DateLife Workflows

Luna L. Sanchez Reyes

2019-04-16

Taxon Spheniscidae

I. Query 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.

```
#> Error in sub(contents[i + 1], new_contents[i], out): invalid regular expression '\\multicolumn\{1\
```

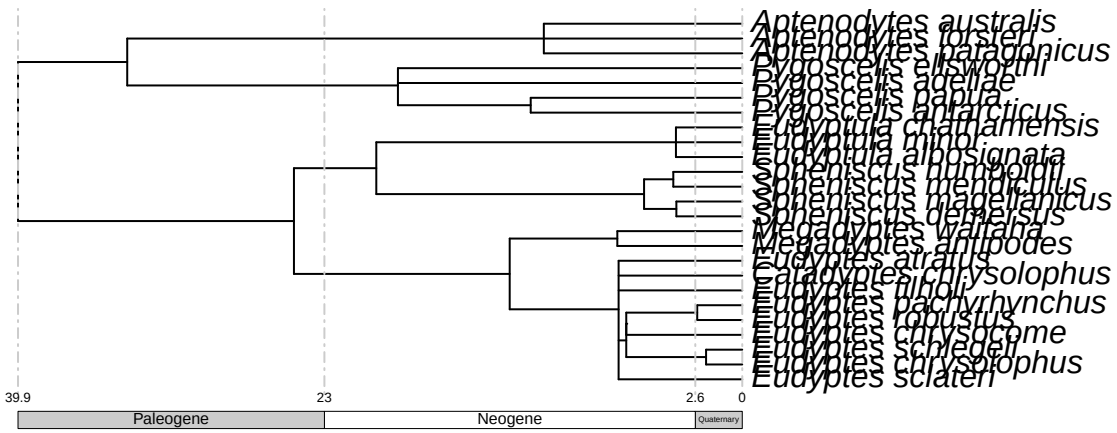

Figure 1: Spheniscidae Species Dated Open Tree of Life Induced Subtree. This chronogram was obtained with `get_dated_otol_induced_subtree()` function.

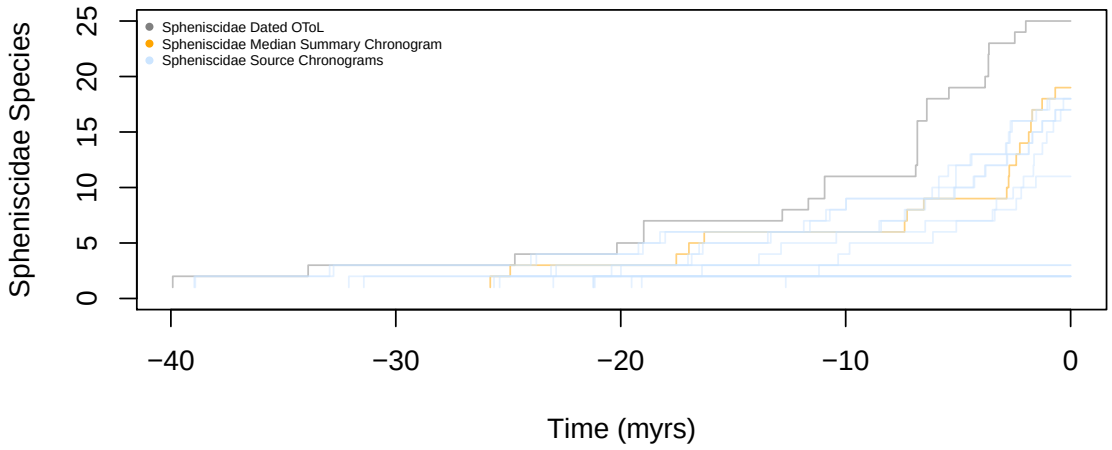

Figure 2: Spheniscidae lineage through time (LTT) plots from source chronograms, summary median chronogram and dated Open Tree of Life chronogram.

II. Summarize results.

II.A. Diagnosing clustering issues.

We identified some issues with chronograms coming from SDM and Median summary matrices. First, 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 Neighbour Joining 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 is a number chosen at random, it was just the number that worked well). NJ returned reasonable ages, but trees are way non ultrametric, as you can see in Fig. 3 and Fig. 4.

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.

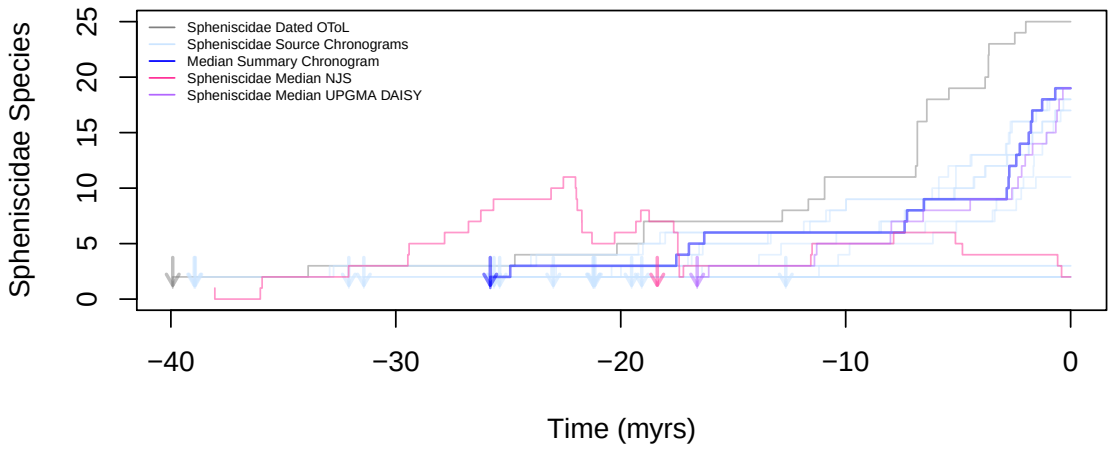

Figure 3: Spheniscidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we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

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

Comparison of summary chronograms reconstructed with min and max ages.

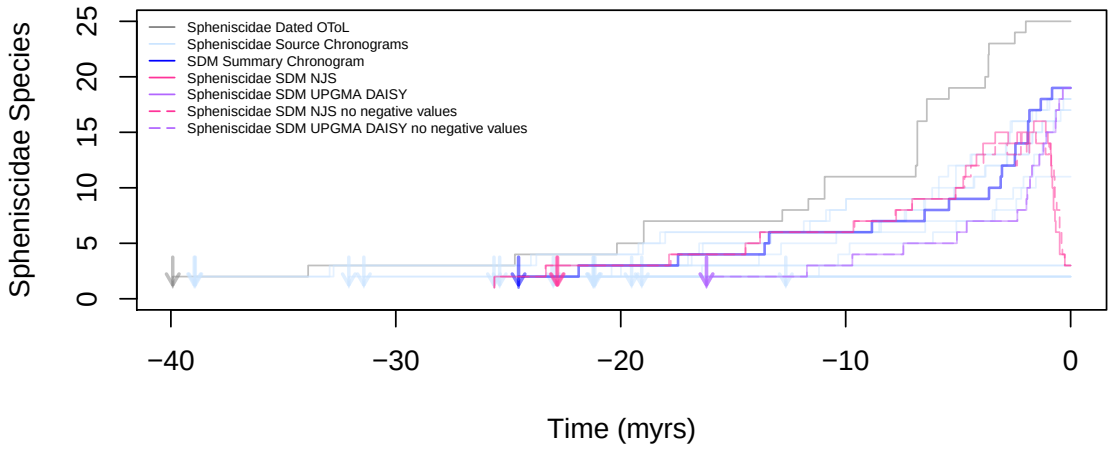

Figure 4: Spheniscidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with different methods (NJ and UPGMA). As you can note, dashed lines and solid lines from trees coming out from both types of clustering algorithms implemented are mostly overlapping. This means that removing negative values does not change results from clustering algorithms much. 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.

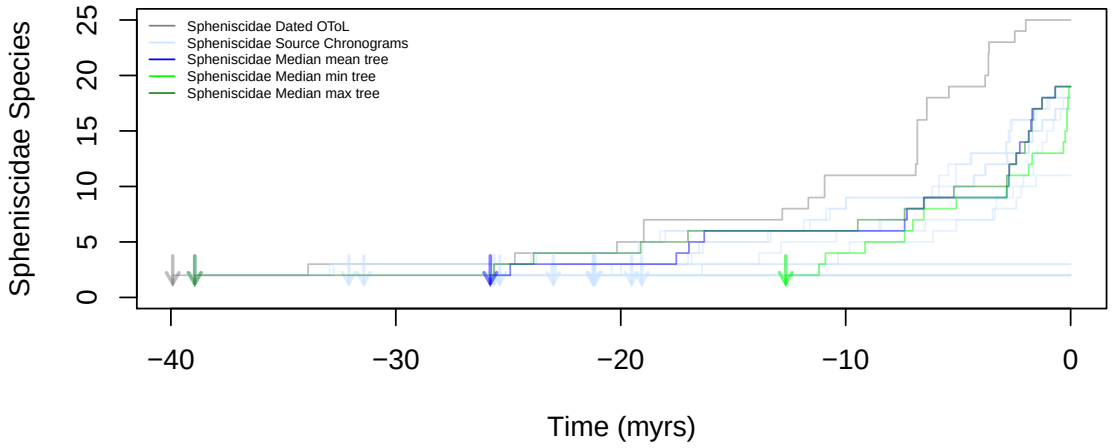

Figure 5: Spheniscidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with `datelife` algorithm.

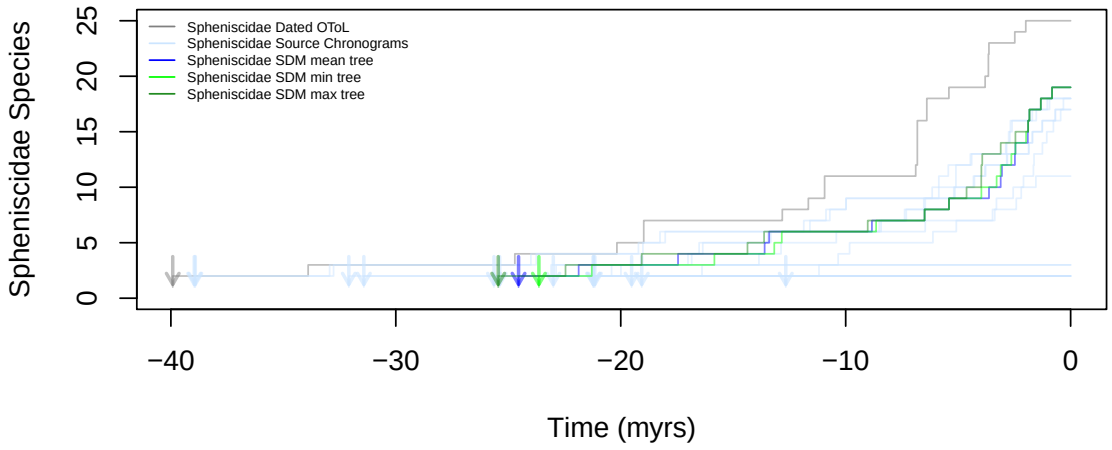

Figure 6: Spheniscidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with `datelife` algorithm.

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 1: 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 'fig
#> 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*