. To the interoperable "catalog of life" with partners. Tsukuba, Japan: National Institute for Environmental Studies. p. 41–47.

Posadas P., Crisci J.V., Katinas L. 2006. Historical biogeography: A review of its basic concepts and critical issues. Journal of Arid Environments. 66:389–403.

Price T.D., Hooper D.M., Buchanan C.D., Johansson U.S., Tietze D.T., Alström P., Olsson U., Ghosh-Harihar M., Ishtiaq F., Gupta S.K., others. 2014. Niche filling slows the diversification of himalayan songbirds. Nature. 509:222.

Rabosky D.L., Chang J., Title P.O., Cowman P.F., Sallan L., Friedman M., Kaschner K., Garilao C., Near T.J., Coll M., others. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature. 559:392.

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. 71:773–779.

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

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

Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Sanderson M.J., Doyle J.A. 2001. Sources of error and confidence intervals in estimating the age of angiosperms

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from rbcL and 18S rDNA data. American Journal of Botany. 88:1499–1516.

Schenk J.J. 2016. Consequences of secondary calibrations on divergence time estimates. PLoS ONE. 11.

Smith S.A., Brown J.W. 2018. Constructing a broadly inclusive seed plant phylogeny. American Journal of Botany. 105:302–314.

Stoltzfus A., Lapp H., Matasci N., Deus H., Sidlauskas B., Zmasek C.M., Vaidya G., Pontelli E., Cranston K., Vos R., Webb C.O., Harmon L.J., Pirrung M., O'Meara B., Pennell M.W., Mirarab S., Rosenberg M.S., Balhoff J.P., Bik H.M., Heath T.A., Midford P.E., Brown J.W., McTavish E.J., Sukumaran J., Westneat M., Alfaro M.E., Steele A., Jordan G. 2013. Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. BMC Bioinformatics. 14.

Title P.O., Rabosky D.L. 2016. Do Macrophylogenies Yield Stable Macroevolutionary Inferences? An Example from Squamate Reptiles. Systematic Biology.:syw102.

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

Webb C.O., Ackerly D.D., Kembel S.W. 2008. Phylocom: Software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics. 24:2098–2100.

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

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

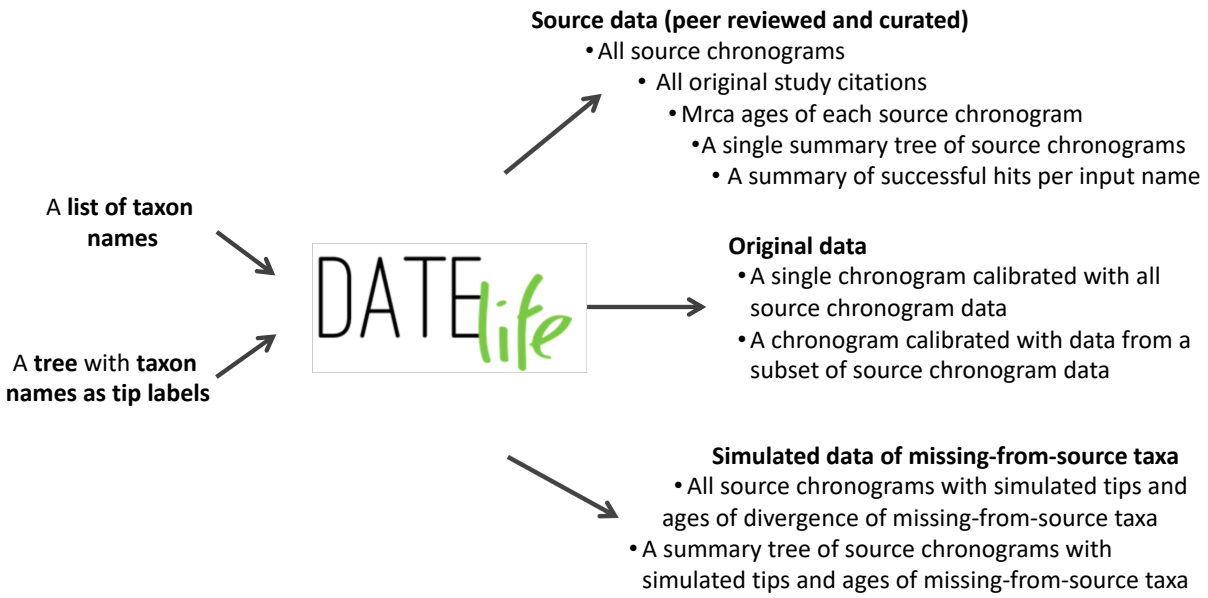

FIGURE 1

Stylized DateLife workflow. This shows the general worflows and analyses that can be performed with `datelife`, via the R package or through the website at http://www.datelife.org/. Details on the functions involved on each workflow are shown in `datelife`'s R package vignette.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

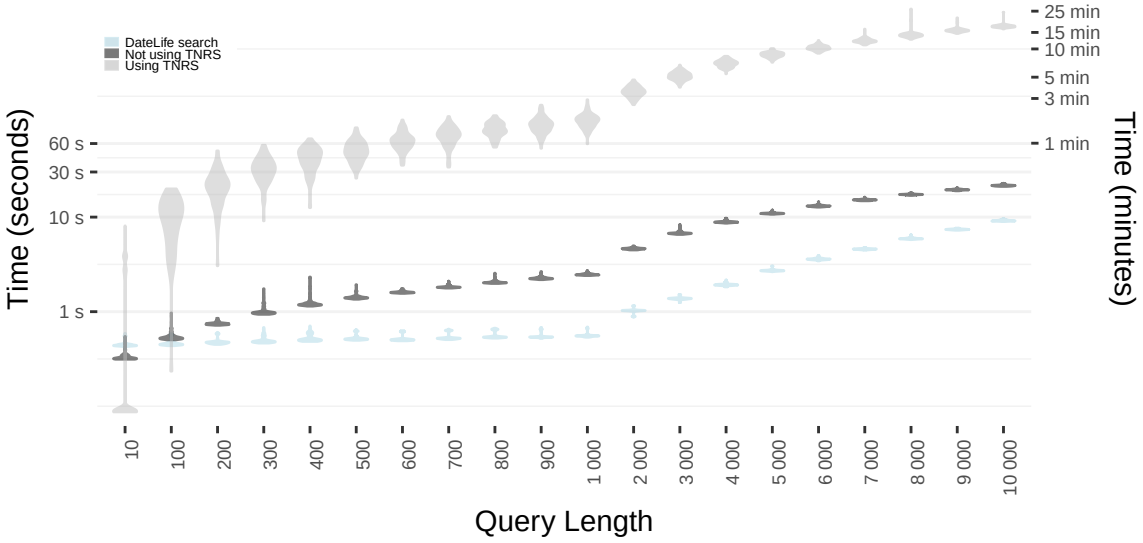

FIGURE 2

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

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

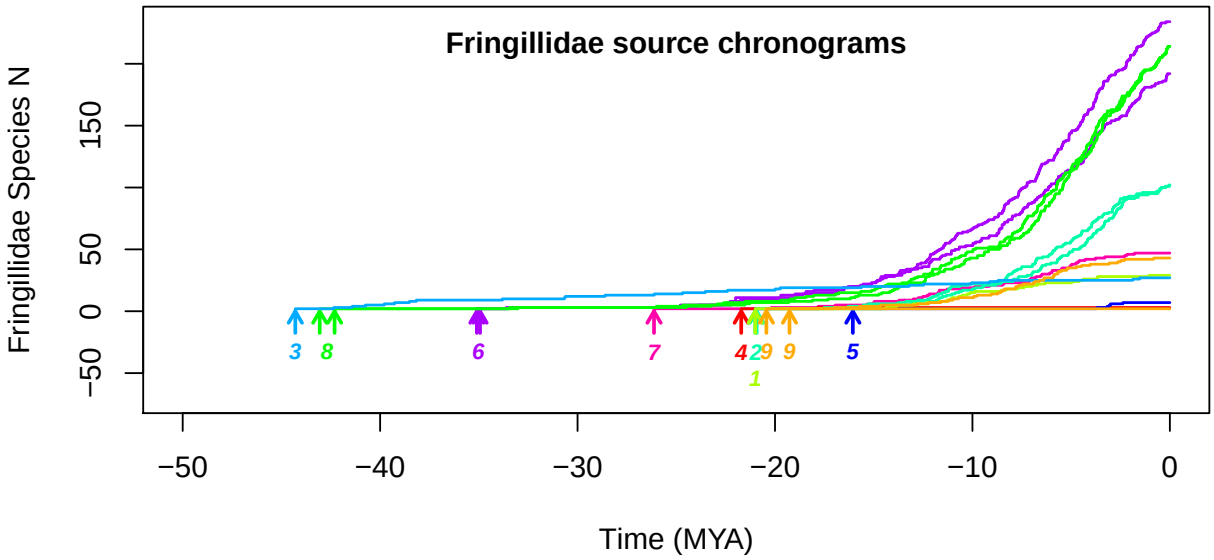

FIGURE 3

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).

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

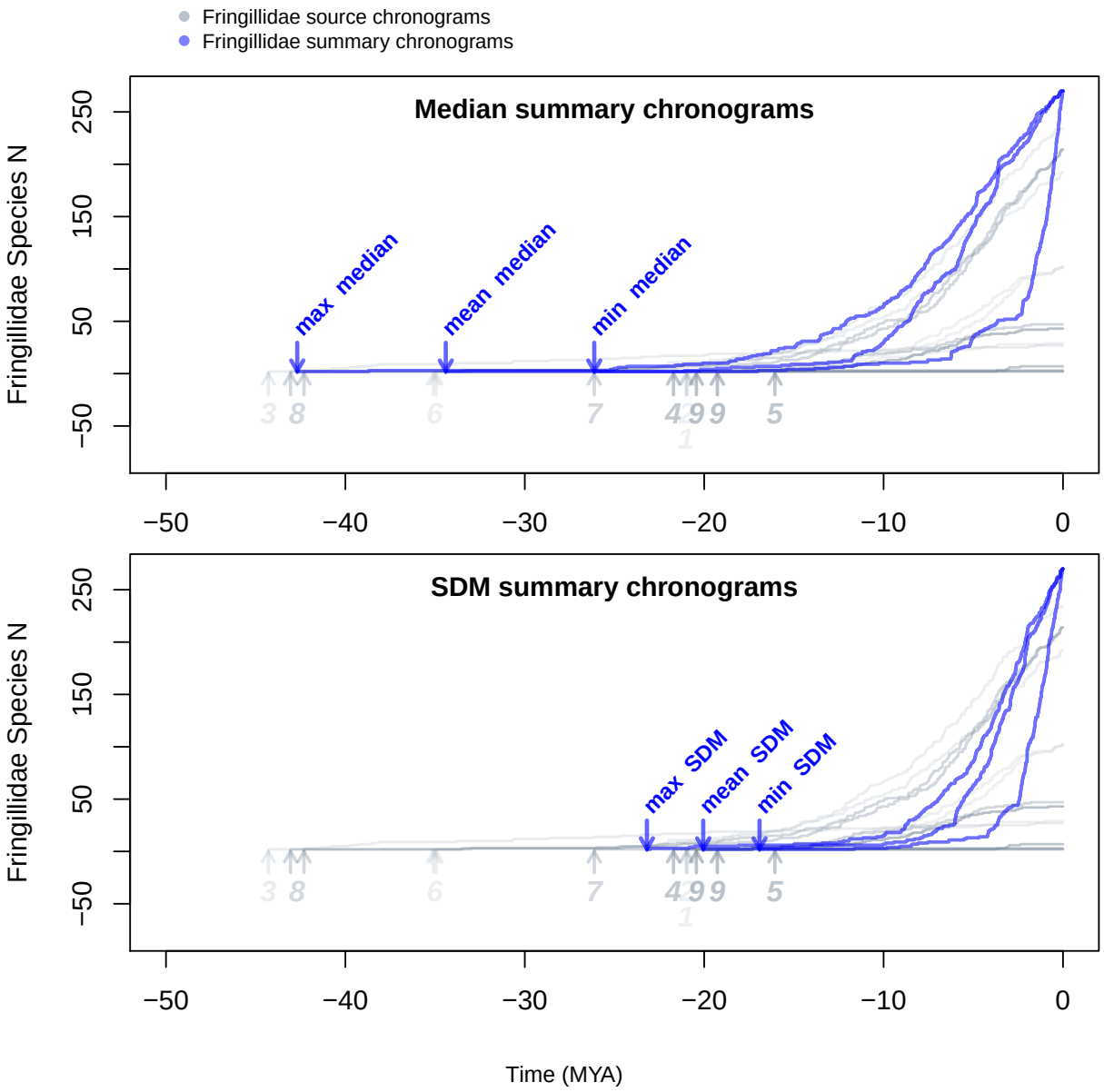

FIGURE 4

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.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

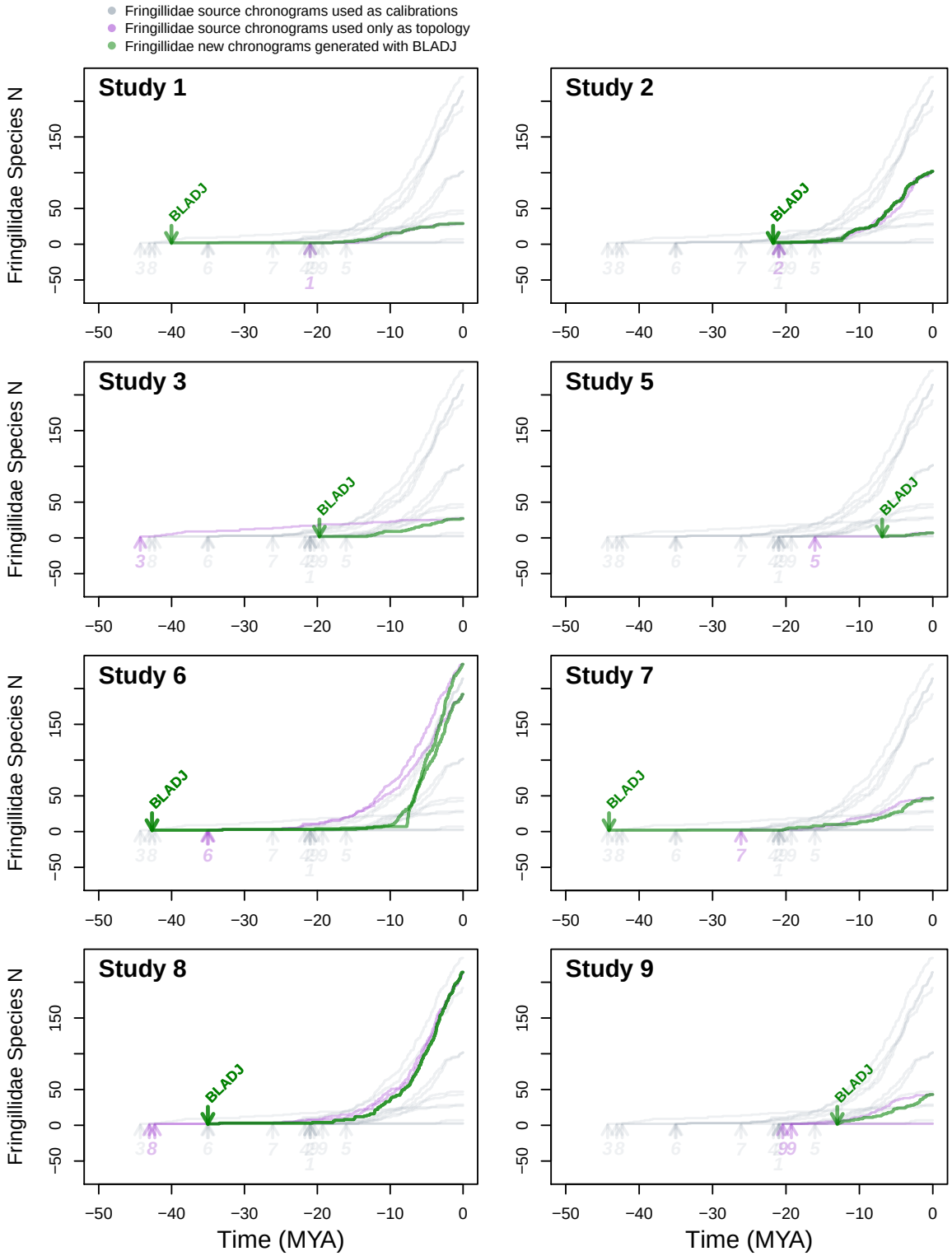

FIGURE 5

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/782094; this version posted October 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

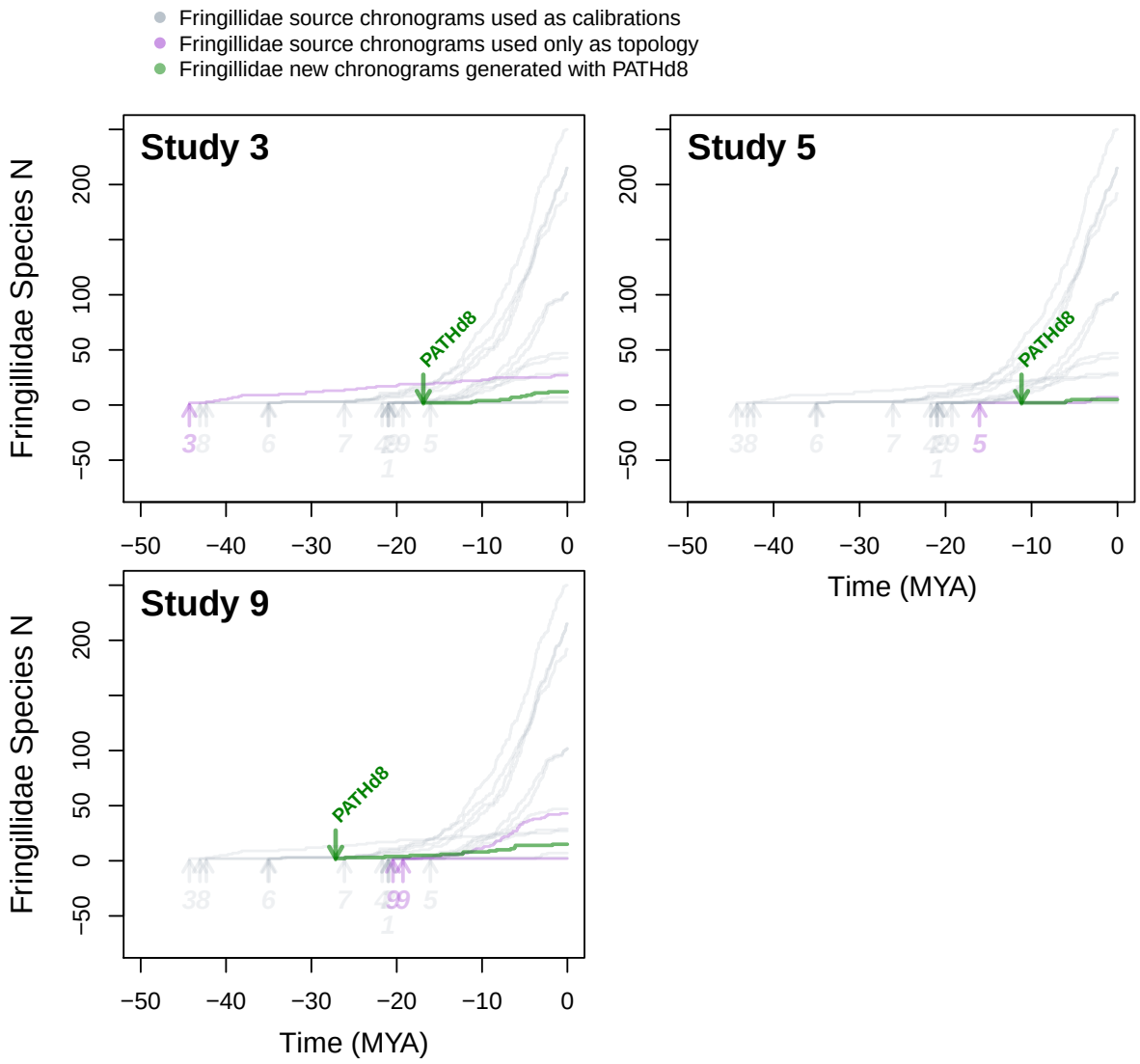

FIGURE 6

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.