DateLife Workflows

Luna L. Sanchez Reyes

2019-04-30

Taxon Hominidae

I. Query source data

There are 7 species in the Open Tree of Life Taxonomy for the taxon Hominidae. Information on time of divergence is available for all of these species across 8 published and peer-reviewed chronograms. Original study citations as well as proportion of Hominidae species found across those source chronograms is shown in Table 1.

All source chronograms are fully ultrametric.

Table 1: Hominidae source chronogram studies information.

	Citation	***Source N***	***Taxon N***
1.	Bininda-Emonds, Olaf R. P., Marcel Cardillo, Kate E. Jones, Ross D. E. MacPhee, Robin M. D. Beck, Richard Grenyer, Samantha A. Price, Rutger A. Vos, John L. Gittleman, Andy Purvis. 2007. The delayed rise of present-day mammals. Nature 446 (7135): 507-512	3	5/7
2.	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	1	7/7
3.	Springer, Mark S., Robert W. Meredith, John Gatesy, Christopher A. Emerling, Jong Park, Daniel L. Rabosky, Tanja Stadler, Cynthia Steiner, Oliver A. Ryder, Jan E. Janečka, Colleen A. Fisher, William J. Murphy. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS ONE 7 (11): e49521.	4	7/7

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 12.075 to 21 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

II. Summarize results.

LTT plots are a nice way to visually compare several trees. But what if you want to summarize information from all source chronograms into a single summary 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 Hominidae have at least two overlapping species. The largest grove has 2 chronograms (out of 8 total source chronograms). Now that we have identified a suitable grove we can go on to summarize it by

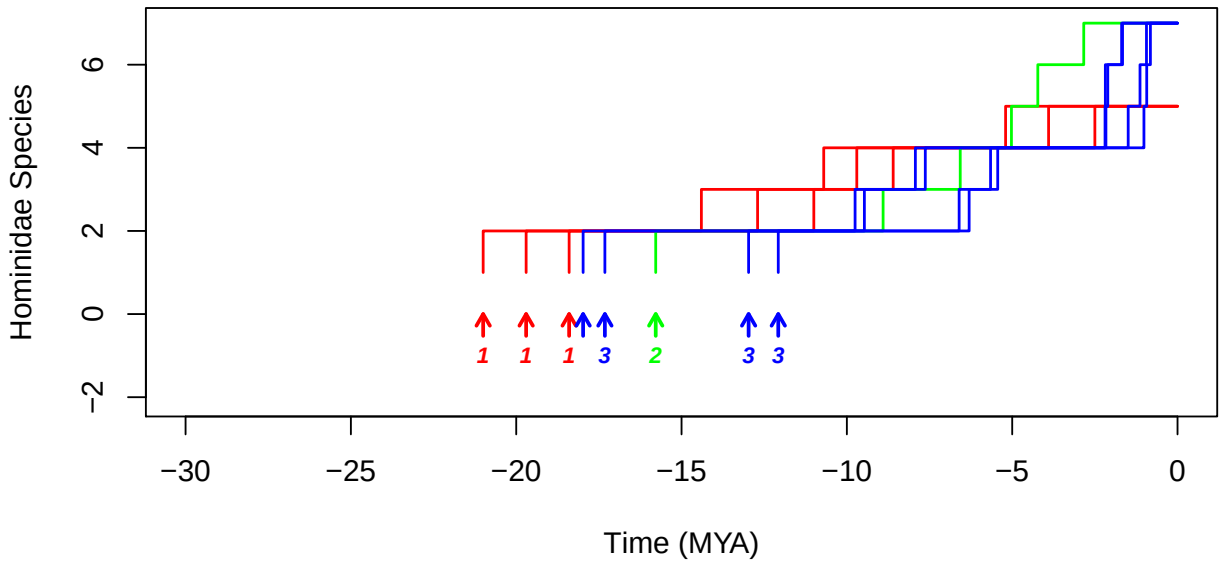

Figure 1: Lineage through time (LTT) plots of source chronograms available in data base for species in the Hominidae. Numbers correspond to original studies in Table 1. Arrows indicate maximum age of each chronogram.

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 methods 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.

Because our summary matrix is basically a distance matrix, a distance-based clustering algorithm could 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, clustering algorithms that have been developed to deal with missing values do not work well, as shown in the following section. This is probbaly because these methods are usually applied to distance matrices that represent evolutionary distance in terms of sunstitution rate nad not absolute time, as is the case in here.

II.A. Detecting clustering issues.

We tested several clustering algorithms on summary distance matrices coming from median and SDM. UPGMA returns ultrametric trees that are considerably older than source chronograms. Even scaling the distance matrix down by a factor of 0.5 would not produce trees with ages that are coherent with the source chronograms. NJ returned trees with reasonable ages, but trees are way non ultrametric, as you can see in Fig. S1 and Fig. 2.

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

Comparison of summary chronograms reconstructed with min and max ages.

```
#> Error in figcap_lttplot_summ[[i]] <- paste(taxon, "lineage through time (LTT) plots from source chron
#> Error in paste0("\n![", figcap_lttplot_summ[[2]], "](plots/", taxon, "_LTTplot_summtrees_Median.pdf)\
#> Error in cat(lttplot_median): object 'lttplot_median' not found
```

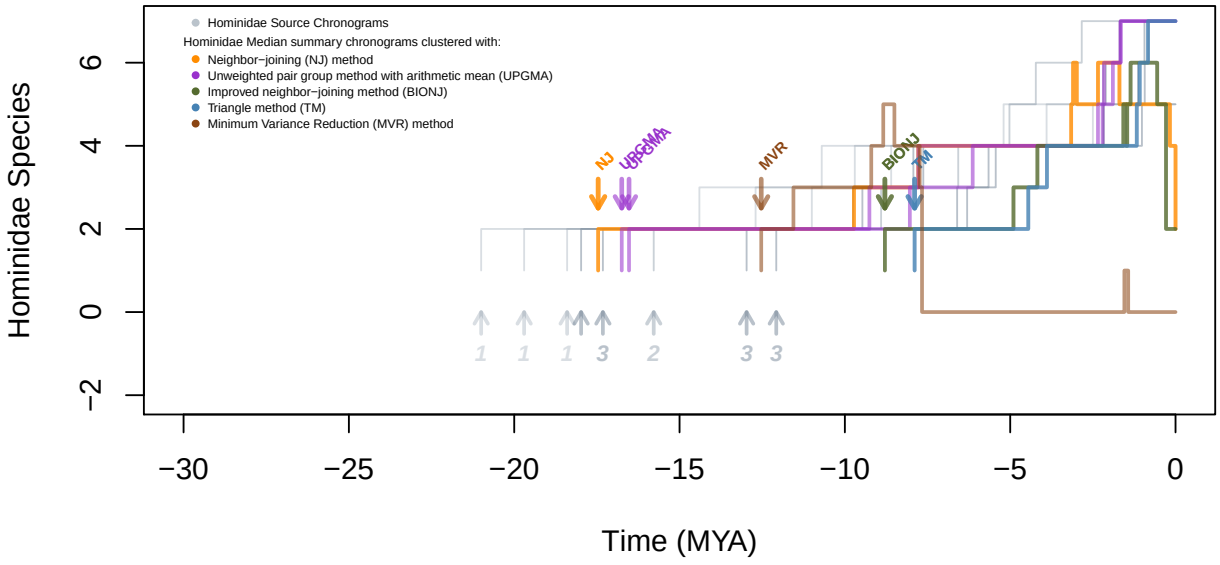

Figure 2: Lineage Through Time plots of Hominidae summary chronograms from median (upper) and SDM (lower) summary matrices obtained with various clustering algorithms.

```
#> Error in paste0("\n![", figcap_lttplot_summ[[1]], "](plots/", taxon, "_LTTplot_summtrees_SDM.pdf)\n")
#> Error in cat(lttplot_sdm): object 'lttplot_sdm' not found
```

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 Hominidae and a tree with 6 tips, 83 % resolved nodes and a MRCA of 9 was constructed. 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	Chrono1	Chrono2	Chrono3	Chrono4	Chrono5	Chrono6	Chrono7	Chrono8
Calibrations1	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations5	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations8	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

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:

Appendix

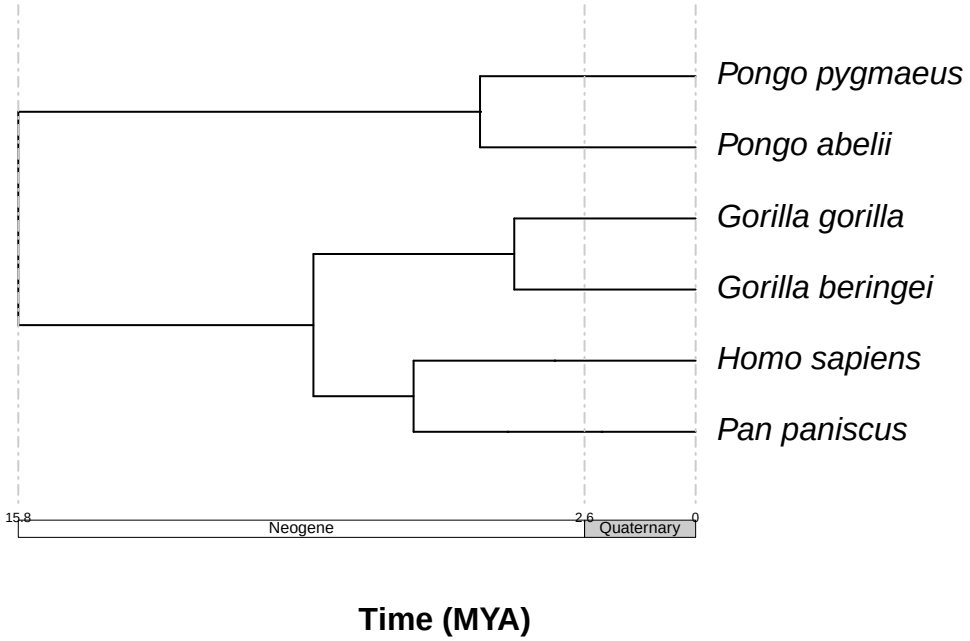

Figure 3: Hominidae Species Dated Open Tree of Life Induced Subtree. This chronogram was obtained with `get_dated_otol_induced_subtree()` function.

This taxon's SDM matrix has NO negative values.This taxon's Median matrix has NO negative values.

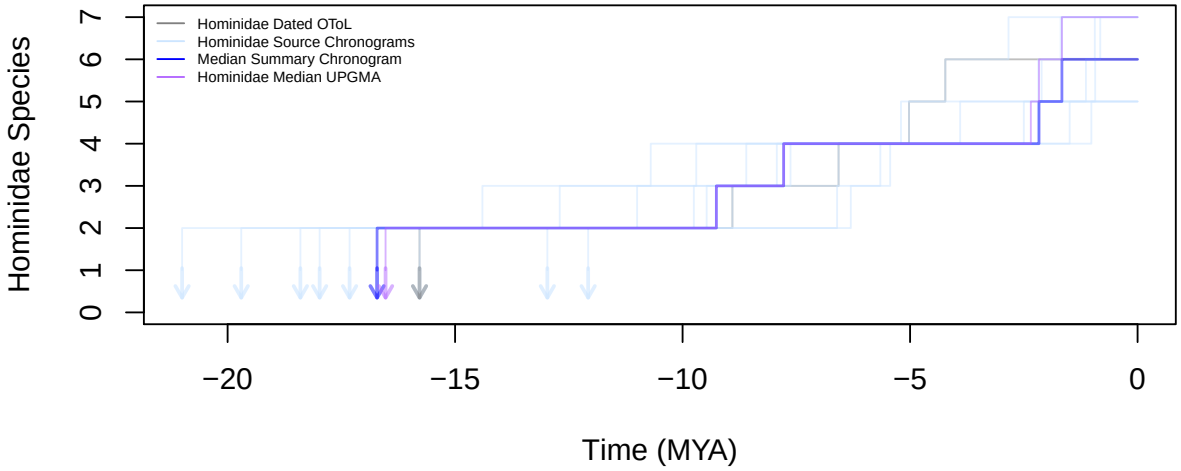

Figure 4: Hominidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developped an alternative algorithm in `datelife` to go from a summary matrix to a fully ultrametric tree.

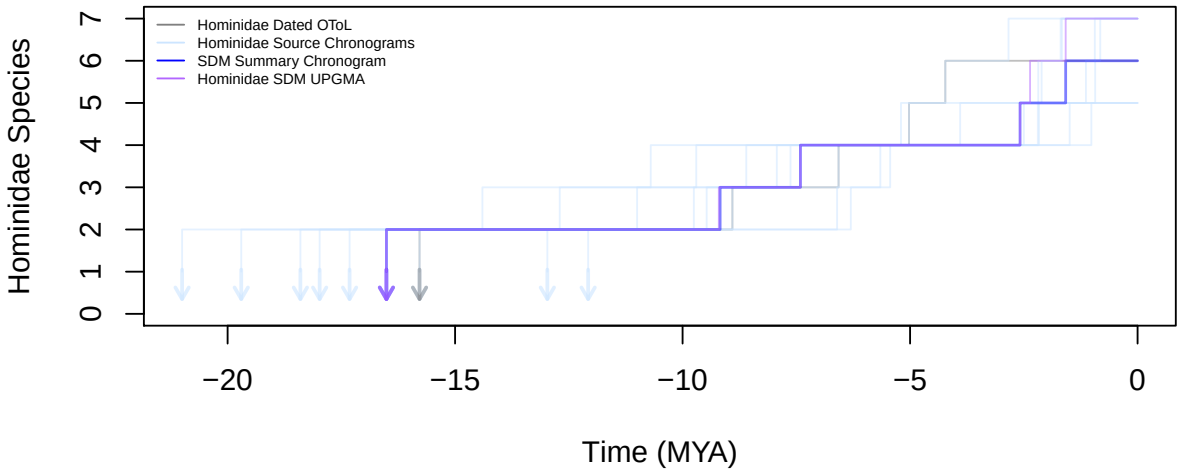

Figure 5: Hominidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developped an alternative algorithm in `datelife` to go from a summary matrix to a fully ultrametric tree.