DateLife Workflows

Luna L. Sanchez Reyes

2019-05-01

Taxon Primates

I. Query source data

There are 526 species in the Open Tree of Life Taxonomy for the taxon Primates. Information on time of divergence is available for 355 of these species across 8 published and peer-reviewed chronograms. Original study citations as well as number of Primates species found across those source chronograms is shown in Table 1.

Table 1: Primates 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	215/526
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	294/526
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	330/526

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 62.766 to 90.4 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
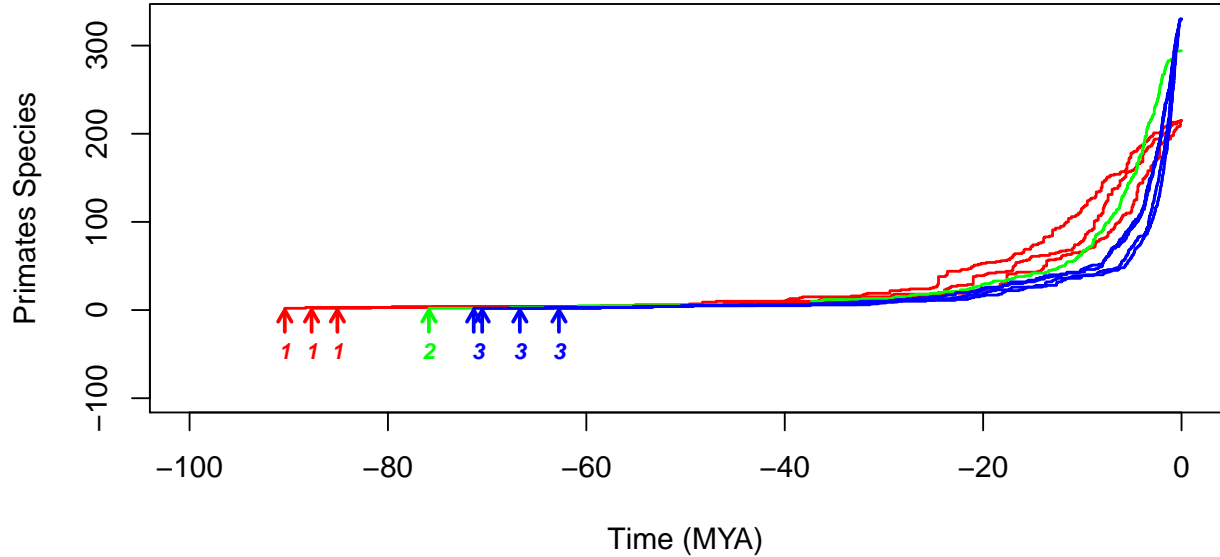
Figure 1: Lineage through time (LTT) plots of source chronograms available in data base for species in the Primates. Numbers correspond to original studies in Table 1. Arrows indicate maximum age of each chronogram.

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 Primates 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

II. A. 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.
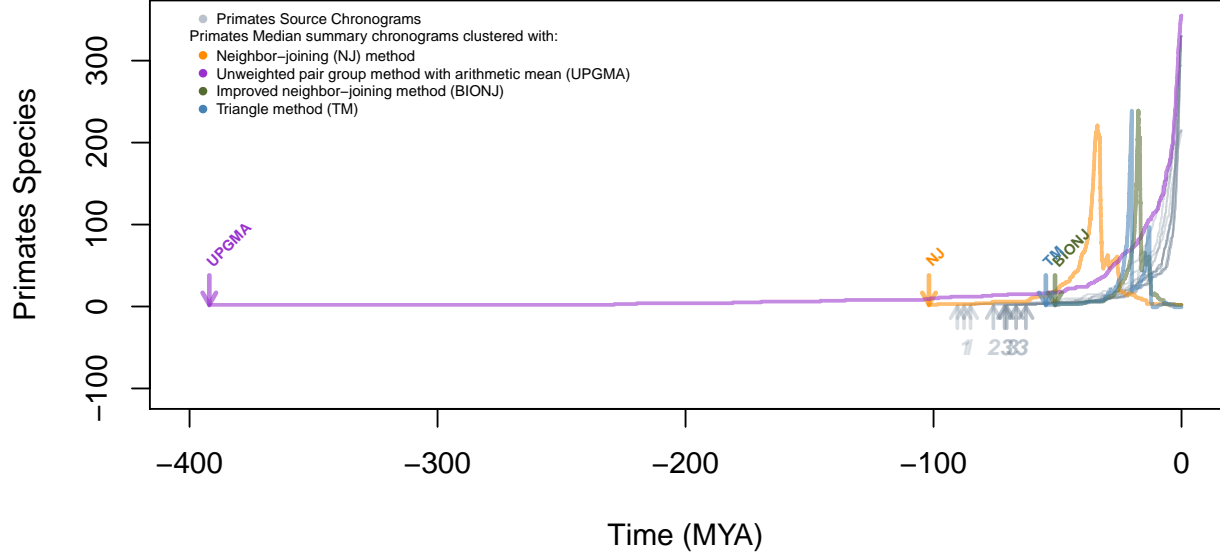

Figure 2: Lineage Through Time plots of Primates median summary chronograms obtained with different clustering algorithms. Not all algorithms worked with this summary matrix and we are only showing here the ones that worked. Chronograms obtained from the SDM summary matrix are very similar to the ones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

II. B. Dating a consensus tree

The trees obtained from a clustering algorithm can be used as consensus tree. A list of ages available for each node is constructed from the matrix. The list and consensus tree can be fed to any dating software that does not require data. We choose bladj because it does not make any evolutionary assumptions regarding the distribution of ages. MrBayes, PATHd8 or other software can be used instead. Chronograms from both summary matrices are very similar.

Use the tree from upgma without branch lengths as target tree

Graph shows source chronograms in gray with arrown top-down, both chronograms obtained with SDM and median in three vertical panes: only from using the min, mean and max of all ages no datedotol tree one legend on the top outside text for MIN, MEAN and MAX on the inner top left of each pane
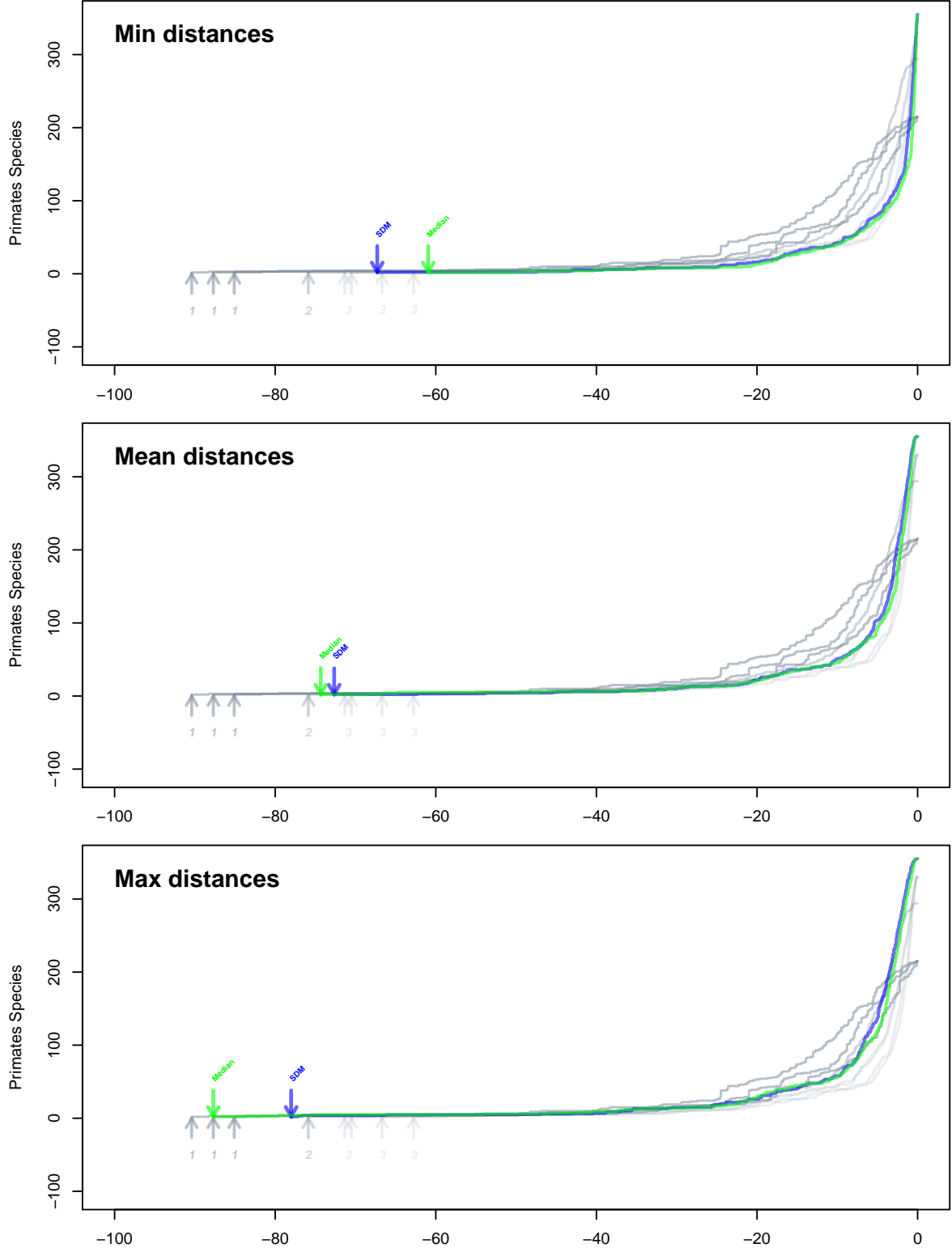
Figure 3: Primates lineage through time (LTT) plots from source chronograms and SDM and median summary matrix converted to phylo with `datelife` algorithm.

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 Primates 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	Chrono1	Chrono2	Chrono3	Chrono4	Chrono5	Chrono6	Chrono7	Chrono8
Calibrations1	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations2	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations3	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations4	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations5	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations6	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations7	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Calibrations8	TRUE	FALSE	FALSE	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

The following species were completely absent from the chronogram data base: *Alouatta arctoidea, Alouatta discolor, Alouatta stramineus, Alouatta ululata, Aotus azarae, Aotus jorgehernandezi, Aotus zonalis, Avahi mooreorum, Avahi ramanantsoavani, Cacajao rubicundus, Callicebus aureipalatii, Callicebus baptista, Callicebus barbarabrownae, Callicebus caquetenesis, Callicebus caquetensis, Callicebus discolor, Callicebus lucifer, Callicebus medemi, Callicebus melanochir, Callicebus miltoni, Callicebus ornatus, Callicebus pallescens, Callicebus regulus, Callicebus stephennashi, Callicebus toppinii, Callicebus urubambensis, Callicebus vieirai, Callithrix cf. emiliae, Callithrix chrysoleuca, Carlito syrichta, Cebus aequatorialis, Cebus brunneus, Cebus cesarae, Cebus cuscinus, Cebus imitator, Cebus leucocephalus, Cebus malitiosus, Cebus polykomos, Cebus unicolor, Cebus versicolor, Cebus yuracus, Cephalopachus bancanus, Cercocebus lunulatus, Cercocebus sanjei, Cercopithecus denti, Cercopithecus doggetti, Cercopithecus kandti, Cercopithecus lomamiensis, Cercopithecus lowei, Cheirogaleus andysabini, Cheirogaleus lavasoensis, Cheirogaleus minusculus, Cheirogaleus thomasi, Cheracebus lugens, Cheracebus purinus, Cheracebus torquatus, Chiropotes utahickae, Chlorocebus djamdjamensis, Daubentonia robusta, Euoticus matschiei, Galagoides cocos, Galagoides orinus, Galagoides rondoensis, Galagoides thomasi, Hylobates entelloides, Hylobates funereus, Lemur indri, Lemur tardigradus, Lemur volans, Lepilemur grewcockorum, Lepilemur hollandorum, Lepilemur jamesorum, Lepilemur mitsinjoensis, Lepilemur scottorum, Lepilemur tymerlachsoni, Lophocebus johnstoni, Lophocebus opdenboschi, Lophocebus osmani, Lophocebus ugandae, Macaca balantak, Macaca leucogenys, Macaca speciosa, Mico acariensis, Mico intermedius, Mico leucippe, Mico marcai, Mico nigriceps, Microcebus lokobensis, Microcebus marohita, Microcebus myonixus, Microcebus tanosi, Nomascus annamensis, Nycticebus bancanus, Nycticebus borneanus, Nycticebus kayan, Papio japonicus, Papio kindae, Phaner electromontis, Phaner parienti, Piliocolobus bouvieri, Piliocolobus epieni, Piliocolobus oustaleti, Piliocolobus parmentieri, Piliocolobus semlikiensis, Piliocolobus temminckii, Piliocolobus waldronae, Pithecia cazuzai, Pithecia chrysocephala, Pithecia hirsuta, Pithecia inusta, Pithecia isabela, Pithecia milleri, Pithecia mittermeieri, Pithecia napensis, Pithecia pissinattii, Pithecia rylandsi, Pithecia vanzolinii, Plecturocebus bernhardi, Plecturocebus brunneus, Plecturocebus caligatus, Plecturocebus cinerascens, Plecturocebus cupreus, Plecturocebus donacophilus, Plecturocebus hoffmannsi, Plecturocebus miltoni, Plecturocebus moloch, Presbytis bicolor, Presbytis canicrus, Presbytis mitrata, Presbytis natunae, Presbytis sabana, Presbytis senex, Presbytis siamensis, Presbytis siberu, Presbytis sumatrana, Propithecus candidus, Pseudopotto martini, Pygathrix cinerea 1 RL-2012, Pygathrix cinerea 2 RL-2012, Rhinopithecus bieti 1 RL-2012, Rhinopithecus bieti 2 RL-2012, Saguinus cruzlimai, Saguinus illigeri, Saguinus lagonotus, Saguinus leucogenys, Saguinus nigrifrons, Saguinus pileatus, Saguinus ursulus, Saguinus weddelli, Saimiri cassiquiarensis, Saimiri macrodon, Sapajus apella, Sapajus cay, Sapajus flavius, Sapajus libidinosus, Sapajus nigritus, Sapajus xanthosternos, Sciurocheirus cameronensis, Sciurocheirus makandensis, Semnopithecus ajax, Semnopithecus hypoleucos, Semnopithecus schistaceus, Tarsius banacanus, Tarsius fuscus, Tarsius pelengensis, Tarsius tarsius, Tarsius tumpara, Trachypithecus ebenus, Trachypithecus mauritius, Trachypithecus selangorensis, Trachypithecus shortridgei*

This taxon's SDM matrix has NO negative values.This taxon's Median matrix has NO negative values.
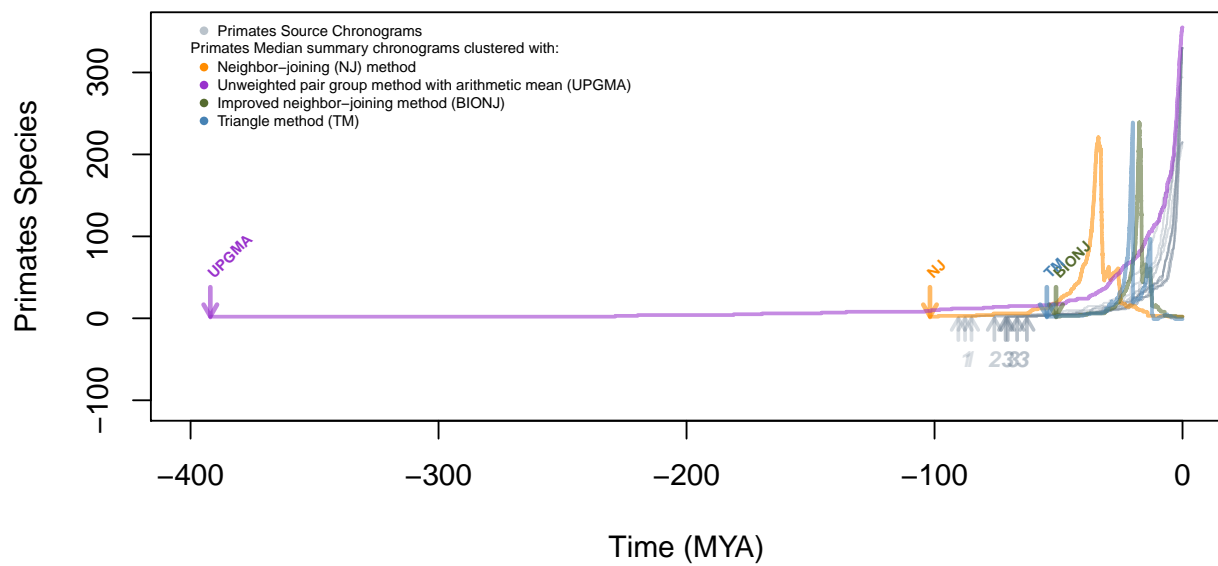
Figure 4: Lineage Through Time plots of Primates SDM summary chronograms obtained with different clustering algorithms. Not all algorithms worked with the SDM summary matrix and we are only showing here the ones that worked. Chronograms obtained from the median summary matrix are very similar to the ones shown here with all algorithms (mainFig. 2).
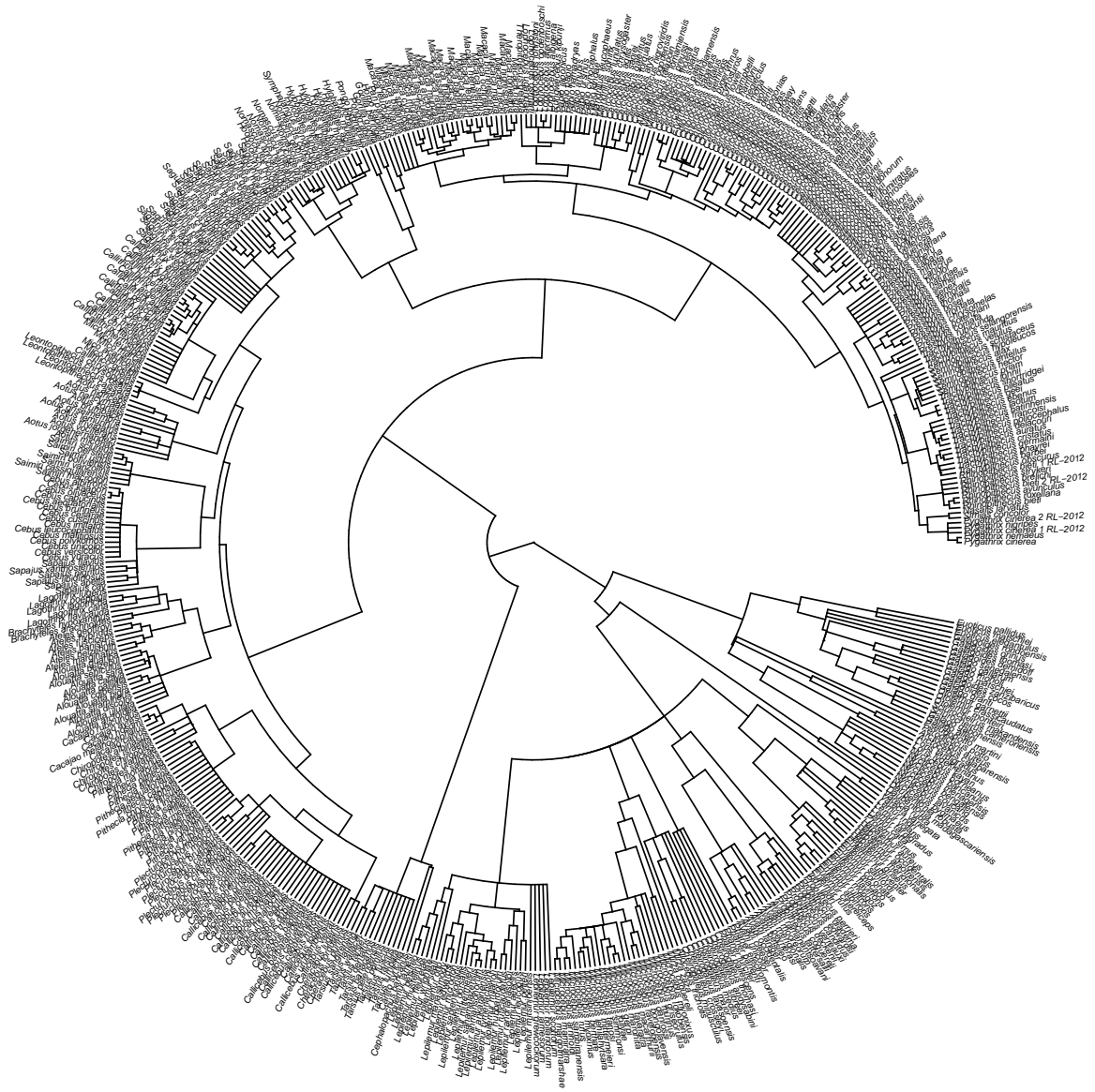
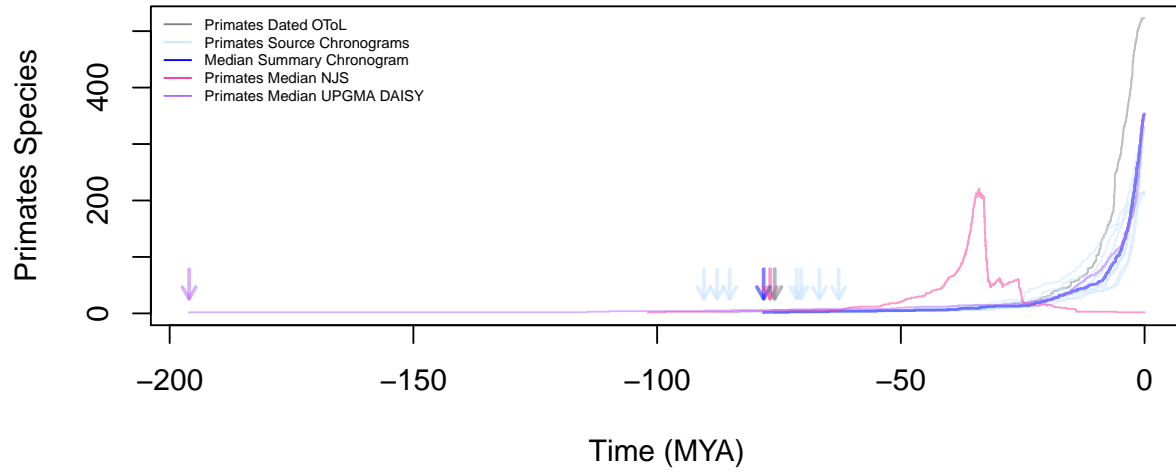Figure 5: Primates Species Dated Open Tree of Life Induced Subtree. This chronogram was obtained with `get_dated_otol_induced_subtree()` function.

Figure 6: Primates 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 developed an alternative algorithm in `datelife` to go from a summary matrix to a fully ultrametric tree.
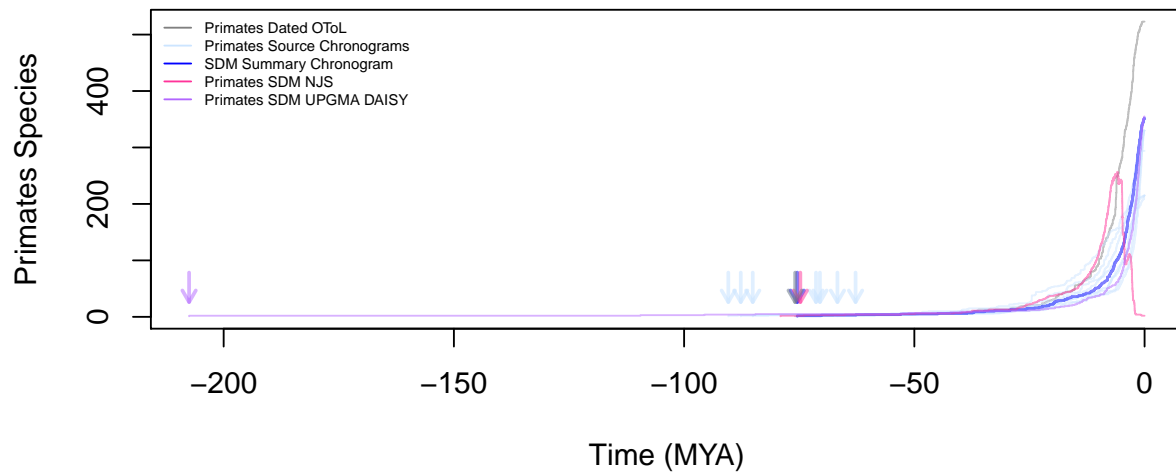

Figure 7: Primates lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developed an alternative algorithm in `datelife` to go from a summary matrix to a fully ultrametric tree.
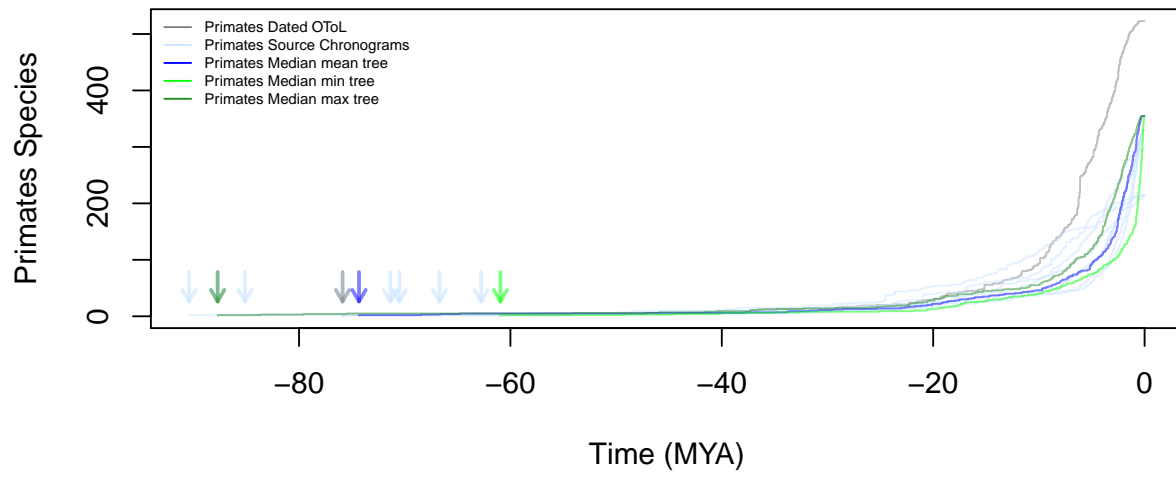
Figure 8: Primates lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with `datelife` algorithm.
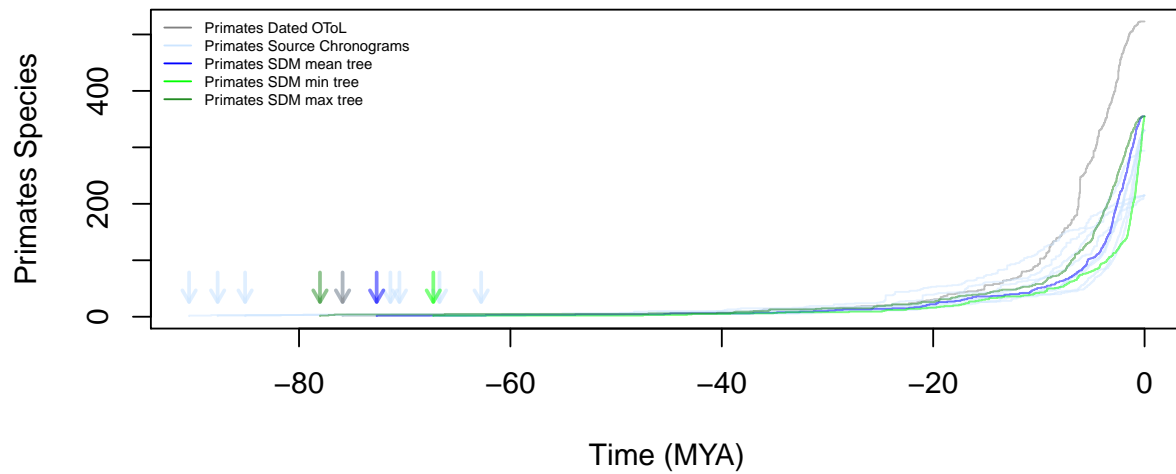

Figure 9: Primates lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with `datelife` algorithm.