DateLife Workflows

Luna L. Sanchez Reyes

2019-05-15

Taxon Cetacea

1. Query source chronograms

There are 198 species in the Open Tree of Life Taxonomy for the taxon Cetacea. Information on time of divergence is available for 89 of these species across 6 published and peer-reviewed chronograms. Original study citations as well as number of Cetacea species found across those source chronograms is shown in Table 1. All source chronograms are fully ultrametric and their maximum ages range from 33.5 to 55.5 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

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 Cetacea have at least two overlapping species. The largest grove has 2 chronograms (out of 6 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. In this case, the SDM summary matrix has some negative values in the following taxa: *Eubalaena japonica*, *Eubalaena glacialis*.

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. None of these clustering algorithms returned

trees matching source chronograms (Fig. 2, Appendix Fig. 5). UPGMA is the only algorithm that returns ultrametric trees, but they are considerably older than expected from ages observed in source chronograms. The other methods returned trees with ages that coincide with those observed in source chronograms. However, they resulting chronograms are not ultrametric. To overcome the issues presented by clustering algorithms, we used all data avilable in the summary matrix as calibrations over a consensus tree to obtain a summary chornogram.

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

Even if the branch lengths coming form the clustered chronograms are not adequate, the topology can still be used as a backbone tree that can be dated using data from the summary matrix as secondry calibrations. A summary of divergence times available for each node can be obtained from the summary matrix, simply by getting the nodes from the backbone tree that correspond to each pair of taxa in the matrix. Finally, this summary of node divergence times can be used with the consensus tree as input in 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 and r8s can be used instead of BLADJ by running them without data. In here, we show summary chronograms obtained using minimum, mean and maximum distances from the summary of node divergence times of the backbone tree as fixed ages in BLADJ (Fig. 3). Summary 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 the variance. In contrast, the median chronograms obtained with minimum, mean and maximum distances have wider variation in their maximum ages, as can be observed in the distance between the green arrows in Fig. 3. This variation simply represents variation in source data.

3. Generate new chronograms

Another way to leverage information from the source chronograms is to use their 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 425 calibrations in total (that basically corresponds to the sum of the number of nodes from 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. Also, some node ages can be conflicting, with descendant nodes being older than parent nodes. We performed a series of cross validation analyses with different dating methods, by dating the topologies of each source chronogram using information from all other source chronograms as calibration points.

3.1. Calibrate a tree without branch length data

To date a tree in the absence of data on relative evolutionary rates (molecular or morphological) we follow the same methodology as the one used to obtain summary chronograms. First, we obtained the nodes that correspond to each pair of taxa in the data set of total calibrations to construct a summary of node calibrations for the backbone tree. Then, we used mean ages as secondary calibrations for the backbone tree with the software BLADJ. In general, the time of divergence information from other source chronograms allows to recover the divergence times from the original study. In some cases, it is evident that information from a particular study really affects the summary of divergence times. In some other cases, the root of the tree is not calibrated. Since BLADJ has no underlying model of evolution, there is no way for the algorithm to calculate this age. To fix this, we simply added a unit of the mean difference across ranked ages from secondary calibrations (Fig. 4).

3.2. Calibrate a tree with data

If you have a tree with branch lengths proportional to relative substitution rates, you can use the source chronogram node ages as secondary calibrations with various algorithms for phylogenetic dating to get

branch lengths proportional to absolute time. To exemplify this, we got DNA markers from the Barcode of Life Database (BOLD) to estimate branch lengths as relative DNA substitution rates on a tree topology of our choosing. In this example we retrieved data from the cytochrome C oxidase subunit I (COI) marker, that is of widespread use in barcoding, providing DNA data for a very wide number of organisms. Unfortunately, a tree with branch lengths could not be constructed for any of the source chronograms available for the Cetacea, so this workflow will not be exemplified here. This can happen for several reasons. If the tree has only two tips, the tree search cannot be performed. If the Please look into other DateLife examples available in here for more infromation about this workflow.

3.2.1. Expanding calibrations

```
### 3.2.2. Summarizing calibrations (congruifying calibrations)
```

3.2.2. Summarizing calibrations (congruifying calibrations)

4. Example with subspecies tree

As an example, we're gonna date the subspecies tree of the group using all approaches for generating new data.

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 Cetacea 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.

Tables and Figures

Table 1: Cetacea 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	78/198
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	79/198
3.	Steeman, M., Hebsgaard M., Fordyce R., Ho S., Rabosky D., Nielsen R., Rahbek C., Glenner H., Sørensen M., & Willerslev E. 2009. Radiation of Extant Cetaceans Driven by Restructuring of the Oceans. Systematic Biology 58 (6): 573-585.	1	86/198
4.	Toljagić O., Voje K.L., Matschiner M., Liow L., & Hansen T.F. 2017. Millions of Years Behind: Slow Adaptation of Ruminants to Grasslands. Systematic Biology, .	1	32/198

Source N: Number of source chronograms reported in study.
Taxon N: Number of queried taxa found in source chronograms.

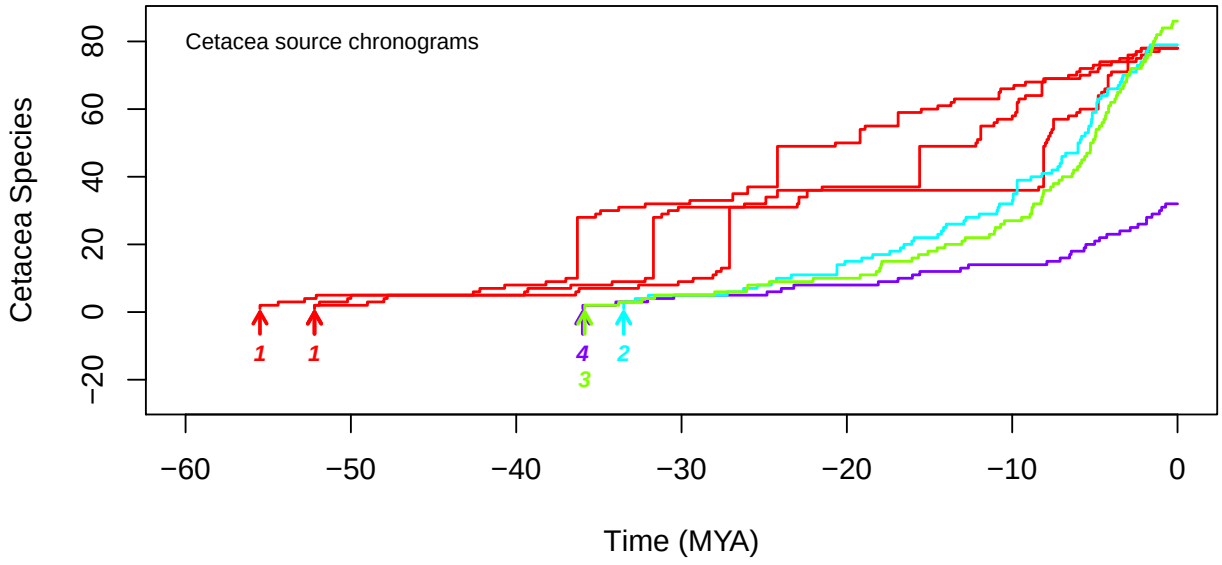


Figure 1: Lineage through time (LTT) plots of source chronograms available in data base for species in the Cetacea. Numbers correspond to original studies in Table 1. Arrows indicate maximum age of each chronogram.

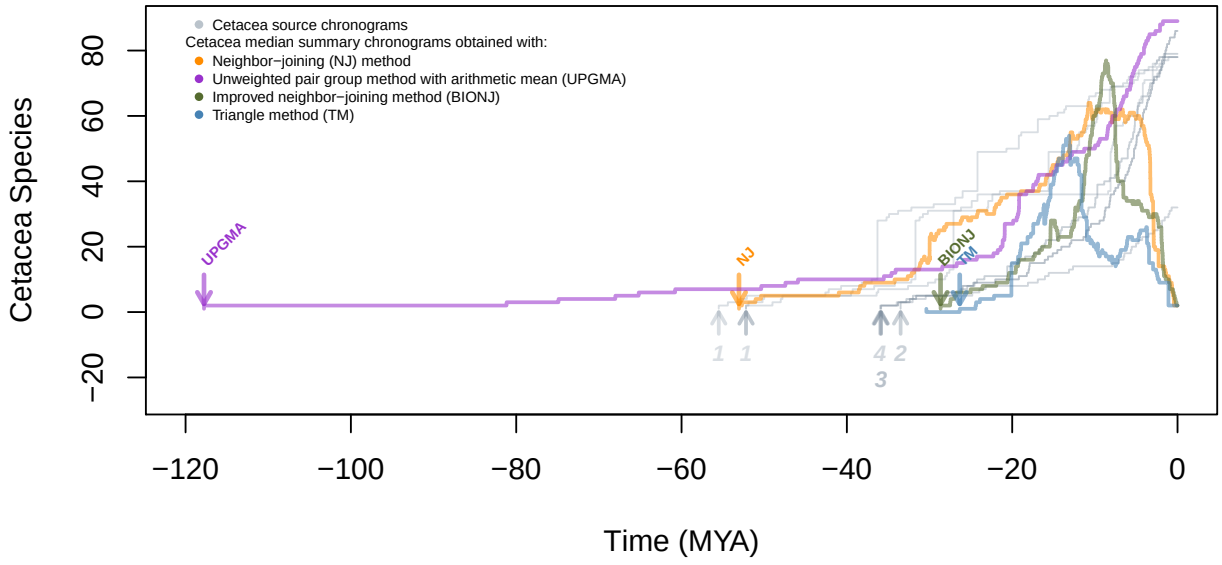
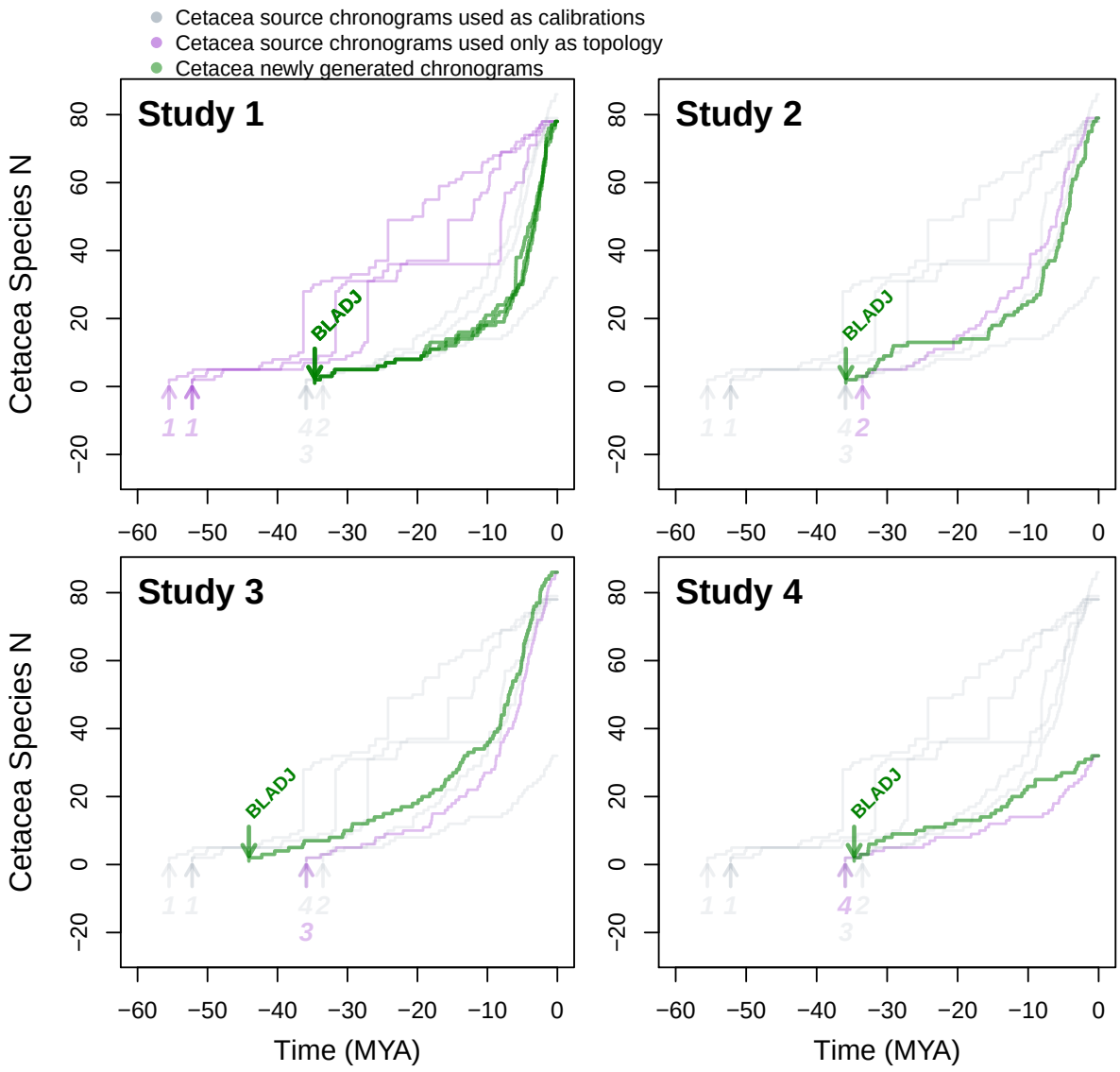


Figure 2: Lineage Through Time plots of Cetacea median summary chronograms obtained with different clustering algorithms. Not all algorithms worked with this summary matrix and we are only showing here the ones that worked. Chronograms obtained from the SDM summary matrix are very similar to the ones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


Cetacea source chronograms used as calibrations
Cetacea source chronograms used only as topology
Cetacea newly generated chronograms
Study 1
Study 2
Study 3
Study 4
BLADJ
Cetacea Species N
Time (MYA)
−60
−50
−40
−30
−20
−10
0
−20
0
20
40
60
80

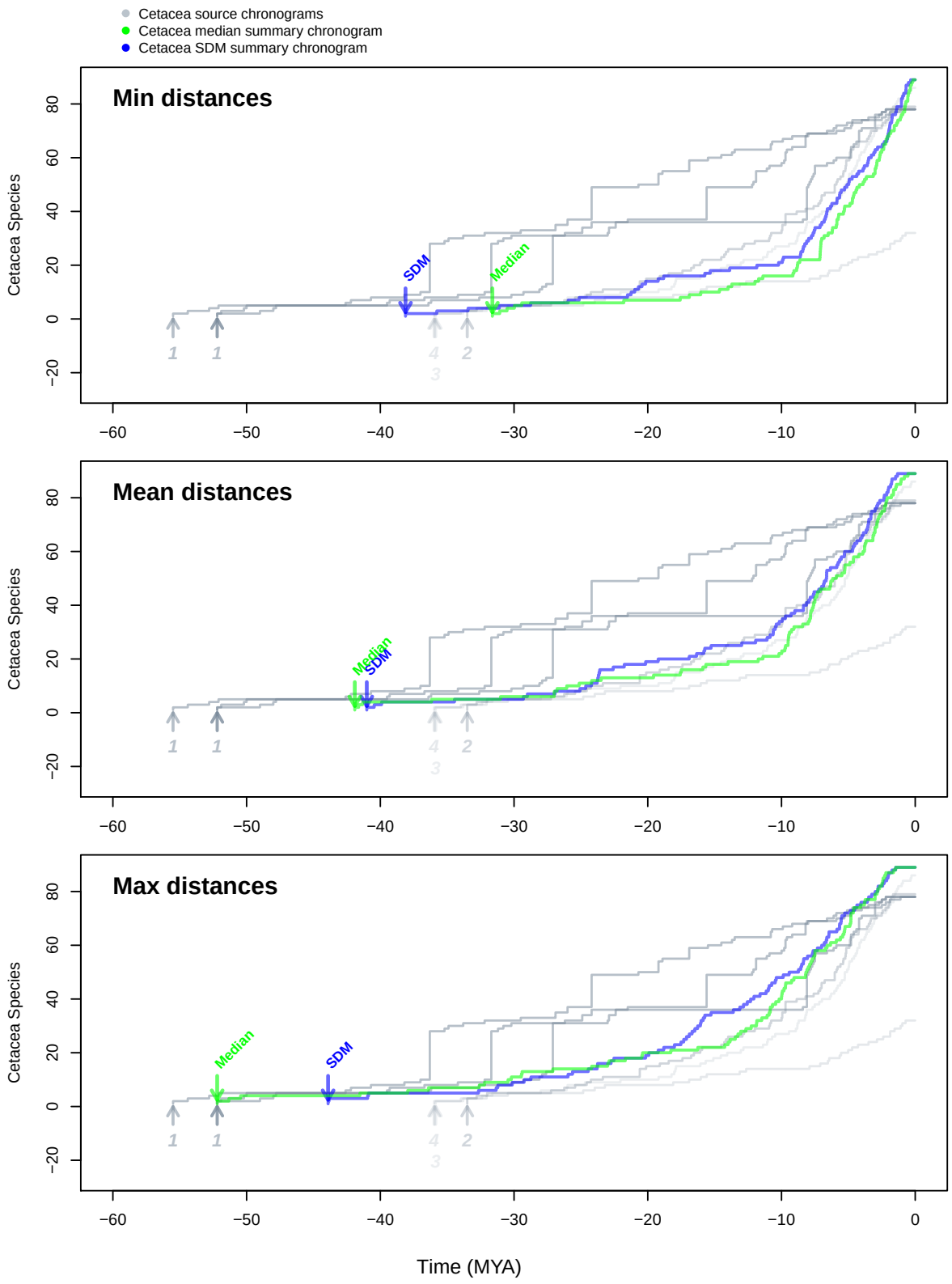


Figure 3: Cetacea lineage through time (LTT) plots from source chronograms (gray), median (green) and SDM (blue) summary chronograms obtained by calibrating a consensus tree tropology with distance data from respective summary matrices and then adjusting branch lengths with BLADJ.

Appendix

The following species were completely absent from the chronogram data base: *Amphiptera pacifica*, *Balaena agamachschik*, *Balaena mangidach*, *Balaenoptera andrejewi*, *Balaenoptera caerulescens*, *Balaenoptera emargenata*, *Balaenoptera grimmi*, *Balaenoptera maculata*, *Balaenoptera nigra*, *Balaenoptera punctulata*, *Catodon polycyphus*, *Catodon polyscyphus*, *Catodon svineval*, *Cephalorhyncus commersonii*, *Cephalorhyncus heavisidii*, *Clymenia gadamu*, *Delphinapterus senedetta*, *Delphinorhynchus maculatus*, *Delphinorhynchus pernettyi*, *Delphinorhynchus santonicus*, *Delphinus abusalam*, *Delphinus anarnacus*, *Delphinus attenuatus*, *Delphinus bertini*, *Delphinus bonnaterrei*, *Delphinus boryi*, *Delphinus caerulescens*, *Delphinus carbonarius*, *Delphinus coronatus*, *Delphinus cymodice*, *Delphinus cymodoce*, *Delphinus epiodon*, *Delphinus eurynome*, *Delphinus fabricii*, *Delphinus feres*, *Delphinus gadamu*, *Delphinus hamatus*, *Delphinus harlani*, *Delphinus leucocephalus*, *Delphinus livittatus*, *Delphinus maculatus*, *Delphinus maculiventer*, *Delphinus minimus*, *Delphinus nesarnac*, *Delphinus niger*, *Delphinus pernettyensis*, *Delphinus pernetyi*, *Delphinus perniger*, *Delphinus rappii*, *Delphinus rhinoceros*, *Delphinus salam*, *Delphinus siculus*, *Delphinus symodice*, *Delphinus walkeri*, *Epiodon rafinesque*, *Epiodon urganantus*, *Eudelphinus tasmaniensis*, *Globicephala macrorhyncus*, *Globicephalus fuscus*, *Globicephalus uneidens*, *Globiocephalus chinensis*, *Inia araguaiaensis*, *Inia boliviensis*, *Lagenodelphis australis*, *Lagenodelphis obliquidens*, *Lagenoelphis hosei*, *Lagenorhynchus bombifrons*, *Lagenorhynchus nilssonii*, *Lagenorhynchus posidonia*, *Lagenorhynchus superciliosus*, *Lagenorhyncus acutus*, *Lagenorhyncus albirostris*, *Lagenorhyncus australis*, *Lagenorhyncus cruciger*, *Lagenorhyncus obliquidens*, *Lagenorhyncus obscurus*, *Mesoplodon hotaula*, *Mesoplodon lazardii*, *Monodon spurius*, *Neophocaena asiaeorientalis*, *Phocaena posidonia*, *Physeter gibbosus*, *Physeter katadon*, *Physeter krefftii*, *Physeter polycephus*, *Physeter polycystus*, *Physeter pterodon*, *Platanista indi*, *Prodelphinus malayanus*, *Sotalia gadamu*, *Sotalia maculiventer*, *Sotalia perniger*, *Sotalia santonicus*, *Sousa gadamu*, *Sousa plumbea*, *Sousa sahulensis*, *Steno fuscus*, *Steno gadamu*, *Steno malayanus*, *Steno perniger*, *Tursio catalania*, *Tursio cymodoce*, *Tursio eurynome*, *Tursiops australis*, *Tursiops catalania*, *Tursiops cymodice*, *Tursiops dawsoni*, *Tursiops fergusoni*, *Tursiops nesarnack*

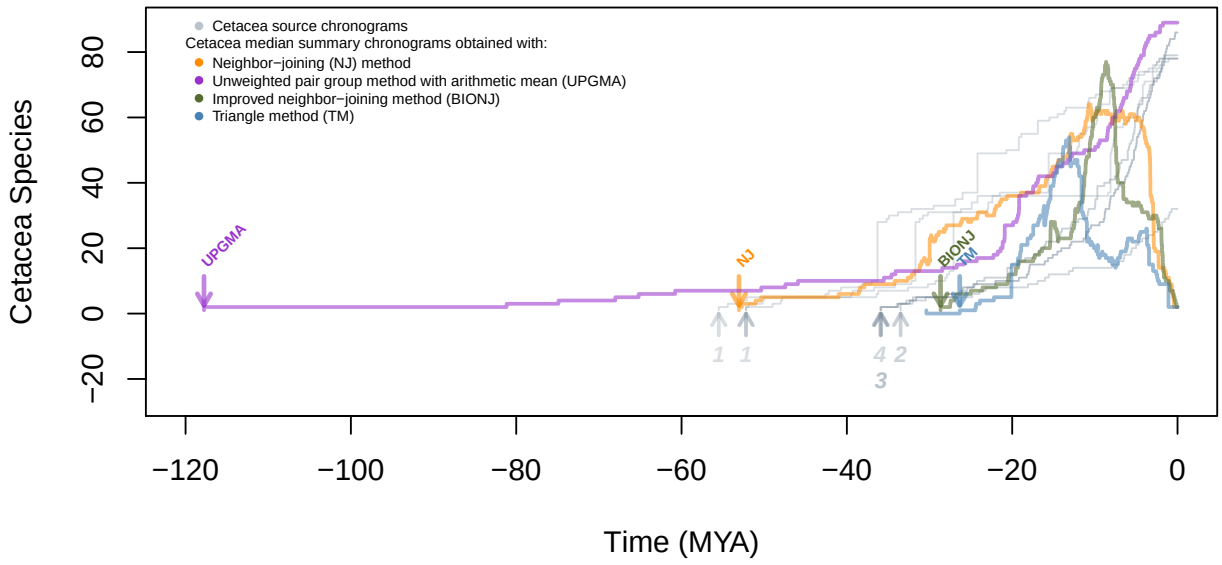


Figure 4: Lineage Through Time plots of Cetacea SDM summary chronograms obtained with different clustering algorithms. Not all algorithms worked with the SDM summary matrix and we are only showing here the ones that worked. Chronograms obtained from the median summary matrix are very similar to the ones shown here with all algorithms (mainFig. 2).

Dated induced subtree could not be obtained for the Cetacea.

This taxon's SDM matrix has some negative values in the following taxa: *Eubalaena japonica*, *Eubalaena glacialis*. This taxon's Median matrix has NO negative values.

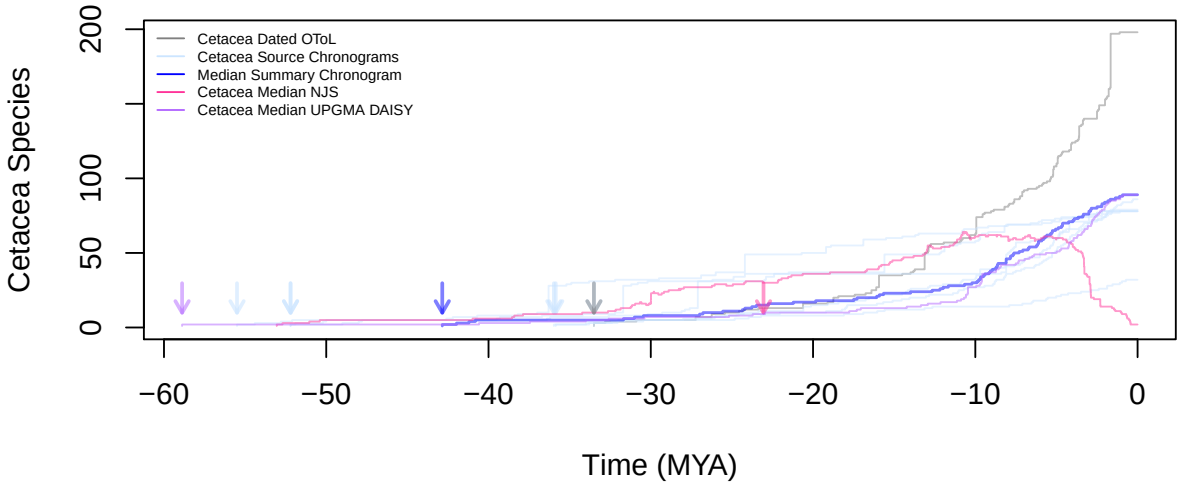


Figure 5: Cetacea lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we developped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

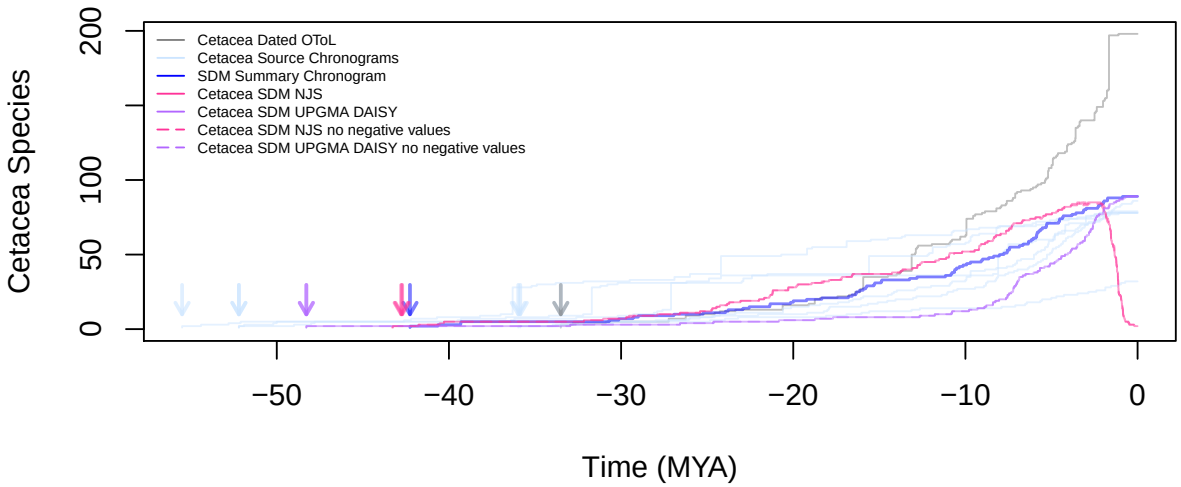


Figure 6: Cetacea lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with different methods (NJ and UPGMA). As you can note, dashed lines and solid lines from trees coming out from both types of clustering algorithms implemented are mostly overlapping. 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.

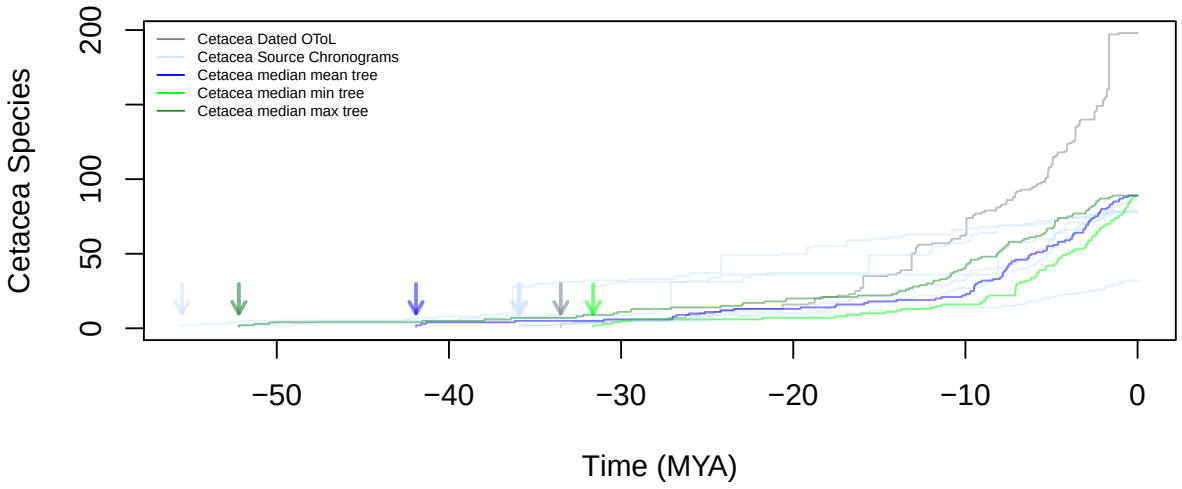


Figure 7: Cetacea lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with `datelife` algorithm.

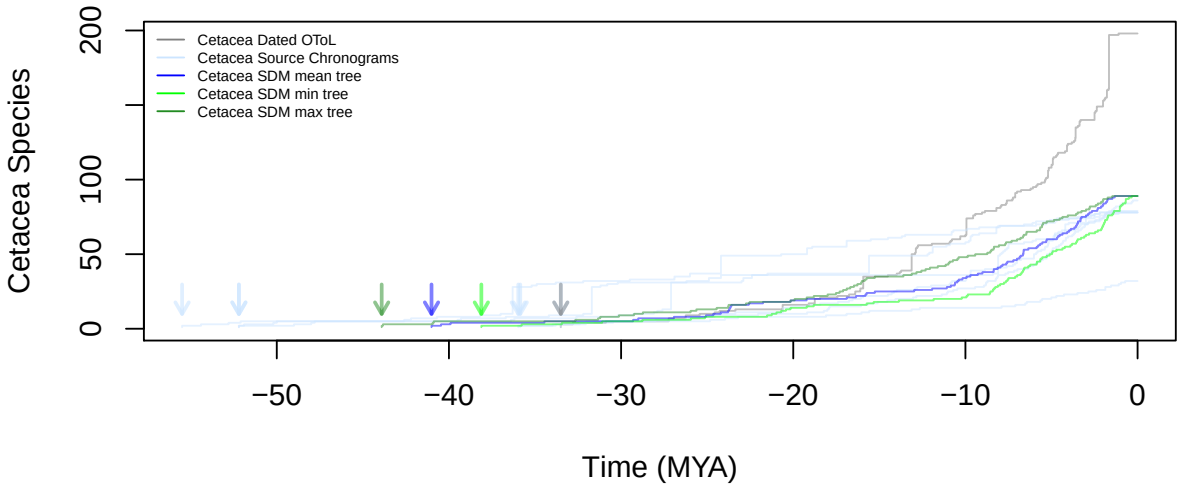


Figure 8: Cetacea lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with `datelife` algorithm.