The impact of tip age distribution on reconstructing trait evolution using phylogenetic comparative methods

William Gearty[1,*], Bethany J. Allen[2,3], Pedro L. Godoy[4], Alfio Alessandro Chiarenza[5]

1 Division of Paleontology, American Museum of Natural History, New York, NY USA
2 Department of Biosystems Science and Engineering, ETH Zürich, Basel Switzerland
3 Computational Evolution Group, Swiss Institute of Bioinformatics, Lausanne, Switzerland
4 Department of Zoology, Institute of Bioscience, University of São Paulo, São Paulo Brazil
5 Department of Earth Sciences, University College London, London UK

*Corresponding email: willgearty@gmail.com

Abstract

Collecting data for use in constructing phylogenies is a valuable but time- and resource-consuming pursuit. As a result, indicators of the potential value of including certain species in a phylogeny *a priori* could prove useful when planning this stage of research. Here, we used a simulation approach to investigate whether there are trends in the ability for phylogenetic comparative methods to recover the correct model of trait evolution based on certain characteristics of the phylogeny. First, we used multiple diversification rates to simulate phylogenies containing varying proportions of fossil and extant tips. We then simulated the evolution of a single trait across each phylogeny using multiple continuous trait evolution models. We then compared the fit of the correct and incorrect models to the simulated traits. This quantitative evaluation allows us to discern whether there are certain tip characteristics associated with identifying the correct trait evolution models. Our results indicate that the inclusion of fossils can be highly beneficial to reconstructing certain trait histories (e.g., Ornstein-Uhlenbeck and ACDC) but not to others (e.g., Brownian motion). In fact, in many cases, increasing the proportion of fossils in a phylogenetic dataset is far more beneficial, and perhaps more time- and resource-efficient, than increasing the number of extant taxa in the dataset. Our results corroborate previous findings that the inclusion of fossil tips can vastly improve the reconstruction of trait histories, but also show that this effect is often stronger for older fossils.

Introduction

Phylogenies describe the hypothesized or inferred evolutionary relationships between biological entities. Phylogenetic comparative methods (PCMs) use phylogenies to investigate the intricate patterns and processes of evolution along the branches of the evolutionary tree (Felsenstein 1985; Cornwell and Nakagawa 2017). Such analyses include estimating ancestral states of discrete or continuous characters (including historical biogeography), inferring the tempo and/or mode of evolution of one or more traits, and identifying diversification dynamics (e.g., Allen et al., 2019; Butler & King, 2004; Gearty et al., 2021; Godoy et al., 2019; Harmon, 2019; Rabosky et al., 2018; Slater, 2013; Stadler, 2011). However, a longstanding conundrum regarding the application of computational methods in evolutionary biology is how to ensure that the results of downstream analyses are valid, particularly as experimental data against which such methods can be verified often does not exist (Barido-Sottani et al. 2020). This is particularly true of PCMs, with many of the most-commonly used methods and models being criticized for their potential inaccuracy (e.g., Boettiger et al., 2012; Cooper et al., 2016; Louca & Pennell, 2020; Pennell et al., 2015).

A major assumption underlying PCMs, which likely further contributes to the inaccuracy of the results they produce, is that the phylogeny being used is a fair representation of the underlying genealogical population, i.e. the clade of interest (Symonds 2002). If the phylogeny is not representative of the clade's evolutionary history (due to, for example, bias in the branches sampled, incorrect relationships, or invalid branch lengths), it may result in incorrect inferences, which may then lead to false evolutionary conclusions. However, as a phylogeny approaches the true history, the inferences of PCMs should likewise approach the true evolutionary dynamics. Choices concerning the data from which the phylogeny is inferred are fundamental to how likely the phylogeny is to truly represent evolutionary history. Collecting this data is a labor-intensive and resource-demanding (e.g., time, choices, money, computation) endeavor. Therefore, it is imperative to find ways to determine which of the decisions made when collecting data for constructing a phylogeny have the largest impact on PCM accuracy, so that researchers can make informed choices based on a cost-benefit approach (Mongiardino Koch and Parry 2020; Mongiardino Koch et al. 2021).

Recent computational developments have allowed for the construction of phylogenies to be conducted via the integration of diverse sources of information, such as genetic sequences, phenotypic characters, fossil ages, and ecological data (e.g., Ronquist et al., 2012; Wright et al., 2022; Zhang et al., 2016). The types of data to include is therefore one of the major choices when planning data collection for phylogenetic inference. One potential choice in the construction of phylogenies for PCMs is whether extinct taxa are included, and, if so, which ones are included. Paleontology has a long history of producing phylogenies based on extinct taxa using morphological character matrices (Vrba 1979; Gauthier 1986). However, the combination of morphological and molecular data is a much more recent phenomenon (Springer et al. 2001; Schrago et al. 2013; Zou and Zhang 2016). Since the inception of these combined data approaches (also known as "total-evidence"), paleontologists have advocated for the inclusion of fossils as tips in phylogenies (Wright et al. 2022).

Previous papers have investigated how well PCMs work when using phylogenies that do or do not include fossils. Several studies have shown that ancestral states estimated based on extant-only data can be extremely biased (Webster and Purvis 2002; Finarelli and Flynn 2006; Royer-Carenzi et al. 2013). Slater et al. (2012) demonstrated that including fossils as tips in phylogenies increases the power of PCMs to detect the true model of trait evolution. This point has been echoed in case studies examining the bird, caniform and monkey fossil records (Finarelli and Goswami 2013; Mitchell 2015; Silvestro et al. 2015). Further, several studies have shown that diversity dynamics cannot be accurately estimated with extant-only phylogenies (Quental & Marshall 2010; Rabosky 2010; Louca & Pennell 2020), although the benefit of including fossils remains unclear (Louca et al. 2021; Cerny et al. 2021; Beaulieu & O'Meara 2023). However, despite these observations and cautions, extant-only trees remain ubiquitous in studies conducting phylogenetic comparative methods (e.g., Álvarez-Carretero et al., 2022; Jetz et al., 2012; Jetz & Pyron, 2018; Magallón et al., 2015; Pie et al., 2023; Rabosky et al., 2018; Smyčka et al., 2023; Upham et al., 2019).

Given the complexity of these issues and piecemeal nature of model adequacy investigation and reporting (Pennell et al. 2015), a comprehensive quantitative assessment of the impact of the inclusion of fossils in phylogenies that are used for PCMs is needed. A central issue, as highlighted by Slater et al. (2012) and Cooper et al. (2016), and further underscored by Grabowski et al. (2023), pertains to the challenge of model traceability, especially in the context of complex (multiparametrised) models such as Ornstein-Uhlenbeck (OU) (Hansen 1997). This challenge primarily emerges when dealing with incomplete and poorly sampled datasets, significantly affecting the ability of the model to accurately track the evolutionary trajectories of specific traits within a phylogenetic framework. This evaluation will not only clarify the suitability of complex but well-defined modes of evolution like Ornstein-Uhlenbeck or Early Burst/ACDC (Blomberg et al. 2003) models in comparative analyses, but also enhance our understanding of the broader challenges related to phylogenetic comparative methods in the face of incomplete datasets. Such constraints are imperative to ensure that researchers can make informed decisions when applying these methods and understand the implications of these decisions on the potential accuracy of their results.

Here, we determine how the inclusion of varying proportions of fossil tips, and their relative ages, affect the ability of PCMs to accurately recover specific and widely-used models of continuous trait evolution. To achieve this, we implemented a multi-faceted strategy: we first simulated phylogenies of varying sizes, which contained varying proportions of old-biased, unbiased, or young-biased fossil (non-extant) tips. We then simulated the evolution of a single continuous trait across each phylogeny using multiple trait evolution models, before comparing the fit of correct and incorrect models to these simulated traits. Our results corroborate previous findings that the inclusion of fossil tips can vastly improve the reconstruction of trait histories, but also show that this effect is stronger for older fossils. The inclusion of fossil tips is also more beneficial for traits which evolved under multi-regime Ornstein-Uhlenbeck or ACDC models, but less impactful for other models, like Brownian motion. Further, it is impossible to correctly infer a shift in the optimum trait value over time in phylogenies only containing extant tips. In some instances, increasing the proportion of extinct taxa within a phylogenetic dataset emerged as more efficient than increasing the overall number of tips by multiple orders of magnitude. These findings are a major step towards developing precise, quantitative guidelines for planning the construction of phylogenetic datasets for PCMs that incorporate both modern and fossil taxa. Further, they quantify how much caution should be taken when interpreting the results of PCMs.

Methods

Simulating Phylogenies

We developed a new workflow to simulate phylogenies that have both a fixed size and a set proportion of tips which are extinct. First, we generated a birth-death tree with a set number of extant taxa and specific birth (λ) and death (μ) rates using the "sim.bd.taxa" function from the *TreeSim* R package (Stadler 2019). We used the `complete = TRUE` option to obtain the full birth-death tree including extinct tips. Then, we randomly sampled the desired number of tips from among these extinct tips. In this sampling process, the odds of sampling a given extinct lineage were scaled with its branch length. By default, the workflow uses a linear scaling (i.e., a branch that is twice as long as another branch has twice the odds of being sampled), but it can also be modified to incorporate a temporal bias by artificially scaling the branch lengths of the tree using the "rescale" function from the *geiger* R package (Harmon et al. 2008). For example, the "early-burst" model (Harmon et al. 2010) can be used to bias the sampling of fossil tips towards the root of the tree (with $a < 0$) or towards the present (with $a > 0$). To prevent computational errors, each branch always has at least a very small chance of being sampled, regardless of the specified sampling bias. Then, for each sampled extinct tip, we sampled a single age for a simulated "fossil" from a uniform distribution spanning the start and end times of the branch. Finally, we used the *FossilSim* R package (Warnock et al. 2022) to adjust the branch lengths of the sampled extinct tips based on these fossil ages and to drop the unsampled extinct tips, resulting in a phylogeny with the desired proportion of extinct tips. This sampling process skews the birth and death rates away from those used to simulate the phylogeny, but is necessary to create strong variation in the distribution of tips over time.

We used this new workflow to simulate a suite of phylogenies using a range of parameters relevant to the evolutionary dynamics of clades and the construction of morphological matrices and phylogenies (Barido-Sottani et al. 2020). First, we used two different death rates (μ; 0.25 and 0.9) in combination with a static birth rate (λ; 1.0) to simulate evolutionary histories with low (0.25) and high (0.9) turnover (μ / λ), respectively. Sufficiently large phylogenies can be computationally difficult to simulate with extremely high turnover values because often the entire clade will go extinct before a sufficient number of lineages have evolved, so we did not attempt turnover values greater than 0.9. Second, we simulated these phylogenies with increasing total numbers of tips to reflect the different sizes of phylogenies that are commonly constructed in evolutionary biology (50, 100, 200, 500, and 1000). We used a set of early-burst parameters (a) to simulate varying temporal fossil sampling biases (-3, 0, and 3). Here, a = -3 (hereafter referred to as "root-biased") represents sampling where fossils are much more likely to be found closer to the root of the tree, a = 0 (hereafter referred to as "random") represents random sampling, where the occurrence of fossils is linearly related to the number of branches at any given time (which increases towards the present), and a = 3 (hereafter referred to as "recent-biased") represents sampling where fossils are much more likely to be found towards the present (than in the random sampling) (Fig. 1). Finally, we generated these phylogenies with varying proportions of fossil tips to reflect the proportions of fossil taxa that are generally included in morphological matrices and paleobiological phylogenies (0, 0.1, 0.25, 0.5, and 0.95). For each combination of these parameters, we simulated 100 different phylogenies. This resulted in a total of 15,000 simulated phylogenies.

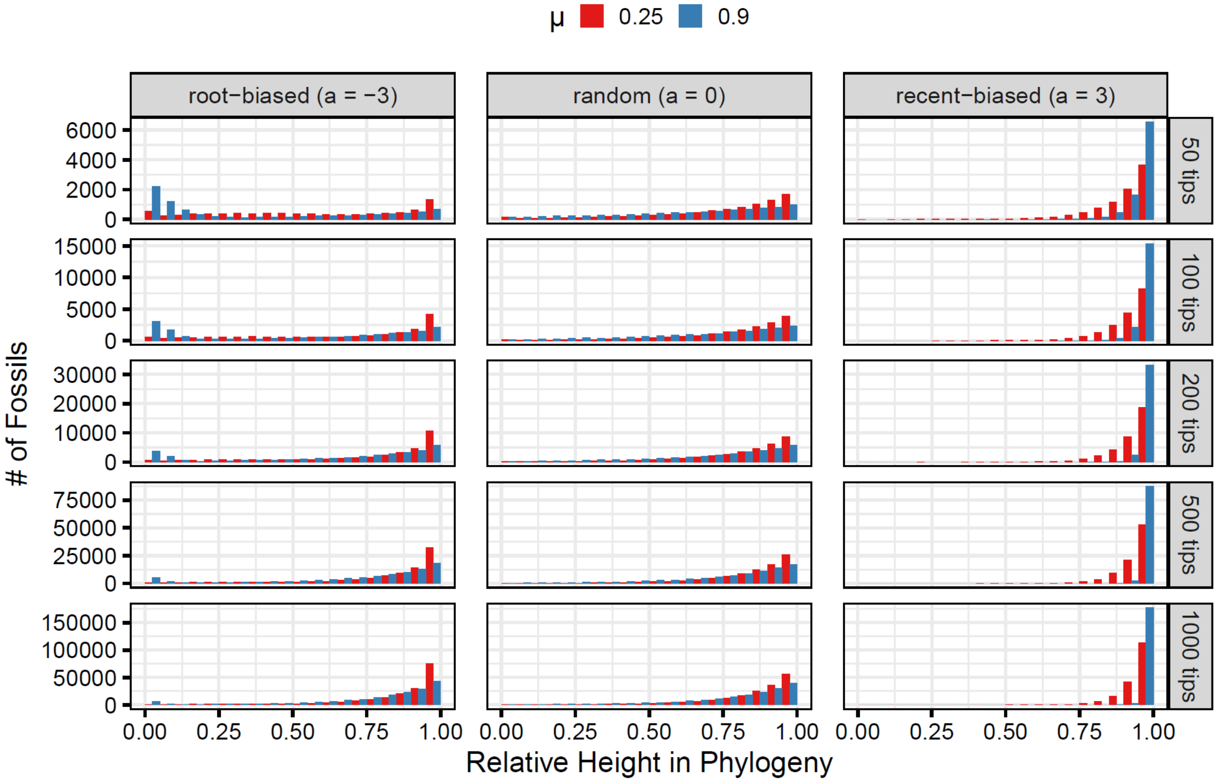


Figure 1: The realized age distributions of the fossils in simulated phylogenies. The results are split out by the temporal bias of the fossils (*a*, column headers), the number of tips in the simulated phylogeny (row headers), and the relative death rate (*μ*, color). Each simulated phylogeny has a different raw height, so the ages of the fossils are presented here as relative heights.

Simulating Traits

For each of our simulated phylogenies, we also simulated the evolution of a single continuous trait under an array of different evolutionary models (Table 1), using the *mvMORPH* R package (Clavel et al. 2015). Brownian motion (see O'Meara et al., 2006), Ornstein-Uhlenbeck (Hansen 1997; Butler and King 2004; Beaulieu et al. 2012), and accelerating and decelerating rate (Blomberg et al. 2003; Harmon et al. 2010) models were all implemented using a “weak” and “strong” strength. In addition to standard BM models, we also implemented BM models with trends (moving average through time). Finally, we implemented OU models with and without a trend in the mean trait value through time. The former type of OU model represents adaptive evolution towards an optimum value which differs from the root state, and the latter represents stabilizing evolution, where evolution is similar to a BM model, but with a spring-like pull towards the optimum/mean (Beaulieu et al. 2012). To account for the varying heights of our simulated trees, instead of using a set alpha (*α*) parameter we calculated *α* individually for each simulated tree based on the desired phylogenetic half-life ($ln(2)/\alpha$) relative to the height of the trees. For the weak OU models, our desired half-life was 100% of the tree height, so we calculated *α* as $\frac{ln(2)}{tree\ height}$. For the strong OU models, our desired half-life was 10% of the tree height, so we calculated *α* as $\frac{ln(2)}{tree\ height\ /\ 10}$.

Table 1. The parameter values used in individual models to simulate trait evolution on each phylogeny. In all models, the ancestral trait state was set to 0.

Model	**Strength**	**Strength of drift (*σ*)**	**Other parameters**	**Abbr.**
Brownian motion (BM)	Weak	0.1		wBM
	Strong	0.5		sBM
BM with trend	Weak	0.1	*trend* = 0.1	wtrend
	Strong	0.1	*trend* = 0.3	strend
Ornstein-Uhlenbeck with identical root and optimum states (centered, OUc)	Weak	0.1	θ = 0; $\alpha = \frac{ln(2)}{tree\ height}$	wOUc
	Strong	0.1	θ = 0; $\alpha = \frac{ln(2)}{tree\ height\ /\ 10}$	sOUc
OU with different root and optimum states (shifting, OUs)	Weak	0.1	θ = 1; $\alpha = \frac{ln(2)}{tree\ height}$	wOUs
	Strong	0.1	θ = 1; $\alpha = \frac{ln(2)}{tree\ height\ /\ 10}$	sOUs
Accelerating (AC)	Weak	0.001	*β* = 0.1	wAC
	Strong	0.001	*β* = 0.3	sAC
Decelerating (DC; "Early Burst")	Weak	0.001	*β* = -0.1	wDC
	Strong	0.001	*β* = -0.3	sDC

Model Fitting

Having simulated the necessary data, we used the mvMORPH R package to fit five evolutionary models to the phylogenies and traits: BM (using the mvBM function), BM with a trend (mvBM with `trend=TRUE`), OU under stabilising selection (mvOU with `root=FALSE`), OU with differing root and optimum states (mvOU with `root=TRUE`), and AC/DC (mvEB with an upper bound of 1). Considering that we had simulated 12 traits across each of 15,000 trees, we ultimately ran 900,000 model fitting analyses in total. In 17 of these analyses, a model was unable to be fit to a particular trait simulation due to computational issues; we ignored these fits for all downstream comparisons and assessments.

For each simulated trait, we calculated the AICc weights of the five fitted models to assess their relative fit to the trait data (Sugiura 1978; Burnham and Anderson 2002). The model with the highest AICc weight was considered the best-fitting model. If the best-fitting model was the same as the simulated model, we considered this a “correct” assessment; otherwise, if the best-fitting model was a different model, we considered this an “incorrect” assessment. We considered the proportion of these assessments that were “correct” as the “accuracy” under a particular set of parameters. We also calculated the ΔAICc between the best-fitting model (regardless of whether it was “correct”) and the other four models to assess whether any other models also had “substantial” empirical support (ΔAICc ≤ 2; Burnham & Anderson, 2002). If all of the other fitted models had ΔAICc > 2, we considered this a “clear” result, otherwise we considered this an “unclear” result. We used these definitions to calculate the proportion of simulations in which 1) the best-fitting model was “correct” and “clear” (i.e., we got the correct model and it was substantially better than the other models), 2) the best-fitting model was “correct” and “unclear” (i.e., we got the correct model, but one or more other model(s) could not be ruled out), 3) the best-fitting model was “incorrect” and “unclear” (i.e., we got the wrong model, but one or more other models could not be ruled out), and 4) the best-fitting model was “incorrect” and “clear” (i.e., we got the wrong model, and it was substantially better than the other models, including the correct model). We used the first two calculations to identify the “accuracy” (the proportion of simulations in categories 1 or 2) and “clear accuracy” (the proportion of simulations in category 1) of our model-fitting approach. We used the latter two calculations to identify the “error rate” of our model-fitting approach (the proportion of simulations in categories 3 or 4). Then, for each type of fitted model (BM, trend, OUc, OUs, ACDC), we also collated all analyses in which it was the best-fitting model and identified which generative model had been used. We then used this information to calculate a “false-positive rate”, or the proportion of simulations with a particular best-fitting model which didn’t match the generative model. We used this to assess whether any particular model had relatively high or low chances of being incorrectly chosen over the simulated model. Finally, we also extracted the estimated parameters (σ, α, β, *trend*, and θ) from the results of the model fitting analyses when the generative and fitted models matched, to assess how well parameters are estimated across the range of simulation variables.

Results

Accuracy

First, we analysed our results based on the model used to generate the trait data. Across all model-fitting analyses, we found that 58.5% of them recover the generative model as the best-fitting model (hereafter referred to as “accuracy”). When only those model-fits that are clearly better than the second best-fitting model are counted, this accuracy (hereafter referred to as “clear accuracy”) is only 35% (Figs. 2 and S1; pink portions only [“correct and clear”]). Across nearly all analyses, the overall proportion of best-fitting models that are “incorrect and clear” (the worst-case scenario) is 9.9%, with no consistent increases associated with any specific simulation parameter (Figs. 2 and S1; green portions). When no fossils are included in the simulated phylogeny, the proportion of analyses in which the best-fitting model matches the simulated model drops to 40.8%, and the clear accuracy drops to 17.1%. However, when fossils are included (10% or more of tips), accuracy almost always increases. Across fossil-included simulations, overall accuracy increases by 22.1% (to 62.9%) and clear accuracy increases by 22.4% (to 39.5%). Even when just 10% of tips are fossils, the overall accuracy is 59.7% and the clear accuracy is 36%. The only exception to this trend is with the smallest set of phylogenies (50 tips), a small proportion (10%) of recent-biased fossils, and the higher relative extinction rate, wherein the accuracy decreases by 1% and the clear accuracy decreases by 0.6% relative to similar simulations without fossils (Fig. 2). However, in this scenario, increasing phylogeny size or the proportion of fossils results in an increase in accuracy over the non-fossil simulations. For random and recent-biased fossils, as the proportion of fossils in the phylogeny increases, so does the accuracy (Figs. 2 and S1; middle and bottom rows). However, for root-biased fossils, a small proportion of fossils is often as beneficial as a large proportion, especially for phylogenies with 200 or more tips (Figs. 2 and S1; top row). There also appears to be an increase in accuracy as phylogeny size increases, with or without fossils. This size effect is largest when fossils are recent-biased, but in most other cases this accuracy increase is fairly minor compared to the accuracy increase due to the inclusion of any fossils. When only model-fits that are clearly better are assessed, the increase in accuracy due to the inclusion of fossils is magnified. This is especially notable when the included fossils are root-biased, with accuracy more than doubled (17% clear accuracy without fossils, 46.7-48.4% clear accuracy with fossils) regardless of the proportion of tips that are fossils.

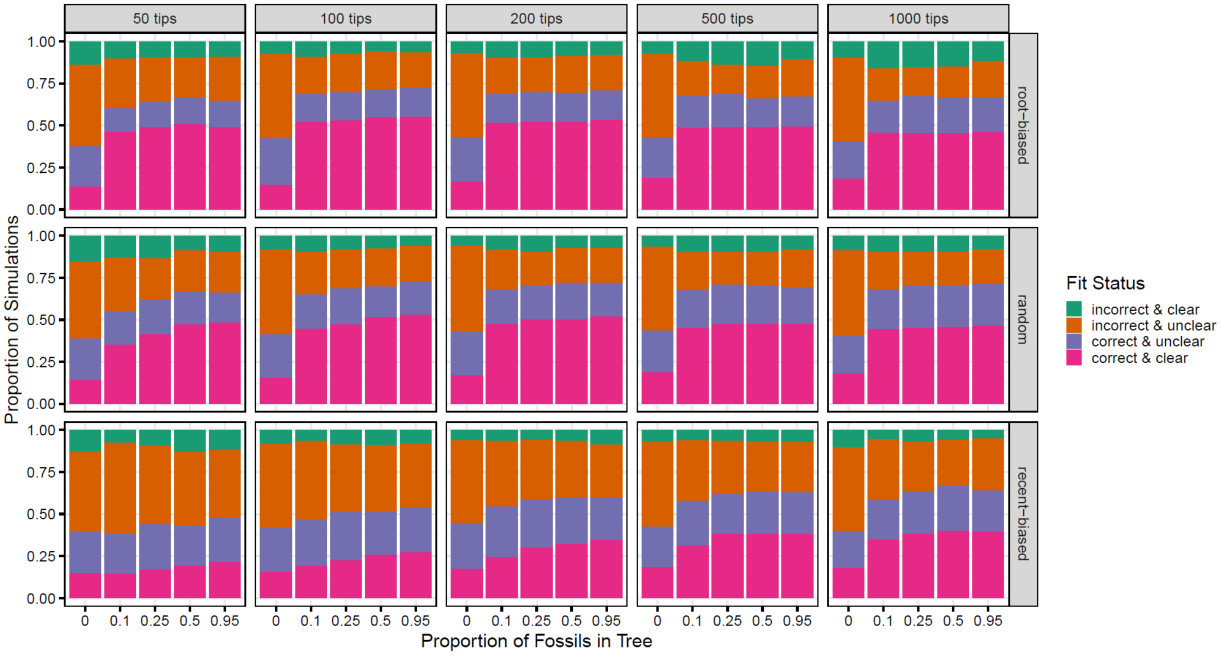


Figure 2: Summary of the best-fitting models across all simulations. The best-fitting model for a given simulation can be either correct (matching the simulated model) or incorrect (not

matching the simulated model) and also either clear (all other models with ΔAICc > 2) or unclear (at least one other model with ΔAICc ≤ 2). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row headers), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Fig. S1.

Model fitting accuracy is highly dependent on the evolutionary model that is being simulated and the parameters of that model (Figs. 3, S2, and S3). When the simulating model was a basic Brownian motion model of character evolution, our analyses returned the correct best-fitting model in 71.2% of all simulations. However, the clear accuracy for these models is much lower, at 7.1% (Fig. S2). These percentages are fairly consistent regardless of fossil distribution, phylogeny size, relative extinction rate, or the strength of the Brownian motion ("wBM" versus "sBM"). There does appear to be a slight decrease in accuracy when there are any fossil tips in the phylogeny, but there doesn't appear to be any relationship between accuracy and the proportion of fossil tips. When a trend is added to the Brownian motion model ("wtrend" and "strend"), phylogenies with no fossil tips have an accuracy of 0% (Figs. 3 and S2). However, when any fossil tips are in the phylogeny, accuracy improves dramatically to 86.9% (with 78.4% clear accuracy). For strong trends ("strend") with any proportion of fossil tips, the overall accuracy is 95% and all sets of parameters have >50% accuracy (Fig. 3). Accuracy generally increases with increasing fossil tip proportion and/or phylogeny size. Furthermore, if the fossil distribution is random, the overall accuracy is 97.4%, and the overall accuracy for simulations with root-biased fossil distributions is 100%. For weak trends ("wtrend"), accuracy is lower, especially for simulations with recent-biased fossil distributions (58.3%). However, with random (87.2% average) or root-biased (90.9% average) fossil distributions, accuracy is >50% across the board, with accuracy again generally increasing with increasing phylogeny size and/or fossil proportion. The BM simulation accuracy results are qualitatively the same for both death rates (μ). One notable difference is that simulations with a weak trend tend to have lower accuracy when simulated with the lower death rate (Fig. S3).

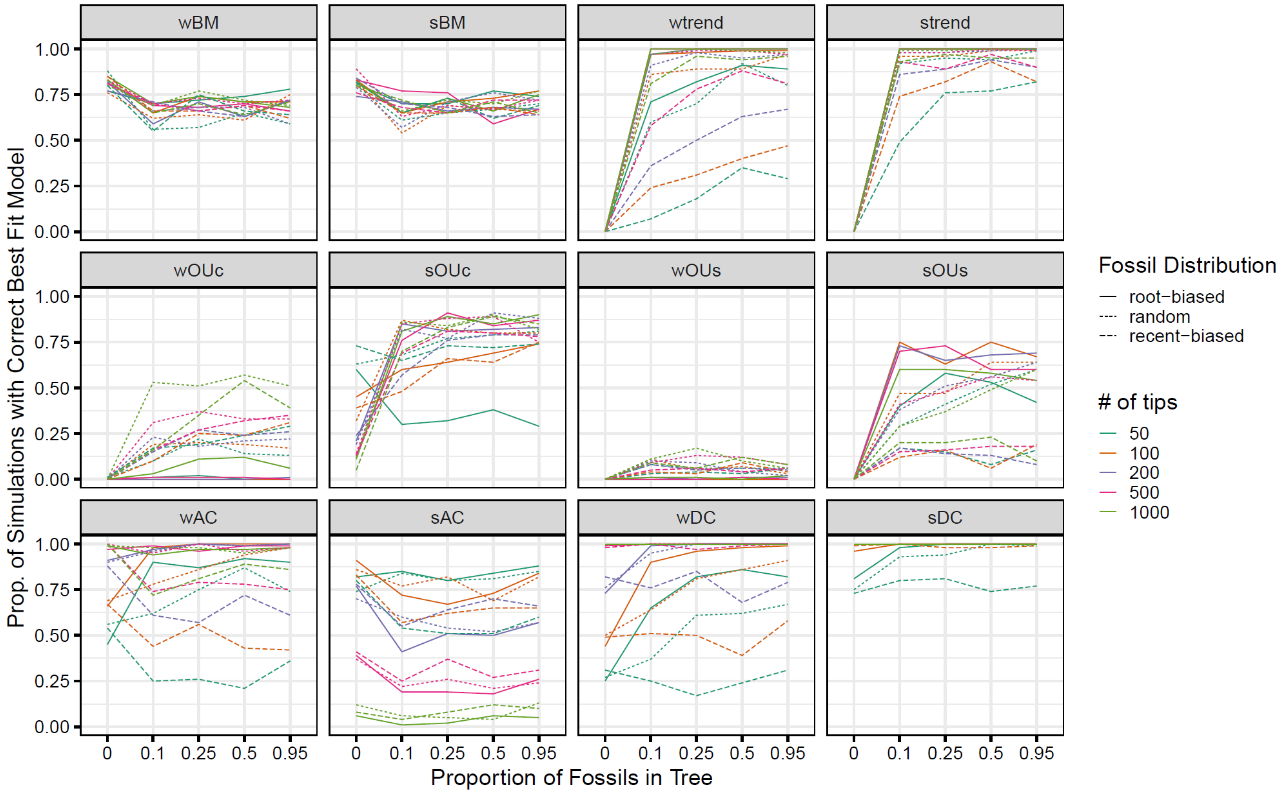


Figure 3: The proportion of simulations for which the best-fit model does match the simulated model (panel headers). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (line type), and the number of tips in the simulated phylogenies (line color). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S3.

Accuracy for data simulated under an Ornstein-Uhlenbeck model of character evolution is somewhat similar to that of trended BM models (Figs. 3, S2, and S3). The overall accuracy for OU simulations when no fossils are included is 16.5%, with a clear accuracy of 12.8%. When any fossils are included, this rises to an overall accuracy of 41.2% and a clear accuracy of 16.8%. Weak OU simulations are particularly difficult to accurately infer. When no fossils are included, weak OU simulations with no shifted optima (“wOUc”) have an accuracy of 1.1%, and weak OU simulations with shifted optima (“wOUs”) have an accuracy of 0%. When any fossils are included, these overall accuracies increase to 24% for wOUc and 9.7% for wOUs. Within both of these sets of simulations, those with root-biased fossil distributions have almost 0% accuracy (Figs. 3 and S2), while simulations with random fossil distributions show the highest relative accuracy (up to 31.8% overall for wOUc simulations and up to 14% overall for wOUs simulations). The initial addition of 10% random or recent-biased fossils to the phylogeny provides the largest increase in accuracy, with additional fossils providing marginal improvements. Accuracy also increases with increasing phylogeny size. Simulations with strong OU signals (“sOUc” and “sOUs”) generally have higher accuracy than their respective weak models. When no fossils are included, strong OU simulations with no shifted optima (“sOUc”) have an overall accuracy of 64.7%. When any fossils are included, this overall accuracy increases to 78.7%. The largest increase in accuracy for strong OU models comes from the initial inclusion of fossils in the phylogeny, with diminishing increases in accuracy with the inclusion of more fossils. The one exception to this, when the relative death rate is high, is a decrease in accuracy when 10% of the tips of a 50 tip tree are replaced with root-biased fossils (Fig. 3). Overall, accuracy appears to increase with increasing phylogeny size, although this increase (5-25%) is much smaller than the increase related to adding fossils (~50-75%). Accuracies are generally lower for strong OU simulations with a shifting optimum ("sOUs"). All simulations of this model without fossils, as for the trended BM and wOUs models, have 0% accuracy. When any fossils are included, this overall accuracy increases to 52.2%. The age of the included fossils plays a major role in the accuracy: models with recent-biased fossils have an average accuracy of 15.7%, models with random fossils have an average accuracy of 60.6%, and models with root-biased fossils have an average accuracy of 80.4%. Fossil tip proportion and phylogeny do not appear to have consistent directional impacts on accuracy. Finally, for all of the OU models, when the simulations were produced with a lower death rate (*μ*), accuracy is generally higher across the board, with "sOUc" accuracies plateauing around 80% and "sOUs" accuracies reaching 100% for some simulations (Fig. S2)

The AC and DC simulations show accuracy patterns that are different from any of those of other simulation models (Fig. 3). The overall accuracy for AC/DC simulations is 68.8%, with a clear accuracy of 53.8%. When no fossils are included, the overall accuracy drops to 65.2%, and when any fossils are included, the overall accuracy increases to 69.7%. The simulations with weak AC/DC signals ("wAC" and "wDC") show wide ranges of accuracies, varying from 15% to 100%. The strong AC ("sAC") simulations show a similar breadth of accuracies, ranging from 0% to 100%, while strong DC ("sDC") simulations show a much narrower range of accuracies (75% to 100%). For "wAC", "wDC", and "sDC" simulations, adding root-biased fossil tips increases accuracy, with some scenarios seeing accuracy increase by more than 50%. On the other hand, adding root-biased fossils to sAC simulations results in a decrease in accuracy. For most simulations, increasing phylogeny size appears to increase accuracy; however, for the strong AC simulations, increasing phylogeny size has the opposite effect, with phylogenies with 1000 tips having nearly 0% accuracy regardless of fossil inclusion. For the strong DC simulations, nearly all simulations have an impressive 100% accuracy except for those based on the smallest phylogeny (50 tips). Generally speaking, when the phylogenies were simulated with a lower death rate (*μ*), the model specification accuracy for AC/DC simulations decreased, except for sAC simulations which showed an accuracy increase (Fig. S2). Most notably, the strong DC simulations showed accuracy drops of up to 75%, except for those using the two largest phylogeny sizes (500 and 1000 tips).

Error Rates

The rate at which particular incorrect models are chosen over the simulated model ("error rate") highly varies depending on fossil inclusion, simulated model type, and phylogeny size (Figs. 4-6, S4-6). The overall error rate, or how often an incorrect model is the best-fitting model is 41.5%, but this number increases to 59.2% when no fossils are included. For simulated

Brownian motion models, when no fossils are included, ACDC models are nearly the only models that are chosen over a BM model, with a 18.2% error rate (Figs. 4 and S4). However, when any fossils are included, the probability of incorrectly choosing an ACDC model stays relatively the same, but the probability of choosing a trend, OUc, or OUs model increases slightly, leading to an overall error rate of 31.5%. The relative proportions of these incorrect models is qualitatively similar regardless of the other simulation parameters. When the BM model includes a trend, there is a 100% error rate when no fossils are included (Figs. 4 and S4). Across these simulations, 18.6% of them were incorrectly identified as ACDC and 81.2% of them were incorrectly identified as BM without a trend. With any fossils included, weak trends have a 21.2% error rate overall, with 14.4% of simulations incorrectly identified as BM, 3.6% of simulations incorrectly identified as ACDC, and 3.2% of simulations incorrectly identified as OUc or OUs. When any fossils are included with a strong trend model, the error rate is 5% overall, with 3.1% of simulations incorrectly identified as OUs and 1.3% of simulations incorrectly identified as BM. Error rates for trend simulations are consistently higher when the fossils are recent-biased, especially when the phylogeny contains 200 or fewer tips (Figs. 4 and S4). In these cases, many more simulations are incorrectly identified as BM without a trend.

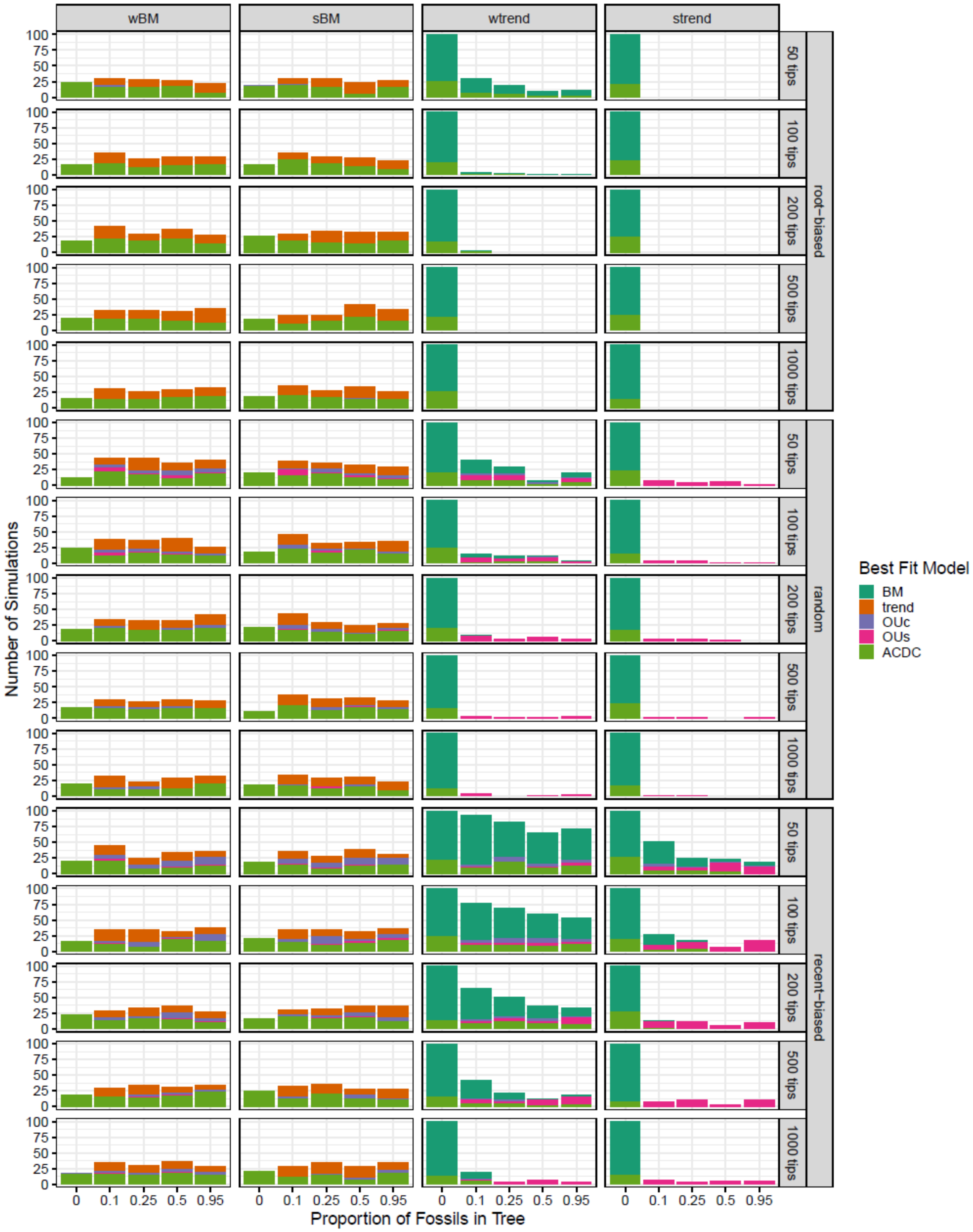


Figure 4: The number of BM/trend simulations for which the best-fit model (color) does not match the simulated model (panel headers). Only BM and trend simulations are included (see Figures 5 and 6 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S4.

For OU models, the overall error rate is 63.8%; however, this is highly dependent on the simulating model and the inclusion of fossils (Figs. 5 and S5). Weak OU models are difficult to infer correctly, with an overall error rate of 86.4%. When no fossils are included the error rate is 99.4%, with 65.7% of simulations incorrectly identified as ACDC and 33.2% of simulations incorrectly identified as BM. When fossils are included, the overall error rate declines to 83.1%, with 29.9% of simulations incorrectly identified as BM, 24.7% of simulations incorrectly identified as trend, and 16.4% of simulations incorrectly identified as ACDC. With increasing phylogeny size, the proportion of simulations that are incorrectly identified as BM or ACDC generally decreases. Further, for wOUs, simulations with root-biased fossils tended to be incorrectly identified as trend (46% of simulations), whereas simulations with recent-biased fossils tended to be incorrectly identified as OUc (29.1% of simulations). Simulations with random fossils showed a gradual transition from tending to a trend model with smaller phylogenies and tending to an OUc model with larger phylogenies. Strong OU models have a much smaller overall error rate of 41.1% (Figs. 5 and S5). When no fossils are included, the overall error rate is 100% for sOUs and 35.2% for sOUc, whereas when any fossils are included the overall error rate drops to 47.8% for sOUs and 21.3% for sOUc. With a low relative extinction rate, when the sOUs model is incorrectly interpreted, it is almost always as the OUc model, whereas when the sOUc model is incorrectly interpreted, it is almost always as the OUs model. However, when a higher relative extinction rate is simulated, more of the simulations for both models are incorrectly identified as an ACDC model (Figs. 5 and S5). The age distribution appears to have little to no effect on the error rate for the sOUc model, but error rates for the sOUs simulations are highly dependent on the age of the included fossils (root-biased: 35.7%, random: 51.5%; recent-biased: 87.5%). One notable exception to these trends is increased error rates for the sOUc and sOUs models when relative extinction is high, fossils are root-biased, and phylogenies are small (50-100 tips). In these cases, there is increased incorrect support for a BM model (Fig. 5).

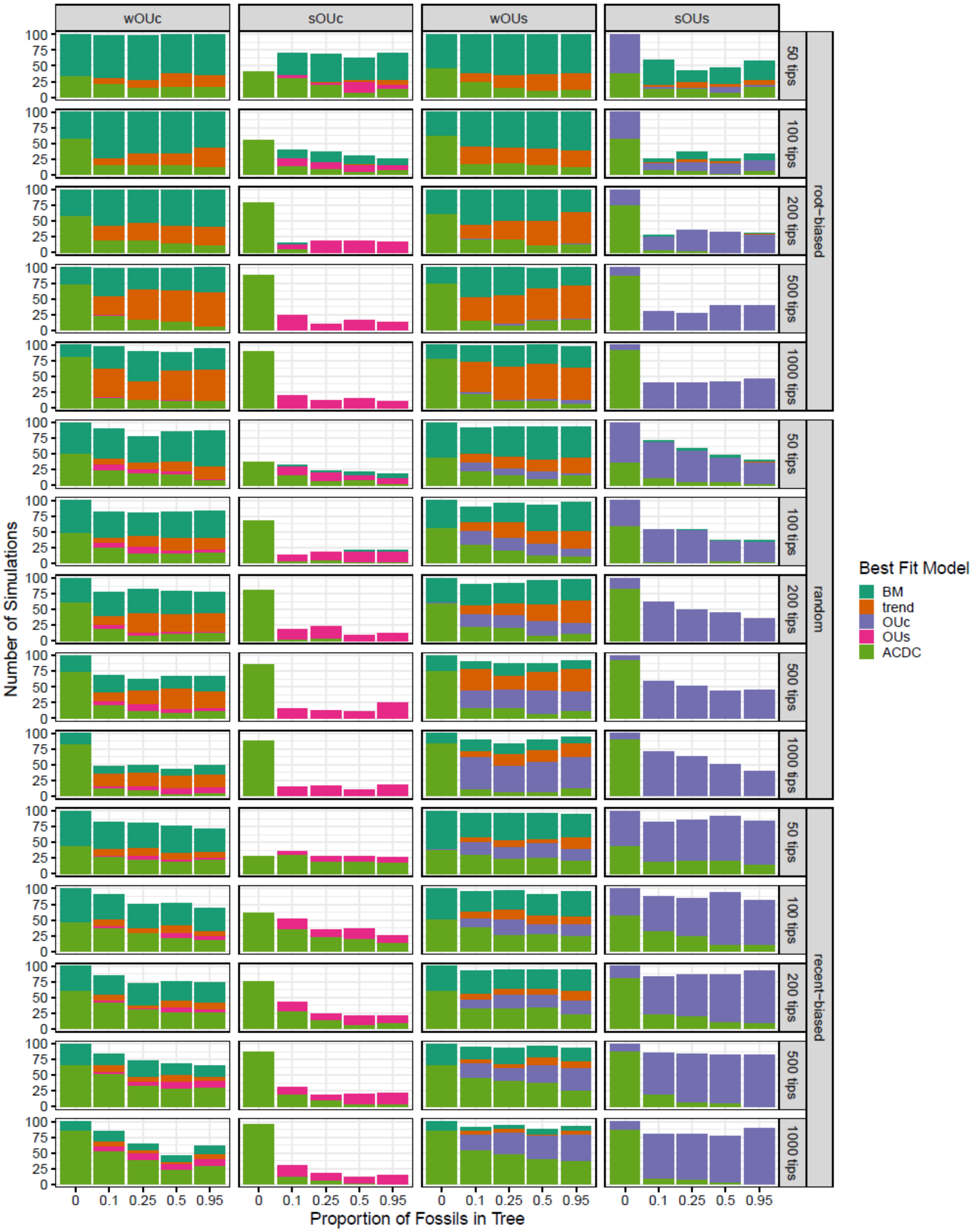


Figure 5: The number of OU simulations for which the best-fit model (color) <u>does not match</u> the simulated model (panel headers). Only OU simulations are included (see Figures 4 and 6 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S5.

The error rate for the ACDC family of simulated models is 31.2% overall (Figs. 6 and S6). When fossils are not used, the overall error rate is 34.8%, whereas simulations with fossils have an overall error rate of 30.2%. When only assessing simulations with a low relative extinction rate, these error rates increase to 45.7% and 38.2%, respectively. In these cases, 32% of simulations are incorrectly identified as BM, 5.5% of simulations are incorrectly identified as trend, and 2.1% of simulations are incorrectly identified as OUc. When a relatively high extinction rate is simulated, the error rates decrease to 24% without fossils and 22.3% with any fossils, with 8.5% of simulations incorrectly identified as BM, 12.1% of simulations incorrectly identified as OUc, and 1.1% of simulations incorrectly identified as trend. Notably, when the strong AC model was simulated with a high relative extinction rate, 42.5% of these simulations were misinterpreted as the OUc model (Fig. 6), with even higher rates when phylogenies with 500 or more tips were used (64.5% for 500 tips, 86.7% for 1000 tips).

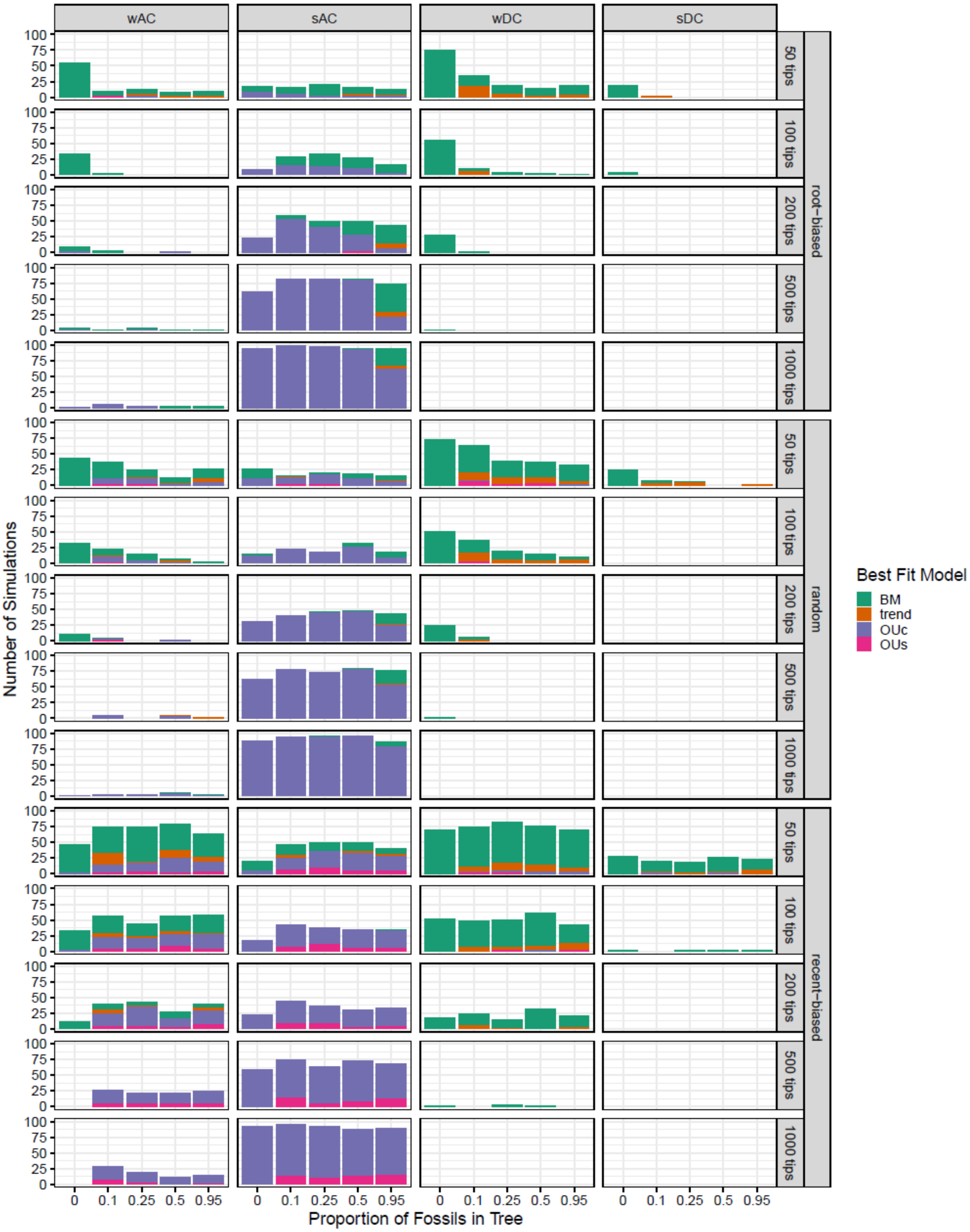


Figure 6: The number of AC/DC simulations for which the best-fit model (color) <u>does not</u> match the simulated model (panel headers). Only AC and DC simulations are included (see Figures 4 and 5 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S6.

False Positive Rates

We also collated our analyses based on the best-fitting model, in order to determine the generative models behind each perceived result. The “false positive rate” is the proportion of all simulations in which a particular model was best-fitting where this model does not match the generative model. As with accuracy and error rate, this metric highly varies depending on fossil inclusion, model type, and phylogeny size (Figs. 7 and S7). When BM is the best-fitting model, this is a false positive about 57.2% of the time. When no fossils are included, the false positive rate is 67.9% overall, whereas when any fossils are included the false positive rate is 52.4%. For smaller phylogeny sizes (50 or 100 tips), there are just as many simulations which misidentified an ACDC model as a BM model as simulations which correctly identified a BM model (Fig. 7). However, with increasing phylogeny size, this false positive rate decreases (e.g., 1000 tips without fossils is 57%; 1000 tips with fossils is 23.2%). For these larger phylogeny sizes, when no fossils are included, the majority of incorrect untrended BM inferences were actually simulated under a trended BM model. When fossils are included in these larger phylogenies (500+ tips), the majority of incorrect untrended BM inferences were simulated under OUc, OUs, or ACDC models. A trended BM model was never the best-fitting model when no fossils were included in the phylogeny. However, when a trended BM model is the best-fitting model, this was also the generating model 65.1% of the time. Phylogeny size plays a minor role in the false positive rate, with 50 tip phylogenies having an overall false positive rate of 42.2% and 1000 tip phylogenies having a rate of 28%. The proportion of fossils in the phylogeny does not appear to have any consistent effect on the false positive rate. These results are qualitatively similar regardless of the simulated death rate (Figs. 7 and S7).

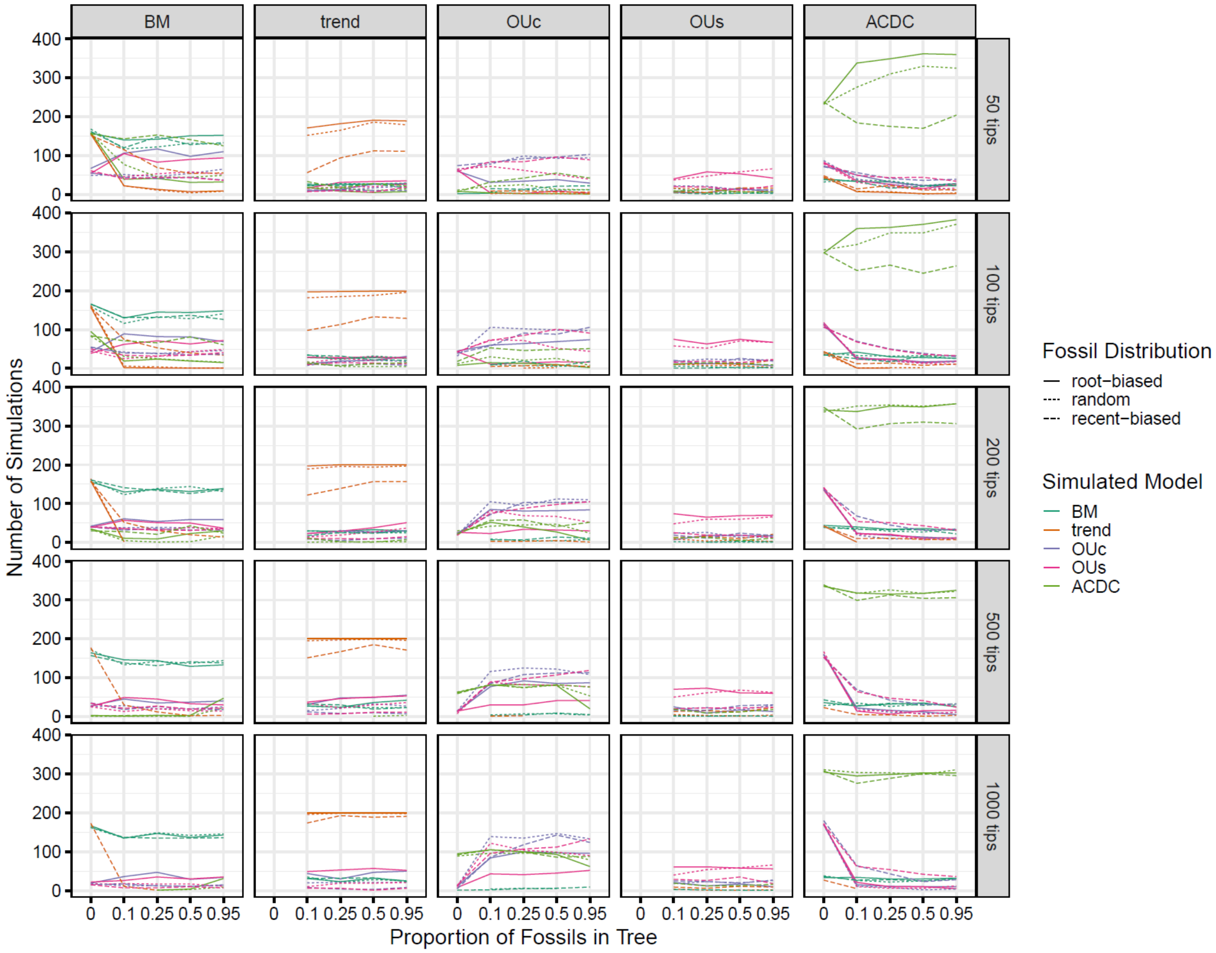


Figure 7: The number of simulations with different generative models (line color) when particular models were best-fitting (column headers). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the age distribution of the simulated fossils (line type), and the number of tips in the simulated phylogenies (row headers). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S7.

For centered OU (OUc) simulations, there is an overall 50.3% false positive rate, with 32.7% of simulations generated under an OUs model, 14.8% of simulations generated under ACDC, and 2.2% of simulations under BM (Figs. 7 and S7). The inclusion of any fossils reduces the false positive rate to 49% overall, with increased fossil proportion only providing marginal decreases. With a high relative extinction rate and no fossils, the false positive rate increases as phylogeny size increases due to an increase in the number of ACDC simulations that are incorrectly identified as OUc. With only 50 tips, 10% of simulations identified as OUc were actually generated under an ACDC model. This increases consistently up to 19.9% with 1000 tips. The inclusion of root-biased fossils results in fewer OUs simulations being misidentified as OUc; however, they also result in fewer OUc simulations being correctly identified as OUc. Of all of the fitted models, the shifting Ornstein-Uhlenbeck (OUs) model was the best-fitting model for the fewest simulations (Figs. 7 and S7). Further, as with trended BM models, the shifting

Ornstein-Uhlenbeck model is never the best-fitting model when no fossils are included in the phylogeny. When fossils are included, the OUs model has an overall false-positive rate of 31%, with 21.4% of the simulations being generated by OUc, 5.7% of simulations being generated by trend, and 3% of simulations being simulated by ACDC. When root-biased fossils are included, the false positive rate for OUs drops to about 15%, regardless of the proportion of fossil tips or the phylogeny size. Phylogenies with randomly-distributed fossils generally have false positive rates of about 25%, except for smaller phylogenies (50-100 tips) which have false positive rates of 30-35%. The inclusion of only recent-biased fossils results in a much larger false positive rate of 59.5%, with 32.8% of best-fitting simulations being generated using centered OU models. False positive rates for OUs models under different relative death rates (μ) are qualitatively similar (Figs. 7 and S7).

The overall false positive rate when the ACDC model is the best-fitting model is 29.4% (Figs. 7 and S7). When no fossils are included, the false positive rate is 51.3%, with 18.8% of simulations actually generated under an OUc model and 18.9% of simulations generated under an OUs model. However, when fossils are included, the false positive rate decreases to 21% overall, with the majority of these false positives generated under a BM model (8.5% of simulations). Recent-biased fossils have a higher false positive rate (29.6%) than random or root-biased fossils (~17%). An increase in phylogeny size leads to lower false positive rates when fossils are included, but has no discernible impact on false positive rates when no fossils are included. The relative extinction rate does not have any consistent effect on false positive rates either.

Parameter Estimation

Along with determining the best-fitting model type, the accuracy of parameter estimation is also an important factor in fitting models. To quantify this, we collated the parameter estimates produced during model fitting for the model matching the generative model (regardless of how well this model fitted). We found that σ estimates for untrended BM simulations tended to be fairly accurate (i.e., average close to the simulated value), regardless of the inclusion of fossils or the size of the phylogeny (Fig. 8). Simulations with weak σ (less Brownian noise) generally had higher precision (i.e., narrower range of estimates) than those with strong Brownian noise. Precision also increased with increasing phylogeny size. However, we found no noticeable relationship between accuracy or precision and the proportion of fossil tips or the age distribution of the fossils. These results are consistent across both relative death rates (Fig. S8).

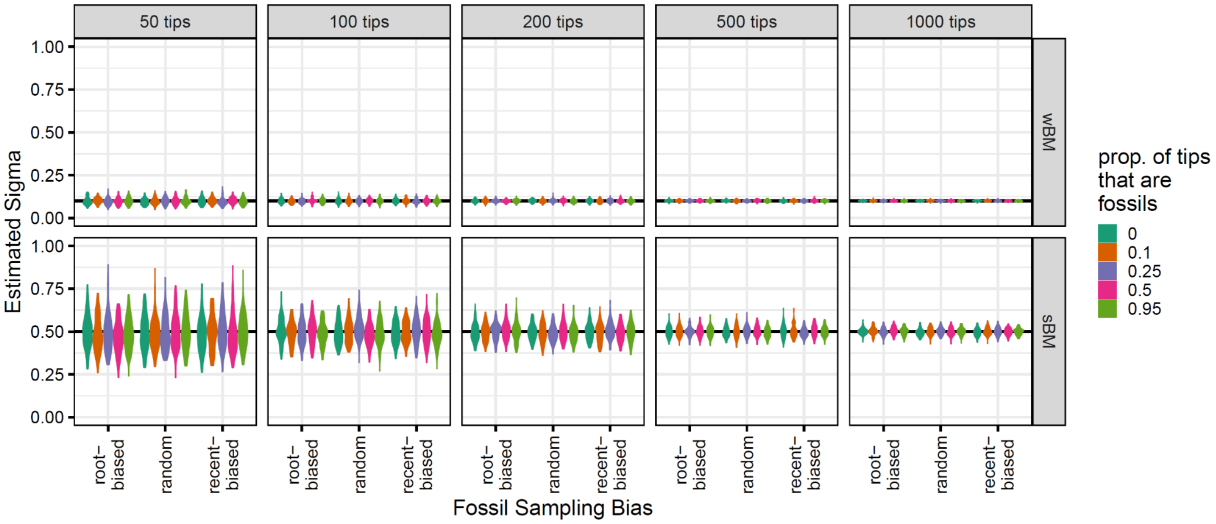


Figure 8: The distributions of the estimates for the σ parameter when a BM model was simulated on a phylogeny and then a BM model was fit to that simulated data. The top row represents simulated trends with a weak σ parameter (0.1) and the bottom row represents a trend model simulated with a strong σ parameter (0.5). The true (simulated) σ parameter is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S8.

When Brownian motion was simulated with a temporal trend, the *trend* parameter was never accurately estimated when no fossils were included (Fig. 9). With the inclusion of any fossils, accuracy appears to be high, although including recent-biased fossils resulted in poor precision compared to root-biased and randomly-distributed fossils. As with untrended BM simulations, increasing phylogeny size is associated with increasing precision. Furthermore, increasing the proportion of fossil tips appears to have a marginal increase in precision, but no effect on accuracy. These results are consistent across both relative death rates, with the higher death rate having marginally higher precision (Fig. S9).

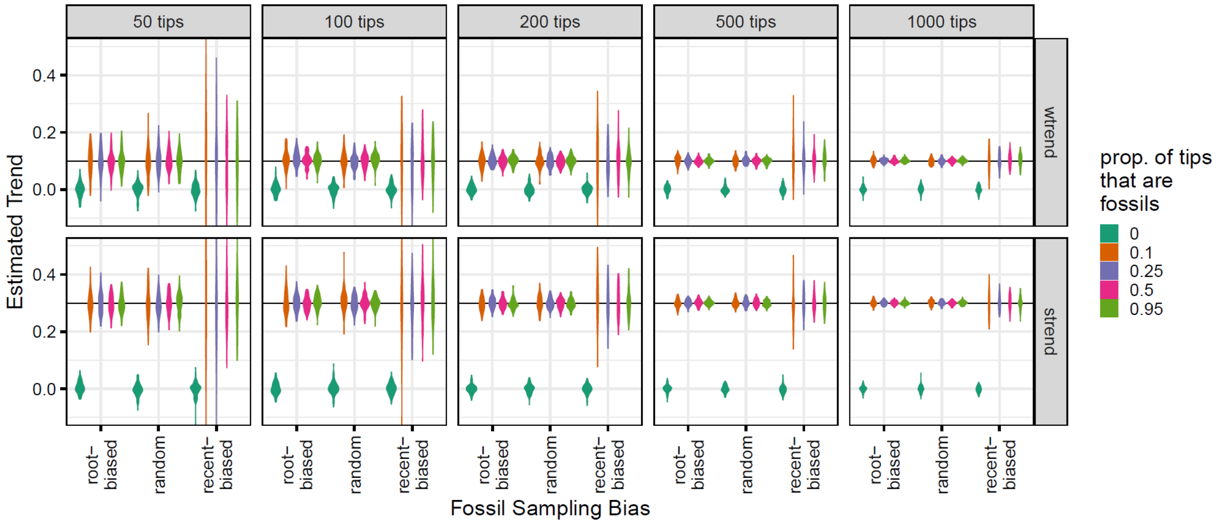


Figure 9: The distributions of the estimates for the *trend* parameter when a trend model was simulated on a phylogeny and then a trend model was fit to that simulated data. The top row represents simulated trends with a weak *trend* parameter (0.1) and the bottom row represents a trend model simulated with a strong *trend* parameter (0.3). The true (simulated) *trend* parameter is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S9.

For the centered Ornstein-Uhlenbeck models (OUc), estimation of the optimum trait value, the theta parameter (*θ*), has high accuracy, regardless of the number of tips, proportion of fossil tips, and fossil distribution (Fig. 10). Precision appears to be steady, again regardless of any of these variables. However, for the weak shifting Ornstein-Uhlenbeck models (wOUs), when no fossils are included, the trait state at the root (root theta) is often overestimated, while the optimum trait state (optimum theta) is often underestimated. The inclusion of fossils stabilizes the estimates of both of these parameters, although recent-biased fossils generally result in less accuracy for these parameters. For the strong shifting Ornstein-Uhlenbeck models (sOUs), the optimum theta has high estimation accuracy regardless of fossil proportion, fossil age, or phylogeny size. The root theta has a similarly high estimation accuracy when root-biased or randomly-distributed fossils are included. However, when the fossils have a recent-biased age distribution, the root theta has notably very low accuracy. In fact, over 95% of model fits resulted in extreme root estimates that were classified as outliers (e.g., 3,000,000). Further, including no fossils results in an unusually accurate and precise root estimate. These theta results are consistent across both relative death rates, with the lower death rate having marginally higher precision (Fig. S10). The halflife parameter generally has high accuracy for strong OU simulations; however, the halflife parameter is often underestimated for weak OU simulations (Fig. 11). Both of these trends appear to be consistent across simulations, regardless of fossil proportion, fossil age, and phylogeny size. These results are consistent for both relative death rates, with slightly higher precision for the lower death rate (Figs. S11).

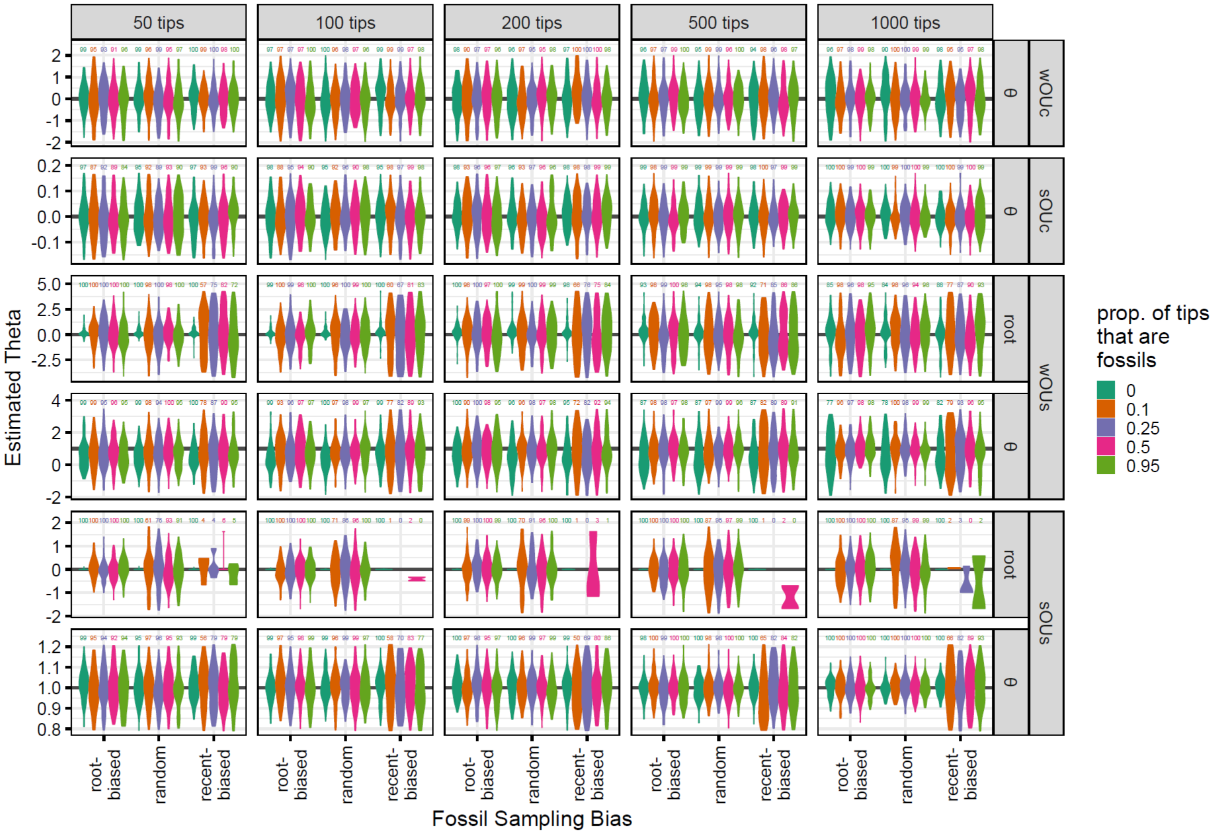


Figure 10: The distributions of the estimates for the *θ* parameter when an OU model was simulated on a phylogeny and then an OU model was fit to that simulated data. The true (simulated) *θ* parameter is represented with a solid horizontal line. Outliers have been removed separately for each row, and the numbers above box plots represent the number of non-outliers that are represented by each boxplot. Note that the scale of the y-axis varies for each row. The results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S10.

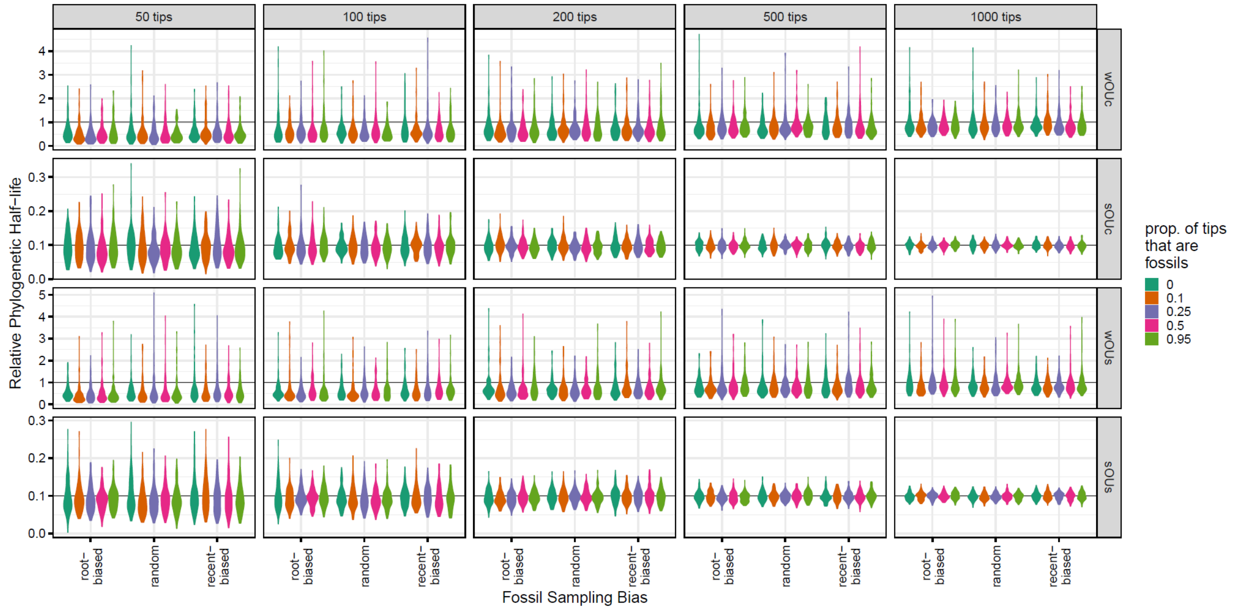


Figure 11: The distributions of the estimates for the relative half-life when an OU model was simulated on a phylogeny and then an OU model was fit to that simulated data. Relative half-life is calculated as $\frac{log(2)\ /\ \alpha}{tree\ height}$. The first and second rows represent centered OU models simulated with weak (1) and strong (0.1) relative half-lives, respectively. The third and fourth rows represent shifting OU models simulated with weak (1) and strong (0.1) relative half-lives, respectively. The true (simulated) relative half-life is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S11.

Generally speaking, beta (β) estimation accuracy is good for accelerating or decelerating simulations (Fig. 12, S12). Accuracy and precision of the beta parameter increases with phylogeny size in most cases. The inclusion of fossils also appears to result in a large increase in accuracy, especially under the higher relative death rate; however the proportion of fossil tips and the ages of the fossils do not seem to matter. The one exception is for the strong AC model under the higher death rate, where increasing phylogeny size results in strongly underestimating the beta parameter. For simulations with phylogenies with at least 100 tips, we see a bimodal distribution of parameter estimates, with one mode centered on the simulated value and one mode centered on 0. As phylogeny size increases, estimates of this parameter trend away from the simulated value of 0.5 and towards the alternative mode centered around 0.

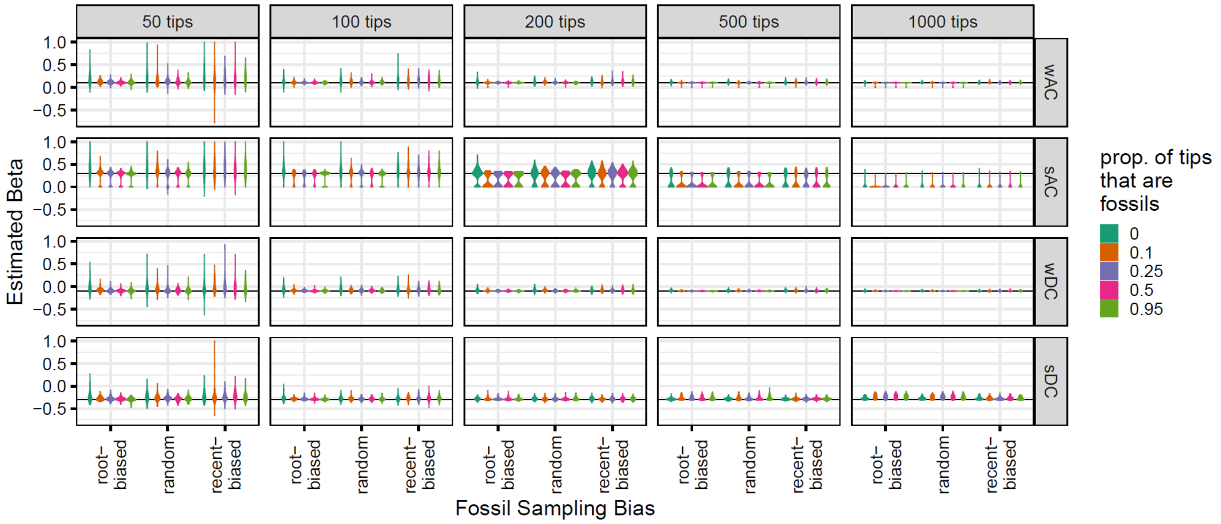


Figure 12: The distributions of the estimates for the *beta* parameter when an ACDC model was simulated on a phylogeny and then an ACDC model was fit to that simulated data. The true (simulated) *beta* parameter is represented with a solid horizontal line. The results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.9, the results for analyses where *mu* = 0.25 are in Figure S12.

Discussion

The ability to accurately and precisely infer how continuous characters have evolved over time is crucial for testing hypotheses in evolutionary biology (Slater 2013; Benson et al. 2014; Gearty et al. 2018). Our results echo past research (Slater et al. 2012; Mitchell 2015), showing that the inclusion of fossil tips in phylogenies can be critical for inferring both the correct tempo and mode of continuous trait evolution. On average, including any fossil taxa in our simulations dramatically increased statistical power (Fig. 2). In fact, generally speaking, the first 10% of fossil tips leads to the largest increase in accuracy, with diminishing returns beyond this proportion. This result should be encouraging for evolutionary paleobiologists who already include fossils in their phylogenetic comparative methods, and should serve as a reiterative warning to evolutionary biologists who use extant-only phylogenies.

Beyond the mere inclusion of fossil taxa in such analyses, we demonstrate that the relative age of these fossil taxa is fundamental to this increase in accuracy. When relatively young (recent-biased) fossil taxa are included, especially in smaller trees, the degree to which accuracy improves can be marginal (Figs. 2-3), indicating that young fossil taxa provide about as much statistical information as extant taxa. This is understandable since they do not provide much novel information in the modelling process with regards to tempo or mode. However, when relatively old (root-biased) or evenly distributed (random) fossil taxa are included, the accuracy improves dramatically (Figs. 2-3). Root-biased fossils provide insight into older parts of evolutionary history which are harder to infer from extant tips alone; it is therefore reasonable to expect that including this information would drastically improve our ability to recover the true (simulated) tempo and mode of evolution.

This improved accuracy when adding root-biased fossils is most notable when the true mode of evolution has an expected value (i.e., mean) that is heterogeneous through time, such as trend and OUs models (Fig. 3). These are two of the most common models used in PCMs to test for adaptive trait evolution, wherein a clade evolves from some less optimal ancestral state towards a different, more optimal, state (Pagel 2002; Butler and King 2004; Cooper et al. 2016; Gill et al. 2017; Gearty et al. 2018; Godoy et al. 2019). For these models, the inclusion of any fossils at all increases accuracy far more than increasing the proportion of fossils or increasing the total number of tips in the phylogeny (Figs. 3-5). Furthermore, when a split OU or trend model is the best-fitting model, and root-biased fossils have been included in the phylogeny, this is rarely an error (Fig. 7). Therefore, it is extremely encouraging that positive results for the split OU model and trend BM model generally can be assumed to be trustworthy, as long as fossils are included.

On the other hand, under simulated models such as trend or OUs, our model fitting approach failed to infer the correct generating model 100% of the time when no fossil tips were included in the phylogeny (Fig. 3). Instead, these simulations were often inferred to have OUc or ACDC evolutionary histories (Fig. 7). Furthermore, when no fossils were included, we never inferred a trend or OUs model as the best-fitting model, even as a error (Fig. 7). Indeed, the documentation for the mvMORPH package states that the trend model is only identifiable with non-ultrametric trees (Clavel et al. 2015) and that the `root=TRUE` option for the mvOU function (which we used across all OUs model fittings) can be problematic with ultrametric trees. Across all of our simulations, not a single analysis failed to converge when attempting to fit a trend or OUs model to an ultrametric phylogeny. Therefore, evolutionary biologists that use this option with ultrametric phylogenies will likely incorrectly infer ACDC or OUc modes of evolutionary history for clades and traits that ultimately have OUs-like evolutionary histories. Alternatively, evolutionary biologists commonly use variations of the OUs model to drop the estimation of the root state (e.g., `root=FALSE` in mvMORPH or `root.station=TRUE` in the OUwie package) when studying ultrametric trees (Jaffe et al. 2011; Price and Hopkins 2015; Gearty et al. 2018; Anelli et al. 2024). This commonly leads to inferences that strongly support a split OU model as the best model. More simulation work is needed to better understand the behavior of the different mathematical models these arguments represent. Overall, evolutionary biologists should be extremely cautious when trying to infer this type of evolutionary trajectory when they are using extant-only trees. Fortunately, the inclusion of even a minimal proportion of fossil tips results in notable increases in the accuracy of inferring such histories.

The inclusion of fossils has mixed results for some other simulated trajectories of evolution. For example, fossils usually improve inferences of strong OUc evolutionary histories (Figs. 3 and 5). However, when the relative death rate is low or the size of the phylogeny is small, including fossils results in lower inference accuracy, with many simulations being misinterpreted as BM or trend (Figs. 3 and 7). Increasing data availability and computational power means that phylogenies are constantly growing in size, and the average relative death rate is close to 1 in the fossil record (Marshall 2017), so these situations are likely unrealistic.

However, further work may be necessary to understand why the inclusion of fossils results in poorer performance in these particular scenarios.

Another example involves weak OUc and OUs modes, which are generally very difficult to correctly infer, regardless of the parameters we tested (Figs. 3 and 5). Including fossils does improve inference accuracy, but these modes are unique in that randomly distributed fossils provide the greatest increase in accuracy, while root-biased fossils provide the smallest increase in accuracy (Figs. 3 and S3). This does not appear to be related to either particular mode, since strong versions of these modes do not behave this way. Instead, this appears to be related to the weak strength of the alpha parameter. In these cases, the phylogenetic half-life is simulated at 1 tree height, which is fairly weak selection (in the case of OUs) or stabilization (in the case of OUc). Most analyses usually misinterpret these evolutionary histories as BM or trend, and, as the proportion of root-biased fossils and the total size of the phylogeny are increased, these simulations are increasingly misinterpreted as trends (Fig. 7). The misinterpretation as BM is expected because OU models with longer phylogenetic half-lives approach a BM process (Uyeda and Harmon 2014). It is also understandable that the wOUs model is misinterpreted as a trend model, as OUs and trend models have similar trajectories with changes in the mean over time. However, this is extremely concerning for wOUc, as this mode represents evolutionary stasis, which is usually interpreted in a fundamentally different manner to a trend model. Further, increasing phylogeny size, and therefore taxonomic coverage, is often a major goal in phylogenetics (Zwickl and Hillis 2002). The fact that this potential for misinterpretation increases with increasing phylogeny size is counterintuitive, and there is great need for further investigation into this pattern in the future.

Finally, we found mixed behavior in the ACDC family of models. The addition of fossils, especially root-biased fossils, dramatically improves the accuracy of inferring models of decelerating trait evolution, regardless of tempo (DC; Figs. 3 and 6). However, the inclusion of fossils almost always has a negative impact on the inference accuracy of rapidly accelerating models (sAC; Figs. 3 and 6). For weak accelerating models (wAC), random or recent-biased fossils have a similar effect. Further, while increases in phylogeny size usually have a positive influence on inference accuracy for other models, larger phylogenies lead to increased misinterpretation of the sAC model as OUc, even when fossils are included (Fig. 7). It has been shown that an AC model is indistinguishable from an OU model when analyzing ultrametric trees (Uyeda et al. 2015). Slater et al. (2012) found that distinguishing between an AC and an OUc-like model ("SSP") was easier when fossils were included, but we do not find that to be the case here when relative death rate is high. This may be due to slight differences between our OUc model and the OUc-like model of Slater et al. (2012) or the fact that Slater et al. (2012) did not consider trees with more than 377 tips. Regardless, under these parameters, the OUc model is apparently easily misspecified for a trait that has been simulated under the AC model regardless of fossil inclusion, especially with increasing phylogeny size (Fig. 7). It should be noted that these strange behaviors do not occur for AC models simulated with low relative death rates (Fig. S7), but this remains an unlikely scenario across the majority of Earth history (Marshall 2017). Overall, we echo Slater et al. (2012) by strongly recommending caution when trying to fit AC models without fossils, particularly when using large, ultrametric phylogenies.

The inference of some particular evolutionary histories is not improved by the inclusion of fossils, regardless of their number or age distribution. Brownian motion is unique in that the model accuracy is relatively consistent regardless of fossil age, fossil abundance, or even phylogeny size (Figs. 3 and 4). This suggests that the information that can be gleaned from individual tips to reconstruct a character history evolving under Brownian motion is quickly saturated. Additional tips and non-extant tips provide little to no benefit once this saturation point (<50 tips in our simulations) is reached. In fact, the inclusion of fossil tips appears to cause a marginal decrease across the board in the ability to infer Brownian motion as the correct model. It appears that including these fossils results in favor of the trend model, which is otherwise never misspecified when no fossils are included (Fig. 7). We intuit that this occurs when the distribution of the simulated values of the included fossils is, purely by chance, significantly different from the distribution of simulated values for the extant tips, providing support for a changing mean through time. Regardless, inference accuracy of histories simulated under Brownian motion remains relatively high. This would be encouraging; however, the error rate when Brownian motion is recovered as the best-fitting model also is relatively high (Fig. 7). The inclusion of fossils in a phylogeny reduces this error rate, especially for larger trees. Therefore, we argue that the marginal decrease in overall inference accuracy due to the inclusion of fossils is well outweighed by this reduction in error rate. However, caution should be taken when Brownian motion is inferred to be the best-fitting model in an empirical analysis of a small phylogeny, even when fossils are included. Finally, our results suggest lower overall accuracy (~75%) for inferring BM histories than those of previous studies (e.g., >90% in Silvestro et al., 2015). This suggests that the method of model comparison (AIC vs. LRT) and the total number and type of tested models are highly influential when conducting both simulation and empirical studies.

For parameter estimation, the size of the phylogeny appears to be the most important factor for determining the average amount of error across simulations, with increasing phylogeny size leading to decreased degrees of overall error (Figs. 8-12, S8-S12). This is especially apparent for the *σ* parameter of the BM model (Figs. 8 and S8). For this parameter, the inclusion of fossils has no discernible effect on accuracy. On the other hand, increasing phylogeny size appears to have an overall negative effect on the accuracy of the *beta* parameter for the strong AC model with a high relative death rate (Figs. 12). For weak OU simulations, many of the estimates of the half-life parameter are underestimated, regardless of the inclusion of fossils or the size of the phylogeny. This likely explains why these weak OU models are often not selected as the best-fitting models when they are the simulating models (Fig. 3). Regardless, given that these parameter estimates are often used to estimate the strength of selection on different traits or for hypothesis testing (Grabowski et al. 2023), this overestimation of the strength of selection is concerning, and caution should be taken when directly interpreting these parameter estimates.

Fossil inclusion is critical for the estimation of the *trend* parameter for the trend model. When fossils are not included, the *trend* parameter is consistently underestimated to be 0 (Figs. 9 and S9). Given that a trend model with a *trend* parameter of 0 is the same as a BM model, this would explain why trend simulations are usually misidentified (75%-80%) as the BM model when no fossils are included (Fig. 7), since the BM model would give the same fit with one fewer parameter, resulting in a *de facto* better AICc value. When fossils are included, the *trend* parameter is estimated fairly accurately, with a decrease in average estimate error with increasing proportion of fossils and/or phylogeny size. However, the average estimate error increases when recent-biased fossils are used. For weak OU models, estimation of the θ parameter is often very poor regardless of the simulated variables. The inclusion of fossils results in less unbiased estimates, but bizarrely also results in more average estimate error. For strong OU models, the estimated θ parameter is relatively much closer to the simulated value, regardless of the inclusion of fossils or their relative age. It is possible that the mixed performance of the parameter estimation for these OU models may be an artifact of the simulation and model-fitting process that we have developed or inherent to the models themselves, although more investigation is necessary.

Finally, we accounted for a wide variety of sources of variability in evolutionary histories and the use of phylogenetic comparative methods in our simulations, including the size of the phylogeny, the relative death rate, the proportion of fossils included, the relative ages of the fossils, and the tempos and modes of character evolution. However, evolution and the implementation of PCMs are both highly complex and we acknowledge that there are various other sources of variability and error that we did not account for. First, we used a standard birth-death procedure with a single relative death rate to simulate phylogenies with roughly similar tree shape. However, empirical phylogenies indicate that there is a large degree of diversification rate heterogeneity among lineages which results in a wide variety of tree shapes, both in terms of the distributions of branch lengths and tree balance (Mooers and Heard 1997; Martins and Housworth 2002; Boettiger et al. 2012). Second, we assumed that the topology and branch lengths of the simulated phylogeny were completely known. However, in empirical studies this is rarely the case, and inaccuracies related to these two sources are known to cause issues with other comparative methods (Purvis et al. 1994; Symonds 2002). In empirical studies, particularly those using the results of Bayesian phylogenetic analyses or gene tree analyses, it it becoming more common to perform analyses over a sample of phylogenies to account for uncertainty in branch length and topology (Gearty et al. 2018; Gearty and Payne 2020; Grossnickle 2020; Soul and Wright 2021; Hibbins et al. 2023). Third, and relatedly, fossils are notoriously incomplete and are often associated with more topological uncertainty than extant taxa (Sansom and Wills 2013; Pattinson et al. 2015). An increase in the proportion of fossils in a phylogenetic inference would therefore also likely increase the degree of uncertainty in the topology. Future work should investigate whether this source of error is outweighed by the benefits of adding fossils which are described above. Finally, we assumed that character states for all taxa were known without error. However, empirical studies have measurement error due to the intraspecific variability of natural populations and from instrumental/human imprecision, and these sources of error are known to introduce biases in the reconstruction of trait evolution (Ives et al. 2007; Felsenstein et al. 2008; Silvestro et al. 2015). Accounting for all of these sources of variability and error was outside the scope of this project, but future simulation studies should address the interplay of the results presented herein and these other factors.

We acknowledge that researchers working exclusively with sequence data may be concerned about the time and effort required to incorporate fossils into their phylogenies as we have recommended. Historical conventions required that all of the taxa were coded for a suite of morphological characters, which can take much more time than is required to sequence the extant taxa. However, recent work has shown that taxonomic constraints and/or backbones can be just as reliable as morphological matrices when integrating fossils into broader phylogenetic contexts (Barido-Sottani et al. 2023) and conducting phylogenetic comparative methods (Soul and Friedman 2015). Furthermore, our results indicate that only a small number of fossils (e.g., 10% of tips) is required to reap their benefits. Ultimately, including this few fossils may have the same statistical effect as would sequencing and adding thousands of additional extant taxa but for far less work.

All-in-all, we find the results of this simulation study to be encouraging. Although we focus on the instances when simulated and fitted trait evolution models do not match, our intention is to equip readers to make informed decisions about how to plan and carry out their entire research pipelines. This includes, but is not limited to, specimen identification, phylogenetic matrix assembly, phylogenetic inference, phylogenetic comparative analysis, and result interpretation. While there is still much more work to be done in understanding trait data and evolutionary models, we hope that our simulations can be used to highlight clear avenues for future research.

Conclusions

The development and use of phylogenetic comparative methods in evolutionary biology and paleobiology has blossomed since the work of Felsenstein (1985). Although the amount of paleontological information available to biologists has increased over this time, the use of PCMs without fossils has remained abundant in the literature. Here we reiterate that including fossils as tips in phylogenetic analyses results in superior downstream reconstructions of the evolution of continuous characters within a phylogenetic context. More importantly, fossils farther in time from the extant tips potentially hold much more information than their extant-adjacent counterparts and can play a profound role in our ability to accurately infer evolutionary tempo and mode. The models that are used to reconstruct these evolutionary histories have several shortcomings, but the inclusion of fossils helps alleviate many of these. Given all of this, we strongly assert that the vast majority of phylogenetic comparative analyses should include fossil tips moving forward, except for the minority of clades where fossils are not (yet) known. This will lead to a more robust field of study and more reliable biological interpretations. While the inclusion of fossils may present a hurdle to some evolutionary biologists, we argue that the extra work is well worth it: a fossil taxon that is included within a phylogeny is, at worst, just as useful as an extant taxon for model inference, and is, at best, far more useful. Ultimately, collaboration between neontologists and paleontologists provides the surest foundation for maximizing the potential of both genetic and fossil data (Hunt and Slater 2016).

Acknowledgments

W.G. was supported by the Lerner-Gray Postdoctoral Research Fellowship from the Richard Gilder Graduate School at the American Museum of Natural History and a Norman Newell Early Career Grant from the Paleontological Society. B.J.A was supported by ETH Zurich. P.L.G. was supported by [ ]. A.A.C. was supported by [ ].

Data Accessibility Statement

All data and code are archived on GitHub at https://github.com/willgearty/PCM_sensitivity.

Conflict of Interest Disclosure

The authors declare no conflicts of interest.

References

Allen B.J., Stubbs T.L., Benton M.J., Puttick M.N. 2019. Archosauromorph extinction selectivity during the Triassic–Jurassic mass extinction. Palaeontology. 62:211–224.

Álvarez-Carretero S., Tamuri A.U., Battini M., Nascimento F.F., Carlisle E., Asher R.J., Yang Z., Donoghue P.C.J., dos Reis M. 2022. A species-level timeline of mammal evolution integrating phylogenomic data. Nature. 602:263–267.

Anelli V., Bars-Closel M., Herrel A., Kohlsdorf T. 2024. Different selection regimes explain morphological evolution in fossorial lizards. Funct. Ecol. n/a.

Barido-Sottani J., Pohle A., De Baets K., Murdock D., Warnock R.C.M. 2023. Putting the F into FBD analysis: tree constraints or morphological data? Palaeontology. 66:e12679.

Barido-Sottani J., Saupe E.E., Smiley T.M., Soul L.C., Wright A.M., Warnock R.C.M. 2020. Seven rules for simulations in paleobiology. Paleobiology. 46:435–444.

Beaulieu J.M., Jhwueng D.-C., Boettiger C., O'Meara B.C. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evol. Int. J. Org. Evol. 66:2369–83.

Benson R.B.J., Frigot R.A., Goswami A., Andres B., Butler R.J. 2014. Competition and constraint drove Cope's rule in the evolution of giant flying reptiles. Nat. Commun. 5:3567.

Blomberg S.P., Garland T., Ives A.R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evol. Int. J. Org. Evol. 57:717–745.

Boettiger C., Coop G., Ralph P. 2012. IS YOUR PHYLOGENY INFORMATIVE? MEASURING THE POWER OF COMPARATIVE METHODS. Evolution. 66:2240–2251.

Burnham K.P., Anderson D.R. 2002. Model Selection and Multimodel Inference. New York, NY: Springer New York.

Butler M., King A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695.

Clavel J., Escarguel G., Merceron G. 2015. mvMORPH: An R package for fitting multivariate evolutionary models to morphometric data. Methods Ecol. Evol. 6:1311–1319.

Cooper N., Thomas G.H., Venditti C., Meade A., Freckleton R.P. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. Biol. J. Linn. Soc. 118:64–77.

Cornwell W., Nakagawa S. 2017. Phylogenetic comparative methods. Curr. Biol. 27:R333–R336.

Felsenstein J. 1985. Phylogenies and the Comparative Method. Am. Nat. 125:1–15.

Felsenstein J., Otto A.E.S.P., Whitlock E.M.C. 2008. Comparative Methods with Sampling Error and Within‐Species Variation: Contrasts Revisited and Revised. Am. Nat. 171:713–725.

Finarelli J.A., Flynn J.J. 2006. Ancestral State Reconstruction of Body Size in the Caniformia (Carnivora, Mammalia): The Effects of Incorporating Data from the Fossil Record. Syst. Biol. 55:301–313.

Finarelli J.A., Goswami A. 2013. Potential pitfalls of reconstructing deep time evolutionary history with only extant data, a case study using the Canidae (Mammalia, Carnivora). Evolution. 67:3678–3685.

Gauthier J. 1986. Saurischian monophyly and the origin of birds. Mem. Calif. Acad. Sci. 8:1--55.

Gearty W., Carrillo E., Payne J.L. 2021. Ecological Filtering and Exaptation in the Evolution of Marine Snakes. https://doi.org/10.1086/716015. 198:506–521.

Gearty W., McClain C.R., Payne J.L. 2018. Energetic tradeoffs control the size distribution of aquatic mammals. Proc. Natl. Acad. Sci. 115:4194–4199.

Gearty W., Payne J.L. 2020. Physiological constraints on body size distributions in Crocodyliformes. Evolution. 74:245–255.

Gill M.S., Tung Ho L.S., Baele G., Lemey P., Suchard M.A. 2017. A Relaxed Directional Random Walk Model for Phylogenetic Trait Evolution. Syst. Biol. 66:299–319.

Godoy P.L., Benson R.B.J., Bronzati M., Butler R.J. 2019. The multi-peak adaptive landscape of crocodylomorph body size evolution. BMC Evol. Biol. 19:167.

Grabowski M., Pienaar J., Voje K.L., Andersson S., Fuentes-González J., Kopperud B.T., Moen D.S., Tsuboi M., Uyeda J., Hansen T.F. 2023. A cautionary note on "A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies." Syst. Biol. 72:955–963.

Grossnickle D.M. 2020. Feeding ecology has a stronger evolutionary influence on functional morphology than on body mass in mammals. Evolution. 74:610–628.

Hansen T.F. 1997. Stabilizing Selection and the Comparative Analysis of Adaptation. Evolution. 51:1341.

Harmon L.J. 2019. Phylogenetic Comparative Methods: Learning From Trees. EcoEvoRxiv.

Harmon L.J., Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings W., Kozak K.H., McPeek M. a, Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte Ii J. a, Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., Mooers A.Ø. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evol. Int. J. Org. Evol. 64:2385–96.

Harmon L.J., Weir J.T., Brock C.D., Glor R.E., Challenger W. 2008. GEIGER: investigating evolutionary radiations. Bioinforma. Oxf. Engl. 24:129–31.

Hibbins M.S., Breithaupt L.C., Hahn M.W. 2023. Phylogenomic comparative methods: Accurate evolutionary inferences in the presence of gene tree discordance. Proc. Natl. Acad. Sci. 120:e2220389120.

Hunt G., Slater G. 2016. Integrating Paleontological and Phylogenetic Approaches to Macroevolution. Annu. Rev. Ecol. Evol. Syst. 47:189–213.

Ives A.R., Midford P.E., Garland T. 2007. Within-Species Variation and Measurement Error in Phylogenetic Comparative Methods. Syst. Biol. 56:252–270.

Jaffe A.L., Slater G.J., Alfaro M.E. 2011. The evolution of island gigantism and body size variation in tortoises and turtles. Biol. Lett. 7:558–561.

Jetz W., Pyron R.A. 2018. The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. Nat. Ecol. Evol. 2:850–858.

Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature. 491:444–448.

Louca S., Pennell M.W. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature. 580:502–505.

Magallón S., Gómez-Acevedo S., Sánchez-Reyes L.L., Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. 207:437–453.

Marshall C.R. 2017. Five palaeobiological laws needed to understand the evolution of the living biota. Nat. Ecol. Evol. 1:1–6.

Martins E.P., Housworth E.A. 2002. Phylogeny Shape and the Phylogenetic Comparative Method. Syst. Biol. 51:873–880.

Mitchell J.S. 2015. Extant-only comparative methods fail to recover the disparity preserved in the bird fossil record. Evolution. 69:2414–2424.

Mongiardino Koch N., Garwood R.J., Parry L.A. 2021. Fossils improve phylogenetic analyses of morphological characters. Proc. R. Soc. B Biol. Sci. 288:20210044.

Mongiardino Koch N., Parry L.A. 2020. Death is on Our Side: Paleontological Data Drastically Modify Phylogenetic Hypotheses. Syst. Biol. 69:1052–1067.

Mooers A.O., Heard S.B. 1997. Inferring Evolutionary Process from Phylogenetic Tree Shape. Q. Rev. Biol. 72:31–54.

O'Meara B.C., Ané C., Sanderson M.J., Wainwright P.C. 2006. TESTING FOR DIFFERENT

RATES OF CONTINUOUS TRAIT EVOLUTION USING LIKELIHOOD. Evolution. 60:922.

Pagel M. 2002. Modelling the evolution of continuously varying characters on phylogenetic trees: the case of Hominid cranial capacity. Morphology, Shape and Phylogeny. CRC Press.

Pattinson D.J., Thompson R.S., Piotrowski A.K., Asher R.J. 2015. Phylogeny, Paleontology, and Primates: Do Incomplete Fossils Bias the Tree of Life? Syst. Biol. 64:169–186.

Pennell M.W., FitzJohn R.G., Cornwell W.K., Harmon L.J. 2015. Model Adequacy and the Macroevolution of Angiosperm Functional Traits. Am. Nat. 186:E33–E50.

Pie M.R., Divieso R., Caron F.S. 2023. Clade density and the evolution of diversity-dependent diversification. Nat. Commun. 14:4576.

Price S.A., Hopkins S.S.B. 2015. The macroevolutionary relationship between diet and body mass across mammals. Biol. J. Linn. Soc. 115:173–184.

Purvis A., Gittleman J.L., Luh H.-K. 1994. Truth or Consequences: Effects of Phylogenetic Accuracy on Two Comparative Methods. J. Theor. Biol. 167:293–300.

Rabosky D.L., Chang J., Title P.O., Cowman P.F., Sallan L., Friedman M., Kaschner K., Garilao C., Near T.J., Coll M., Alfaro M.E. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature. 559:392–395.

Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012. A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera. Syst. Biol. 61:973–999.

Royer-Carenzi M., Pontarotti P., Didier G. 2013. Choosing the best ancestral character state reconstruction method. Math. Biosci. 242:95–109.

Sansom R.S., Wills M.A. 2013. Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. Sci. Rep. 3:2545.

Schrago C.G., Mello B., Soares A.E.R. 2013. Combining fossil and molecular data to date the diversification of New World Primates. J. Evol. Biol. 26:2438–2446.

Silvestro D., Kostikova A., Litsios G., Pearman P.B., Salamin N. 2015. Measurement errors should always be incorporated in phylogenetic comparative analysis. Methods Ecol. Evol. 6:340–346.

Slater G.J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary. Methods Ecol. Evol. 4:734–744.

Slater G.J., Harmon L.J., Alfaro M.E. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution. 66:3931–3944.

Smyčka J., Toszogyova A., Storch D. 2023. The relationship between geographic range size and rates of species diversification. Nat. Commun. 14:5559.

Soul L.C., Friedman M. 2015. Taxonomy and Phylogeny Can Yield Comparable Results in Comparative Paleontological Analyses. Syst. Biol. 64:608–620.

Soul L.C., Wright D.F. 2021. Phylogenetic Comparative Methods: A User’s Guide for Paleontologists. Elem. Paleontol.

Springer M.S., Teeling E.C., Madsen O., Stanhope M.J., de Jong W.W. 2001. Integrated fossil and molecular data reconstruct bat echolocation. Proc. Natl. Acad. Sci. 98:6241–6246.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. 108:6187–6192.

Stadler T. 2019. TreeSim: Simulating Phylogenetic Trees. .

Sugiura N. 1978. Further analysts of the data by akaike’ s information criterion and the finite corrections. Commun. Stat. - Theory Methods. 7:13–26.

Symonds M.R.E. 2002. The Effects of Topological Inaccuracy in Evolutionary Trees on the Phylogenetic Comparative Method of Independent Contrasts. Syst. Biol. 51:541–553.

Upham N.S., Esselstyn J.A., Jetz W. 2019. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLOS Biol.

17:e3000494.

Uyeda J.C., Caetano D.S., Pennell M.W. 2015. Comparative Analysis of Principal Components Can be Misleading. Syst. Biol. 64:677–689.

Uyeda J.C., Harmon L.J. 2014. A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. Syst. Biol. 63:902–918.

Vrba E.S. 1979. Phylogenetic analysis and classification of fossil and recent Alcelaphini Mammalia: Bovidae. Biol. J. Linn. Soc. 11:207–228.

Warnock R., Barido-Sottani J., Pett W., O'Reilly J. 2022. FossilSim: Simulation of Fossil and Taxonomy Data. .

Webster A.J., Purvis A. 2002. Testing the accuracy of methods for reconstructing ancestral states of continuous characters. Proc. R. Soc. Lond. B Biol. Sci. 269:143–149.

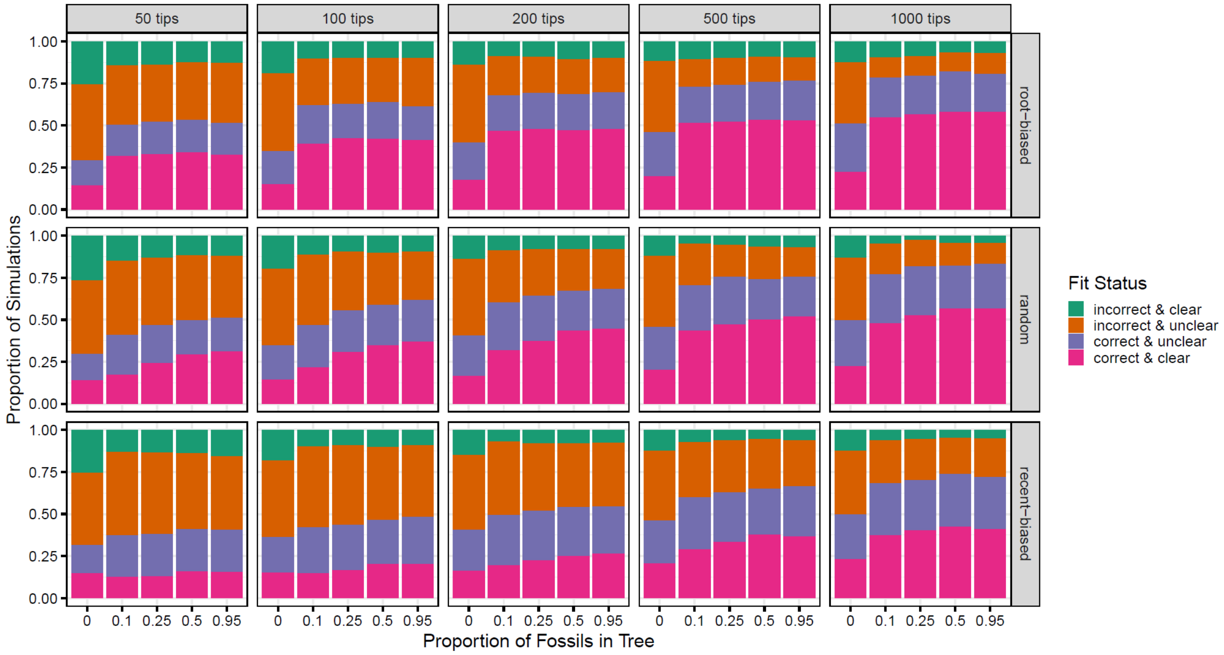
Wright A.M., Bapst D.W., Barido-Sottani J., Warnock R.C.M. 2022. Integrating Fossil Observations Into Phylogenetics Using the Fossilized Birth–Death Model. Annu. Rev. Ecol. Evol. Syst. 53:251–273.

Zhang C., Stadler T., Klopfstein S., Heath T.A., Ronquist F. 2016. Total-Evidence Dating under the Fossilized Birth–Death Process. Syst. Biol. 65:228–249.

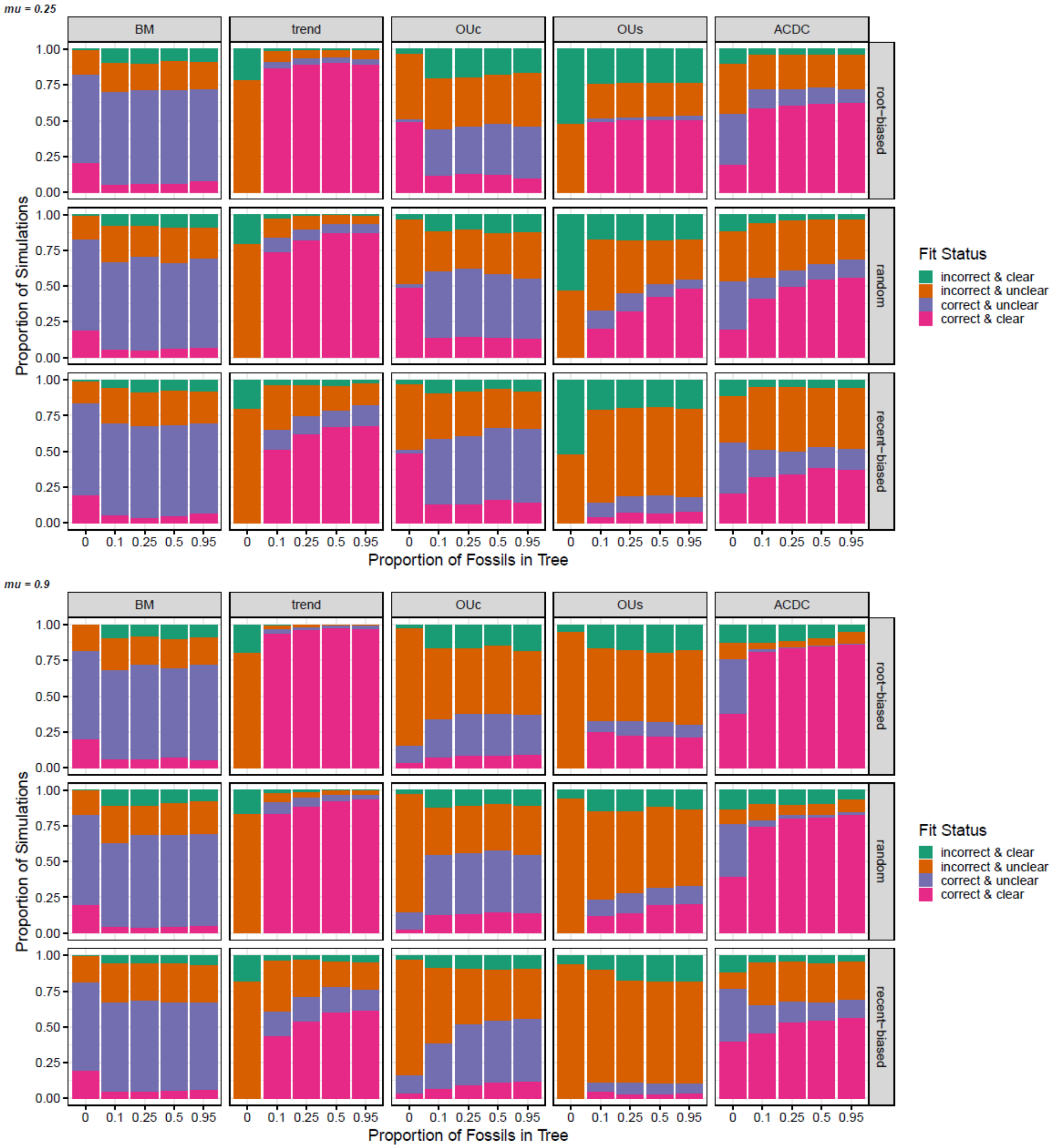
Zou Z., Zhang J. 2016. Morphological and molecular convergences in mammalian phylogenetics. Nat. Commun. 7:12758.

Zwickl D.J., Hillis D.M. 2002. Increased Taxon Sampling Greatly Reduces Phylogenetic Error. Syst. Biol. 51:588–598.

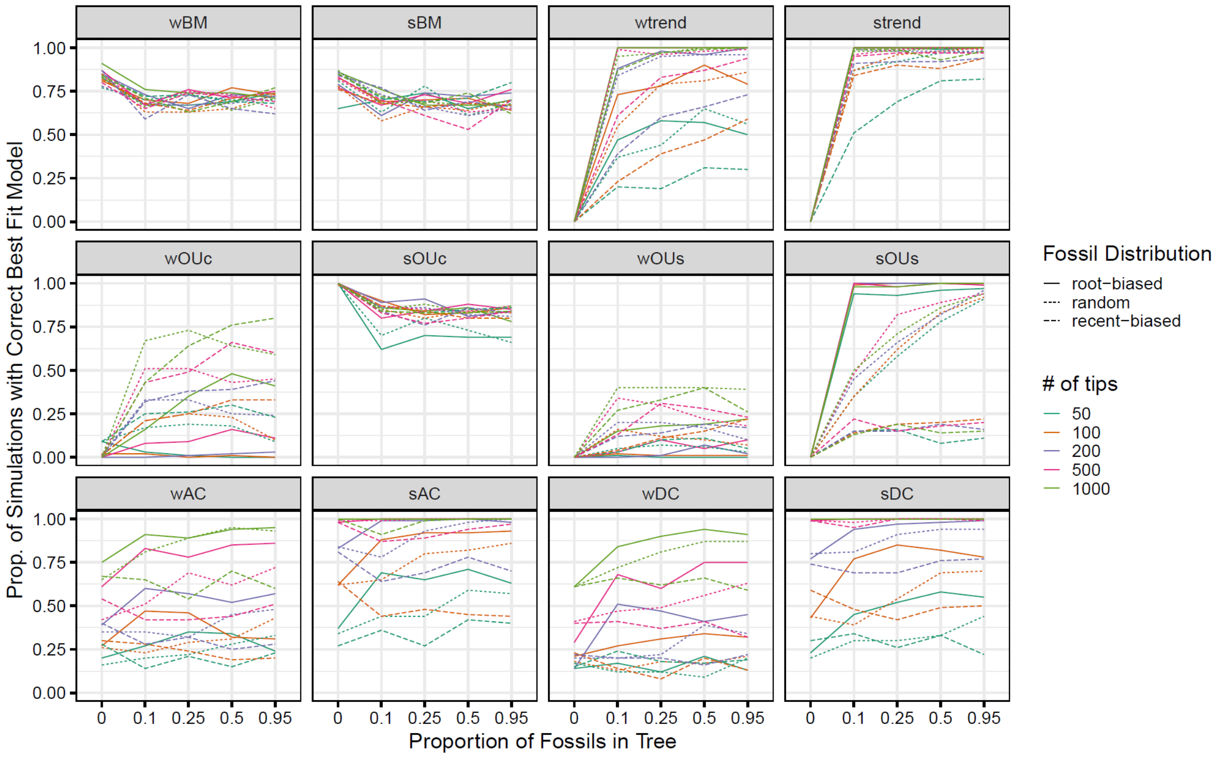
Supplement



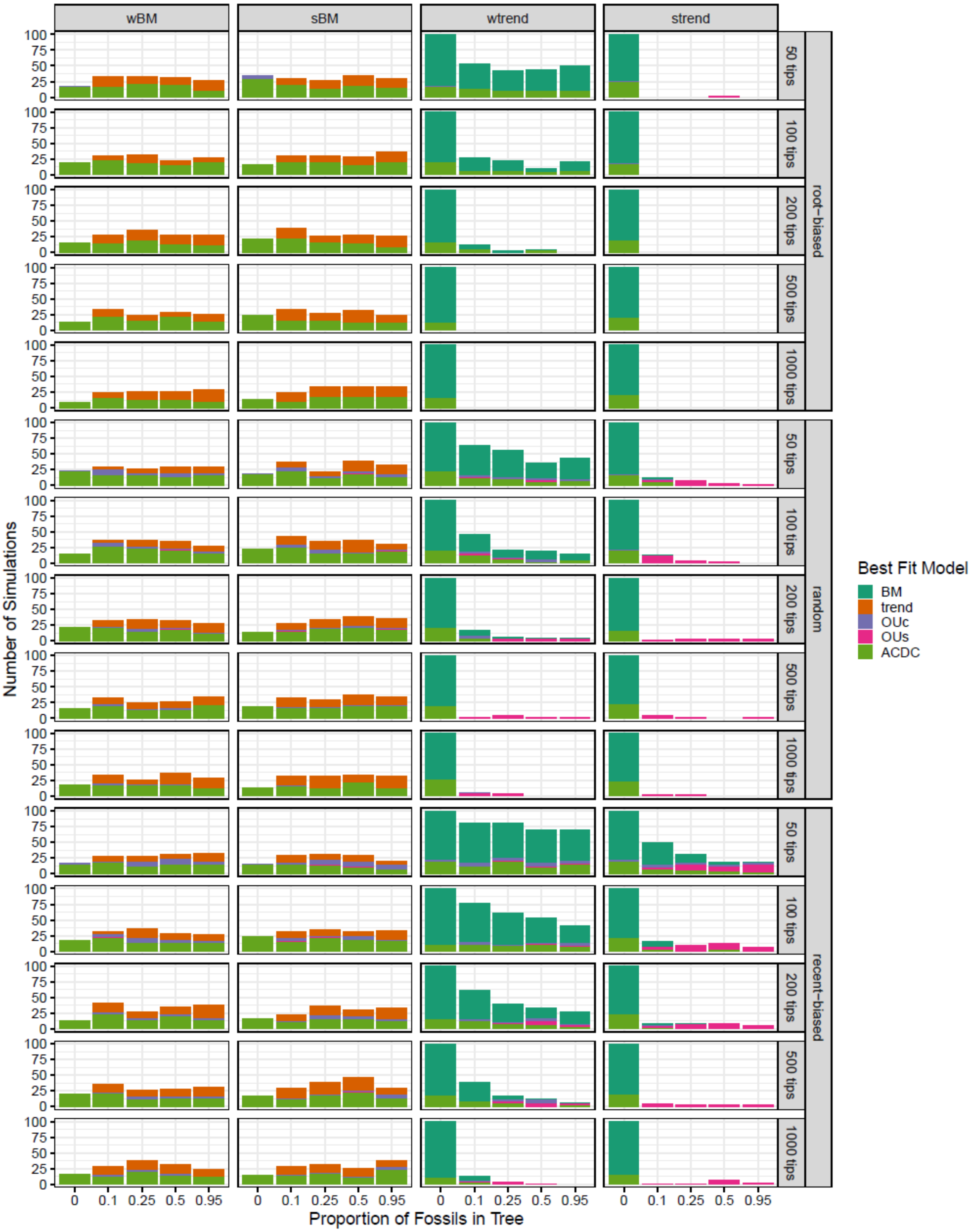
Supplemental Figure 1: Summary of the best-fitting models across all simulations. The best-fitting model for a given simulation can be either correct (matching the simulated model) or incorrect (not matching the simulated model) and also either clear (all other models with ΔAICc > 2) or unclear (at least one other model with ΔAICc ≤ 2). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row headers), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 2.



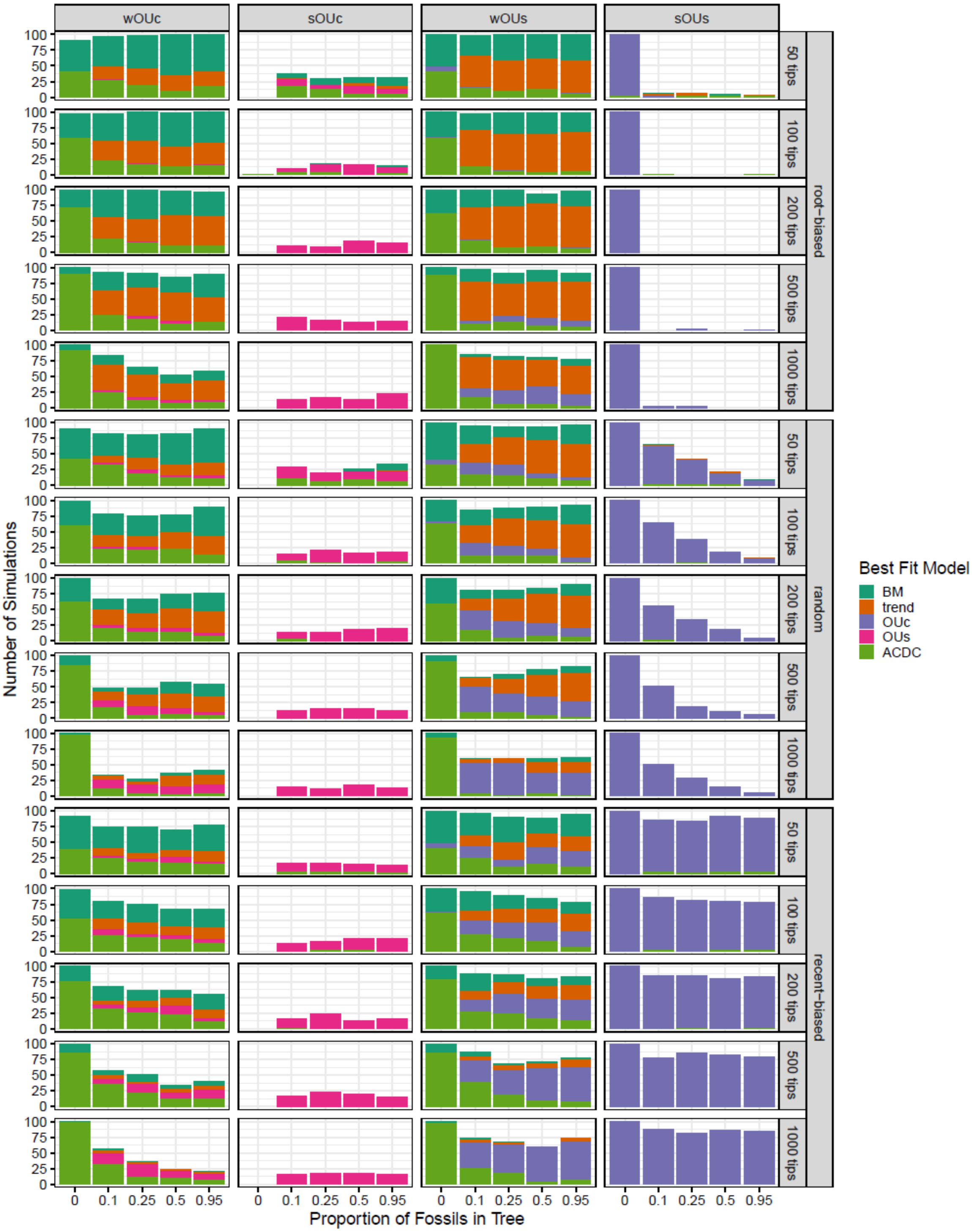
Supplemental Figure 2: Alternate summary of the best-fitting models across all simulations. The best-fitting model for a given simulation can be either correct (matching the simulated model) or incorrect (not matching the simulated model) and also either clear (all other models with ΔAICc > 2) or unclear (at least one other model with ΔAICc ≤ 2). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row headers), and the type of simulated model (column headers). The upper panel shows the results for analyses where *mu* = 0.25, and the lower panel shows the results for analyses where *mu* = 0.9.



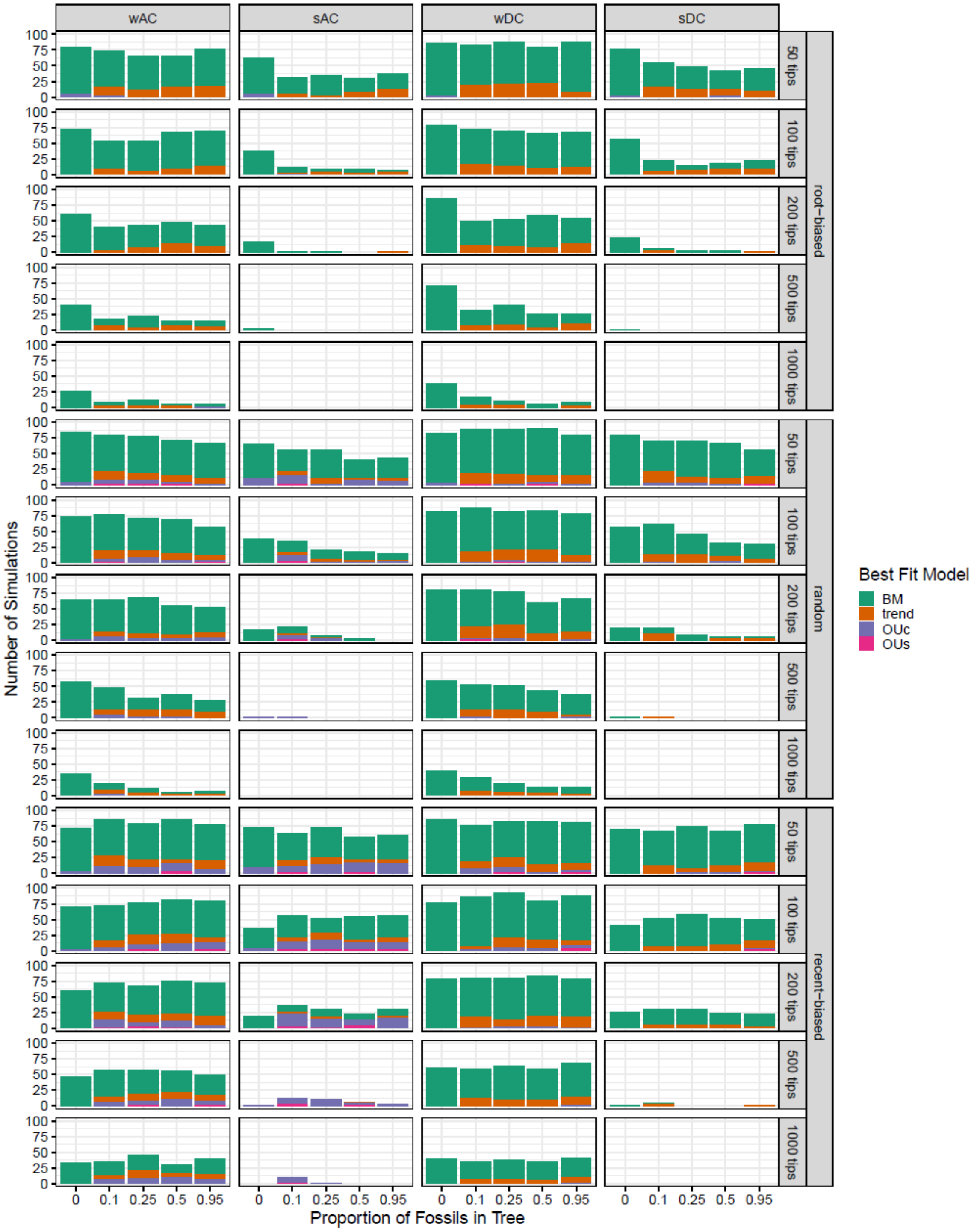
Supplemental Figure 3: The proportion of simulations for which the best-fit model does match the simulated model (panel headers). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (line type), and the number of tips in the simulated phylogenies (line color). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 3.



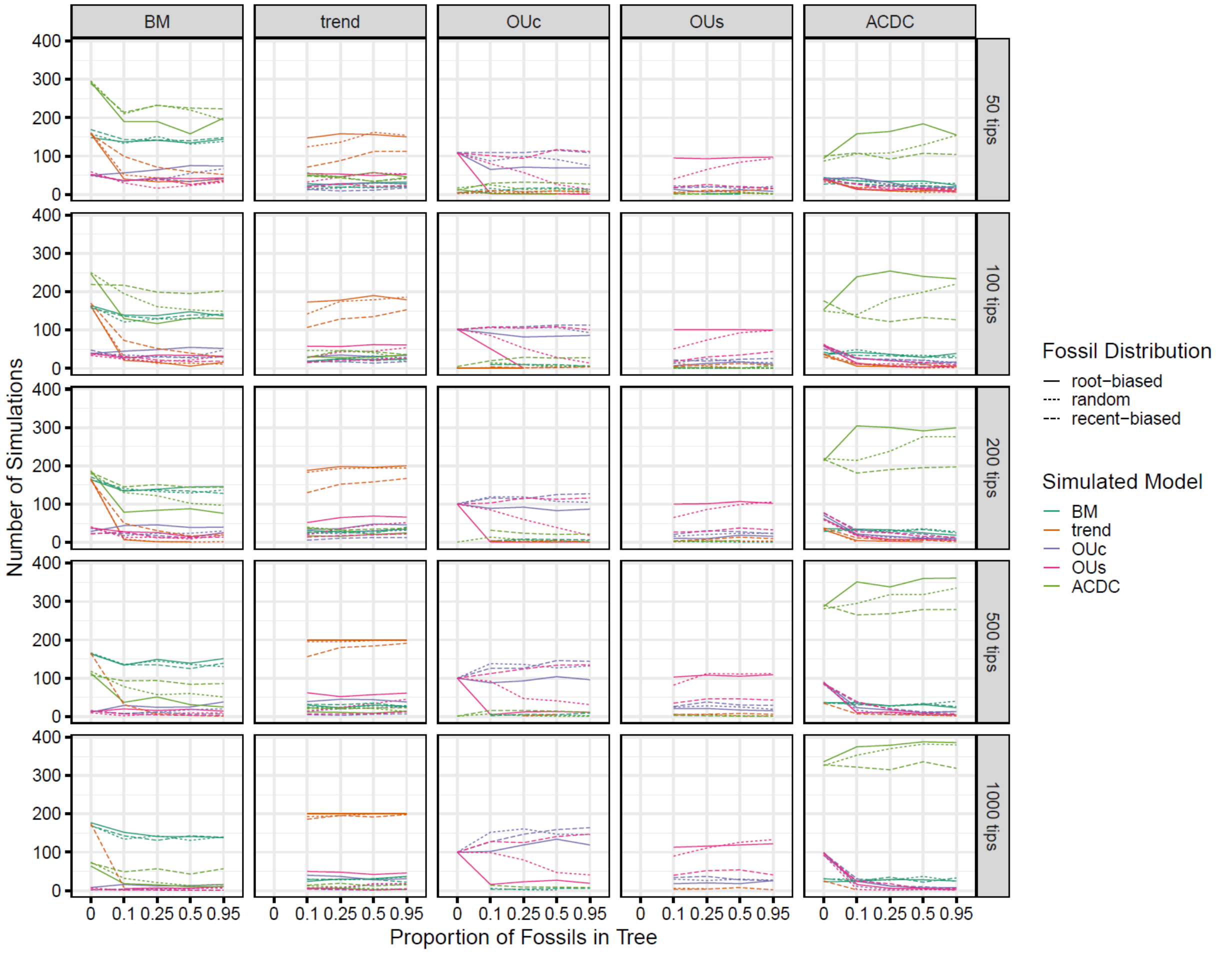
Supplementary Figure 4: The number of BM/trend simulations for which the best-fit model (color) does not match the simulated model (panel headers). Only BM and trend simulations are included (see Figures S5 and S6 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 4.



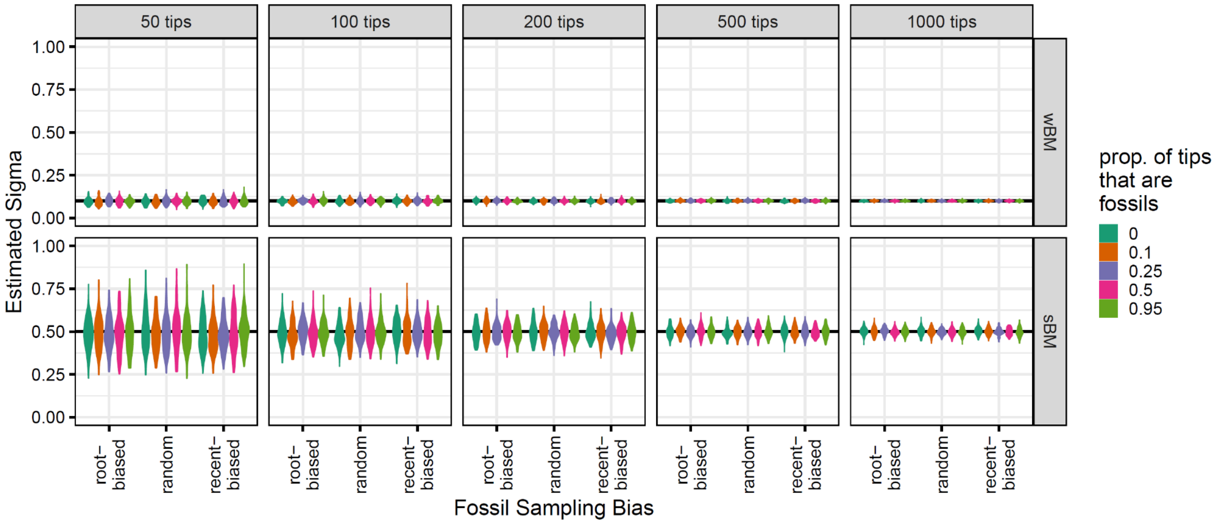
Supplementary Figure 5: The number of OU simulations for which the best-fit model (color) does not match the simulated model (panel headers). Only OU simulations are included (see Figures S4 and S6 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 5.



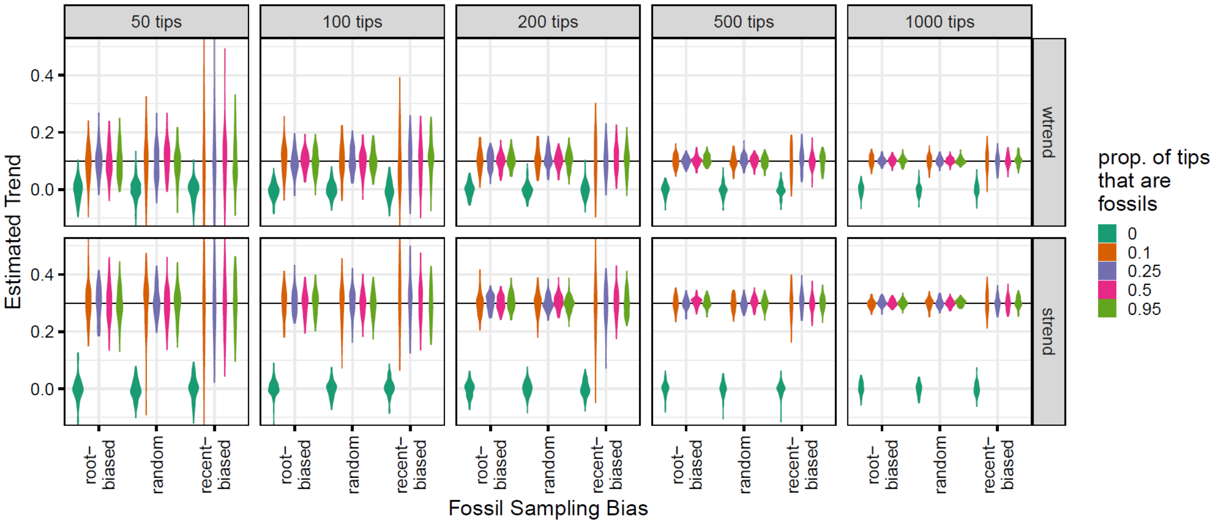
Supplementary Figure 6: The number of BM/trend simulations for which the best-fit model (color) does not match the simulated model (panel headers). Only BM and trend simulations are included (see Figures S4 and S5 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 6.



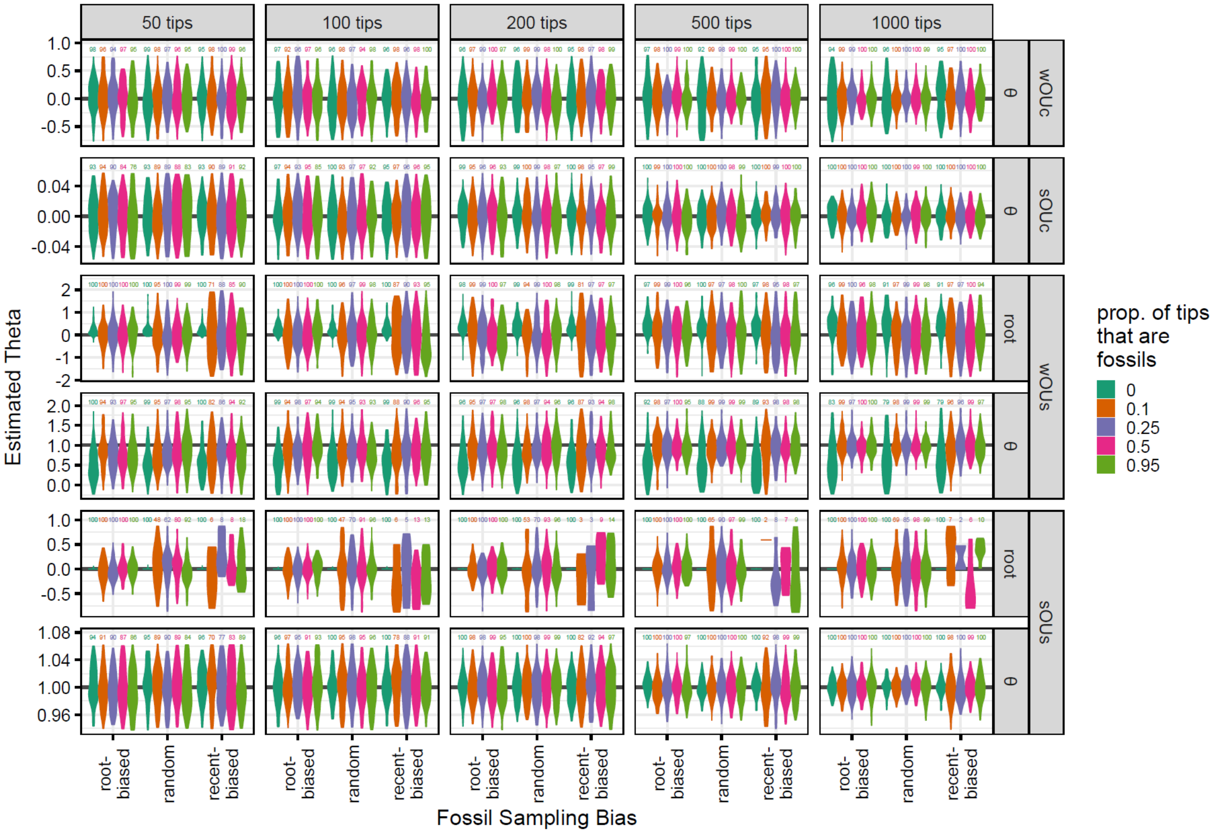
Supplemental Figure 7: The proportions of different generative models (line color) when particular models were best-fitting (column headers). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the age distribution of the simulated fossils (line type), and the number of tips in the simulated phylogenies (row headers). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 7.



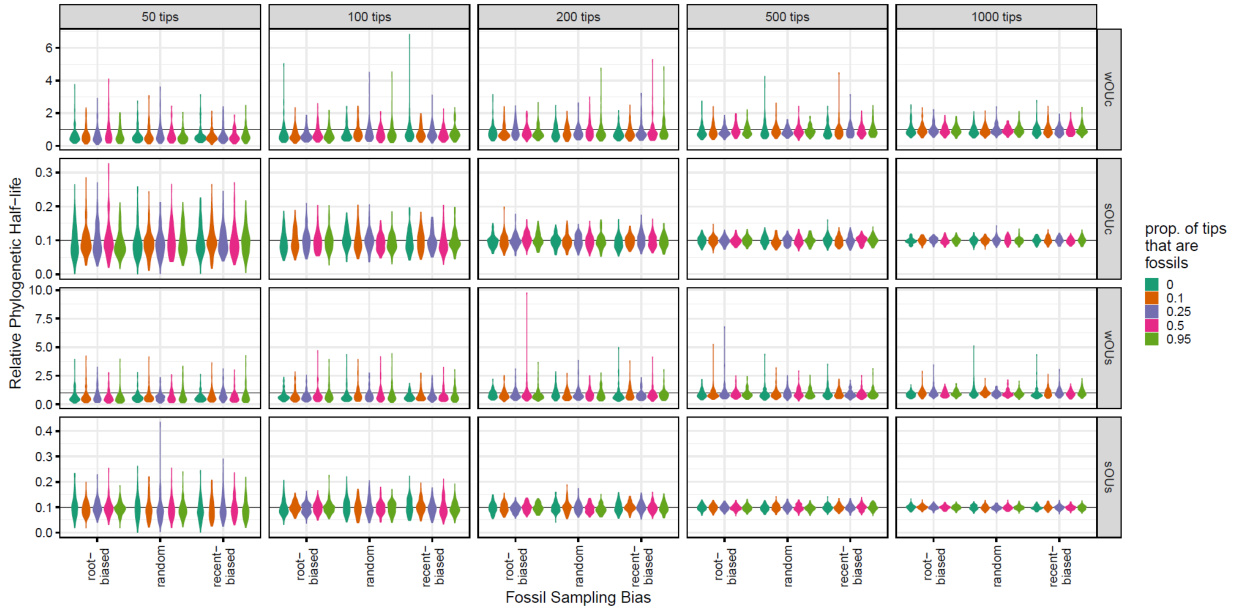
Supplemental Figure 8: The distributions of the estimates for the *σ* parameter when a trend model was simulated on a phylogeny and then a trend model was fit to that simulated data. The top row represents simulated trends with a weak *σ* parameter (0.1) and the bottom row represents a trend model simulated with a strong *σ* parameter (0.5). The true (simulated) *σ* parameter is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 8.



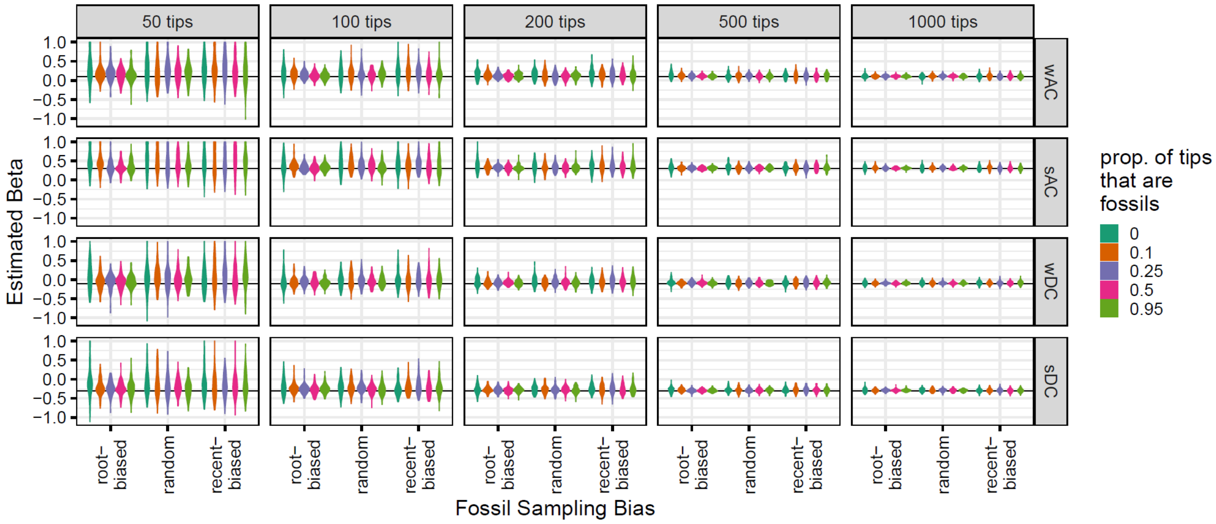
Supplemental Figure 9: The distributions of the estimates for the *trend* parameter when a trend model was simulated on a phylogeny and then a trend model was fit to that simulated data. The top row represents simulated trends with a weak *trend* parameter (0.1) and the bottom row represents a trend model simulated with a strong *trend* parameter (0.3). The true (simulated) *trend* parameter is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 9.



Supplemental Figure 10: The distributions of the estimates for the θ parameter when an OU model was simulated on a phylogeny and then an OU model was fit to that simulated data. The true (simulated) θ parameter is represented with a solid horizontal line. Outliers have been removed separately for each row, and the numbers above box plots represent the number of non-outliers that are represented by each boxplot. Note that the scale of the y-axis varies for each row. The results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 10.



Supplemental Figure 11: The distributions of the estimates for the relative half-life when an OU model was simulated on a phylogeny and then an OU model was fit to that simulated data. Relative half-life is calculated as $\frac{log(2)\ /\ \alpha}{tree\ height}$. The first and second rows represent centered OU models simulated with weak (1) and strong (0.1) relative half-lives, respectively. The third and fourth rows represent shifting OU models simulated with weak (1) and strong (0.1) relative half-lives, respectively. The true (simulated) relative half-life is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 11.



Supplemental Figure 12: The distributions of the estimates for the *beta* parameter when an ACDC model was simulated on a phylogeny and then an ACDC model was fit to that simulated data. The true (simulated) *beta* parameter is represented with a solid horizontal line. The results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where *mu* = 0.25, the results for analyses where *mu* = 0.9 are in Figure 12.