# DateLife Workflows

Luna L. Sanchez Reyes

2019-04-16

## Taxon Hominidae

### I. Query 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.

```
#> Error in gsub("\\\\", "\\\\textbackslash", x): object 'Col1' not found
```

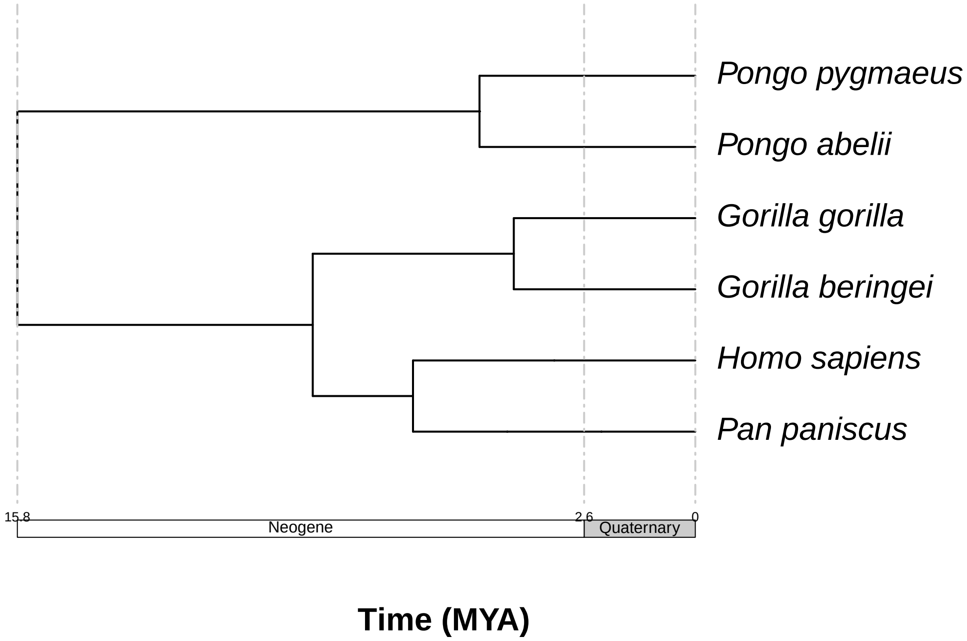

Figure 1: Hominidae Species Dated Open Tree of Life Induced Subtree. This chronogram was obtained with `get_dated_otol_induced_subtree()` function.

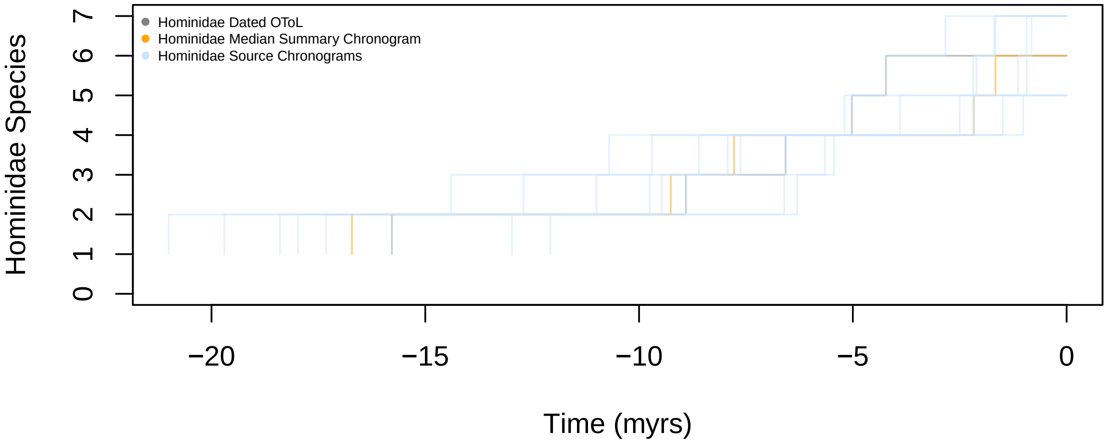

Figure 2: Hominidae lineage through time (LTT) plots from source chronograms, summary median chronogram and dated Open Tree of Life chronogram.

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

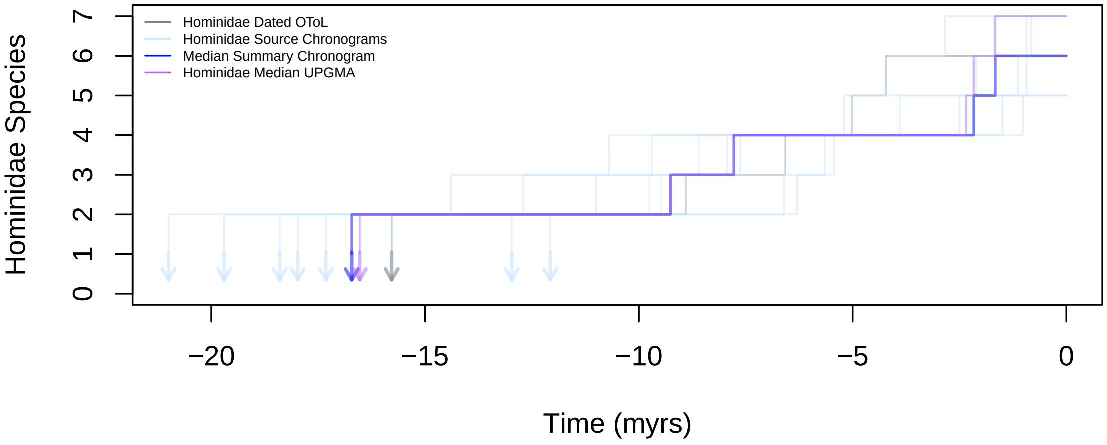

Figure 3: Hominidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developped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

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

Comparison of summary chronograms reconstructed with min and max ages.

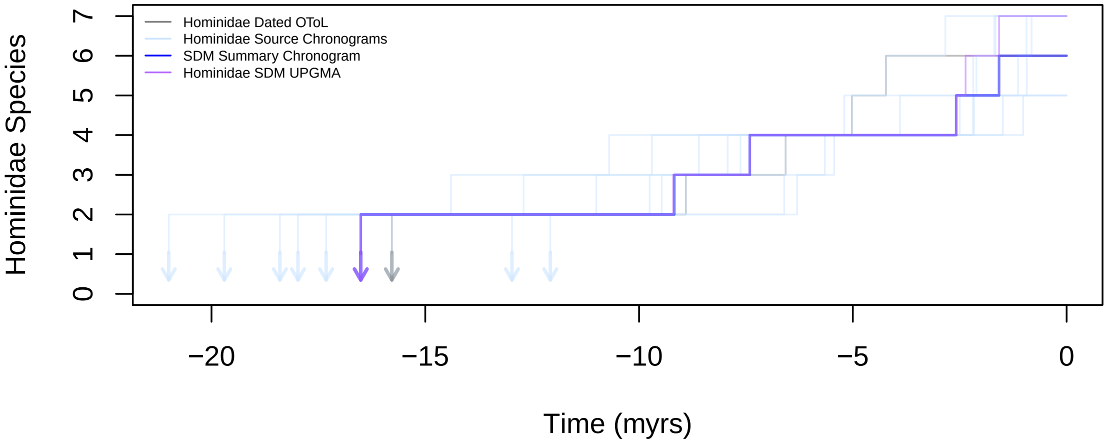

Figure 4: Hominidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developped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

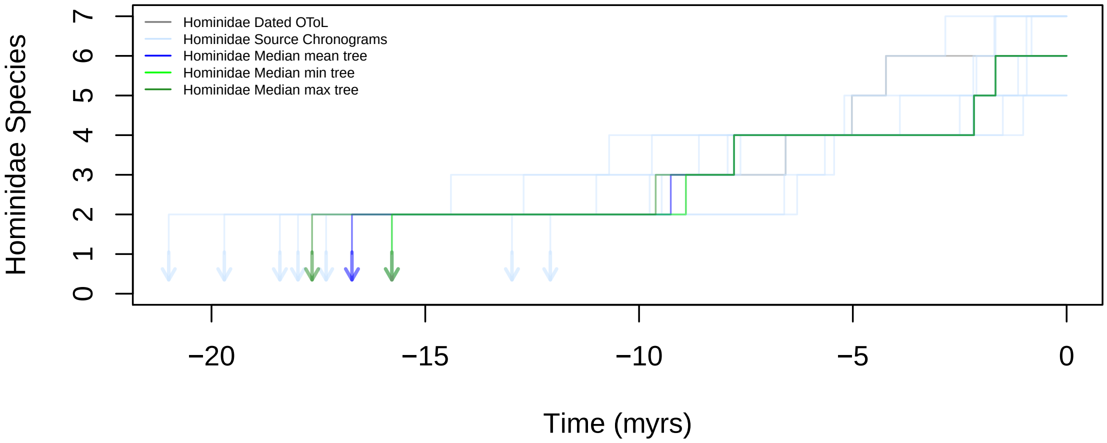

Figure 5: Hominidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with `datelife` algorithm.

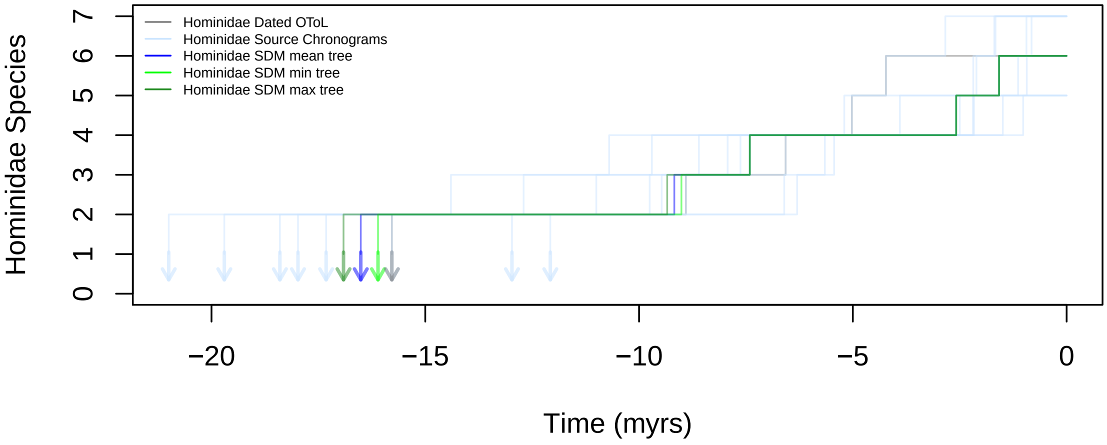

Figure 6: Hominidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with `datelife` algorithm.

## 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 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 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 | 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:

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

## Appendix