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.- Here goes the abstract.

Keywords: Tree; Phylogeny; Scaling; Open; Ages; Congruify; Supertree;

Supersmart is an open tool but not easy to use, still requires a lot of curation and knowledge.

Time of lineage divergence constitutes in many ways the fundamental/main knowledge necessary for evolutionary understanding. Coupled to species number and distribution, it is the main information necessary for the study of diversification processes (i.e., the tempo and mode of speciation and extinction), central for the understanding of how biodiversity patterns are shaped across space, time and clades (Morlon 2014). Evolutionary understanding also relies on comparative studies, for which knowing the time context for all life is crucial. Efforts to have a whole tree of life have been great and here are some examples. In the past two decades, the possibility to obtain good quality DNA sequences coupled to methodological developments in phylogenetic and dating inference, allowed the application of molecular dating methods on a very large amount and diversity of organisms, greatly increasing the quantity of data on taxon ages across the tree of life. To date, there is a large amount of both fossil and molecular-based data on taxon ages and phylogenetic relationships in public repositories such as Dryad, TreeBASE and Open Tree of Life (OToL). OToL alone holds more than 200 chronograms. Methods to include living and fossil lineages are in continued development and increased usage by the community, which coupled to better sharing data practices, are greatly contributing to the accumulation in number and type of available data on taxon ages.

The TimeTree project (Hedges et al. 2006, 2015; Kumar et al. 2017) has aggregated chronograms from 3,163 studies, encompassing 97,085 species (Kumar et al. 2017), and continues to grow. However, even in this gold standard resource, the included taxa only encompass between 0.097 and 3.236% of total species diversity (following taxonomic expert opinion on the global, extant species numbers, which ranges from 3 to 100 million species [Mayr2010; Moran2011]). One advantage of TimeTree is that it includes taxa from across the tree of life, versus more specialized chronograms focusing on plants [PHYLOMATIC], birds [JETZ ET AL BIRDTREE.ORG], and other groups. Users can choose between a web interface or a mobile app to receive information on divergence times for the evolutionary history of a lineage, pairs of taxa, all lineages within a taxon, or a list of taxa. As a science communication tool, TimeTree project is very powerful: it has a friendly graphical interface, with informative and colorful outputs, that allows the general public to satisfy curiosity regarding a particular organism of interest or group of them. It is of limited utility for scientific

studies, however. The thousands of trees that have been entered are unavailable for examination or reuse; according to the creators (see TimeTree web FAQ), methods for allowing data downloading have been under discussion for the past several years yet the primary data remain closed. Moreover, there is no Application Programming Interface (API) allowing programmatic access to any data, greatly impairing the possibility of large-scale, automated data-mining, which is not allowed under TimeTree website's terms of use. The nearly hundred thousand taxon summary chronogram generated from TimeTree resources is not available with its publication (Kumar et al. 2017) or the TimeTree website, though the still substantial chronogram from a previous publication (Hedges et al. 2015) was made available at OToL.

Despite its great importance, analytical tools to summarize available information on taxon ages for the scientific community are still lacking. We identified several aspects that might have so far delayed the exploitation of existing data. First, original chronograms available publicly are scattered across various repositories (otol tree store, dryad, treebase, journals supplementary data) usually with different formats too. Second, lineage names due to taxonomic idiosincracy can be different among studies and manual curation of that is usually necessary. Third, data curation Recent advances on this area (e.g., supersmart) aim to: Generate new dates using all available DNA sequence information; Perform one global analysis using all available information; Problems or downsides: This might be time consuming for large groups and a lot of data curation and knowledge on the group of interest is still necessary. For example, choosing correct fossils for calibration requires a lot of expertise and knowledge on the group. An incorrect use of fossils can generate severe bias in dating results (Sauquet et al. 2012). Hence, data curation is still an important part of any biological study. The research community considers it as an important or even crucial step before data analysis. Hence, automated processes for large data analysis are frequently received with skepticism.

DateLife palliates this by only using information available from already published studies, which are ideally constructed using robust information, such as sequence data and thoughtfully curated fossil calibrations.

Rapidly increasing data on time of lineage divergence both from molecular and paleontological studies; the increasing importance of use of these data in distant areas of research, often not specialized enough to rapidly obain data on their own; and the lack of an open (both the data sources and the code underlying

the analyses) easy to use tool inspired the development of a prototype `DateLife` service over a series of phylotastic hackathons (Stoltzfus et al. 2013) at the National Evolutionary Synthesis Center. In this paper we present the first formal description of `DateLife`, featuring an improved database of chronograms, more methods to summarize trees, and new functions to visualize data, as well as comparisons of summary trees. DateLife is the main service for scaling phylogenetic trees in Phylotastic! system (Stoltzfus et al. 2013) It can be used through an R package , a web interface (http://www.datelife.org/query/) and an API.

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) `DateLife` then performs a search across its database of peer reviewed and curated chronograms; identifies and gets source trees with at least two matching input names; drops unmatching taxa from positively identified source trees; and finally transforms each source tree to a patristic matrix named by the citation of the original study. This format facilitates and greatly speeds up all further analyses and summarization algorithms.
3) The user can obtain different types of summaries from the source data including: a) all source chronograms, b) mrca 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 source chronograms, and f) a report of succesful matches per input taxon name across source chronograms.
4) At this point, users can choose to use all or some source data as calibration points to date a tree of their own making or choosing.
5) Users can also simulate age and/or phylogenetic data of input taxa not found in the database. A variety of algorithms are available for this purpose.
6) Finally, users can easily view results graphically as well as construct their own graphs using inbuilt `DateLife` graphic generators.

`DateLife`'s chronogram database is currently built from Open Tree of Life (OToL)'s (Hinchliff et al. 2015) tree repository. Among currently existing repositories (e.g., TreeBase, Dryad), OToL's metadata rich tree store is the only one meeting the requirements for proper/accurate automatized handling of trees. Input taxon names accepted by `DateLife` are binomial species names or clades. Taxon searches are performed at the species level, so when input names correspond to higher clades, `DateLife` pulls all accepted species names within the clade from OToL's reference taxonomy to perform the search. Currently, searches at the infrapsecies level are not allowed, so input names belonging to subspecies or any other infraspecific category are treated as species. `DateLife` also processes input names with the taxon name resolution service (TNRS), which corrects potentially misspelled names and typos, and standardizes variation in spelling and synonyms (Boyle et al. 2013), increasing the probability to correctly find the queried taxa in `DateLife`'s chronogram database.

Source chronogram summary tree can be assembled using the Super Distance Matrix (SDM) supertree construction approach (Criscuolo et al. 2006) or using the median of branch lengths and the hierarchical clustering method. Tree dating and simulation options 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. treePL, is a semi-parametric, rate-smoothing, penalized likelihood dating method (Smith and O'Meara 2012). 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, th econgruification algorithm described in (Eastman et al. 2013) is used to find shared nodes between trees (congruent nodes).

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), phylocomr (Ooms and Chamberlain 2018) and rphylotastic (O'Meara et

al. 2019).

Details on each step are further developed in `DateLife`'s R package documentation `datelife workflow` vignette at (https://LINK).

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 relative to number of input names and `DateLife` service. Input processing increases roughly linearly with number of input taxon names, and increases considerably if tnrs service is activated (Fig. 2). Results show that searching time increases linearly with number of input names and number of chronograms in database.

Summarizing DateLife results processing times

Adding dates processing time

get_bold_otol_tree running time

`DateLife`'s code performance was evaluated with a set of unit tests designed and implemented with the R package testthat (R Core Team 2018). These tests were run both locally –using the devtools package (R Core Team 2018)– and on a public server –via GitHub– using the continuous integration tool Travis CI (<//travis-ci.org>). At present, unit tests cover more than 50% (for now) of `DateLife`'s code (https://codecov.io/gh/phylotastic/datelife).

Biological example

In this section we demonstrate the types of outputs that can be obtained with `datelife`, using the bird family Fringillidae of true finches as example. We performed a higher-taxon search to obtain all data on lineage divergence available from `datelife`'s database for all recognised species within the Fringillidae (475 spp. according to the Open Tree of Life taxonomy). We found 13 trees across 9 studies (Fig. 3.

Conclusions

Taxon ages are key to many areas of evolutionary studies: trait evolution, species diversification, biogeography, macroecology and more. Obtaining these ages is difficult, especially for those who want to use phylogenies but who are not systematists, or do not have the time to develop the necessary knowledge and data curation skills to produce new chronograms. Knowledge on taxon ages is also important for non-biological studies and the non-academic community. 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 or 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.

Here, we have shown that DateLife allows an easy and fast obtention of all publicly available information on taxon ages, which can be used to generate new data. This information can be used to account for the effect of phylogenetic signal in studies of trait evolution; to explore potential speciation and extinction dynamics of interest within a clade; to obtain a time frame of biogeographical events; for science communication and outreach, amongst others. Compared to similar platforms such as time tree of life and supermart, it offers several advantages. It is fast; source data is completely open; it requires no expert biological knowledge from users for any of its functionalities; 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; it facilitates reproducibility of analyses;

Improvements, short and long-term: * 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 2016). * bayesian congruification * topological congruification

Problems and caveats: Not many databases, only OToL Why TreeBase is not very useful for us? Be precise. Are these chronograms reliable to study evolutionary patterns, such as species diversification? DateLife can be seen as an open resource to know the current state of knowledge on lineage divergence times. Whether

chronograms obtained using this original data can be used reliably to study complicated patterns of evolution is still uncertain. If all, la facilidad para obtener hipotesis de tiempo de divergencia nos ayudará a evaluar la capacidad de los cronogramas para estudiar otros fenomenos evolutivos. Por ahora, no podemos aseverar que estos cronogramas puedan usarse para todo tipo de analisis.

Availability

DateLife is free and open source and it can be used through its current website http://www.datelife.org/query/, or through Phylotastic!'s web portal http://phylo.cs.nmsu.edu:3000/. RStudio's Shiny Server and the shiny package open infrastructure are used to maintain the former. Also Docker. DateLife can also be used locally through its R package. The 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 R. Development versions are available from GitHub repository (https://github.com/phylotastic/datelife) and can be installed using the devtools R package command `install_github("phylotastic/datelife")`.

Supplementary Material

Supplementary material, including code files and online-only appendices and vignettes, can be found in the GitHub repository of the paper at (https://github.com/LunaSare/datelife_paper1) as well as in the package vignettes and are also available from the Dryad Digital Repository at LINK.

Funding

Funding was provided by NSF grant 1458603

NESCent

Open Tree of Life

University of Tennessee, Knoxville

Acknowledgements

We thank colleagues (students and postdocs) at 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 US National Science Foundation (NSF) for funding nearly all the above, in addition to the ABI grant that funded this project itself.

References

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

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.

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.

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.

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. 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.

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.

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.

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.

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.

Sauquet H., Ho S.Y.W., Gandolfo M. a, Jordan G.J., Wilf P., Cantrill D.J., Bayly M.J., Bromham L., Brown G.K., Carpenter R.J., Lee D.M., Murphy D.J., Sniderman J.M.K., Udovicic F. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). Systematic Biology. 61:289–313.

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

Smith S.A., O'Meara B.C. 2012. TreePL: Divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics. 28:2689–2690.

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.

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.

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.

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.

Claramunt S., Cracraft J. 2015. A new time tree reveals earth history's imprint on the evolution of modern birds. Science advances. 1:e1501005.

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.

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.

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

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.

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.

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 input processing and search across `datelife`s chronogram database.

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 summarizing information from source chronograms found for the Fringillidae. Arrows indicate maximum age.

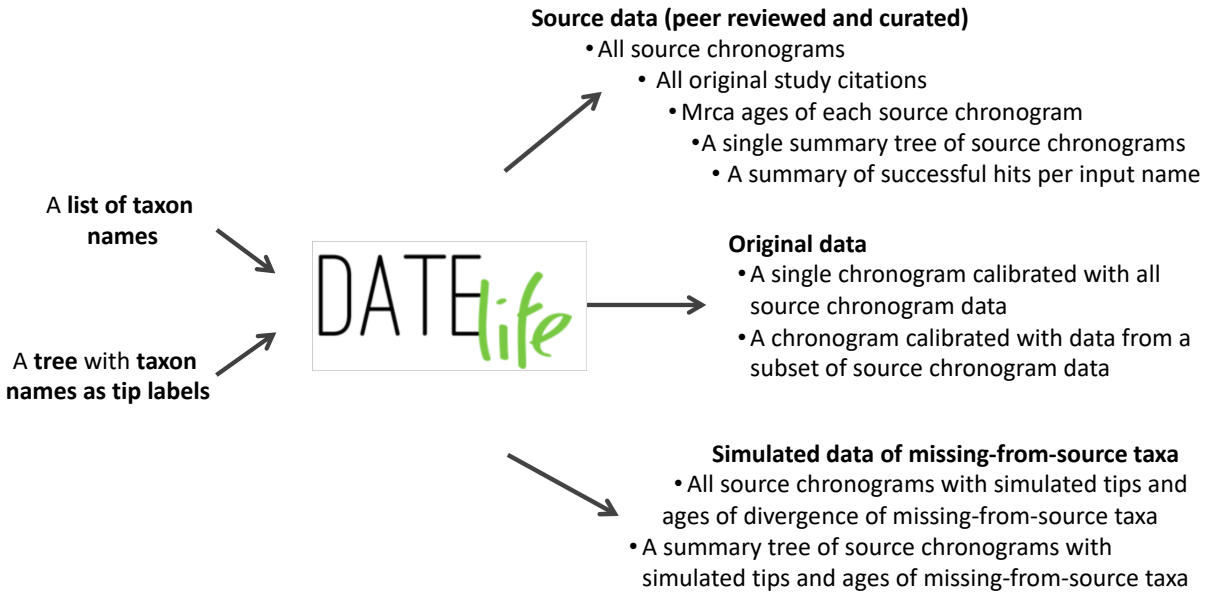
Source data (peer reviewed and curated)
• All source chronograms
• All original study citations
• Mrca ages of each source chronogram
• A single summary tree of source chronograms
• A summary of successful hits per input name
A list of taxon names
A tree with taxon names as tip labels
DATElife
Original 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:

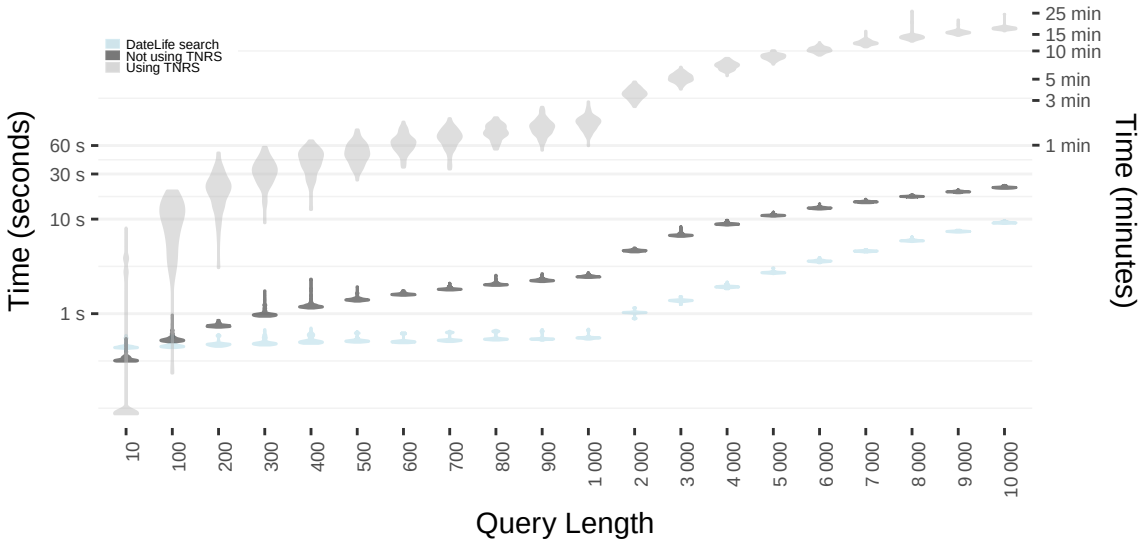
DateLife search
Not using TNRS
Using TNRS
Time (seconds)
60 s
30 s
10 s
1 s
25 min
15 min
10 min
5 min
3 min
1 min
Time (minutes)
10
100
200
300
400
500
600
700
800
900
1 000
2 000
3 000
4 000
5 000
6 000
7 000
8 000
9 000
10 000
Query Length

Figure 2:

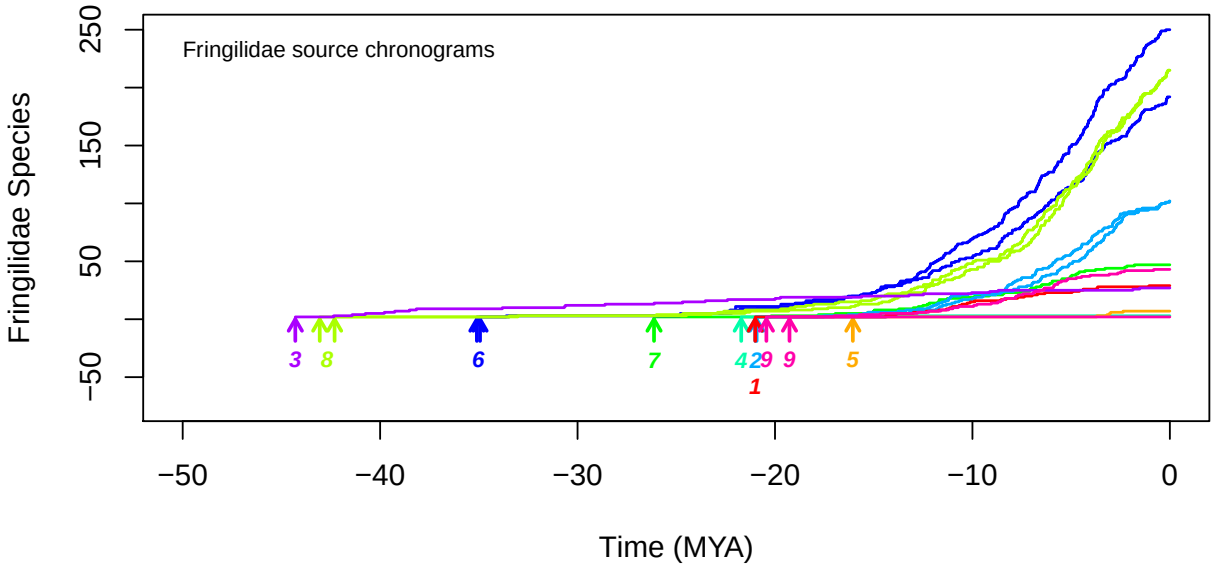
Fringilidae source chronograms
Fringilidae Species
250
150
50
−50
−50
−40
−30
−20
−10
0
Time (MYA)
3
8
6
7
4
2
9
9
1
5

Figure 3:

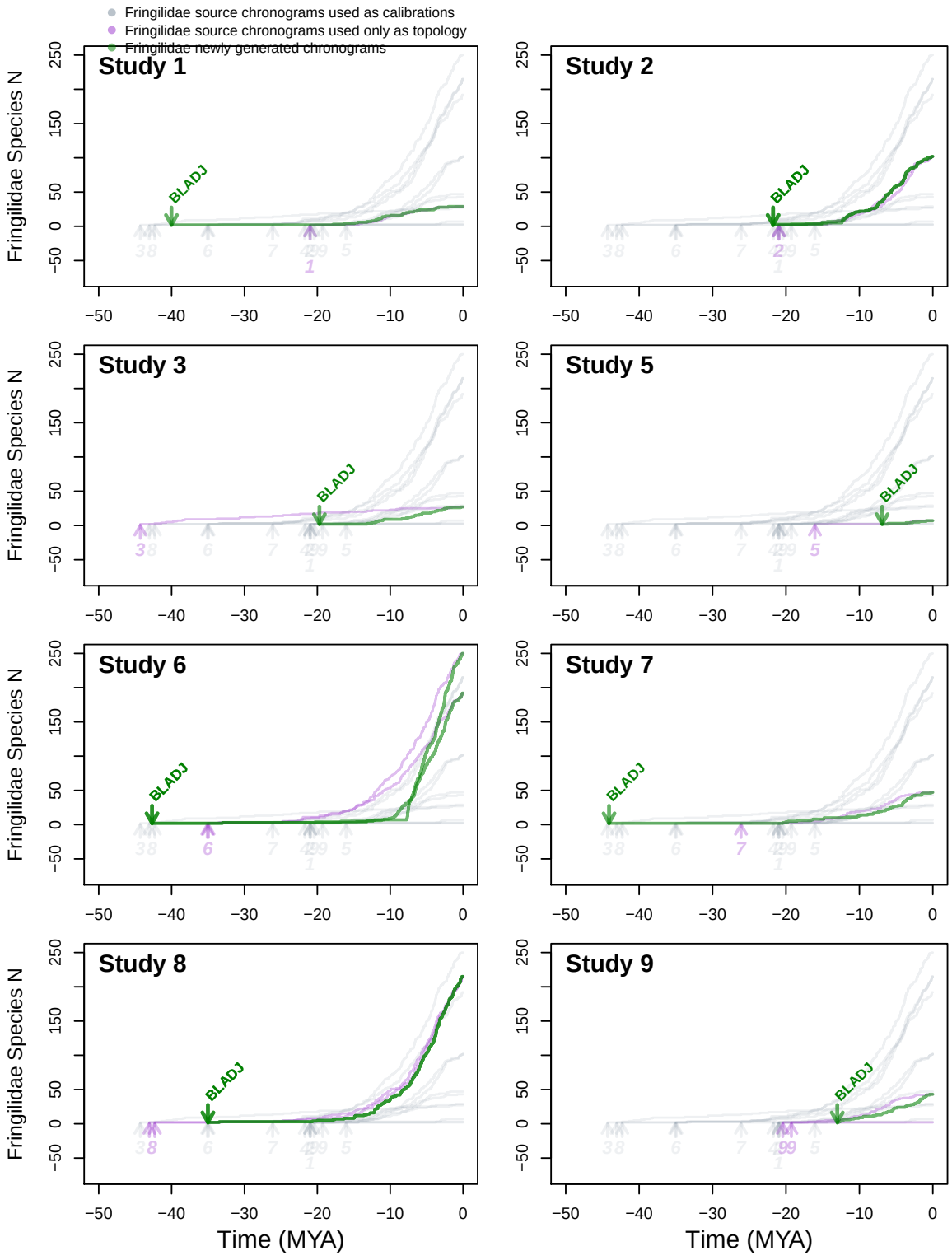
Fringilidae source chronograms used as calibrations
Fringilidae source chronograms used only as topology
Fringilidae newly generated chronograms
Study 1
Study 2
Study 3
Study 5
Study 6
Study 7
Study 8
Study 9
BLADJ
Fringilidae Species N
Time (MYA)

Figure 4: