DateLife Workflows

Luna L. Sanchez Reyes

2019-04-23

Taxon Spheniscidae

I. Query source data

There are 25 species in the Open Tree of Life Taxonomy for the taxon Spheniscidae. Information on time of divergence is available for 19 of these species across 13 published and peer-reviewed chronograms. Original study citations as well as proportion of Spheniscidae species found across those source chronograms is shown in Table 1.

All source chronograms are fully ultrametric.

Table 1: Spheniscidae source chronogram studies information.

	Citation	*Source N*	*Taxon N*
1.	Claramunt, Santiago, Joel Cracraft. 2015. A new time tree reveals Earth historys imprint on the evolution of modern birds. Science Advances 1 (11): e1501005-e1501005	2	2/25
2.	García–R, Juan C., Gillian C. Gibb, Steve A. Trewick. 2014. Eocene diversification of crown group rails (Aves: Gruiformes: Rallidae). PLoS ONE 9 (10): e109635	1	3/25
3.	Gavryushkina, Alexandra, Tracy A. Heath, Daniel T. Ksepka, Tanja Stadler, David Welch, Alexei J. Drummond. 2016. Bayesian Total-Evidence Dating Reveals the Recent Crown Radiation of Penguins. Systematic Biology, p. syw060	1	18/25
4.	Gibb, Gillian C., Martyn Kennedy, David Penny. 2013. Beyond phylogeny: pelecaniform and ciconiiform birds, and long-term niche stability. Molecular Phylogentics and Evolution 68 (2): 229-238.	1	3/25
5.	Hedges, S. Blair, Julie Marin, Michael Suleski, Madeline Paymer, Sudhir Kumar. 2015. Tree of life reveals clock-like speciation and diversification. Molecular Biology and Evolution 32 (4): 835-845	2	18/25

6.	Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, M. S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldon, S. Capella-Gutierrez, J. Huerta-Cepas, B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E. Green, M. W. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry, M. F. Bertelsen, F. H. Sheldon, R. T. Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M. P. C. Schneider, F. Prosdocimi, J. A. Samaniego, A. M. V. Velazquez, A. Alfaro-Nunez, P. F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou, P. Perelman, A. C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldsa, L. Orlando, F. K. Barker, K. A. Jonsson, W. Johnson, K.-P. Koepfli, S. O'Brien, D. Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alstrom, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow, W. Jun, M. T. P. Gilbert, G. Zhang. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346 (6215): 1320-1331.	1	2/25
7.	Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, A. O. Mooers. 2012. The global diversity of birds in space and time. Nature 491 (7424): 444-448	2	17/25
8.	Johnson, Jeff A., Joseph W. Brown, Jérôme Fuchs, David P. Mindell, 2016, 'Multi-locus phylogenetic inference among New World Vultures (Aves: Cathartidae)', Molecular Phylogenetics and Evolution, vol. 105, pp. 193-199	2	2/25
9.	Subramanian, S., G. Beans-Picon, S. K. Swaminathan, C. D. Millar, D. M. Lambert. 2013. Evidence for a recent origin of penguins. Biology Letters 9 (6): 20130748-20130748.	1	11/25

Source N: Number of source chronograms reported in study.
Taxon N: Number of queried taxa found in source chronograms.

Source chronograms maximum age range from 12.663 to 38.961 million years ago (MYA). As a means for comparison, lineage through time plots of all source chronograms available in data base are shown in Fig. 1

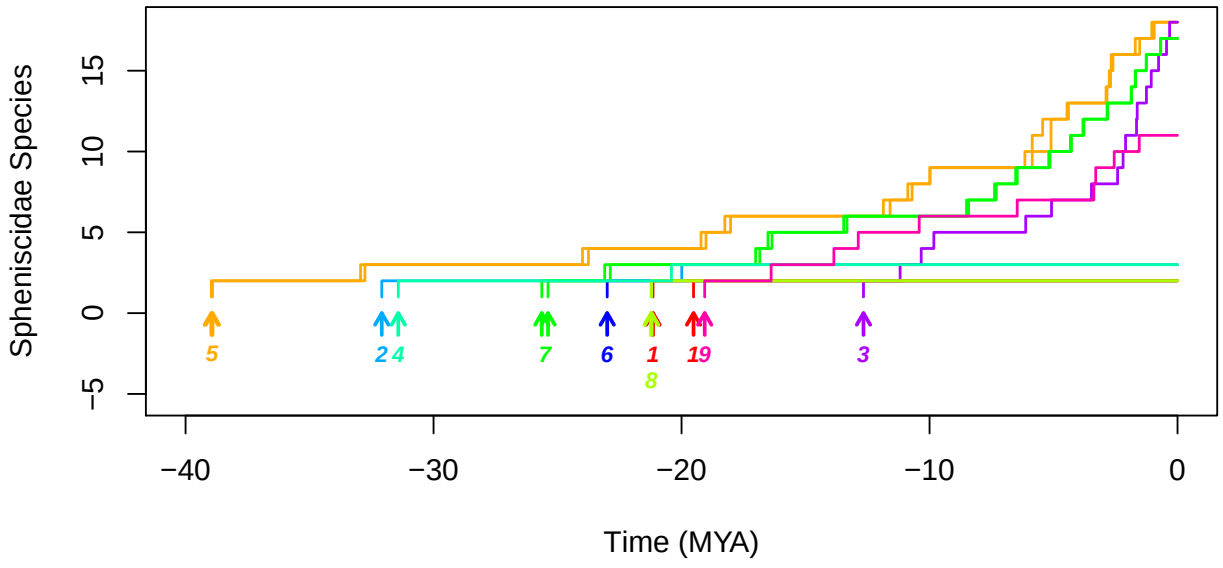

Figure 1: Lineage through time (LTT) plots of source chronograms available in data base for species in the Spheniscidae. Numbers correspond to original studies in Table 1. Arrows indicate maximum age of chronograms.

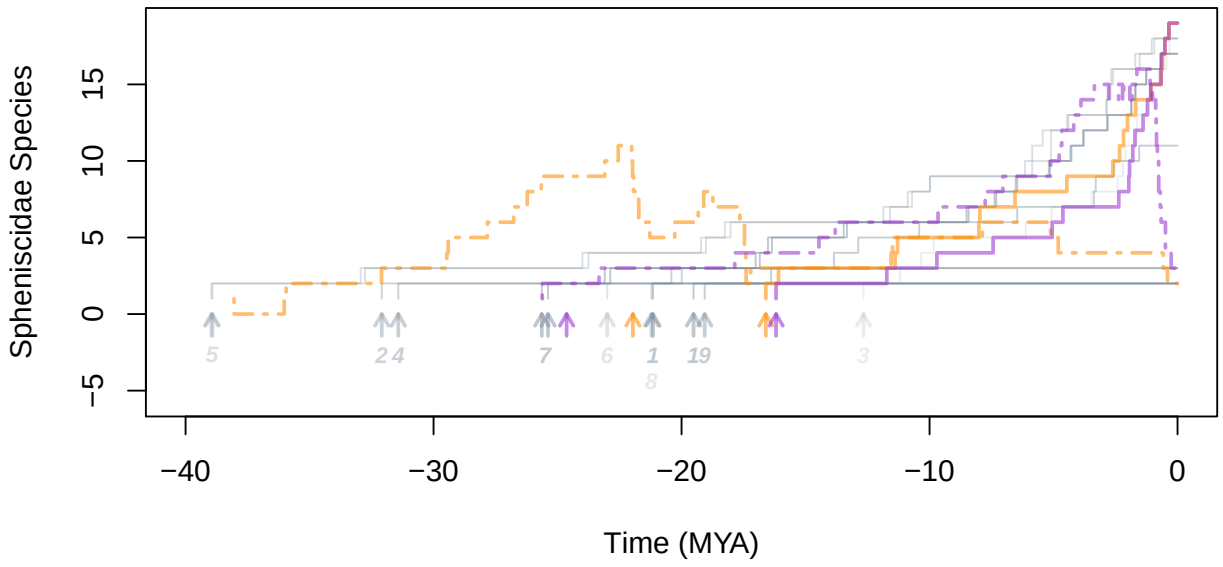

Figure 2: Test of make_lttplot_summ2 function

II. Summarize results.

LTT plots are a nice way to visually compare several trees. But what if you want to summarize all that information into a single chronogram?

The first step is to identify the degree of species overlap among your source chornograms: if each source chronogram has a unique sample of species, it will not be possible to combine them into a single summary chronogram. To identify the set of trees or *grove* with the most source chronograms that have at least two overlapping taxa, we followed Ané et al. 2016. In this case, not all source chronograms found for the Spheniscidae have at least two overlapping species. The largest grove has 2 chronograms (out of 13 total source chronograms). Now that we have identified a suitable grove we can go on to summarize it by translating the source chronograms into patristic distance matrices and then averaging them into a single summary matrix; yes, this first step is *that* straightforward. We can average the source matrices by simply using the mean or median distances, or we can use more complicated approaches that involve transforming the original distance matrices –such as the super distance matrix (SDM) approach of Criscuolo et al. 2006– by minimizing the distances across source matrices.

Once with a summary matrix, a distance-based clustering algorithm can be used to reconstruct the tree. Algorithms such as neighbour joining (NJ) and unweighted pair group method with arithmetic mean (UPGMA) are fast and work well when there are no missing values in the matrices. However, summary matrices coming from source chronograms usually have several NAs and missing rows. When this happens, even available variants of NJ and UPGMA algorithms that are designed to deal with missing data do not work well, as shown in the next section. Other methods designed to deal with missing data are BIONJ*, MVR*, and the triangle method, but we have not tried them yet.

II.A. Diagnosing clustering issues.

Clustering algorithms used to go from a summary distance matrix to a tree return trees that are too old (generally with UPGMA algorithms) or non-ultrametric (generally with NJ algorithms). In most studied cases, UPGMA returns fully ultrametric trees but with very old ages (we had to multiply the matrix by 0.25 to get ages approximate to source chronograms ages, however this number is not justified, it is just the number that approximates ages to source maximum ages the most). NJ returned reasonable ages, but trees are way non ultrametric, as you can see in Fig. S1 and Fig. 2.

This taxon's SDM matrix has some negative values in the following taxa: *Eudyptes chrysocome*, *Eudyptes filholi*. This taxon's Median matrix has NO negative values.

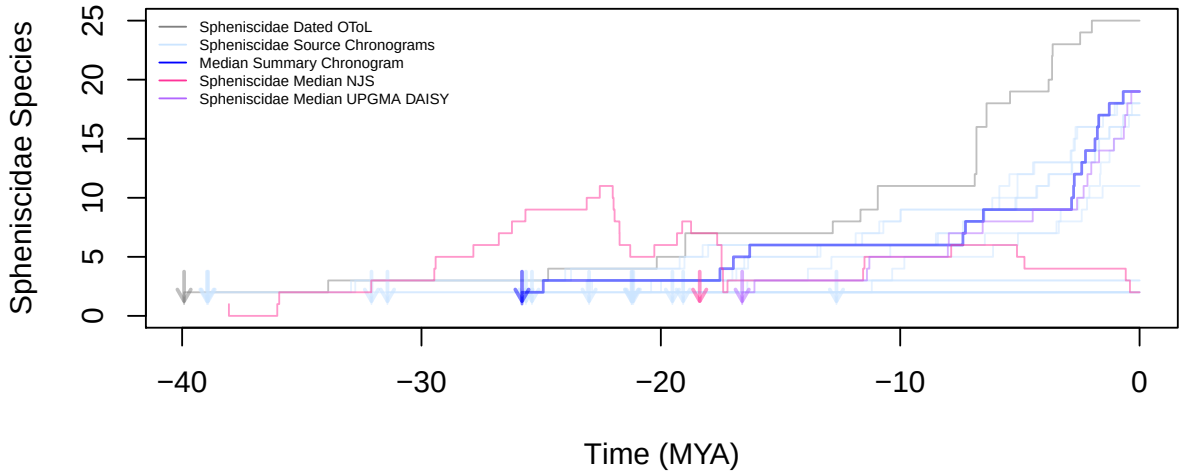

Figure 3: Spheniscidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developped an alternative algorithm in `datelife` to go from a summary matrix to a fully ultrametric tree.

II.B. Age distributions from Median and SDM summary trees.

Comparison of summary chronograms reconstructed with min and max ages.

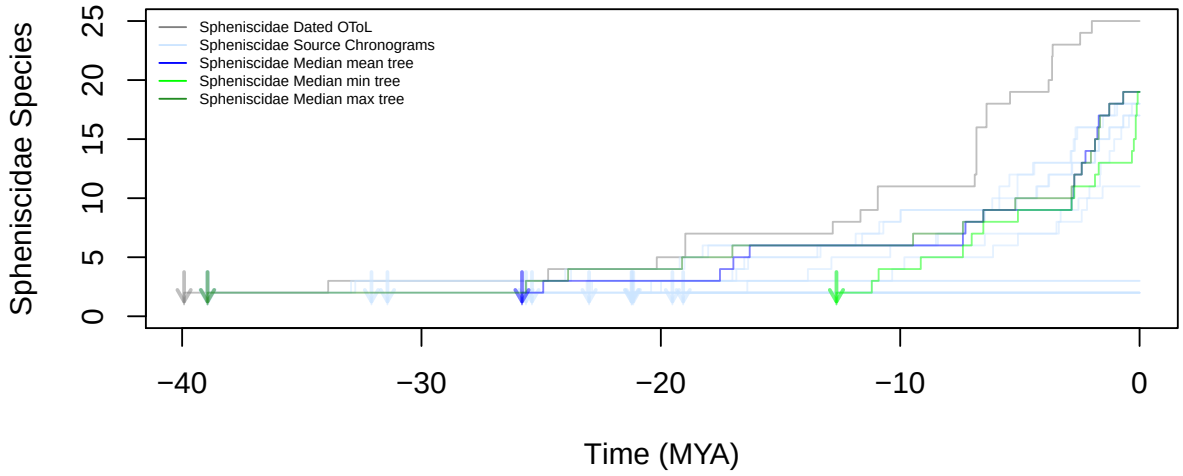

Figure 4: Spheniscidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with `datelife` algorithm.

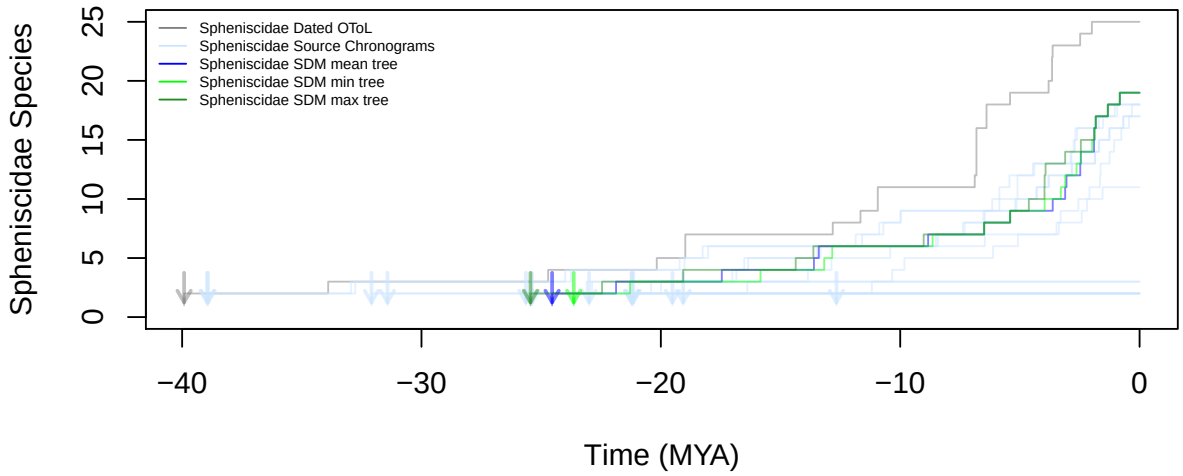

Figure 5: Spheniscidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with `datelife` algorithm.

III. Create new data

As an example, we're gonna date the Open Tree Synthetic tree (mainly because the taxonomic tree is usually less well resolved.)

Now, let's say you like the Open Tree of Life Taxonomy and you want to stick to that tree. Dates from available studies were tested over the Open Tree of Life Synthetic tree of Spheniscidae and a tree was constructed, but all branch lengths are NA. We also tried each source chronogram independently, with the Dated OToL and with each other, as a form of cross validation in Table 2. This is not working perfectly yet, but we are developping new ways to use all calibrations efficiently.

Table 2: Was it successful to use each source chronogram independently as calibration (CalibN) against the Dated Open Tree of Life (dOToL) and each other (ChronoN)?

	dOToL	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Chr12	Chr13
Calib1	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib2	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib3	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib4	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib5	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib6	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib7	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib8	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib9	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib10	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib11	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib12	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Calib13	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

III. Simulate data

An alternative to generate a dated tree from a set of taxa is to take the available information and simulate into it the missing data. We will take the median and sdm summary chronograms to date the Synthetic tree of Life:

```
#> Error in paste0("\n![", figcap_lttplot_sdm, "](plots/", taxon, "_LTTplot_sdm.pdf)\n"): object 'figcap
#> Error in cat(lttplot): object 'lttplot' not found
```

Appendix

The following species were completely absent from the chronogram data base: *Aptenodytes australis*, *Catadyptes chrysolophus*, *Eudyptes atratus*, *Eudyptula chathamensis*, *Megadyptes waitaha*, *Pygoscelis ellsworthi*

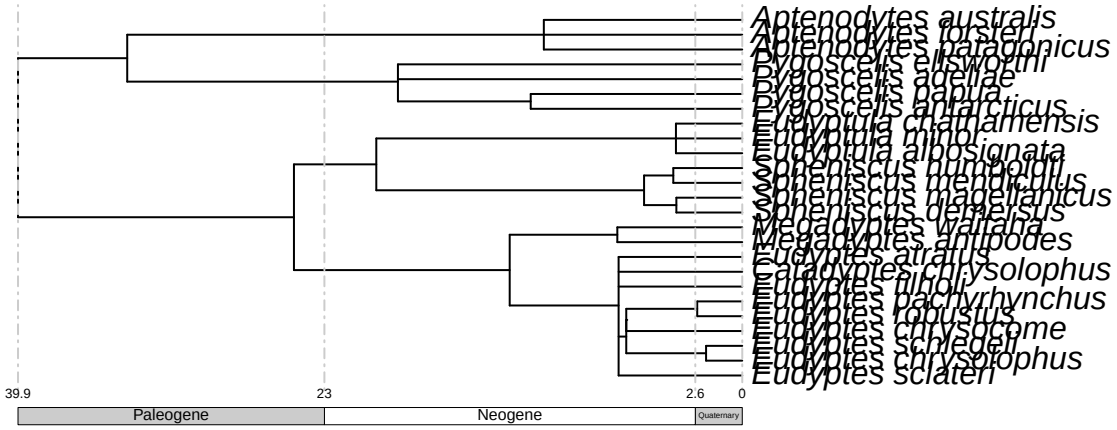

Figure 6: Spheniscidae Species Dated Open Tree of Life Induced Subtree. This chronogram was obtained with `get_dated_otol_induced_subtree()` function.

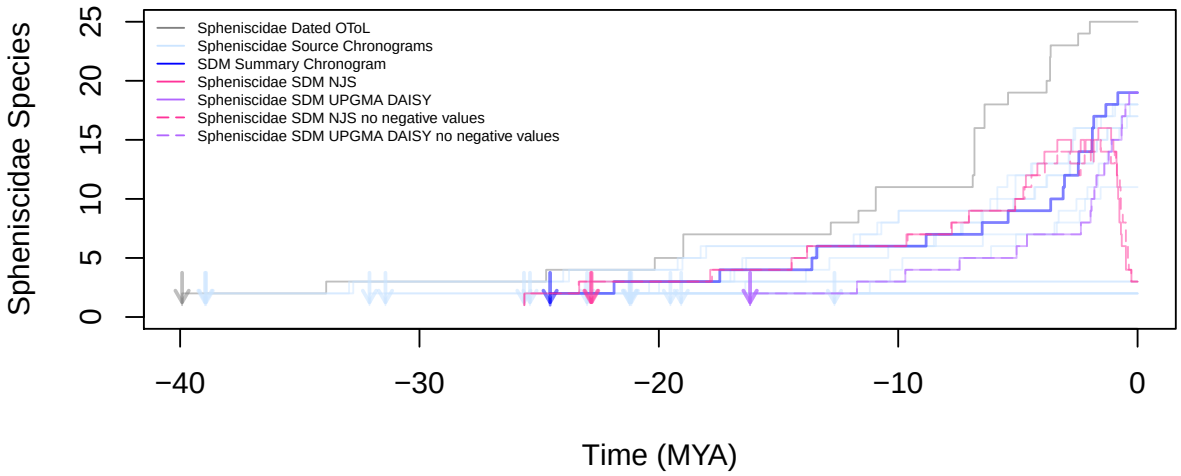

Figure 7: Spheniscidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with `datelife` algorithm.