Running head: DATELIFE: REVEALING THE DATED TREE OF LIFE

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

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 Dryad, TreeBASE and Open Tree of Life, for a large quantity and diversity of organisms. However, this information remains difficult to synthesize for many biologists and the non-academic community.

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

Time of lineage divergence constitutes 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 time of lineage divergence estimation comes from the fossil record. Coupled to molecular phylogenies (and more recently to morphological data), the time of divergence of extant and extinct lineages is reconstructed with molecular dating methods. Probably encouraged by the great developments in DNA sequencing techniques, phylogenetic inference and molecular dating methods, the number of studies publishing phylogenies with branch lengths proportional to geologic 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. That's 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) platforms store and make available published phylogenies and chronograms for easy scientific reuse. All of them can be queried using automatised web procedures, which permit personalised, large scale queries 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 does not store branch length information 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.

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). However, the thousands of chronograms they have compiled for synthesis are only publicly available for visual examination in their website or for download as images, but not for scientific reuse or reanalysis. 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 existent expert chronograms.

A tool for efficient reuse of expert, published data on time of lineage divergence should have an open source 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. R packages for benchmarking of functionalities and demonstrating services were also developed and made available.

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 summarising algorithms.

3. The user can obtain different summary information from the source chronograms including: a) all source chronogram ages, 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, and f) a report of succesful matches of input taxon names across source chronograms. Summary information can be used to make decisions on the next steps of the workflow.

4. 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. <!–, a taxonomic tree–>

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

6. Finally, users can easily view and compare results graphically, as well as construct their own graphs using inbuilt ‘datelife‘ graphic generation functions.

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. Chronogram 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 to perform the search. 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 to correctly find the queried taxa in `datelife`'s chronogram database.

The chronogram search 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, 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 using the Super Distance Matrix (SDM) approach for supertree reconstruction (Criscuolo et al. 2006) or using a summary statistic of branch lengths (such as median or mean). Tree dating and simulations are performed with various algorithms: Branch Length Adjuster (BLADJ) is a simple algorithm to distribute ages of undated nodes evenly, which minimizes age variance in the chronogram (Webb et al. 2008). PATHd8 is a non-clock, rate-smoothing method (Britton et al. 2007) to date trees that requires initial branch lengths and a set of calibrations.

MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) can be used when adding taxa at random, following a reference taxonomy or a topological constraint. It draws ages from a pure birth model, as implemented by Jetz and collaborators (2012).

To apply calibrations to a tree, the congruification algorithm described by Eastman et al. (2013) is implemented to find shared nodes between trees (congruent nodes).

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`, data from source chronograms was first summarised into a single distance matrix (using either the median or the SDM method) and then the available node ages were used as calibrations points over a consensus tree topology, to obtain a 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 coherent with source ages. It is not certain if these chronograms can be used to perform downstream evolutionary analyses.

There is currently wide interest in determining this. However, we know that these chronograms are useful for...

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). Branch lengths were successfully generated using the BOLD database for all source chronograms. However, dating with PATHd8 (using congruified calibrations) was only successful in three cases (studies 3, 5, and 9; 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 details 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.

At the moment, `datelife`'s chronogram database is not very large. It stores 231 chronograms up to the time the manuscript was written. This represents 5.79% of the largest existing chronogram database, which is unfortunately not open for scientific reuse and 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.

Other open repositories are not suitable for time of divergence data mining because they either do not store that data (i.e., TreeBASE), or their metadata does not have enough information to allow automatised chronogram searches (i.e., Dryad).

A previous version of TimeTree's synthetic chronogram (Hedges et al. 2015) was made available in OToL repository, hence the amount of lineages represented in datelife's database is at least as substantial as TimeTree's. This ensures that at least 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 gien lineage.

Thus, incorporation of the many chronograms that are available in data repositories such as Dryad and also from supplementary material of journals to `datelife`'s database is crucial to improve `datelife`'s services. Methods to automatically mine chronogram data from the Dryad repository could be designed and implemented. However, the unit of branch lengths would need to be determined by "hand". Consequently, we would like to emphasize on the importance for the scientific community of sharing chronogram data in repositories that require expert input and manual curation, such as OToL's tree repository.

We introduce different ways of summarizing this data, but are these summary chronograms reliable to study evolutionary patterns? The short answer is that we do not know to what extent the summary trees can be used for reanalyses and in which areas they can be used more reliably than others. Surely they might be useful for ecological analyses, but further studies should be conducted.

As implemented now, we currently by default summarize all source data avilable. Users can subset source data if they have reasons to favor some or one source chronogram over others. Strictly speaking, a good chronogram reflects the real time of 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 sampling and the numer of calibrations used. Scientists usually also favor chronograms coming from studies with primary calibrations to ones from secondary calibrations. `datelife`'s database includes all chronograms derived from expert knowledge, and this includes those derived from secondary calibrations. It has been shown with simulations that in analyses performed with bayesian inference methods,

divergence times inferred with secondary calibrations are significantly younger than those inferred with primary calibrations (using secondary calibrations in the same way as primary calibrations are used Schenk 2016a). But there are different ways to apply secondary calibrations and the bias might not be encountered with other dating methods that do not require prior assumptions (such as ML methods). This information should be included as metadata to facilitate subset of source chronograms. This will need to be manually curated as is done in OToL.

As observed from the comparison of source chronograms in the example above, even chronograms obtained with primary data can be very different. There is some evidence that different chronograms will result in very different macroevolutionary histories inferred from them (Title and Rabosky 2016). In other areas of research, such as ecological and conservation biology, it has been demonstrated that the presence of data on lineage divergence represents an improvement for testing alternative hypothesis, compared to testing in a total absence of such time data. However, we do not know how using wrong or different kinds of age data will affect the outcomes of ecological tests. More work has to be performed in this area. A large discrepancy in time of lineage divergence across expert knowledge is well known for different groups of organisms (e.g., angiosperms; Magallón et al. 2015). Comparison of available chronogram data for a wide range of organisms shown in the examples here suggest that this is a widespread phenomenon that requires further attention. In the case of the example presented here, most chronograms are new, meaning that we expect that data was generated with high quality and the software used should be very advanced. However, this is just an assumption. Scores on source data used (quality of alignement: missing data, GC content; tip sampling strategy; phylogenetic inference method; dating inference method; numer of fossils used; support for nodes and ages; age confidence intervals). But we should also be able to choose on

Even if all source chronograms have been generated by excellent standards and using same practices, the evolutionary context they depict might be very different. The ideal would be to use dates that were obtained from the same DNA context, which should reflect a more homogeneous evolutionary history or hypothesis (Antonelli et al. 2017). Summarizing trees implies summarizing evolutionary hypothesis. It is unsure if this is a good or bad thing.

For all these reasons, the ideal would be to use fossil data as primary sources of calibrations. This will be an upscale for datelife, taking fossils that have already been curated as calibrations to date other trees.

Improvements: * fossils as calibrations: Using secondary calibrations can generate biased ages when using bayesian methods, mainly because we don't know what prior to give to secondary calibrations (Schenk 2016b). * bayesian congruification * topological congruification

Conclusions

Time of lineage divergence 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 taxon 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 timeframe of different regions of the tree of life, allowing identification of regions that need more work or that have conflicting information. It can also summarize public data, or use it to generate new chronograms based on trees provided by the user. Both summary and newly generated trees have the potential to be useful to evaluate evolutionary hypothesis in different areas of research. How well this trees work for this purpose still needs to be tested. `datelife` helps with awereness on the variation of expert data, and the actual effect of the underlying data on the results of analyses, nurturing a culture of more cautious interpretation of evolutionary results.

`datelife` is also part of the reproducibility of science movement. It strives to facilitate usage of public data sets as well as encouraging the making of information publicly available. Acknowledging source information and describing workflows and analyses.

it allows exploration of alternative taxonomic and phylogenetic schemes; it allows rapid exploration of the

effect of alternative divergence time hypothesis; it allows rapid synthesis in a number of different formats;

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 with RStudio's shiny server and the shiny package open infrastructure, as well as Docker. `datelife`'s R package stable version is 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 can be found in GitHub repositories at https://github.com/LunaSare/datelife_paper1, https://github.com/LunaSare/datelife_examples, and https://github.com/LunaSare/datelife_benchmark, respectively.

Funding

Funding was provided by US National Science Foundation (NSF) grant 1458603

NESCent

Open Tree of Life

University of Tennessee, Knoxville

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 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 OpenTree. The NSF for funding nearly all the above, in addition to the ABI grant that funded this project itself.

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.

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..

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., 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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. Details on the functions involved on each workflow are shown in `datelife`'s R package vignette.

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).

Figure 3

Lineage through time (LTT) plots of source chronograms containing all or a subset of species from the bird family Fringillidae of true finches. Arrows indicate maximum age of each chronogram. Numbers reference to chronograms' original publications 1: Barker et al. (2012), 2: Barker et al. (2015), 3: Burns et al. (2014), 4: Claramunt and Cracraft (2015), 5: Gibb et al. (2015), 6: Hedges et al. (2015), 7: Hooper and Price (2017), 8: Jetz et al. (2012), 9: Price et al. (2014).

Figure 4

LTT plots of median and Supermatrix Distance Method (SDM) chronograms summarising information from source chronograms found for the Fringillidae. Arrows indicate maximum age.

Figure 5

LTT plots showing results from the cross-validation analyses of trees without branch lengths dated using BLADJ. The dating analysis can only be performed in trees with more than 2 tips, thus excluding chronogram from study 4; its data was still used as calibration for the other source chronograms.

Figure 6

LTT plots showing results from the cross-validation analyses of trees with branch length reconstructed with data from the Barcode of Life Database (BOLD) dated using PATHd8. We could construct a tree with branch lengths for all source chronograms. However, dating with PATHd8 was only successful in three source chronograms shown here.

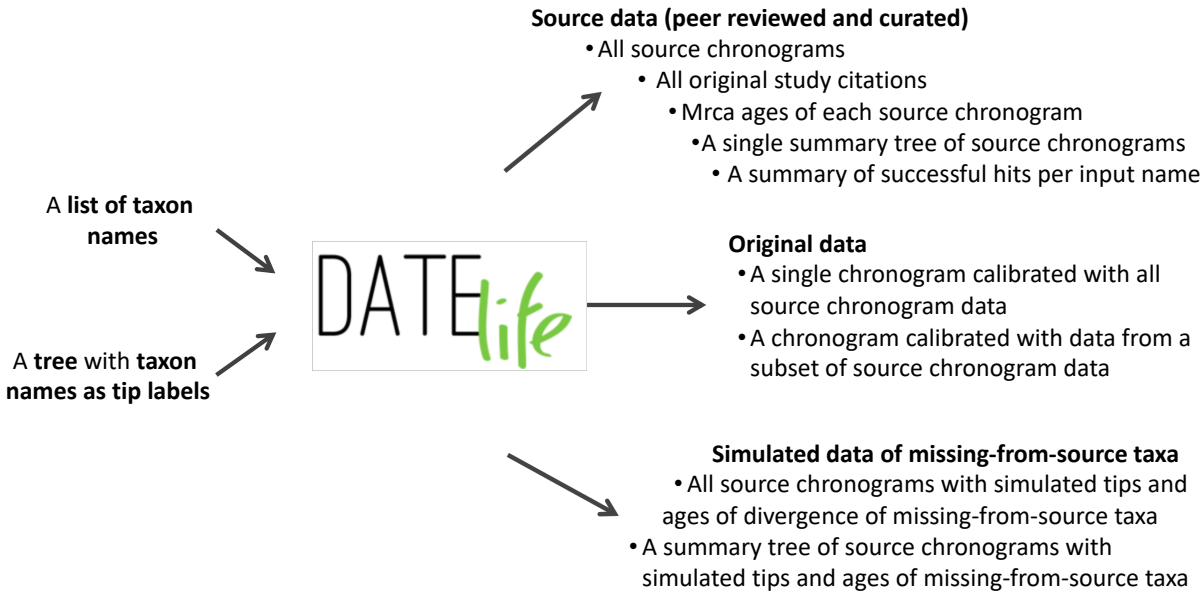
Source data (peer reviewed and curated)
• All source chronograms
• All original study citations
• Mrca ages of each source chronogram
• A single summary tree of source chronograms
• A summary of successful hits per input name
A list of taxon names
A tree with taxon names as tip labels
DATElife
Original data
• A single chronogram calibrated with all source chronogram data
• A chronogram calibrated with data from a subset of source chronogram data
Simulated data of missing-from-source taxa
• All source chronograms with simulated tips and ages of divergence of missing-from-source taxa
• A summary tree of source chronograms with simulated tips and ages of missing-from-source taxa

Figure 1:

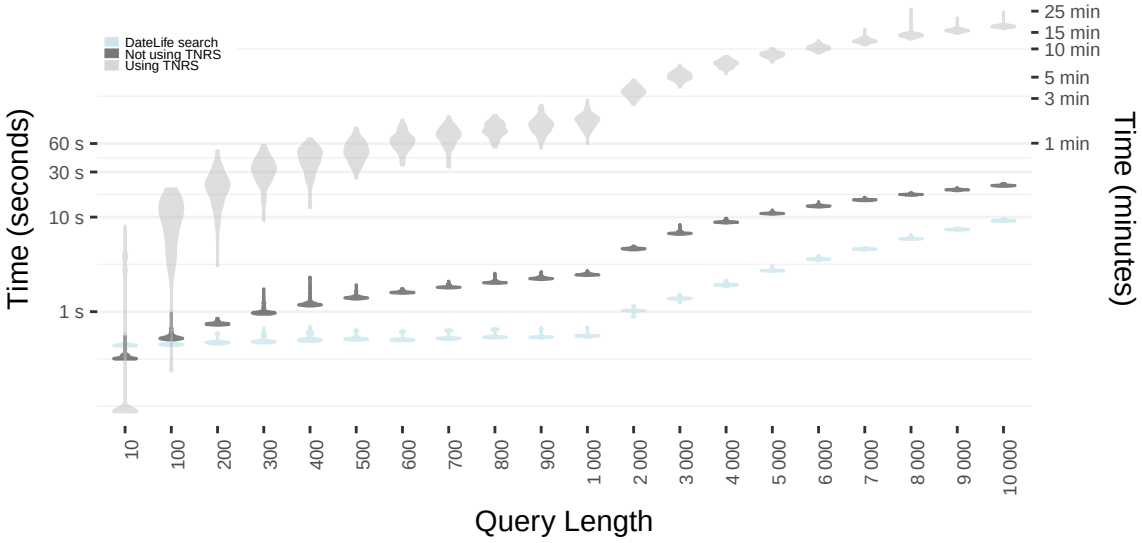
DateLife search
Not using TNRS
Using TNRS
Time (seconds)
60 s
30 s
10 s
1 s
25 min
15 min
10 min
5 min
3 min
1 min
Time (minutes)
10
100
200
300
400
500
600
700
800
900
1 000
2 000
3 000
4 000
5 000
6 000
7 000
8 000
9 000
10 000
Query Length

Figure 2:

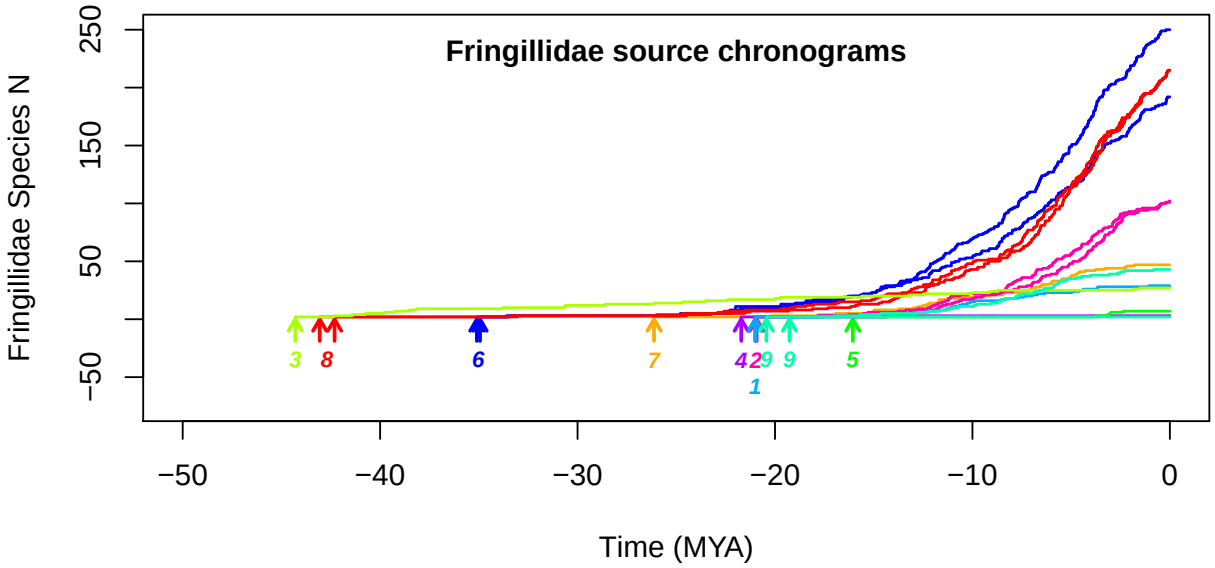
Fringillidae source chronograms
Fringillidae Species N
250
150
50
−50
−50
−40
−30
−20
−10
0
Time (MYA)
3
8
6
7
4
2
9
9
1
5

Figure 3:

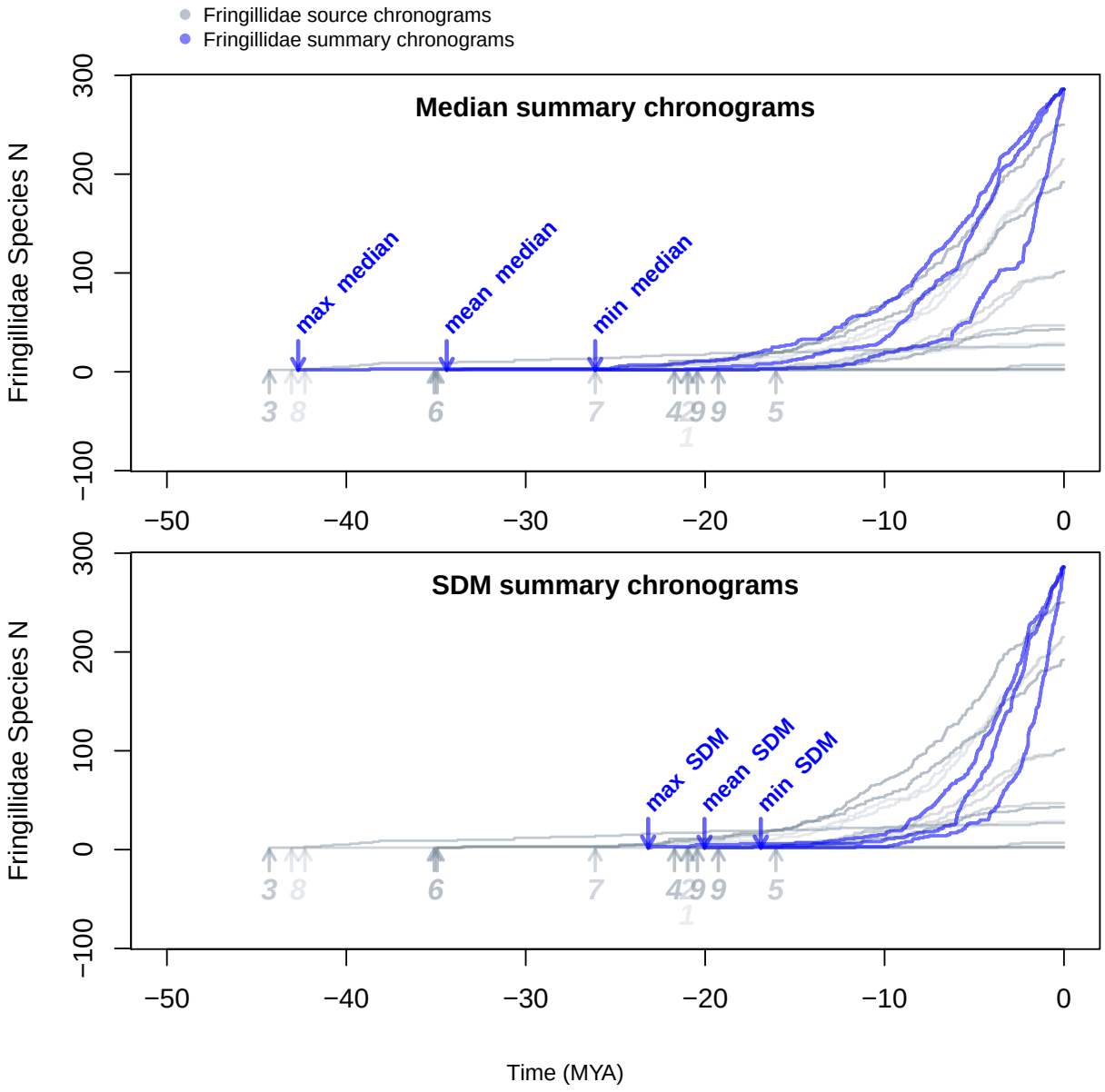
Fringillidae source chronograms
Fringillidae summary chronograms
Median summary chronograms
max median
mean median
min median
SDM summary chronograms
max SDM
mean SDM
min SDM
Fringillidae Species N
Time (MYA)

Figure 4:

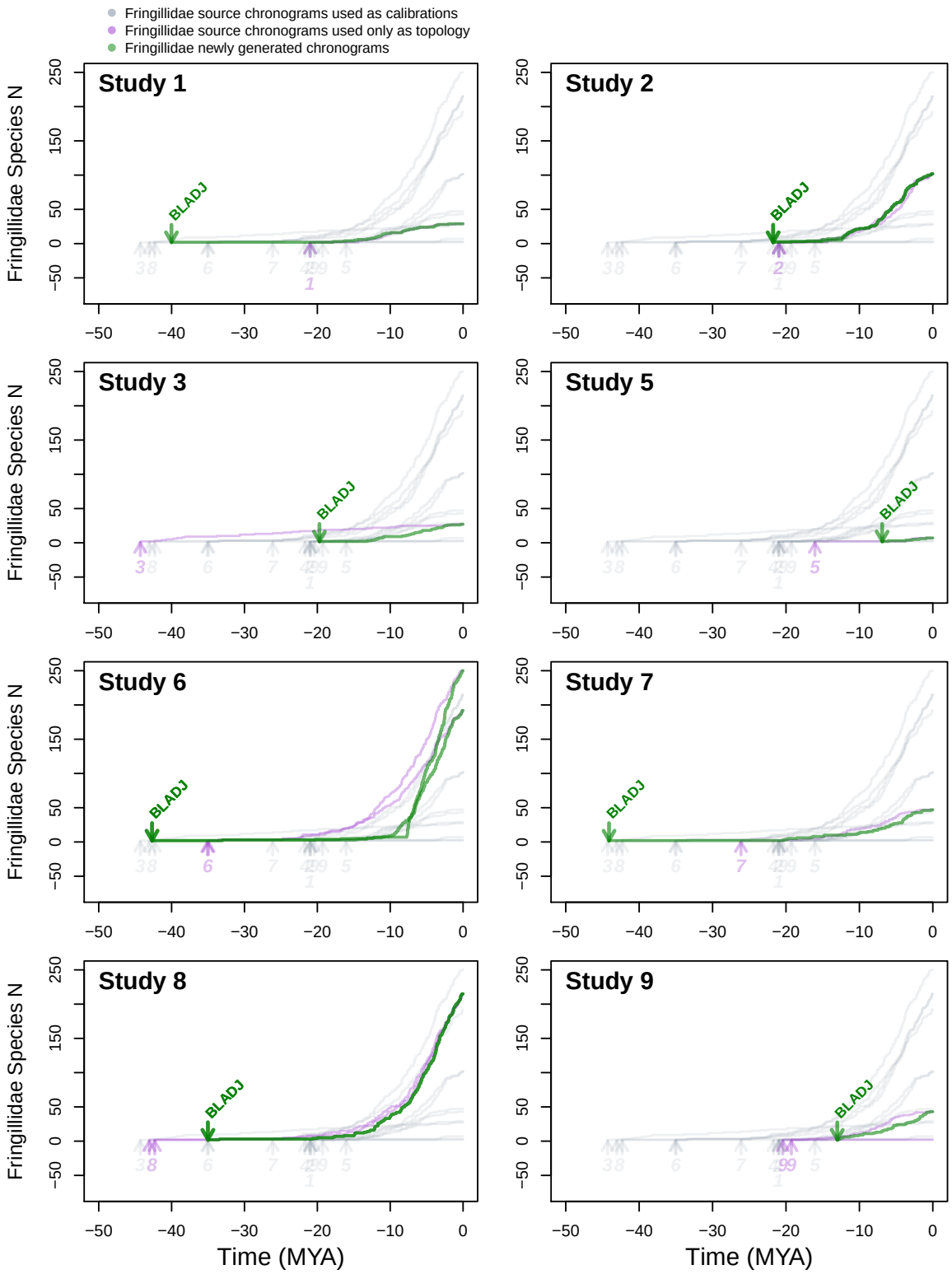
Fringillidae source chronograms used as calibrations
Fringillidae source chronograms used only as topology
Fringillidae newly generated chronograms
Study 1
Study 2
Study 3
Study 5
Study 6
Study 7
Study 8
Study 9
BLADJ
Fringillidae Species N
Time (MYA)

Figure 5:

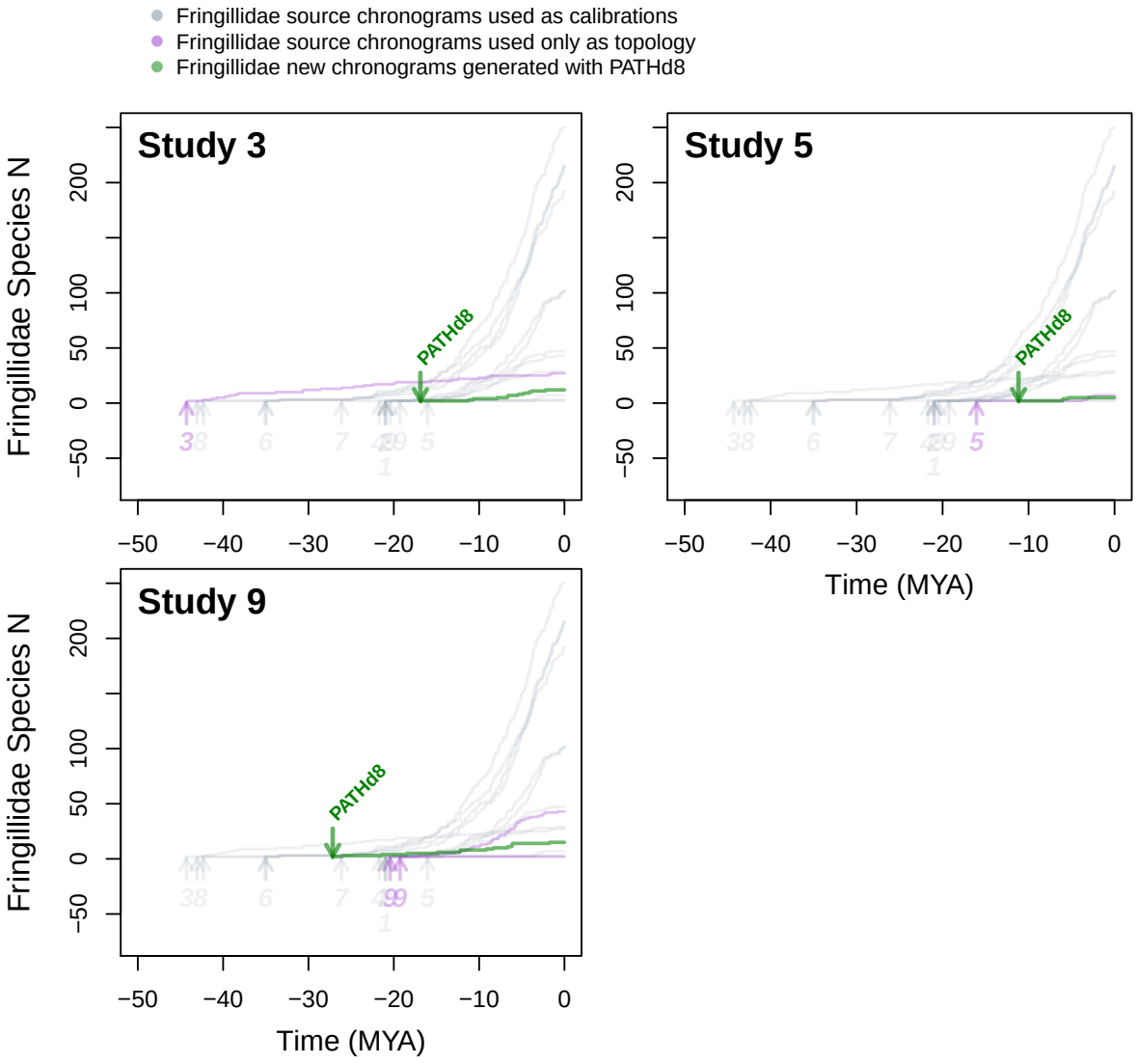
Fringillidae source chronograms used as calibrations
Fringillidae source chronograms used only as topology
Fringillidae new chronograms generated with PATHd8
Study 3
Study 5
Study 9
PATHd8
Fringillidae Species N
Time (MYA)
−50
−40
−30
−20
−10
0
−50
0
50
100
200

Figure 6: