

1 Running head: DATELIFE: REVEALING THE DATED TREE OF LIFE

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

3 Authors: Luna L. Sánchez-Reyes¹, Brian C. O'Meara¹

4 Correspondence address:

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

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

8 **abstract.-** Here goes the abstract.

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

Divergence time of lineages constitutes in many ways the main knowledge necessary for evolutionary understanding. Coupled to species number and distribution, it is the basic information for the study of diversification processes, such as the tempo and mode of speciation and extinction, crucial for the understanding of how biodiversity patterns are shaped across space and time (Morlon 2014).

When organisms are preserved in a fossil form, a time frame of taxon origin can be obtained directly from the age of rock strata. However, not all organisms fossilize well or at all. Fossilization success alone is highly circumstantial, and varies depending on a number of parameters including the nature of the habitat, population size, species range breadth and physical characteristics of the organism. Thus, relying only on the fossil record to obtain a time frame of lineage divergence for all life is not possible. Relative rate of DNA or aminoacid substitution constitutes another important source of information on lineage divergence. It is usually obtained from hypothesis of character homology (alignments) when reconstructing phylogenetic relationships. Molecular dating techniques use external data such as absolute time calibrations (e.g., fossils, geologic events) or absolute substitution rates to generate dated phylogenies (chronograms) which contain information on absolute times of node divergence and taxon ages.

In the past decades, the possibility to obtain DNA sequences in large quantities from a wide variety of organisms became a reality, which, coupled to methodological development 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.

One inspiration for the DateLife project was the TimeTree project (Hedges et al. 2006, 2015; Kumar et al. 2017). This project 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

include 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 a very powerful tool: it has a friendly graphical interface, with informative while colorful outputs, that allows the general public to satisfy curiosity regarding their favorite organism 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 their FAQ, methods for allowing download of the data have been under discussion since 2015. Also, there is no Application Programming Interface (API) allowing programmatic access to any data. Moreover, large-scale, automated, data-mining is not allowed under the site's terms of use. The nearly hundred thousand taxon chronogram is not available with (Kumar et al. 2017) or the TimeTree website, though the still substantial chronogram from (Hedges et al. 2015) is available at OTOL.

With this inspiration, a prototype DateLife service was developed over a series of phylotastic hackathons [CITE] at the National Evolutionary Synthesis Center. NSF funding allowed for the further development leading to this paper. A core goal has been openness of both the data sources and the code underlying the analyses.

Despite its importance, analytical tools to summarize available information on taxon ages for the scientific community are still lacking. Data might not have been exploited because: Data is in different repositories and formats; Lineage names are different among studies and difficult to reconcile; Taxonomy is also different among studies and difficult to reconcile. Also, data curation is 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. Recent work on this area (e.g., supersmart) aims 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 is still necessary. Choosing correct fossils for calibration requires a lot of expertise and knowledge on the group. Incorrect use of fossils can generate severe bias in dating results (Sauquet et al. 2012). DateLife palliates this by using only information available from already published studies, which are ideally constructed using robust information, such as sequence data and curated fossil calibrations. DateLife can summarize this information in several formats that can be easily inspected by users. This allows rapidly obtaining a time frame of lineage divergence for a wide number of taxa. DateLife can also generate chronograms for taxa with little available information, by using the available data as calibration points. 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

DateLife is a service for searching and processing information on ages of any taxon of interest, available in chronograms from public data repositories of published peer reviewed studies. With this data, it can also generate new taxon age information by calling a variety of external services and proved methods. It only requires as input a set of taxon names, in the form of a comma separated listing or vector, or of a phylogeny with taxon names on the tips. Taxon names can correspond to binomial species names or clades. When taxon names are clades, DateLife pulls all accepted species names within the clade (up to OToL's limit of _____ species) from OToL's reference taxonomy using a service of rphylotastic R package. Names belonging to subspecies or any other infraspecific category are treated as species. DateLife can process input names with the taxon name resolution service (TNRS), which corrects misspelled names or typos, and standardizes variation in spelling and synonyms (Boyle et al. 2013), increasing the probability to correctly find the queried taxa in the chronogram database. DateLife uses TNRS to compare names against OToL's reference taxonomy using a service from the R package rotl (Michonneau et al. 2016).

DateLife's main function searches taxon names across the chronogram database specified by the user. At the moment, it queries chronograms from the OToL (Hinchliff et al. 2015) repository. DateLife identifies chronograms having at least two taxon names, and subset them to contain only the taxa of interest. It

then stores taxon age information from each chronogram individually as a patristic matrix, named with the citation of the original study. This format allows a rapid summary in a number of different ways, including: 1) citations of the original studies containing the subset chronograms, 2) a list of subset chronograms' mrca ages, 3) a list of subset chronograms in newick or phylo format, 4) a table containing all information in html or R's data frame format, or 5) as a single chronogram summarized from subset chronograms using the supertree Super Distance Matrix (SDM) approach (Criscuolo et al. 2006) or using the median of branch lengths.

DateLife also stores information on input taxon presence/absence across subset chronograms. Users can choose to add ages of missing taxa to subset chronograms in different ways, depending on the amount of knowledge they want to input or how much they want to be involved in the steps of the addition process. If users have no access to biological information (i.e., a character, DNA or protein matrix), missing taxa can be added to any chronogram simply at random, or by following taxonomic or phylogenetic knowledge from expert sources. There are a wide number of open reference taxonomies available, such as the Catalogue of Life (Roskov et al. 2017) or the NCBI taxonomy database (Federhen 2012). Expert phylogenies (with or without branch lengths) to be used as topological constraint (backbone) can also be obtained from a number of public repositories, such as OTOL (Hinchliff et al. 2015), TreeBASE (Piel et al. 2002) and Dryad (<https://www.datadryad.org/>). At the moment, DateLife only uses OTOL's synthetic tree and reference taxonomy as expert knowledge to automatically add missing taxa to chronograms. Alternatively, users can input a reference taxonomy or topological constraint of their choosing or making. If OTOL's synthetic tree is not satisfactorily resolved for the taxa of interest, DateLife can construct a sequence data matrix from DNA markers available from the Barcode of Life Database (BOLD; Ratnasingham and Hebert (2007)), to attempt to further resolve polytomies. It will follow OTOL's synthetic tree as backbone. To use information from a topological constraint, DateLife calls the congruification method described in (Eastman et al. 2013) to find shared nodes between trees (congruent nodes). It then fixes their ages, and add ages to remaining nodes with a dating method that can be specified by the user. If users have access to biological data, they can input a tree with branch lengths proportional to relative substitution rates as topological constraint. In this case, age data from congruent nodes will be used as calibration points. Age data from several chronograms can be combined and congruified to be used as calibration points in a single analysis.

Several dating methods are implemented in DateLife. 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). DateLife implements BLADJ from the development R version of phylocom’s R package (Webb et al. 2008), phylocomr (<https://github.com/ropensci/phylocomr>). It can only be used when there is a topological constraint with no branch lengths. PATHd8 is a non-clock, rate-smoothing method (Britton et al. 2007) to date trees. It is also called through R. treePL, is a semi-parametric, rate-smoothing, penalized likelihood dating method (Smith and O’Meara 2012). It is called through R. MrBayes program (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). DateLife calls MrBayes through an R function.

DateLife can also correct negative branch lengths in several ways.

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. 1). 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 (<https://travis-ci.org>). At present, unit tests cover around 30% (for now) of DateLife’s code (<https://travis-ci.org>).

139 [//codecov.io/gh/phylostatic/datelife](https://codecov.io/gh/phylostatic/datelife)).

140 BIOLOGICAL EXAMPLE

141 Find a clade with at least one chronogram containing all clade's species. (Penguins look good, but they are
142 giving weird results in SDM)

143 Remove this chronogram from datelife Results.

144 Make sdm and median trees and Compare

145 add taxa with different methods and Compare

146 Use ltt to compare for now. Fig. X2 shows comparison of available chronograms for Felidae species and
147 chronograms generated through DateLife

148 think of a test to compare trees, topology- and date-wise

149 CONCLUSIONS

150 Taxon ages are key to many areas of evolutionary studies: trait evolution, species diversification, biogeography,
151 macroecology and more. Obtaining these ages is difficult, especially for those who want to use phylogenies
152 but who are not systematists, or do not have the time to develop the necessary knowledge and data curation
153 skills to produce new chronograms. Knowledge on taxon ages is also important for non-biological studies and
154 the non-academic community. The combination of new analytical techniques, availability of more fossil and
155 molecular data, and better practices in data sharing has resulted in a steady accumulation of chronograms
156 in public and open databases such as Dryad, TreeBASE or Open Tree of Life, for a large quantity and
157 diversity of organisms. However, this information remains difficult to synthesize for many biologists and the
158 non-academic community.

159 Here, we have shown that DateLife allows an easy and fast obtention of all publicly available information on
160 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:

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/>. 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, can be found in the GitHub repository

FUNDING

Funding was provided by NSF grant 1458603

NESCent

Open Tree of Life

University of Tennessee, Knoxville

ACKNOWLEDGEMENTS

REFERENCES

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.

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.

Federhen S. 2012. The NCBI Taxonomy Database. *Nucleic Acids Research*. 40:D1086–D1098.

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.

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.,

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

207 Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*.
 208 17:754–755.

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

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

213 Michonneau F., Brown J.W., Winter D.J. 2016. rotl: an R package to interact with the Open Tree of Life
 214 data. *Methods in Ecology and Evolution*. 7:1476–1481.

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

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

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

221 Ratnasingham S., Hebert P.D.N. 2007. BARCODING, BOLD : The Barcode of Life Data System
 222 (www.barcodinglife.org). *Molecular Ecology Notes*. 7:355–364.

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

225 Roskov Y., Abucay L., Orrell T., Nicolson D., Bailly N., Kirk P., Bourgoin T., DeWalt R., Decock W., De
 226 Wever A., Nieukerken E. van, Zarucchi J., Penev L. 2017. Species 2000 & ITIS Catalogue of Life. Digital

227 resource at www.catalogueoflife.org/col. Species 2000: Leiden, the Netherlands: Naturalis.

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

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

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

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

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