On the estimation of species age from phylogenetic trees

Carlos Calderón del Cid

Torsten Hauffe

Juan D. Carrillo

Daniele Silvestro

# Abstract

The estimation of the age of extant species (the elapsed time from origination to the present) can give an insight into how species longevity might influence the diversification dynamics and ecological evolution. Traditionally, species ages have been measured in the fossil record; however, recently, researchers have estimated the ages of extant species from the branch lengths of time-calibrated phylogenies. This approach poses problems because most phylogenies do not contain information about extinctions and speciation modes. When phylogenies do not include extinct lineages, the branch lengths of sister extant species are likely to appear longer than they truly are. The speciation modes (bifurcating, budding, or anagenetic) result on different topological representation and branch lengths of a phylogeny and, thus, likely affect the estimation of species age. We expected that bifurcating speciation would cause a lower mismatch between phylogenetic and true age than budding and anagenetic speciation. The information about extinctions (provided by fossils) is incomplete, and the mode of speciation in most cases is unknown. Therefore, computer simulations are valuable tool to evaluate the accuracy of estimating species ages from time-calibrated phylogenies, as the true age of species, extinctions and mode of speciation is known. This study aims to evaluate, through simulations, the potential sources of error (extinctions and modes of speciation) on the estimation of phylogenetic age. Furthermore, we developed a geometric function to partially correct phylogenetic ages. We found that extinction rate and speciation modes greatly influence the estimation of phylogenetic age. Bifurcating speciation presented low mismatch between true and estimated age, which increases with extinction rates. Budding speciation showed a high mismatch, which increases slightly with extinction rate. Anagenetic speciation (combined with bifurcating and budding) presented a very high mismatch not affected by extinction rates. An extinction sensitivity analysis showed many incorrect inferences when using the phylogenetic age. The geometric function did not have a good accuracy but presented a high confidence intervals coverage. Altogether our results indicate that ……. We advise a critical evaluation on speciation modes and extinction risk when estimating species ages from phylogenetic trees.

**Keywords**: Evolutionary history, extinction rate, simulations, speciation modes, phylogenies.

# Introduction

Species age expresses evolutionary history; thus, its association with ecological high-level processes sensu *Vellend (2016)* gives an insight into the interplay between evolution and ecology. For speciation, species age directly spurs from diversification rates, and therefore it relates to the history and taxa’s relationships of the clade itself (Hagen et al. 2015; Marshall 2017). Additionally, species age has been related to the range overlap between sister species to detect the role of geography in speciation (Barraclough, Vogler, and Harvey 1998; Barraclough and Vogler 2000). In the case of selection, it has been postulated that species age, by its influence over traits associated with fitness, would affect survivorship both negatively and positively (Benton 2016; Hagen et al. 2018). Biotic and abiotic changes could make older species that originated in a different environment than were they later live more prone to extinction than younger species (Pearson 1995; Eldredge et al. 2005). In contrast, strong competition from older species could inhibit recent species from establishing, leading to a high early extinction (Balmford 1996; Januario and Quental 2021). For dispersal, species age could be a measure of colonization time, especially in island systems or biotic invasions triggered by geological events, such as the formation of the Central American Isthmus for the Great American Interchange (Lee, Tanentzap, and Heenan 2012; Carrillo et al. 2015), (Leopold et al. 2015). In the case of drift, species age could reflect population size. At its inception, species start with a small population being more prone to drift (Davies et al. 2011; Januario and Quental 2021). On the other hand, if species age is positively related to specialization, old species would tend to have smaller populations and spatial ranges, experiencing greater drift (Pearson 1995; Poisot et al. 2011). Thus, when associated with ecological processes, species age can offer insights into the history that molded past and extant biodiversity (McGill et al. 2019).

Traditionally, species age is measured in the fossil record through different statistical and probabilistic approaches based mostly on taxa’s stratigraphic duration (Mike Foote and Raup 1996; Michael Foote et al. 2007). Several of these approaches consider the differences in fossil sampling and temporal resolution (Hagen et al. 2015; Silvestro et al. 2019). Species ages estimated from paleobiological data offer a reliable measure of species’ temporal duration which can be used in macroevolutionary studies (Van Valen 1973; Benton 2016; Silvestro et al. 2020). More recently, some studies have interpreted that the age of extant species can be measured from the branch lengths of time-calibrated phylogenies (Gaston et al. 1997; Johnson, Delean, and Balmford 2002; Verde Arregoitia, Blomberg, and Fisher 2013; Tanentzap et al. 2020). In these studies species age is measure as the since the species’ origination to the present. In this sense, there are two age types for living species. Crown age is the interval between species origination, or the divergence from its nearest sister, to the present (Harmon 2019). Similarly, stem age is the lapse between the origination of the most recent common ancestor to the present (Magallon and Sanderson 2001). However, estimating species’ age from phylogenetic trees (hereafter called "phylogenetic age") can lead to over or underestimations due to three different shortfalls.

First, incomplete sampling of extant species, especially in extreme habitats and the Global South, may not only reduce the accuracy of the tree topology and increase the uncertainty in any phylogenetic estimation (Heath, Hedtke, and Hillis 2008; Cai et al. 2022; Mynard et al. 2023). Inevitable, incomplete taxon sampling results in longer branch lengths and therefore an overestimation of species ages than when all known extant species are included in a time-calibrated phylogeny.

Second, extinct lineages are not included in time-calibrated phylogenies of extant species only and, due to incomplete fossil sampling, underrepresented in tip-dated phylogenies. Not including unsampled extant and extinct species from the phylogeny results in branch lengths of extant species appearing larger than they indeed are. As an example, when the extinct species of our own genus *Homo* genus are not included in the phylogeny; the phylogenetic of *Homo sapiens* is approximately 6 million years (i.e. our last common ancestor with the chimpanzee), much older than the oldest fossil record for our species of 300 thousand years (Figure [1](#fig:fossils_vs_phylo); Mann and Weiss 1996; Callaway 2017).

The third shortfall of phylogenies of extant species is that the tree alone does not contain information about the speciation mode (Mike Foote 1996; Silvestro et al. 2018). Three different speciation modes are discussed in the paleobiological literature to reflect various biological processes generating divergence between species and different species concepts.

Explain here the modes of speciation, give an example of the biological process causing the speciation, the corresponding species concept, and whether the phylogenetic age is an over- or underestimation given the respective mode of speciation.

Budding

In the case of budding speciation, a speciation event gives rise to one new species while the ancestral lineage persists. The phylogenetic age for the ancestral species will therefore be an underestimation of its true age because by the phylogeny alone it is impossible to tell what the ancestral lineage of two sister species is (Raup 1985; Caetano and Quental 2022; Stadler et al. 2018).

bifurcating

Under the assumption of bifurcating speciation, phylogenetic ages equal the age of extant species when there is no extinction. With higher extinction rates, extinct sister species would cause the branch lengths of the surviving sister to appear larger than they are (Figure [2](#fig:phylo_raccoons)), resulting in an overestimation of species age.

Anagenetic

For anagenetic speciation, a diversification event where one species arises, replacing the ancestral species, the phylogenetic age always overestimates the true age, and the problem gets worst regarding missing species (Bapst 2012; Stadler et al. 2018; Barido-Sottani et al. 2019).

These speciation modes determine species longevities (Wagner, Erwin, and Anstey 1995; Rosenblum et al. 2012). Nowadays, most phylogenies are depicted in the rectangular shape where the node representing the ancestral lineage is placed inbetween two descending one (e.g., Figure 2), implying the bifurcating mode of speciation.

Besides, due to undersampling and/or extinct relatives, phylogenetic age can both under and overestimate true age (Figure [2](#fig:phylo_raccoons)). Thus, sampling biases and speciation modes are likely to affect the estimation of extant species’ ages from calibrated phylogenies.

While the effect of incomplete taxon sampling is mitigated by the ever growing phylogenies and advances in augmenting trees by missing species (e.g., Thomas et al., 2013, Change et al., 2020), there is little hope of tracking down most of the extinct species and revealing how species emerged. This paper focuses on the latter two. Given the theoretical predictions on the consequences of ignoring extinctions and not knowing the mode of speciation on the reliability of taking branch lengths as surrogate for species ages, it aims to quantify how much an error is actually introduced by doing so and whether it affects our conclusions drawn by comparing species ages. Specifically, we performed simulations where we know the true age of species to: (1) quantify the introduced error in estimated ages under various scenarios of different speciation modes and diversification rates, (2), examine whether this error affects our ability to make qualitative decisions on the order of species ages, and (3) explore whether the signal of a simulated age-related conservation status is preserved in the phylogenetic age of species. We note that we quantified relative error ins species age to be comparable among phylogenies and we therefore also use turnover instead of extinction as this measure expresses the relative species replacement over time. Finally, we propose a way to efficiently quantify the uncertainty in species age for the most error-prone mode of speciation.

This paper aims to demonstrate the multiple sources of error associated with phylogenetic age. First, we used simulations to show how error measures vary regarding speciation modes and diversification rates. Next, we proposed a geometric function to estimate a distribution of probable ages based on budding speciation for each species within a phylogeny to partially account for the error sources. Finally, we performed a set of simulations to associate different levels of extinction signals to species’ true age and assessed whether the phylogenetic age and the ages estimated by our function capture these patterns. We hope this paper will stimulate discussions about the evolutionary information contained by phylogenetic trees and the pertinence of using them for infering phylogenetic ages.

# Methods

The following two methodological sections regarding error measures follow the same workflow:

* Phylogenetic trees simulation through the "sim.bd.taxa" function from the *TreeSim* package (Stadler 2011), controlling the number of extant species, speciation, and extinction rates.
* Taxonomy simulation through "sim.taxonomy" function from the *FossilSim* package (Barido-Sottani et al. 2019), using as input the trees generated previously and controlling the speciation modes.
* The "sim.taxonomy" function allows us to save the true age of species. We pruned the extinct species from the phylogenies using the "prune.fossil.tips" function from the *FossilSim* package (Barido-Sottani et al. 2019). Then, we estimated the Phylogenetic ages from the branch lengths of the pruned trees. Both, phylogenetic and true ages, were scaled to the root age of the corresponding tree.

*Simulating species ages*

For all simulation experiments, we generated complete phylogenies of extant and extinct species under the stochastic birth-death process (i.e. without cladogenetic extinction) using the package TreeSim 2.4 (Stadler 2011) for the R 4.3.0 statistical programing environment (R Core Team, 2023). We always simulated 100 extant species, while the number of extinct species was stochastically controlled by a given turnover. We mapped species to the complete phylogenies using the R package FossilSim 2.3.1 (Barido-Sottani et al. 2019) according to the three different speciation modes, providing the true species ages. We then pruned all extinct species from the phylogeny and obtained the length of terminal branches, which we took as the phylogenetic age of the extant species. We rescaled all phylogenetic trees to an root age of 1, obtained relative ages, and where therefore able to compare results among different simulation scenarios.

*Error in equating phylogenetic and species age*

To explore whether there is a consistent over or underestimation of species ages and how large the error is by equating phylogenetic ages with species age, we simulated various scenarios of different speciation modes and diversification rates. First, we simulated 100 phylogenetic trees with birth rates set to 0.1, 0.5, and 1, combined with 100 death rates ranging from 0 to 0.99 in equal increments. We defined turnover as birth/death which is sometimes called extinction fraction (Beaulieu and O’Meara 2016). Second, on each of these phylogenies, we then mapped species on these 300 phylogenies according to different scenarios of speciation: (1) strictly budding speciation at branch nodes, (2) strictly bifurcating speciation at branch nodes, (3) a combination of strictly budding speciation at branch nodes and anagenetic origination of species with a rate set to half the birth rate that was used to generate the phylogeny, and (4) bifurcating speciation combined with anagenetic speciation equal to half the birth rate.

We obtained in total 120,000 extant species, 30,000 for each speciation scenario. We calculated the mean absolute percentage error (MAPE) (put formula) as measure of the deviation of the phylogenetic ages from the true age. Next, we plotted the MAPE for different speciation modes against the simulated turnover.

*Impact of age error on comparing species ages*

To explore whether the error introduced by equating phylogenetic age with species age impacts our ability to make qualitative judgements such as which of two extant species is the younger one, we simulated 1,000 phylogenetic trees with values of extinction rates of 0.9, 0.5, and 0, combined with a fixed speciation rate of 1 (3000 trees). Second, on each of these phylogenies, we mapped species according to budding and bifurcating speciation. Thus, we simulated 300,000 extant species for each speciation mode. Next, we calculated the proportion of cases where the younger of two species, according to its phylogenetic age, is, in fact, the older one given the true age of the two species. We performed this comparison from the perspective of an empirical research that can only obtain the phylogenetic ages. Two types of comparisons were made for each phylogeny: (1) between the youngest and oldest species in the phylogeny, and (2) two randomly selected species.

*Quantifying uncertainty in species age under budding speciation*

For the scenario of budding speciation, which is the assumption of the birth-death process, we implemented an R function to quantify the uncertainty in species ages from their phylogenetic ages and the tree topology. As the probability of a branch to be the ancestral species belonging to a new or the ancestral species are the only two possible outcomes for the two branches descending from a budding speciation event, and the probability of both is always 50%. Therefore, the probability for each ancestral node to be the origin of a species is given by the geometric distribution:

(1 – p)(k – 1)p,

with p = 0.5 and k equal to the corresponding number of ancestral nodes. We then used these probabilities to calculate a weighted mean and 95% confidence interval of the node heights, which is the time until the present for extant species or extinction for non-living species, respectively.

When using this approach to quantify the uncertainty in species ages from a phylogeny of only extant species, it will be result in an overestimation because the number of ancestral nodes of a species is biased due to the missing extinct species.

To evaluated the consequences of this bias, we simulated 100 phylogenetic trees with 0.25 (high), 0.15 (intermediate), and 0.05 (low) extinction rates, combined with a fixed speciation rate of 0.2 (i.e. 300 trees). We evaluated the accuracy of the geometric function by quantifying the function’s coverage or the proportion of true ages inside the 95% confidence interval. Additionally, we compared the the true age with the modal and the mean age, respectively obtained by the function.

*Extinction signal*

To evaluate whether the often-assumed relationship between species age and extinction risk (e.g. 2-3 REFs) is preserved the in phylogenetic ages, we simulated 1,000 phylogenetic trees with extinction rates of 0.25, 0.15, and 0.05, and a fixed speciation rate of 0.3. Then, using the phylogenies as inputs, we mapped species on each tree considering only budding speciation. Next, we binned the true ages in five categories corresponding to the IUCN conservation categories: Least Concern (LC), Near Threatened (NT), Vulnerable (VU), Endangered (EN), and Critically Endangered (CR; Internation Union for the Conservation of Nature et al. (2001). With this, we generated two scenarios: (1) a positive effect with older species being at higher extinction risk), and (2) a null effect where species age and extinction risk are uncorrelated. Then, we plotted for each phylogeny a line between the five conservation categories and the mean true and phylogenetic ages, respectively. Finally, we quantified how many trees, for the phylogenetic and geometric ages, were estimated right regarding the relationship between true age and conservation status categories.

# Results

*Quantifying the error in equating phylogenetic and species age*

True age vs Phylogenetic age

As expected for bifurcating speciation (Figure [3](#fig:true_vs_phylo)), there was no underestimation of the true age. At low turnover (<0.25), 96% of the phylogenetic age estimations are congruent with the true age. At high turnover (>0.75), this was also the case for most species (73%). However, overestimation increased with turnover, and in some cases the phylogenetic age suggests that the species is as old as the root age. For budding speciation, there were both under and over-estimations (Figure [3](#fig:true_vs_phylo)). At low turnover, most ages were underestimated (96%) and only few had correct (0.5%). With higher turnover, phylogenetic ages were overestimating the true age more often (17%), and in few cases phylogenetically old species were indeed young.

As in the strictly bifurcating speciation scenario, there is also no underestimation for a mixture of anagenetic andbifurcating speciation (Figure SM1),. In a low turnover scenario, fewer phylogenetic ages equaled the true ages (68%) than in strictly bifurcating speciation. The phylogenetic ages varied more their degree of overestimation than for strictly bifurcating speciation, almost approaching the variability under high turnover. For the anagenetic-budding scenario, there was no clear difference between high or low turnover, as it was the case in the budding speciation scenario. Thus, anagenetic speciation, in addition to bifurcating or budding speciation, does not systematically affected the error in species age due to extinction in comparison to the latter two modes of speciation in isolation. This is not a due to a low rate of anagenesis because we specified a stochastically equal number of anagenetic speciation events as there were bifurcating and budding events, respectively.

Mean Absolute Percentage error vs. Turnover

For the two speciation modes evaluated (budding and bifurcating) , there was no error variation regarding the relative speciation rate (Figure [4](#fig:mape_bud_bif)). That implies that MAPE increases only due to the relative extinction rates of each phylogenetic tree. For strictly bifurcating speciation, there was no error when there is no extinction but it increased exponentially with the increase of the relative extinction rate, starting from. In contrast, under budding speciation, the baseline error, which is the error when there is no extinction, was already around 25%, from where on it increased linearly with the relative extinction. For the combination of these speciation modes with anagenetic speciations, there was no effect regarding relative extinction or speciation rates (Figure SM2). However, the MAPE error reached as high as 500% in some replicates of our simulation.

*Impact of age error on comparing species ages*

Distribution of True and Phylogenetic ages

For bifurcating speciation (Figure [5](#fig:distributions_bif)), there is no great shape difference between the true and phylogenetic age distributions. However, for the phylogenetic age distribution in both High and Intermediate extinction scenarios, there are fewer young and more old species than for the true age distributions. In the no-extinction scenario, both distributions are identical. For budding speciation (Figure [6](#fig:distributions_bud)), the phylogenetic age distributions show more young species accumulation in all the extinction scenarios than the true age distributions.