DateLife Workflows

Luna L. Sanchez Reyes

2019-06-18

Taxon Anolis

1. Query source chronograms

There are 273 species in the taxon Anolis, following the NCBI taxonomy database. Information on time of divergence is available for 252 of these species across 6 published and peer-reviewed chronograms. Original study citations as well as number of Anolis species found across those source chronograms is shown in `Table` 1. All source chronograms are fully ultrametric and their maximum ages range from 47.843 to 72.578 million years ago (MYA). As a means for comparison, lineage through time plots of all source chronograms available in data base are shown in `Figure` 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 chronograms: 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 the case of the Anolis all source chronograms have at least two overlapping species.

Now that we know that the best grove has all source chronograms 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. 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 (`Figure` 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 `Figure` 3. This variation simply represents variation in source data.

3. Generate new chronograms

Another way to take advantage of the information available in 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 the topology of interest. In this data set we have 1132 calibrations in total (that basically corresponds to the sum of the number of nodes from each source chronogram). Once we have chosen or generated 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 apply minimum, mean or maximum node ages as secondary calibrations over the backbone tree using the software BLADJ. In general, the time of divergence information from other source chronograms allows to recover the divergence times from the original study (`Figure` 4). 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.

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 other algorithms for phylogenetic dating to get branch lengths proportional to absolute time such as PATHd8, treePL and MrBayes. 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 backbone tree topology. For 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 wide number of organisms. A tree with branch lengths could be constructed for 6 source chronograms (out of 6) available for the Anolis. To date these trees we use the software PATHd8 for tree dating without a molecular clock model, using calibrations from all other source chronograms. Sometimes, calibrations conflict between them. 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.

Tables and Figures

Table 1: Anolis source chronogram original studies information.

	Citation	***Source N***	***Taxon N***
1.	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	191/273
2.	Mahler, D. L., T. Ingram, L. J. Revell, J. B. Losos. 2013. Exceptional Convergence on the Macroevolutionary Landscape in Island Lizard Radiations. Science 341 (6143): 292-295.	1	98/273
3.	Pyron, R. Alexander, Frank T. Burbrink. 2013. Early origin of viviparity and multiple reversions to oviparity in squamate reptiles. Ecology Letters 17 (1): 13-21	1	203/273
4.	Steven Poe, Adrián Nieto-montes de oca, Omar Torres-carvajal, Kevin De Queiroz, Julián A. Velasco, Brad Truett, Levi N. Gray, Mason J. Ryan, Gunther Köhler, Fernando Ayala-varela, Ian Latella, 2017, 'A Phylogenetic, Biogeographic, and Taxonomic study of all Extant Species of Anolis (Squamata; Iguanidae)', Systematic Biology, vol. 66, no. 5, pp. 663-697	1	242/273
5.	Wright, April M., Kathleen M. Lyons, Matthew C. Brandley, David M. Hillis. 2015. Which came first: The lizard or the egg? Robustness in phylogenetic reconstruction of ancestral states. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 324 (6): 504-516	1	203/273
6.	Yuchi Zheng, John J. Wiens, 2016, 'Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species', Molecular Phylogenetics and Evolution, vol. 94, pp. 537-547	1	201/273

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

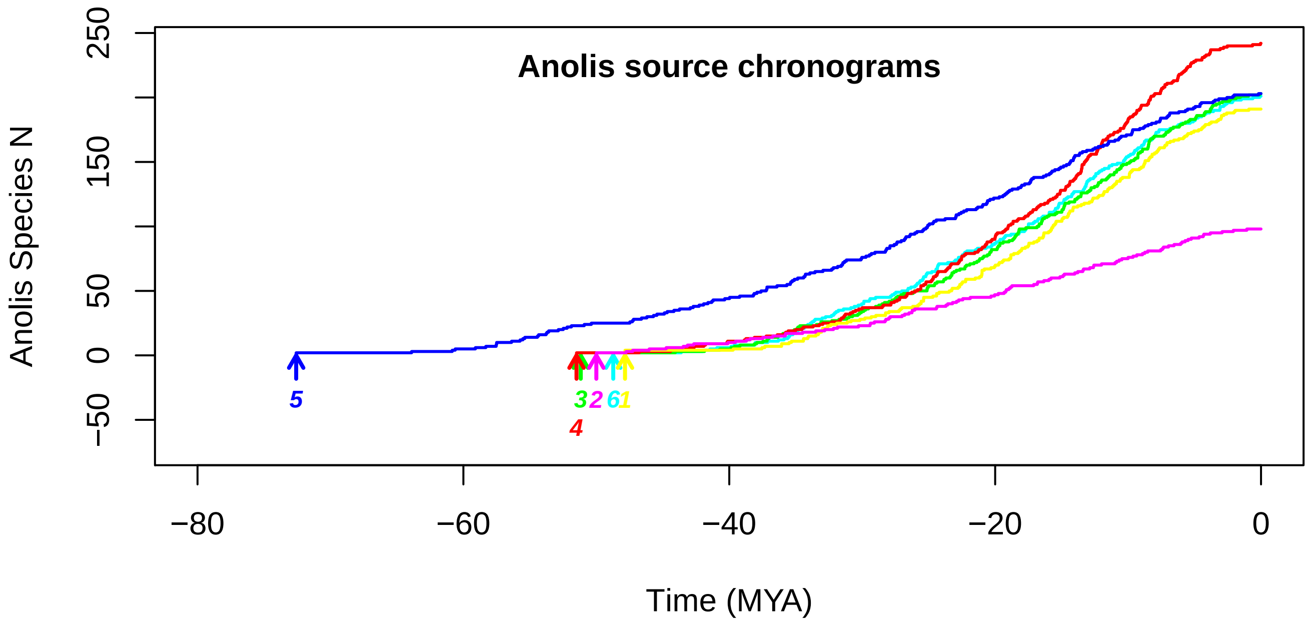

Figure 1: Lineage through time (LTT) plots of source chronograms available in database for species in the Anolis. Numbers correspond to original studies in Table 1. Arrows indicate maximum age of each chronogram.

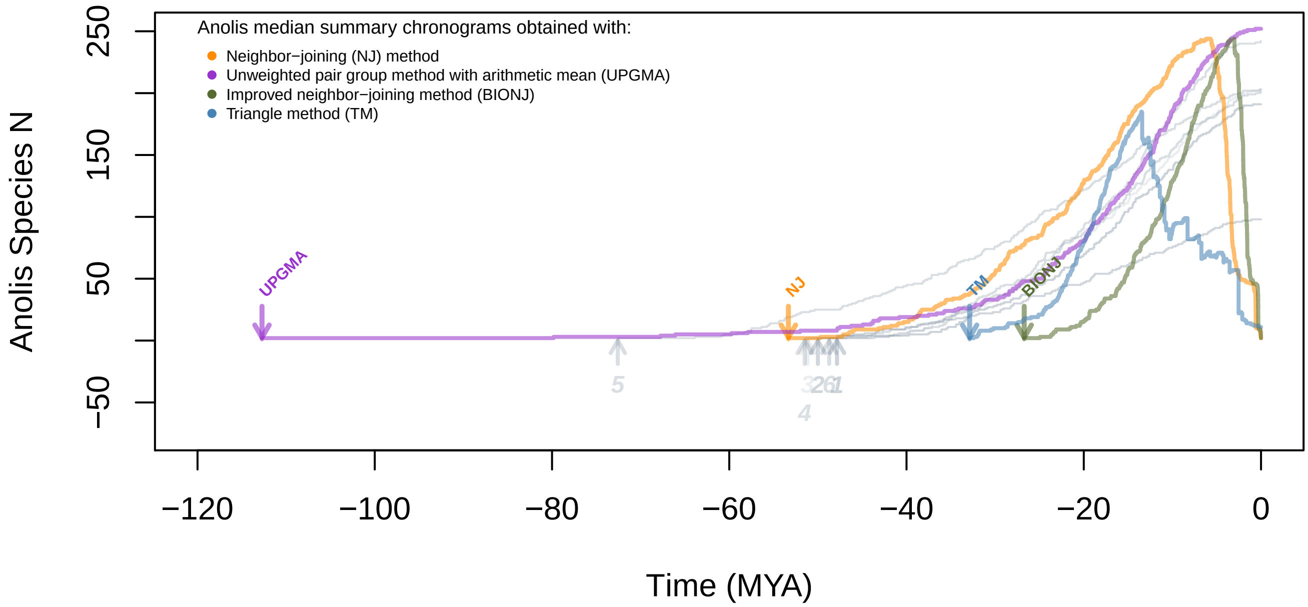

Figure 2: Lineage Through Time plots of Anolis median summary chronograms obtained with different clustering algorithms. Not all algorithms worked with this summary matrix and we are only showing here the ones that worked. Chronograms obtained from the SDM summary matrix are very similar to the ones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

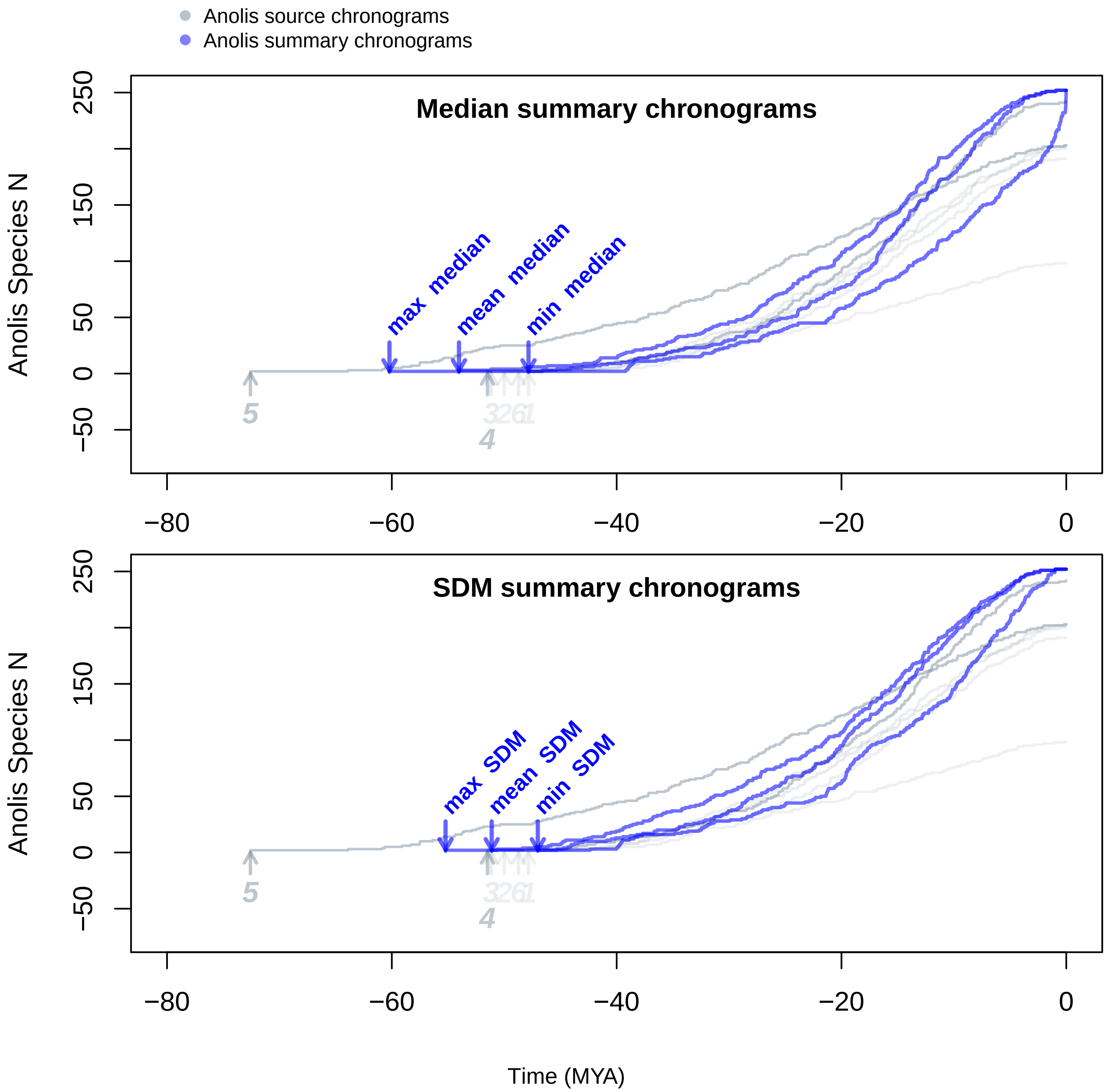

Figure 3: Anolis lineage through time (LTT) plots of summary chronograms obtained by calibrating a consensus tree tropology with distance data from median (upper) and SDM (lower) summary matrices and then adjusting branch lengths with BLADJ. Source chronograms are shown in gray for comparison.

Appendix

The following species were not found in the chronogram database: *Anolis alocomyos, Anolis carlliebi, Anolis cf. alocomyos GK-2015, Anolis cf. humilis JJK-2013, Anolis cf. polylepis, Anolis chrysops, Anolis desiradei, Anolis dissimilis, Anolis immaculogularis, Anolis landestoyi, Anolis leditzigorum, Anolis nietoi, Anolis osa, Anolis philopunctatus, Anolis phyllorhinus, Anolis sacamecatensis, Anolis stevepoei, Anolis tenorioensis, Anolis terueli, Anolis wattsii, Anolis zapotecorum*

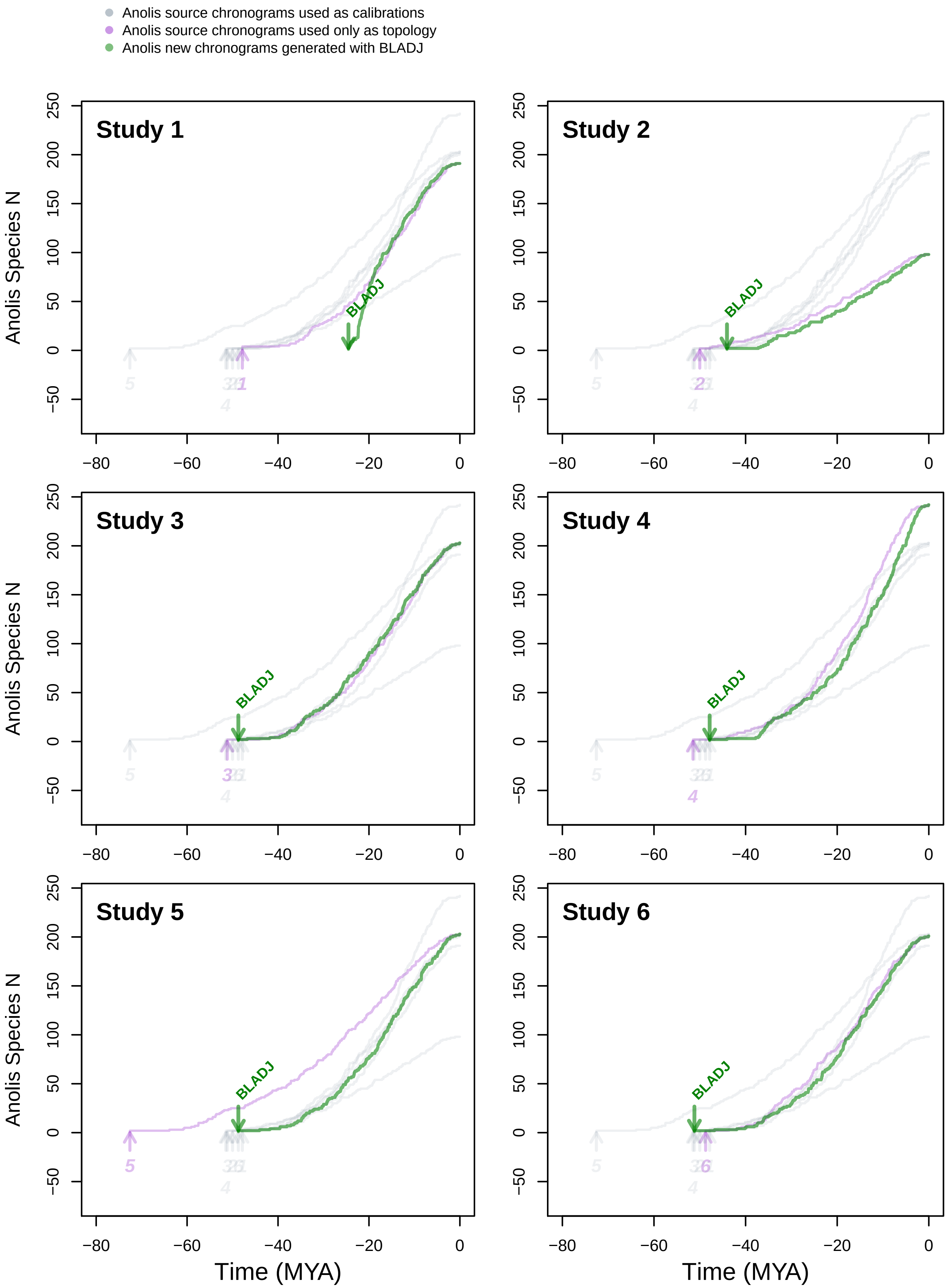

Figure 4: Anolis lineage through time (LTT) plots from source chronograms used as secondary calibrations (gray), source chronograms used as topology (purple) and chronograms resulting from calibrating the latter with the former, using BLADJ (green).

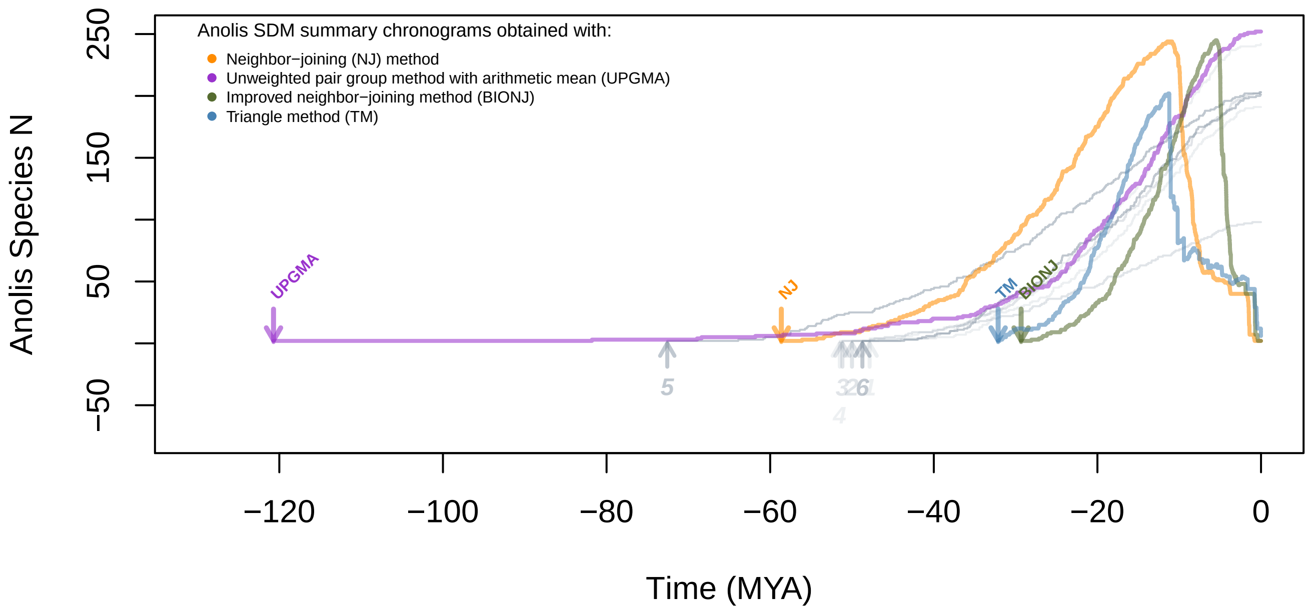

Figure 5: Lineage Through Time plots of Anolis SDM summary chronograms obtained with different clustering algorithms. Not all algorithms worked with the SDM summary matrix and we are only showing here the ones that worked. Chronograms obtained from the median summary matrix are very similar to the ones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