DateLife Workflows

Luna L. Sanchez Reyes

2019-05-15

Taxon Fringilidae

1. Query source chronograms

There are 475 species in the Open Tree of Life Taxonomy for the taxon Fringilidae. Information on time of divergence is available for 286 of these species across 13 published and peer-reviewed chronograms. Original study citations as well as number of Fringilidae species found across those source chronograms is shown in Table 1. All source chronograms are fully ultrametric and their maximum ages range from 16.057 to 44.296 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 Fringilidae have at least two overlapping species. The largest grove has 2 chronograms (out of 13 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: *Carduelis uropygialis*, *Spinus crassirostris*.

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 1221 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. A tree with branch lengths could be constructed for 13 source chronograms (out of 13) available for the Fringilidae.

To date these trees we are using the software PATHd8 for tree dating without a molecular clock model, using calibrations from all other source chronograms. Sometime, calibrations conflict. To deal with conflicting calibrations, we can either expand them to make them agree, or we can congruify them to the topology of the tree to be dated. Results from both approaches are shown in the two following sections.

3.2.1. Expanding calibrations

crossval_pathd8_exp <- use_calibrations(phy = source_chronogram_bold_tree, calibrations = all_other_calibrations, dating_method = "pathd8", expand = default) crossval_pathd8_exp_notc <- use_calibrations(phy = source_chronogram_bold_tree_notc, calibrations = all_other_calibrations, dating_method = "pathd8", expand = default)

3.2.2. Summarizing calibrations (congruifying calibrations)

crossval_pathd8_summ <- use_calibrations(phy = source_chronogram_bold_tree, calibrations = all_other_calibrations, dating_method = "pathd8", expand = 0) crossval_pathd8_summ_notc <- use_calibrations(phy = source_chronogram_bold_tree_notc, calibrations = all_other_calibrations, dating_method = "pathd8", expand = 0)

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 Fringilidae 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: Fringilidae source chronogram studies information.

	Citation	***Source N***	***Taxon N***
1.	Barker, F. K., K. J. Burns, J. Klicka, S. M. Lanyon, I. J. Lovette. 2013. Going to extremes: contrasting rates of diversification in a recent radiation of New World passerine birds. Systematic Biology 62 (2): 298-320.	1	29/475
2.	Barker, F. Keith, Kevin J. Burns, John Klicka, Scott M. Lanyon, Irby J. Lovette. 2015. New insights into New World biogeography: An integrated view from the phylogeny of blackbirds, cardinals, sparrows, tanagers, warblers, and allies. The Auk 132 (2): 333-348.	2	102/475
3.	Burns, Kevin J., Allison J. Shultz, Pascal O. Title, Nicholas A. Mason, F. Keith Barker, John Klicka, Scott M. Lanyon, Irby J. Lovette. 2014. Phylogenetics and diversification of tanagers (Passeriformes: Thraupidae), the largest radiation of Neotropical songbirds. Molecular Phylogenetics and Evolution 75: 41-77.	1	27/475
4.	Claramunt, Santiago, Joel Cracraft. 2015. A new time tree reveals Earth historys imprint on the evolution of modern birds. Science Advances 1 (11): e1501005-e1501005	1	3/475
5.	Gibb, Gillian C., Ryan England, Gerrit Hartig, P.A. (Trish) McLenachan, Briar L. Taylor Smith, Bennet J. McComish, Alan Cooper, David Penny. 2015. New Zealand passerines help clarify the diversification of major songbird lineages during the Oligocene. Genome Biology and Evolution 7 (11): 2983-2995.	1	7/475
6.	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	2	250/475
7.	Hooper, Daniel M., Trevor D. Price. 2017. Chromosomal inversion differences correlate with range overlap in passerine birds. Nature Ecology & Evolution 1 (10): 1526-1534	1	47/475
8.	Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, A. O. Mooers. 2012. The global diversity of birds in space and time. Nature 491 (7424): 444-448	2	215/475
9.	Price, Trevor D., Daniel M. Hooper, Caitlyn D. Buchanan, Ulf S. Johansson, D. Thomas Tietze, Per Alström, Urban Olsson, Mousumi Ghosh-Harihar, Farah Ishtiaq, Sandeep K. Gupta, Jochen Martens, Bettina Harr, Pratap Singh, Dhananjai Mohan. 2014. Niche filling slows the diversification of Himalayan songbirds. Nature 509: 222-225.	2	2/475

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

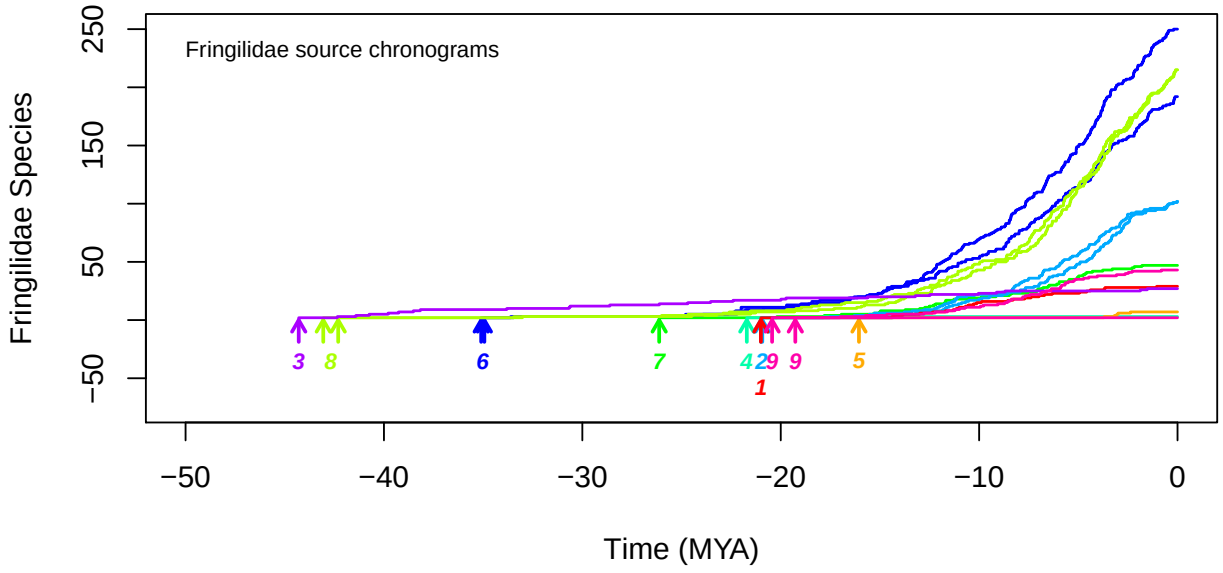


Figure 1: Lineage through time (LTT) plots of source chronograms available in data base for species in the Fringilidae. Numbers correspond to original studies in Table 1. Arrows indicate maximum age of each chronogram.

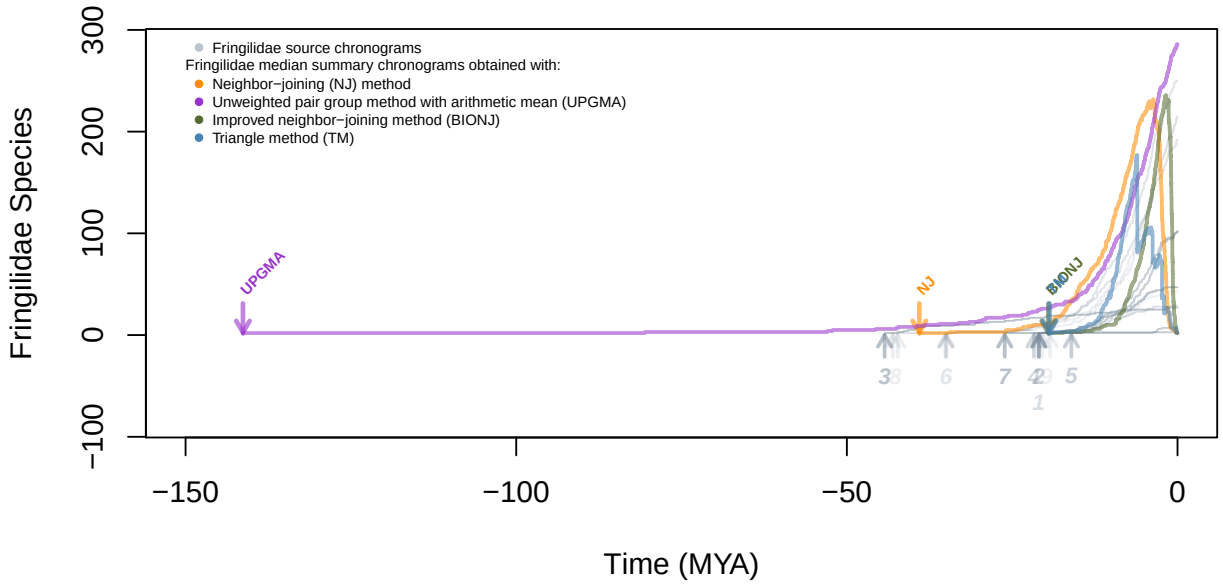
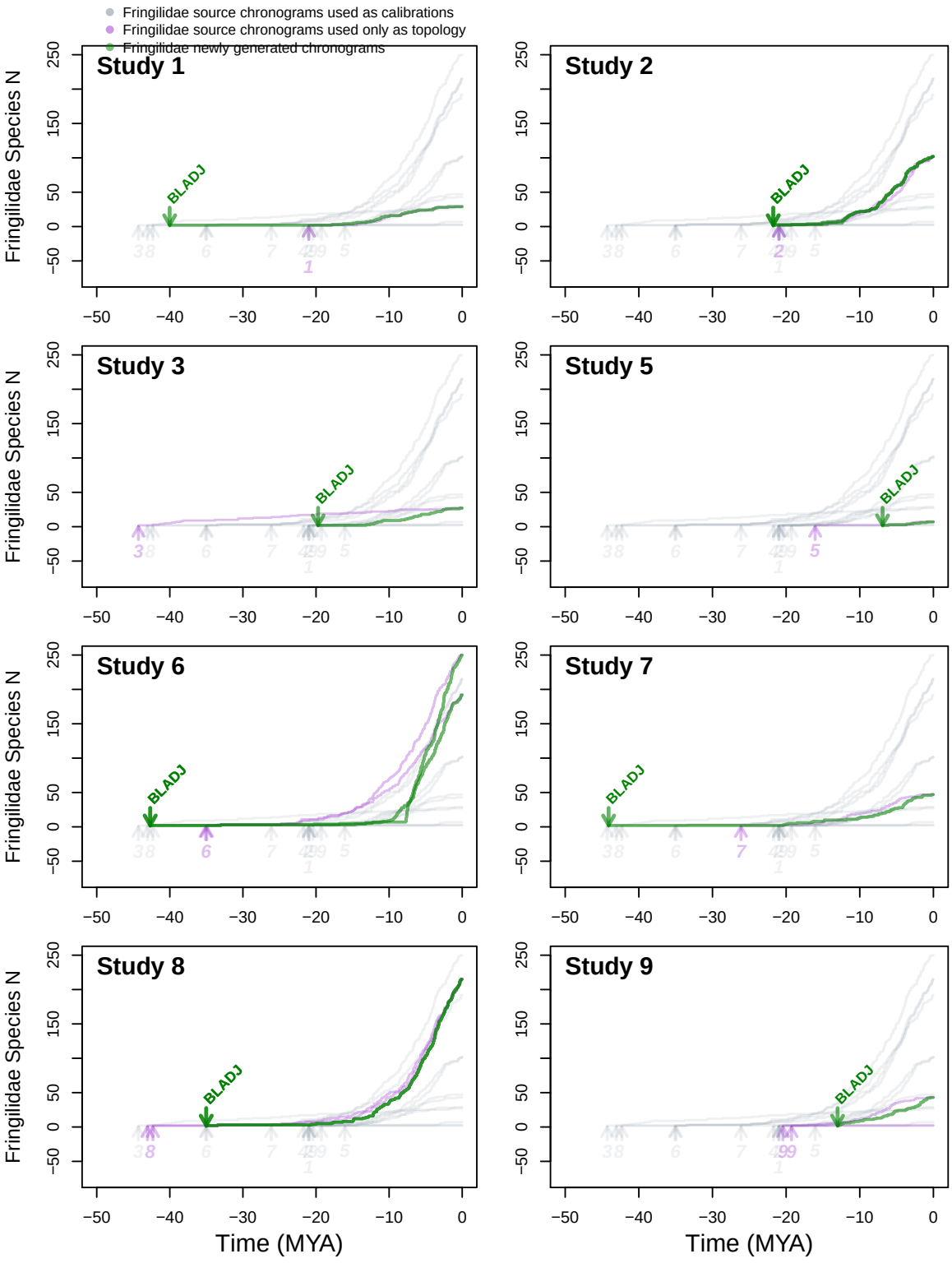


Figure 2: Lineage Through Time plots of Fringilidae median summary chronograms obtained with different clustering algorithms. Not all algorithms worked with this summary matrix and we are only showing here the ones that worked. Chronograms obtained from the SDM summary matrix are very similar to the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


Fringilidae source chronograms used as calibrations
Fringilidae source chronograms used only as topology
Fringilidae newly generated chronograms
Study 1
Study 2
Study 3
Study 5
Study 6
Study 7
Study 8
Study 9
BLADJ
Fringilidae Species N
Time (MYA)

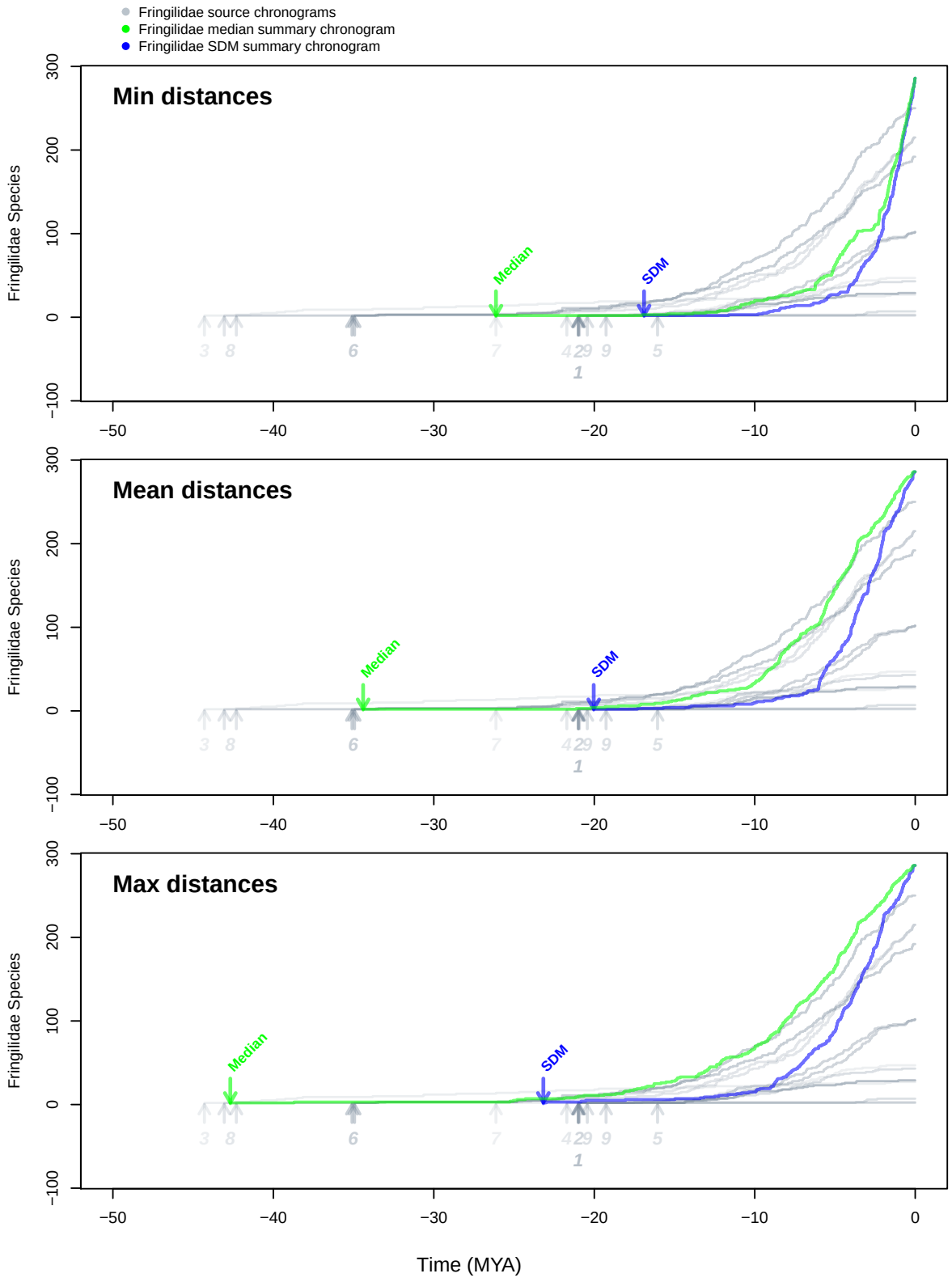


Figure 3: Fringilidae lineage through time (LTT) plots from source chronograms (gray), median (green) and SDM (blue) summary chronograms obtained by calibrating a consensus tree tropology with distance data from respective summary matrices and then adjusting branch lengths with BLADJ.

Appendix

The following species were completely absent from the chronogram data base: *Acanthis cabaret*, *Acanthis rostrata*, *Akialoa ellisiana*, *Akialoa lanaiensis*, *Akialoa obscura*, *Buarremon apertus*, *Bucanetes crassirostris*, *Calcarius coloratus*, *Cardinalis carneus*, *Cardinalis peninsulae*, *Carduelis ankoberensis*, *Carduelis elegans*, *Carduelis ultima*, *Carpodacus beicki*, *Carpodacus davidianus*, *Carpodacus deserticolor*, *Carpodacus dubius*, *Carpodacus formosanus*, *Carpodacus henrici*, *Carpodacus longirostris*, *Carpodacus lucifer*, *Carpodacus portenkoi*, *Carpodacus rhodopeplus*, *Carpodacus roseatus*, *Carpodacus rubicundus*, *Carpodacus stoliczkae*, *Carpodacus verreauxii*, *Carpodacus waltoni*, *Caryothraustes brasiliensis*, *Caryothraustes scapularis*, *Chaunoproctus ferreorostris*, *Chloridops kona*, *Chloris heinrichi*, *Chloris turkestanica*, *Chrysocorythus mindanensis*, *Ciridops anna*, *Coccothraustes japonicus*, *Coccothraustes migratorius*, *Corytus rhenana*, *Crithagra albifrons*, *Crithagra ankoberensis*, *Crithagra buchanani*, *Crithagra canicapilla*, *Crithagra capistrata*, *Crithagra concolor*, *Crithagra deserti*, *Crithagra donaldsoni*, *Crithagra flavigula*, *Crithagra frontalis*, *Crithagra granti*, *Crithagra hewitti*, *Crithagra hildegardae*, *Crithagra kikuyensis*, *Crithagra koliensis*, *Crithagra leucoptera*, *Crithagra marshalli*, *Crithagra menachensis*, *Crithagra montanorum*, *Crithagra mozambica*, *Crithagra reichenowi*, *Crithagra rothschildi*, *Crithagra rufobrunnea*, *Crithagra symonsi*, *Crithagra thomensis*, *Crithagra tristriata*, *Crithagra xantholaema*, *Crithagra xanthopygia*, *Cyanerpes holti*, *Cyanerpes isthmicus*, *Cyanerpes microrhynchus*, *Cyanocompsa argentina*, *Cyanocompsa rothschildii*, *Drepanis coccinea*, *Drepanis funerea*, *Drepanis pacifica*, *Dysmorodrepanis munroi*, *Emberiza buturlini*, *Emberiza ciodes*, *Emberiza ciopsis*, *Emberiza continentalis*, *Emberiza elegantula*, *Emberiza erythrogenys*, *Emberiza flemingorum*, *Emberiza fronto*, *Emberiza kuatunensis*, *Emberiza lydiae*, *Emberiza meridionalis*, *Emberiza militaris*, *Emberiza musica*, *Emberiza neobscura*, *Emberiza nivenorum*, *Emberiza omissa*, *Emberiza omoensis*, *Emberiza orientalis*, *Emberiza ornata*, *Emberiza pyrrhulinus*, *Emberiza rufibarba*, *Emberiza rufigularis*, *Emberiza sahari*, *Emberiza semenowi*, *Emberiza sloggetti*, *Emberiza sordida*, *Emberiza vincenti*, *Emberiza zaidamensis*, *Embernagra gossei*, *Eophona magnirostris*, *Eophona sowerbyi*, *Erythrospiza phaenicoptera*, *Euphonia aurantiicollis*, *Euphonia carnegiei*, *Euphonia flavifrons*, *Euphonia gnatho*, *Euphonia nitida*, *Euphonia olivacea*, *Euphonia praetermissa*, *Euphonia purpurascens*, *Euphonia rufivertex*, *Euphonia serrirostris*, *Euphonia tavarae*, *Fringilla albicollis*, *Fringilla bella*, *Fringilla brissonii*, *Fringilla nortoniensis*, *Fringilla palmae*, *Fringilla polatzeki*, *Fringilla syriaca*, *Fringillaria goslingi*, *Fringillaria poliopleura*, *Haemorhous californicus*, *Haemorhous griscomi*, *Hemignathus affinis*, *Hemignathus hanapepe*, *Hemispingus castaneicollis*, *Hemispingus macrophrys*, *Hemispingus ochraceus*, *Hemispingus urubambae*, *Hesperiphona abeillei*, *Hesperiphona cobanensis*, *Hesperiphona montana*, *Himatione fraithii*, *Leucosticte brunneonucha*, *Leucosticte wallowa*, *Leucosticte walteri*, *Linaria harterti*, *Linaria johannis*, *Linaria rufostrigata*, *Linaria yemenensis*, *Linurgus kilimensis*, *Loxia cardinalis*, *Loxia cyanea*, *Loxia dominica*, *Loxia mesamericana*, *Loxops ochraceus*, *Loxops wolstenholmei*, *Melopyrrha taylori*, *Mycerobas melanoxanthos*, *Passerina lazula*, *Passerina pallidior*, *Passerina purpurascens*, *Peucaea cohaerens*, *Peucaea ibarrorum*, *Peucaea vulcanica*, *Peucedramus micrus*, *Pheucticus aurantiacus*, *Pinicola eschatosa*, *Plectrophenax townsendi*, *Psittirostra psittacea*, *Pyrrhula cineracea*, *Pyrrhula owstoni*, *Pyrrhula rosacea*, *Pyrrhula steerei*, *Pyrrhula uchidai*, *Rhodacanthis flaviceps*, *Rhodacanthis palmeri*, *Rhodopechys alienus*, *Rhodopechys sanguineus*, *Rhynchostruthus louisae*, *Rhynchostruthus percivali*, *Rhynchostruthus socotranus*, *Serinus huillensis*, *Spinus atriceps*, *Spinus colombiana*, *Spinus dominicensis*, *Spinus longirostris*, *Spinus nigricauda*, *Spinus oleacea*, *Spinus perplexa*, *Spinus stejnegeri*, *Viridonia sagittirostris*

Dated induced subtree could not be obtained for the Fringilidae.

This taxon's SDM matrix has some negative values in the following taxa: *Carduelis uropygialis*, *Spinus crassirostris*. This taxon's Median matrix has NO negative values.

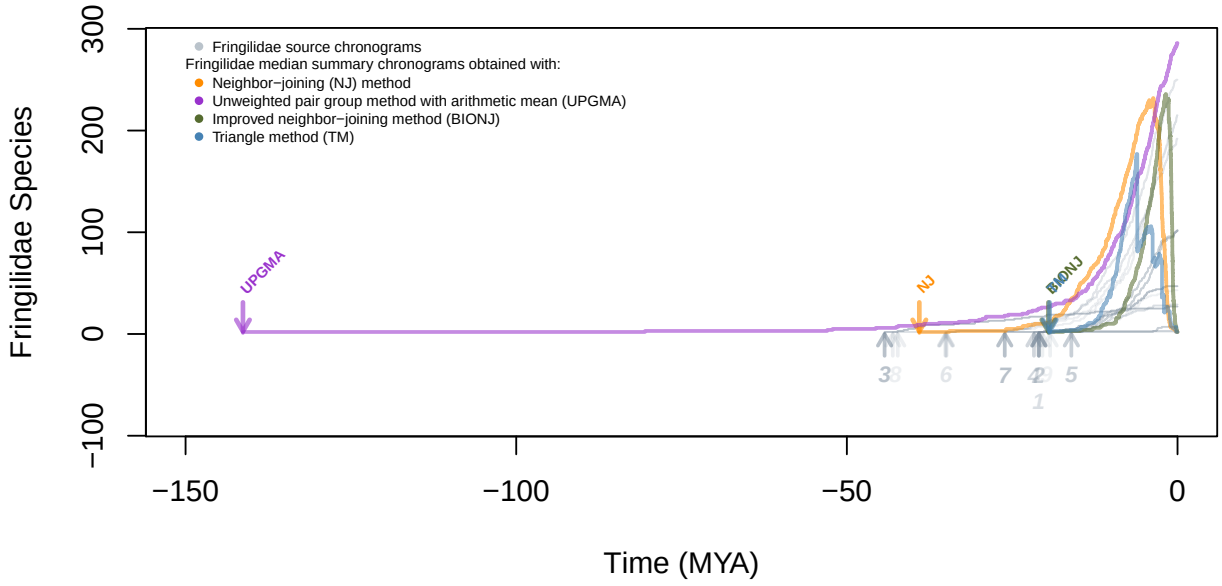


Figure 4: Lineage Through Time plots of Fringilidae SDM summary chronograms obtained with different clustering algorithms. Not all algorithms worked with the SDM summary matrix and we are only showing here the ones that worked. Chronograms obtained from the median summary matrix are very similar to the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

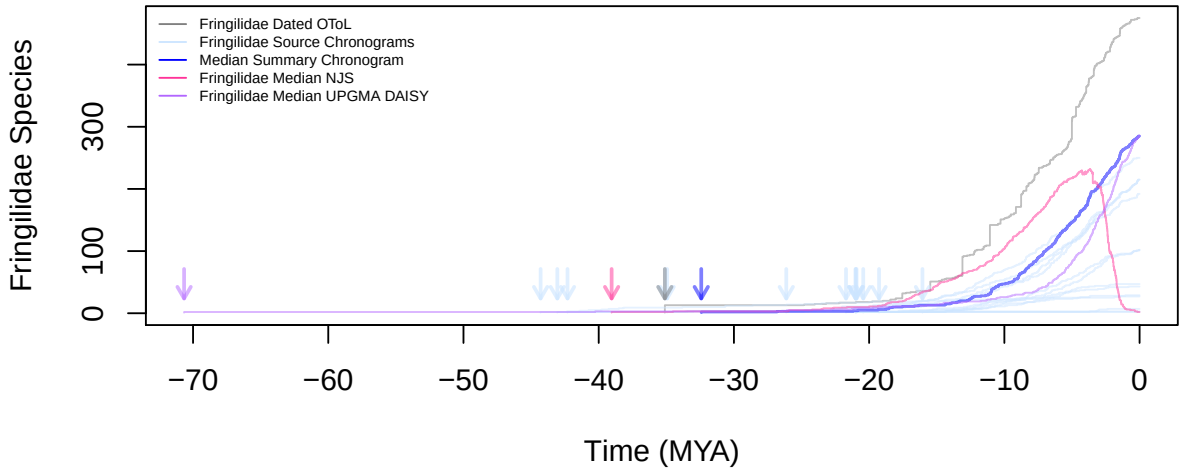


Figure 5: Fringilidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with different methods (NJ and UPGMA). Clustering algorithms used often are returning non-ultrametric trees or with maximum ages that are just off (too old or too young). So we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

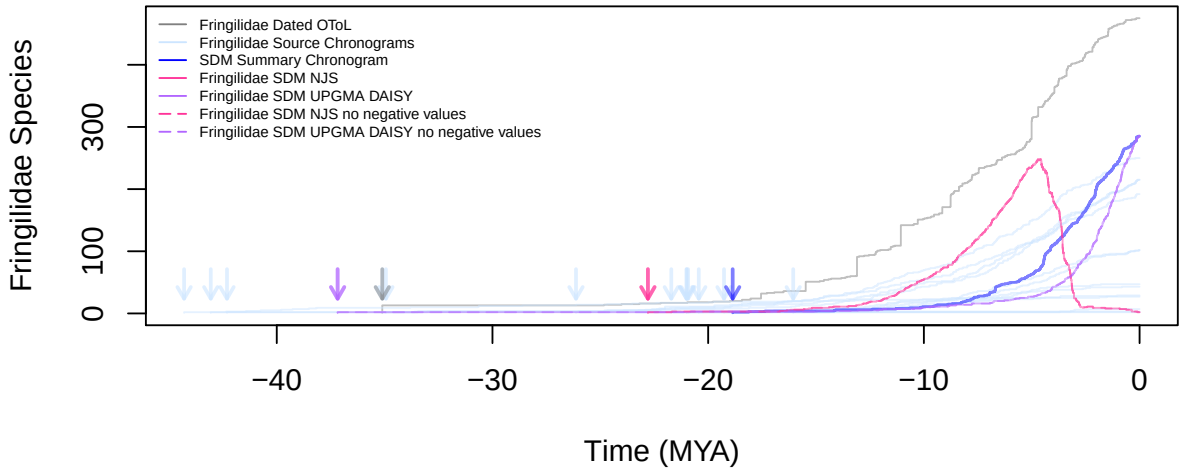


Figure 6: Fringilidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with different methods (NJ and UPGMA). As you can note, dashed lines and solid lines from trees coming out from both types of clustering algorithms implemented are mostly overlapping.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

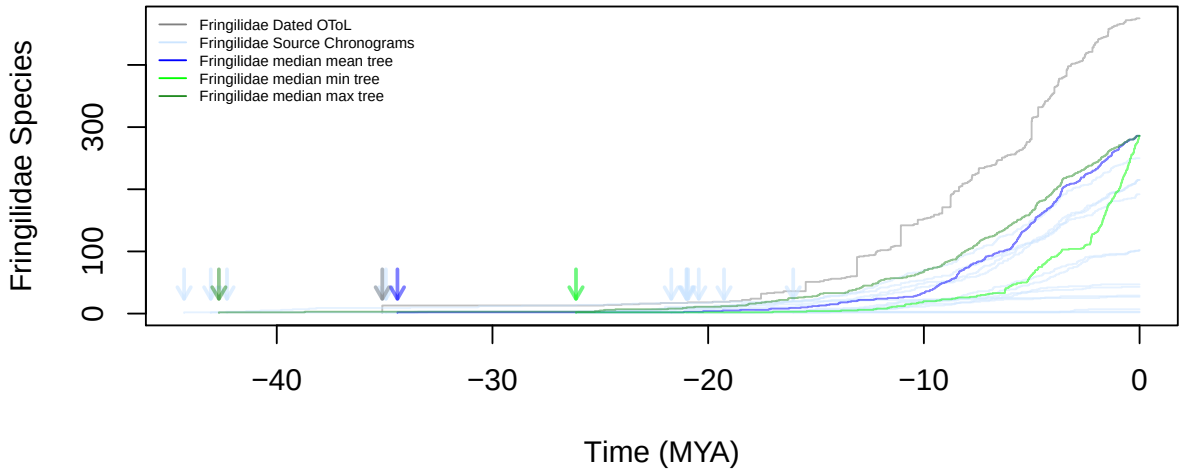


Figure 7: Fringilidae lineage through time (LTT) plots from source chronograms and Median summary matrix converted to phylo with `datelife` algorithm.

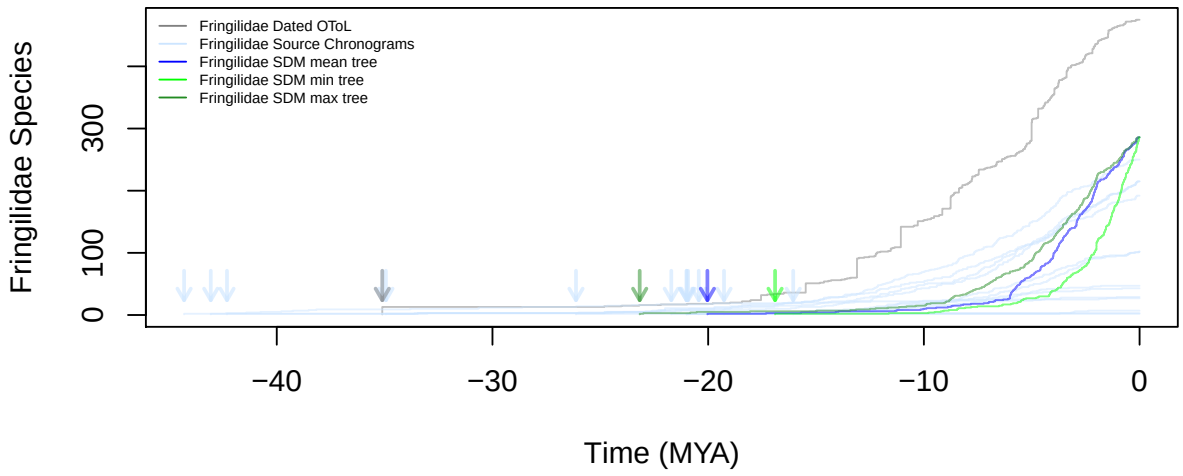


Figure 8: Fringilidae lineage through time (LTT) plots from source chronograms and SDM summary matrix converted to phylo with `datelife` algorithm.