DateLife Workflows

Luna L. Sanchez Reyes

2019-05-09

Taxon Hominidae

1. Query source chronograms

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 number of Hominidae species found across those source chronograms is shown in Table 1.

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.

All source chronograms are fully ultrametric and their maximum ages 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

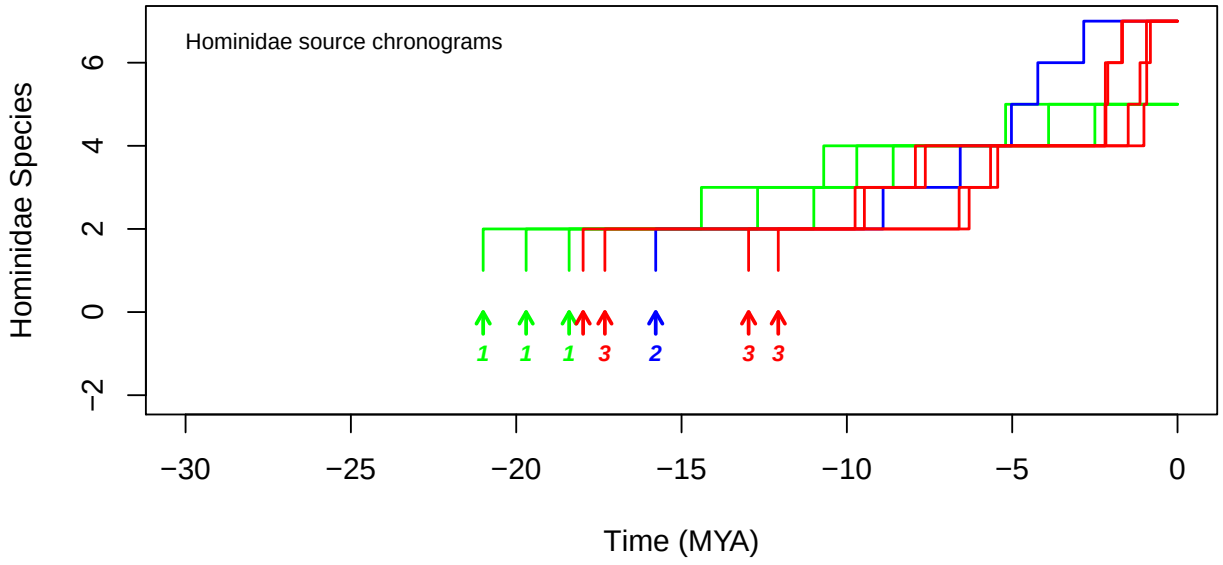

Figure 1: Lineage through time (LTT) plots of source chronograms available in data base for species in the Hominidae. Numbers correspond to original studies in Table 1. Arrows indicate maximum age of each chronogram.

2. Summarize results from query

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 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 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. As a result of such transformation, an SDM summary matrix can contain negative values. But, the SDM summary matrix of this taxon has no negative values.

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 very 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, variants of traditional clustering algorithms have been developed to deal with missing values. However, even these methods do not work well with our summary matrices, as shown in the following section. We should note that these clustering methods are usually applied to distance matrices representing substitution rates and not absolute time.

2.1. Clustering a summary matrix

NJ, UPGMA, BIONJ, minimum variance reduction (MVR) and the triangle method (TM) algorithms were used to cluster median and SDM summary distance matrices. All clustering algorithms returned very similar trees with both types of summary matrices (Fig. 2, Appendix Fig. 5). UPGMA is the only algorithm that returns ultrametric trees, but they are considerably older than expected from source chronograms. The other methods returned trees with reasonable ages, but that are not ultrametric.

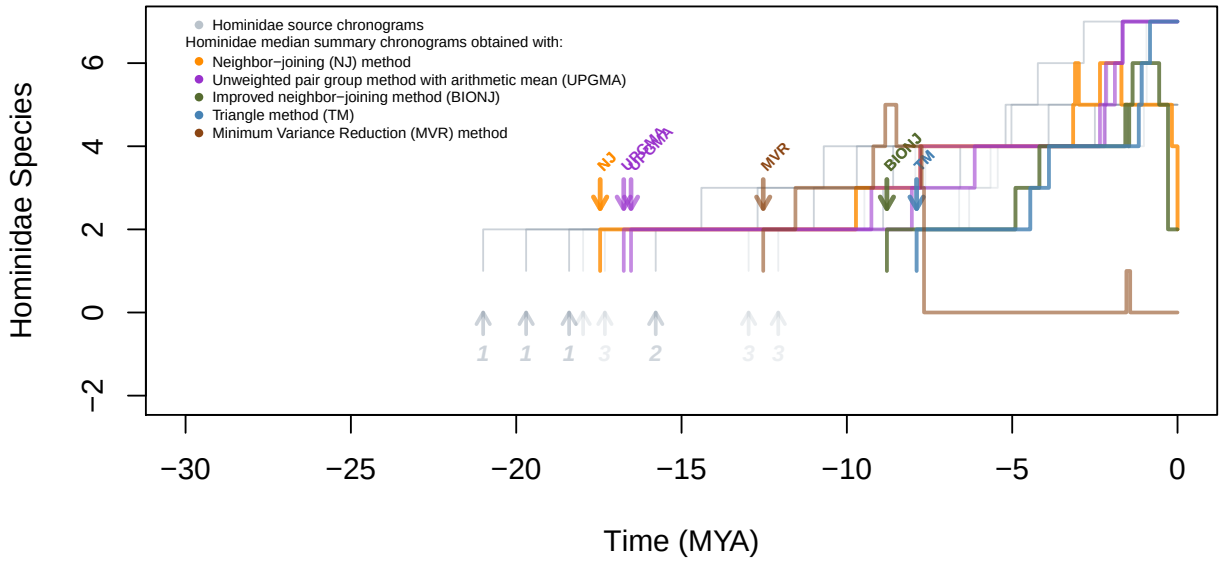

Figure 2: Lineage Through Time plots of Hominidae median summary chronograms obtained with different clustering algorithms. Not all algorithms worked with this summary matrix and we are only showing here the ones that worked. Chronograms obtained from the SDM summary matrix are very similar to the ones from the median summary matrix with all clustering algorithms (Appendix Fig. 5).

An alternative to clustering algorithms is to use all data avilable in the summary matrix as calibrations over a consensus tree. The advantage of this is that we can get a distribution of ages for the nodes and that we can essentially use this summary matrix to date any topology containing at least some of the nodes, as shown in the `Create new data` section.

2.2. Calibrating a consensus tree with data from a summary matrix

Even if the branch lengths coming from the clustered chronograms are not adequate, the topology can still be used as a consensus tree of the taxa with time data available. Then, a list of divergence times available for each node can be constructed from the summary matrix, simply by matching it to the node that corresponds to each pair of taxa in any given tree. Finally, the list and consensus tree can be fed to any dating software that does not require data. The branch length aduster (BLADJ) algorithm [@Webb2000] is really fast and does not make any evolutionary assumptions on age distribution. Other software such as MrBayes or r8s can be used without data instead of BLADJ. In here, we show summary chronograms obtained with BLADJ, using minimum, mean and maximum distances (from node age summary matrices) as fixed ages on the consensus tree (Fig. 3). Chronograms from both types of summary matrices are quite similar. As expected, SDM chronograms using minimum, mean and maximum distances do not vary much in their maximum age, because ages are transformed to minimize variance across them. In contrast, median chronogram obtained with minimum, mean and maximum distances have wider variation in their maximum ages, as can be observed from the separation between green arrows in Fig. 3.

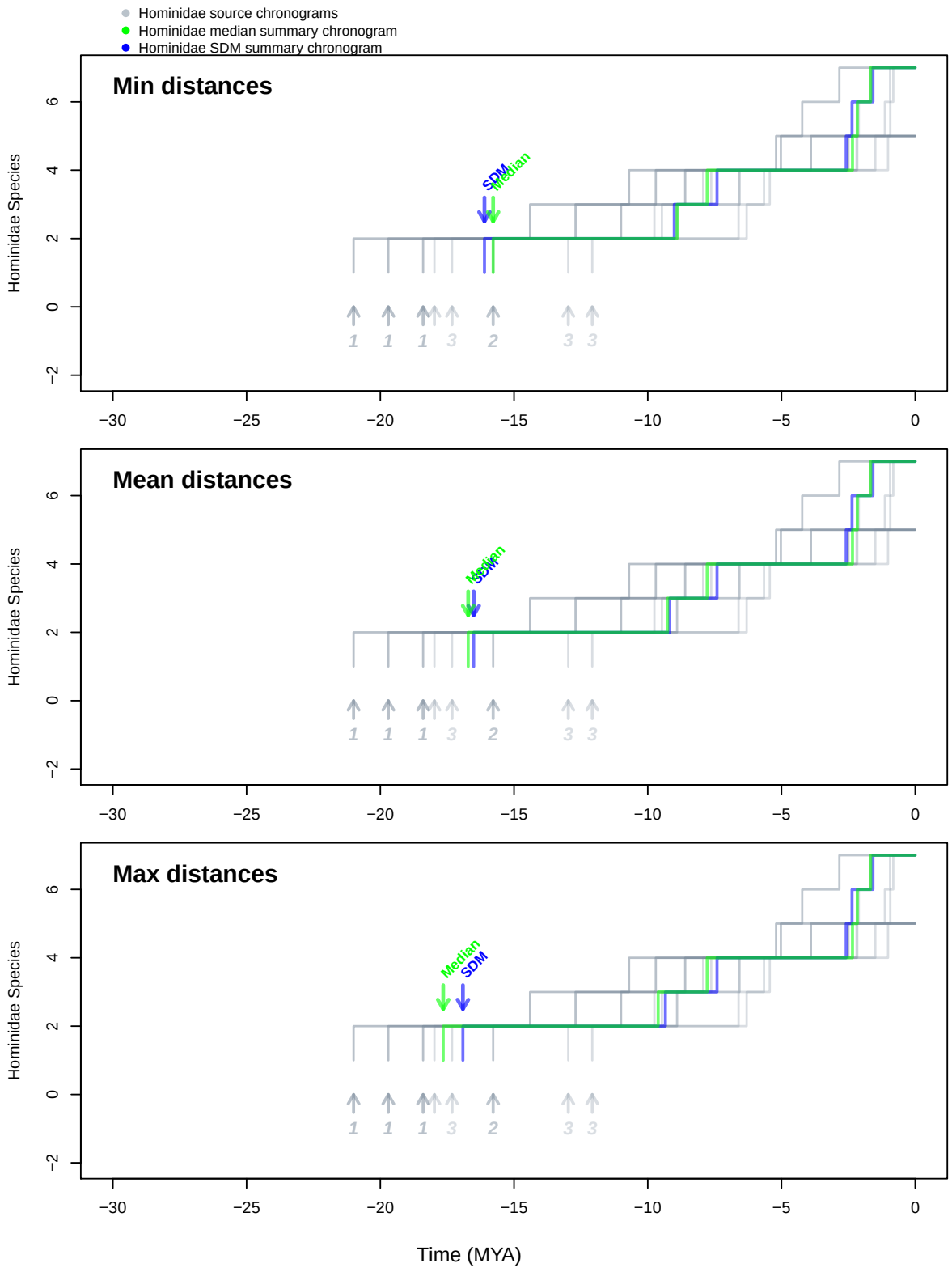

Figure 3: Hominidae lineage through time (LTT) plots from source chronograms (gray), median (green) and SDM (blue) summary chronograms obtained by calibrating a consensus tree tropology with distance data from respective summary matrices and then adjusting branch lengths with BLADJ.

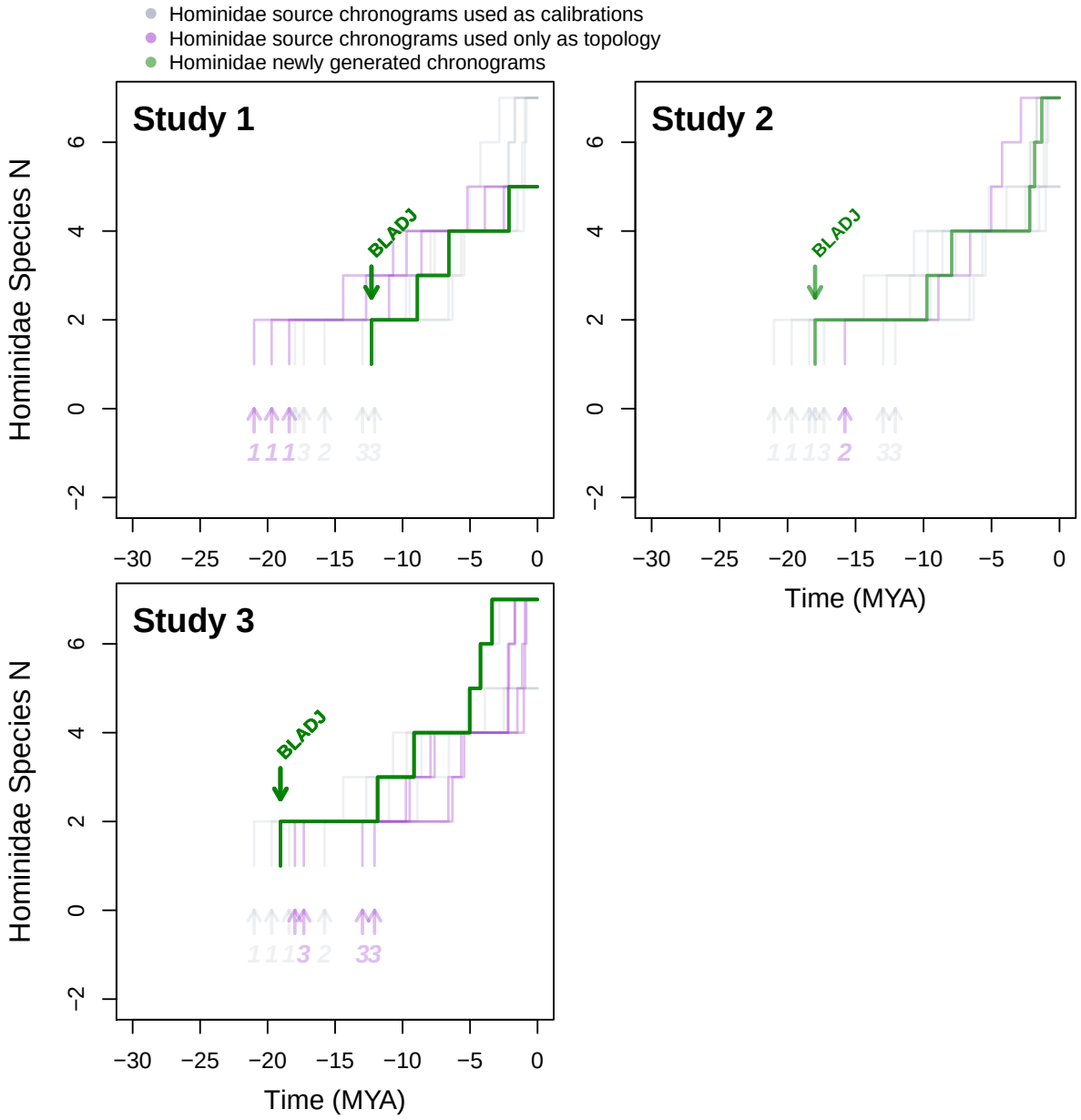

Figure 4: Hominidae lineage through time (LTT) plots from source chronograms used as secondary calibrations (gray), source chronograms used as topology (purple) and chronograms resulting from calibrating the latter with the former using BLADJ (green).

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

3. Generate new chronograms.

Another way to leverage information from the source chronograms is to use the node ages as secondary calibration points to date any tree topology (with or without branch lengths) given that at least two taxa from source chronograms are in the tips of that topology. In this data set, we have 42 calibrations in total (that basically correspond to the sum of the number of nodes in each source chronogram). Once we have a target tree topology, we can map the calibrations to the target tree. Some nodes will have several calibrations and some others might have none. To deal with this, we can expand the calibrations to make them agree, or we can summarize them. To exemplify each method we performed a series of cross validation analyses by using the information from all other source chronograms to date the topology of source chronograms from each study

3.1. Calibrate a tree without branch lengths

3.2. Calibrate a tree with data (from BOLD).

4.1. Expanding calibrations

show cross validation of LTTs from chronograms obtained by dating the topology of each study with data from any other study.

4.2. Summarizing calibrations

4.3. Example with subspecies tree

As an example, we're gonna date the subspecies tree of the group (coming from otol).

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 10 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.

4. Simulate data/ Add missing taxa

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:

References

Appendix

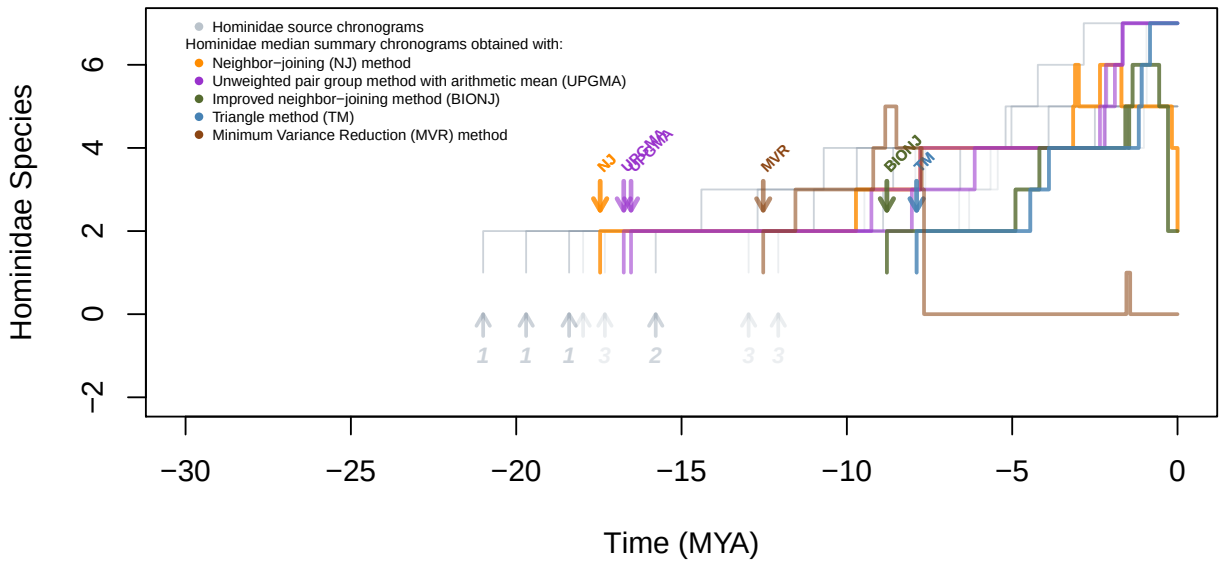

Figure 5: Lineage Through Time plots of Hominidae SDM summary chronograms obtained with different clustering algorithms. Not all algorithms worked with the SDM summary matrix and we are only showing here the ones that worked. Chronograms obtained from the median summary matrix are very similar to the ones shown here with all algorithms (mainFig. 2).

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

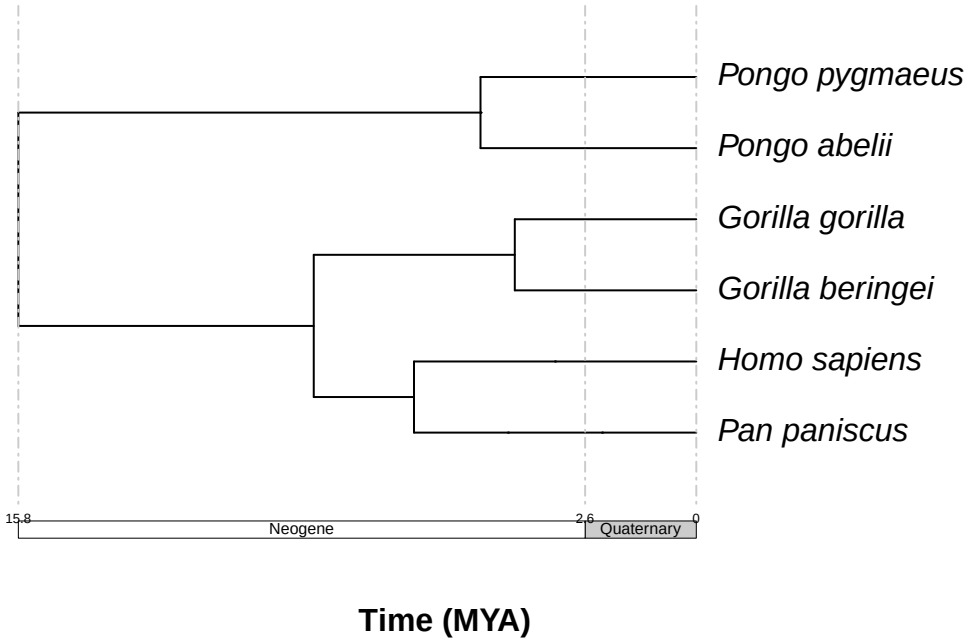

Figure 6: Hominidae Species Dated Open Tree of Life Induced Subtree. This chronogram was obtained with `get_dated_otol_induced_subtree()` function.

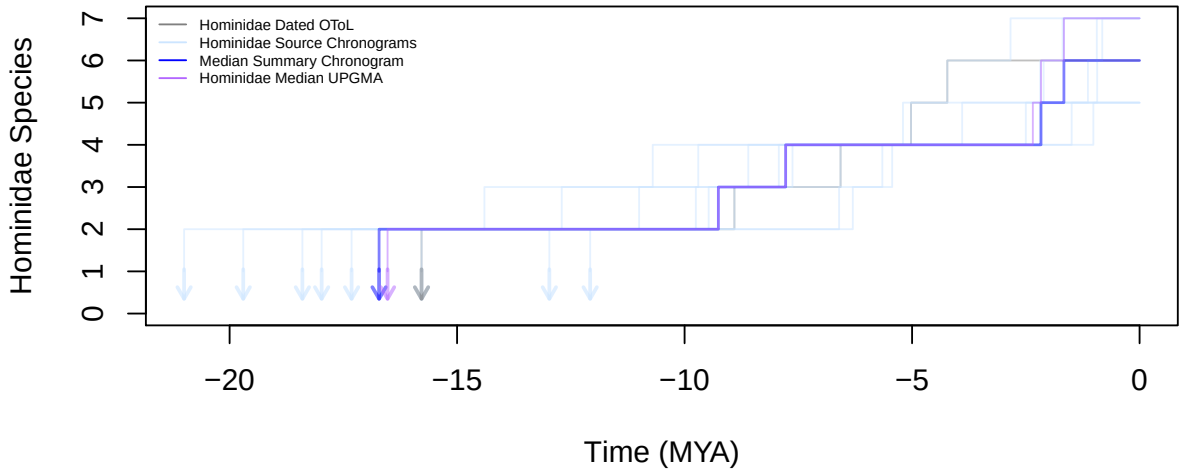

Figure 7: Hominidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developped an alternative algorithm in `datelife` to go from a summary matrix to a fully ultrametric tree.

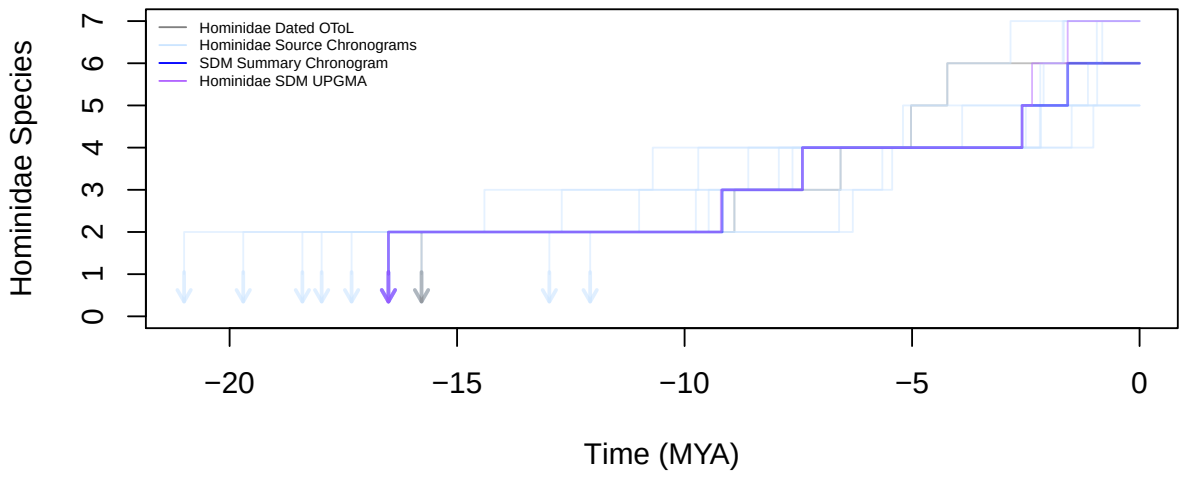

Figure 8: Hominidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developped an alternative algorithm in `datelife` to go from a summary matrix to a fully ultrametric tree.

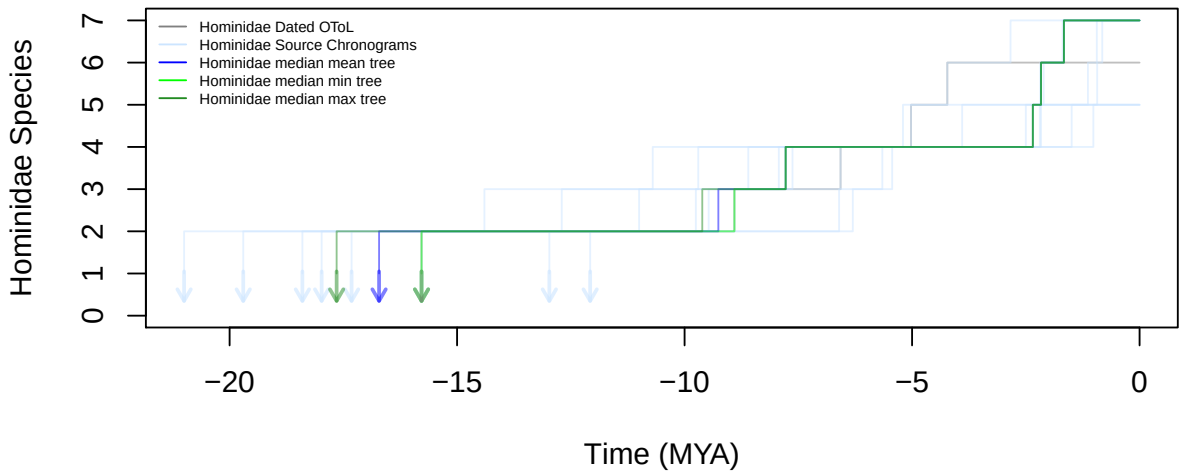

Figure 9: Hominidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with `datelife` algorithm.

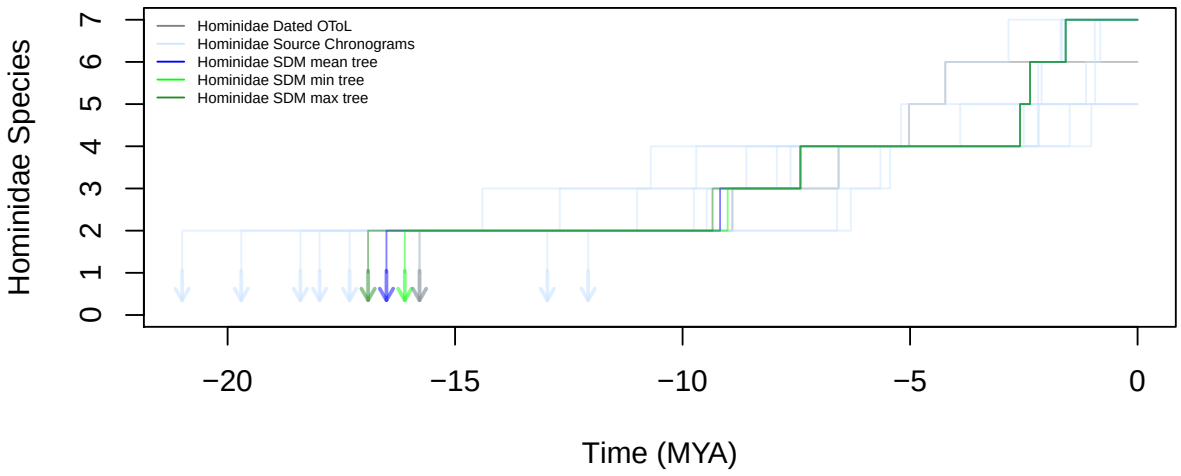

Figure 10: Hominidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with `datelife` algorithm.