**On the estimation of species age from phylogenetic trees**

Carlos Calderón del Cid1,2, Torsten Hauffe2, Juan D. Carrillo2, Daniele Silvestro2,3

1Laboratório de Ecologia Espacial, Instituto de Biologia, Universidade Federal da Bahia, CEP 40170-110, Salvador, Bahia, Brasil

2Department of Biology, University of Fribourg, Switzerland and Swiss Institute of Bioinformatics, Fribourg, Switzerland

3Department of Biological and Environmental Sciences and Gothenburg Global Biodiversity Centre, University of Gothenburg, Sweden

**Abstract**

Estimating species age (the elapsed time since origination) can give an insight into how species longevity might influence eco-evolutionary dynamics. 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) cause different topological representations of phylogeny and, thus, affect the branch lengths and the estimation of species age. The information about extinctions (provided by fossils) is incomplete for most phylogenies, and the mode of speciation in most cases is unknown. This study aims to evaluate, through simulations, the potential sources of error (extinctions and modes of speciation) in estimating the phylogenetic age. Furthermore, we developed a geometric function to partially correct phylogenetic ages when considering budding speciation. We found that extinction rate and speciation modes greatly influence the estimation of phylogenetic age. Bifurcating speciation presented a low mismatch between true and phylogenetic age, which increases with extinction rates. Budding speciation showed a high mismatch, which increases slightly with the 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 was not accurate but presented a high confidence interval coverage. Altogether our results indicate that we should be cautious when inferences are made from the relationship between phylogenetic ages and eco-evolutionary variables, and, when possible, we should perform a critical evaluation of speciation modes, through spatial analysis and extinction events, through fossil calibrations.

**Key words:** Evolutionary history, Extinction rates, Simulations, Speciation modes.

**Introduction**

The estimation of species age, or the time elapse since species origin, is fundamental to evaluate mechanisms that link species longevity with eco-evolutionary processes (Benton 2013; Swenson 2019). As an example, Age-dependent extinction hypotheses test the relationship between age and extinction probability, arguing that species longevity is correlated with the time needed to evolve certain traits conferring fitness or extinction selectivity (Balmford 1996; Eldredge et al. 2005; Pearson 1995). Likewise, species age could be a measure of colonization time, especially in island systems or during biotic invasions triggered by geological events, such as the formation of the Central American Isthmus for the Great American Interchange (Carrillo et al. 2015; Tanentzap et al. 2015). Traditionally, species age is measured in the fossil record through different statistical and probabilistic approaches based mostly on taxa’s stratigraphic duration (Foote 1996; Foote and Raup 1996). Several of these approaches consider the differences in fossil sampling and temporal resolution (Alroy et al. 2001; 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 (Benton 2016; Silvestro et al. 2020; Van Valen 1973). More recently, some studies have assumed that the age of extant species can be measured from the branch lengths of time-calibrated phylogenies (Gaston and Blackburn 1997; Christopher N. Johnson et al. 2002; Tanentzap et al. 2020; Verde Arregoitia et al. 2013). 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 reduces the accuracy of the tree topology and increases the uncertainty in any phylogenetic estimation (Heath et al. 2008; Mynard et al. 2023). Second, due to incomplete fossil sampling, extinct lineages are not considered to calibrate phylogenies of extant species. 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 the *Homo* genus are not included in the phylogeny, the phylogenetic age of *Homo sapiens* is approximately 6 million years (i.e., the age of our last common ancestor with the chimpanzee), much older than the oldest fossil record for our species of 300 thousand years (Figure 1; Callaway 2017; Mann and Weiss 1996).

The third shortfall of extant species phylogenies is that the tree alone does not contain information about the speciation mode (Foote 1996; Silvestro et al. 2018). Similarly, most phylogenies are depicted in the rectangular shape where the node representing the ancestral lineage is placed in between two descending ones, suggesting a bifurcating speciation mode where two new species replace the ancestral lineage (Baum et al. 2005; Caetano and Quental 2022). Three different speciation modes are discussed in the paleobiological literature and each reflect a particular biological process and different species concepts. Likewise, because they define the relationship between the ancestral and the descendant species, these speciation modes partly determine species longevities (Rosenblum et al. 2012; Wagner et al. 1995).

As described before, bifurcating speciation is a diversification event that gives rise to two species while the ancestral lineage perishes; it is associated with vicariance or allopatric speciation (Willmann and Meier 2000). The Hennigian species concept tends to consider all speciation events as bifurcating (Henning 1999). Under the assumption of bifurcating speciation and with no extinction events, phylogenetic ages equal the true age of extant species. However, with extinction events, extinct sister species would cause the branch lengths of the surviving sister to appear larger than they are (Figure 2), resulting in an overestimation of species age.

Budding speciation is a diversification event that gives rise to one species while the ancestral lineage persists; it is associated with peripatric speciation (Anacker and Strauss 2014; Caetano and Quental 2022). Budding speciation is related to the evolutionary species concept, which considers any species as descendant of an ancestral lineage and that possesses own evolutionary identity regarding other lineages (Simpson 1951). In the case of budding speciation, the phylogenetic age for the ancestral species would be an underestimation of its true age because a phylogeny alone does not reveal which is the ancestral lineage of two sister species. Also, due to incomplete fossil or extant species sampling, phylogenetic age can overestimate true age, given that recent species, if the ancestral species is missing, could seem older than they genuinely are.

Anagenetic speciation is a diversification event where one species arises, replacing the ancestral species; it is not related with any particular geographical mode of speciation (Roopnarine et al. 1999). This process suggests phenotypic change and not speciation per se; thus, it is associated with the morphospecies concept (Emerson and Patiño 2018). For anagenetic speciation, the phylogenetic age always overestimates the true age, given that the ancestral and new species branches are fused in the phylogeny.

While the effect of incomplete taxon sampling is mitigated by the ever-growing phylogenies and advances in augmenting trees by missing species (e.g., Chang et al. 2020; Thomas et al. 2013), there are significant issues regarding tracking down most of the extinct species and revealing the predominant speciation mode in phylogenies. This paper focuses on the latter two. Given the theoretical predictions on the consequences of ignoring extinctions and not knowing the speciation mode, it aims to quantify how much an error is introduced by assuming branch lengths as a surrogate for species ages and whether it affects our inferences when linking species longevity with eco-evolutionary processes. 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 (speciation – extinction) 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. In addition, we note that we quantified relative error in species ages to be comparable among phylogenies; thus, we used 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. We hope this paper will stimulate discussions about the evolutionary information in phylogenetic trees and the pertinence of using them to infer phylogenetic ages.

**Methods**

*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 2010) for the R 4.3.0 statistical programing environment (R Core Team 2023). For all trees, we 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. Finally, we rescaled all phylogenetic trees to a root age of 1, obtained relative ages, and compared 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 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. For each speciation mode, we plotted the True scaled age against the Phylogenetic scaled age, using a discrete color gradient for low and high turnover rates. We calculated the mean absolute percentage error (MAPE) as measure of the deviation of the phylogenetic ages from the true age.

MAPE =

Next, for each speciation mode, we plotted the MAPE 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.

*Imprint of extinction risk on phylogenetic ages*

To evaluate whether the assumed relationship between species age and extinction risk (e.g., Johnson et al. 2002; Tanentzap et al. 2020; Verde Arregoitia et al. 2013) is preserved in the 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 2016). 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 linking the five conservation categories and the mean true and mean phylogenetic ages, respectively. Finally, we quantified how many trees, for the phylogenetic and geometric ages, were estimated right regarding the relationship between the mean true age and the conservation status categories.

*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. The two branches descending from a budding speciation event are always the new and the ancestral species and the probability of both outcomes 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.

For quantifying the uncertainty in species ages from a phylogeny of only extant species, this approach will present phylogenetic age 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 true age with the modal and the mean age, respectively obtained by the function.

**Results**

*Error in equating phylogenetic and species age*

As expected for bifurcating speciation (Figure 3), 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 overestimations (Figure 3). At low turnover, most ages were underestimated (96%). With higher turnover, phylogenetic ages overestimated true age more often (17%), and in few cases phylogenetically old species were indeed young.

As in the strictly bifurcating speciation scenario, there was also no underestimation for the mixture of anagenetic and bifurcating speciation (Figure SM1). In a low turnover scenario, fewer phylogenetic ages equaled the true ages (68%) than in a 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.

For the MAPE, the two speciation modes evaluated (budding and bifurcating), there was no error variation regarding the relative speciation rate (Figure 4; given that the points follow the same trend regardless of their color which represents the speciation rates). That implies that MAPE increased only due to the relative extinction rates of the phylogenetic tree. For strictly bifurcating speciation, there was no error when there is no extinction but it increased exponentially with the augment of the relative extinction rate. In contrast, under budding speciation, the baseline error, which is the error when there is no extinction, was already around 25% (Figure 4), from where it increased linearly with the relative extinction rate. For the combination of these speciation modes with anagenetic speciation, the MAPE error reached as high as 500% in some replicates of our simulation (Figure SM2).

*Impact of age error on comparing species ages*

For the combination of strictly bifurcating speciation and all extinction scenarios, selecting the phylogenetically youngest and oldest species never resulted in a case where the presumed older species has been in fact the younger of the two species according to their simulated age (Figure 5). Thus, for this speciation mode, there is no risk of committing a qualitative error when comparing species at the extremes of the age range of a time-calibrated phylogeny. In contrast, for budding speciation, even under no extinction, in 2.2% of all cases the phylogenetically oldest species had a true age younger than the phylogenetically youngest species. The rate of committing a qualitative error reached 7.5% and 12.2% for intermediate and high extinction, respectively.

When we compare the ages of two random species instead of the phylogenetically youngest and oldest species, the error rate for confusing the older with the younger one generally increased. While there is no such risk under bifurcating speciation with no extinction, the error rate increased 6% and 8%, respectively, for intermediate and high extinction (Figure 6). Under budding speciation, irrespectively of the extinction level, the younger species was confused with the older in more than 25% of all cases (Figure 6).

*Imprint of extinction risk on phylogenetic ages*

The mean phylogenetic age showed a completely correct estimation (i.e., 5 of 5 conservation status reflected by the mean phylogenetic age of the species in each category) in 75% of the trees in a low, 69% in an intermediate, and 56% in a high extinction scenario (Figure 7)

When acknowledging partial correct estimation (3 or more conservation status estimated corrected), the mean phylogenetic age showed the same pattern with 97% in a low, 94% in an intermediate, and 91% in a high extinction scenario. Moreover, the mean probable age slightly improved with 99% in a low, 97% in an intermediate, and 87% in a high extinction scenario.

*Quantifying uncertainty in species age under budding speciation*

The modal age from the geometric function has almost the same accuracy as the phylogenetic age for estimating true age in a budding speciation scenario (Figure SM5). Moreover, it also shows the tendency of a slightly lower accuracy with increasing extinction rates. The mean probability age from the function is inaccurate regarding any true age but possesses a symmetrical error distribution in all extinction scenarios (Figure SM6).

The coverage of true ages by the 95% confidence interval obtained with our geometric age function is considerably high, with 94% in a low, 90% for an intermediate, and 76% for a high extinction scenario (Figure 8). However, the width of the confidence intervals is large, ranging often from the root of the phylogeny to the present.

**Discussion**

Our study shows that extinction rates and speciation modes greatly influence the mismatch between true and phylogenetic age. Moreover, the error patterns are different regarding the combination of both processes. As we expected, a phylogeny whose taxonomy was derived from a bifurcating speciation process presents no or low mismatch between true and phylogenetic age due to the bifurcating node structure of phylogenies and the way phylogenetic age is measured (Meier and Willmann 2000). Our study has also shown that under budding speciation phylogenetic age shows a high error even with no extinction rates (Figure 4). Likewise, anagenetic speciation, combined with budding or bifurcating speciation, resulted in high errors. The anagenetic error is not affected by extinction rates, suggesting that if a clade experiences a certain degree of anagenesis, it should be discouraged to estimate ages from phylogenies. The geometric function did not improve the accuracy regarding the true age estimation (Figure SM5). Its large confidence intervals and the high coverage, suggest that there is little signal of the species age in a phylogeny under budding speciation. Since the coverage is also relatively high with low extinction, the main source of uncertainty is indeed the budding speciation mode. Thus, our study points out that budding and anagenetic speciation modes are the principal source of inaccuracy regarding estimating species ages from time-calibrated phylogenies branch lengths.

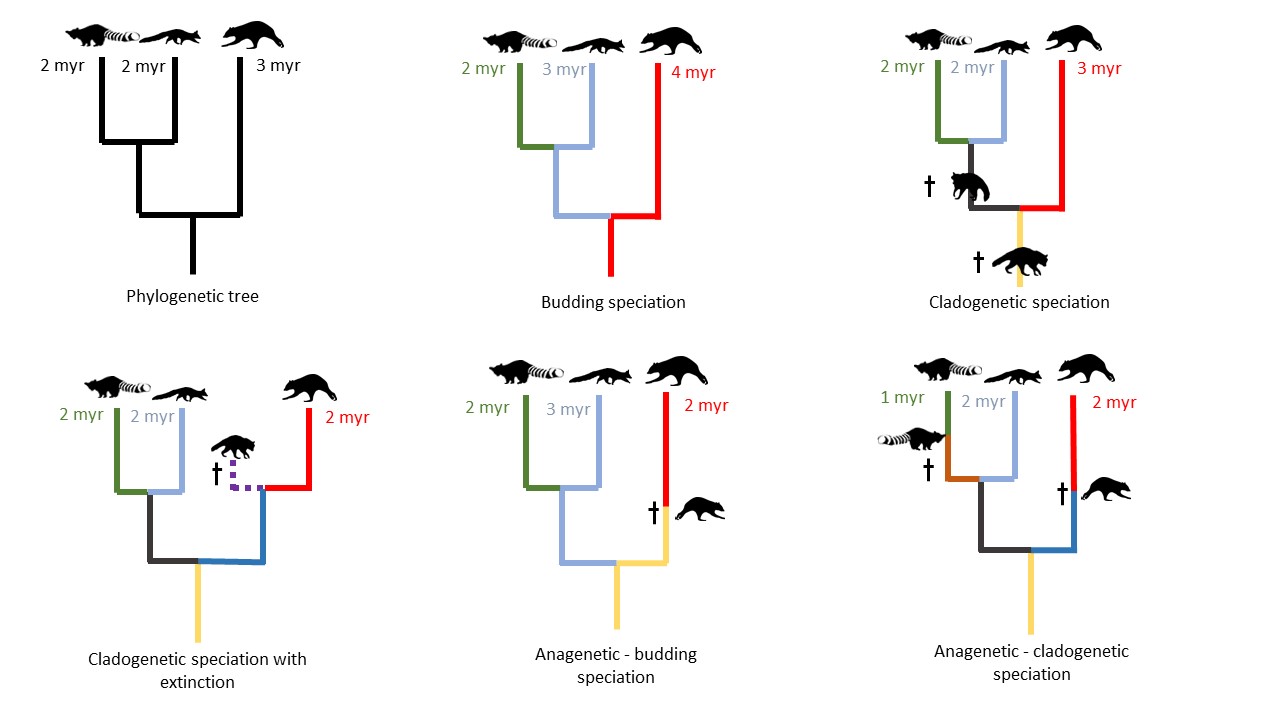
For bifurcating speciation, the mean error between true and phylogenetic age was low, even in scenarios of high extinction. This is because taking the most recent ancestral node as phylogenetic age results in sister species to be identical in age, the same as implied by bifurcating speciation. Due to only introducing bias by extinction, there was a low risk of confusing the oldest with the youngest species. Therefore, under a bifurcating speciation scenario, the phylogenetic age indeed captures the signal of the true species age. For budding speciation, there is a high mismatch between ages, even in no-extinction scenarios; however, the overall error does not increase considerably with increasing extinction rates. Half of the phylogenetic age estimations in no-extinction scenarios always underestimate the true age, while when extinction rates increase, overestimation appears and rises until balancing with underestimation at high extinction scenarios. This pattern happens due to the phylogenetic estimation assumption of considering sister species as identical in age, which, in no extinction scenario, only correctly estimates the youngest species of any bifurcation. Moreover, when extinction rates increase, the youngest species of a clade evolving through bifurcation speciation can be wrongly taken as phylogenetically old (i.e., having a long terminal branch) if all ancestral species to that branch are extinct. Due to these misinterpretation patterns, the qualitative error is high and increases considerably with extinction rates in the example of the oldest vs. youngest species (Figure 5). Thus, for clades diversifying predominantly through budding speciation, the phylogenetic age is not equal to the true age for half of the species, and some of these mismatches would be time substantial regarding the true age.

Given the large inaccuracy of phylogenetic age, especially for phylogenies diverging under budding speciation, the question is whether this affects the inferences made from the relationship between species longevities and eco-evolutionary variables, such as extinction risk, range size, or clade’s diversity (Gaston and Blackburn 1997; Johnson et al. 2002; Tanentzap et al. 2015). If phylogenetic ages capture the true age signal, it could be argued that phylogenetic age is a valuable measure of species longevity. Some authors acknowledge the problems associated with measuring species age from phylogenies, and have proposed approaches to account for them by combining phylogenetic, biogeographic, and ecological information (Swenson 2019). For example, Tanentzap et al. (2020) used the stem age, given that it requires only one species to be estimated within each clade to reflect its evolutionary history (Scholl and Wiens 2016), to explore the relationship between taxon age, range size and extinction risk in plants. Additionally, Sonne et al. (2022) determined young and old Andean hummingbirds by generating 1000 trees and listing the species that fell into the first and fourth branch length quartile, respectively, and then used a sensitivity analysis to evaluate the consequences of different species composition in their results (Fjeldså et al. 2012). Nevertheless, neither strategy acknowledges the real problems of age uncertainty: the unknown extinction rates, which for most groups is probably high (Pimm et al. 2014), and speciation modes of phylogenies. Our evaluation of the imprint of extinction risk on phylogenetic ages makes this point clear because it showed that, even with the strong signal we assigned on purpose to the true ages, the number of incorrect inferences is considerable high among extinction scenarios (Figure 7).

Our results point out the errors associated with budding and bifurcating speciation; however, how do we know which is the predominant speciation mode of a clade to contemplate a likely percentage error in the phylogenetic age? Anacker and Strauss (2014) proposed, for example, that budding speciation leaves some signatures in sister species: they should have overlapping or adjacent ranges, their range sizes should be asymmetrical, and specific ecological traits should differ between them. These signatures are associated with sympatric and peripatric speciation (Barraclough, Vogler, and Harvey 1998). In this sense, bifurcating speciation can be associated with allopatric speciation and the signatures it leaves in the range of sister species: ranges should not be overlapping or adjacent, range sizes should not necessarily be asymmetrical, and ecological traits should not necessarily differ between them (Barraclough and Vogler 2000; Fitzpatrick and Turelli 2006). Thus, assessing a clade’s spatial dynamics before performing species age analyses could give us insights into the speciation modes and an approximate error of phylogenetic ages. In the case of the errors associated with different extinction scenarios, combining phylogenetic information with fossil data is recommended, although the fossil record is scarce in some groups (Rabosky 2010; Silvestro, Salamin, and Schnitzler 2014; Brée, Condamine, and Guinot 2022; López-Martı́nez et al. 2023).

Using simulations, we showed the problems in estimating species age from phylogenetic trees and quantify the error in estimations under different scenarios. The challenges of using phylogenetic age as a proxy for species longevity are not merely methodological. Time-calibrated phylogenetic trees of extant species do not contain information on the error sources, namely speciation modes and extinction rates. Thus, the phylogenetic age provides some information about true age, but it is a weak proxy to measure evolutionary history. Assessing the speciation modes with the methods suggested here or others and estimate extinction rates by combining paleontological and neontological data, might help to estimate the ages percentage error. Approaches combining phylogenies, biogeography (to inform speciation modes), diversification (extinction rates), and traits might provided a more comprehensive understanding on the relationship of species longetivity and eco-evolutionary processes.

**Figure 1**



**Figure 2**