Impact of Phylogenetic Tree Completeness and Misspecification of Sampling Fractions on Trait Dependent Diversification Models

Mynard, Poppy; Algar, Adam ; Lancaster, Lesley T.; Bocedi, Greta; Fahri, Fahri; Gubry-Rangin, Cecile; Lupiyaningdyah, Pungki; Nangoy, Meis; Osborne, Owen; Papadopulos, Alexander S. T.; Sudiana, I Made; Juliandi, Berry; Travis, Justin; Herrera-Alsina, Leonel

Systematic Biology

DOI:
10.1093/sysbio/syad001

E-pub ahead of print: 16/01/2023

Peer reviewed version

Impact of Phylogenetic Tree Completeness and Misspecification of Sampling Fractions on Trait Dependent Diversification Models

Poppy Mynard[1], Adam Algar[2], Lesley Lancaster[1], Greta Bocedi[1], Fahri Fahri[3], Cécile Gubry-Rangin[1], Pungki Lupiyaningdyah[4], Meis Nangoy[5], Owen Osborne[6], Alexander Papadopulos[6], I Made Sudiana[7], Berry Juliandi[8], Justin Travis[1], Leonel Herrera-Alsina[1]

[1] *School of Biological Sciences, University of Aberdeen, Aberdeen, UK*

[2] *Department of Biology, Lakehead University, Thunder Bay, Ontario, Canada*

[3] *Department of Biology, Tadulako University, Palu, Indonesia*

[4] *Museum Zoologicum Bogoriense, Research Center for Biology, Indonesian Institute of Sciences, Cibinong, 16911, Indonesia*

[5] *Faculty of Animal Husbandry, Sam Ratulangi University, Kampus Bahu Street, Manado, 95115 Indonesia*

[6] *School of Natural Sciences, Bangor University, Environment Centre Wales, Deiniol Road, Bangor, LL57 2UW, UK*

[7] *Microbial Ecology Research Group, Research Center for Biology, Indonesian Institute of Sciences, Cibinong 19611, Indonesia*

[8] *Department of Biology, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, 16680, Indonesia*

Corresponding Author

Poppy Mynard: poppymynard@hotmail.com

Abstract

Understanding the origins of diversity and the factors that drive some clades to be more diverse than others are important issues in evolutionary biology. Sophisticated SSE (state-dependent speciation and extinction) models provide insights into the association between diversification rates and the evolution of a trait. The empirical data used in SSE models and other methods is normally imperfect, yet little is known about how this can affect these models. Here, we evaluate the impact of common phylogenetic issues on inferences drawn from SSE models. Using simulated phylogenetic trees and trait information, we fitted SSE models to determine the effects of sampling fraction (phylogenetic tree completeness) and sampling fraction misspecification on model selection and parameter estimation (speciation, extinction, and transition rates) under two sampling regimes (random and taxonomically biased). As expected, we found that both model selection and parameter estimate accuracies are reduced at lower sampling fractions (i.e., low tree completeness). Furthermore, when sampling of the tree is imbalanced across subclades and tree completeness is $\leq 60\%$, rates of false positives increase and parameter estimates are less accurate, compared to when sampling is random. Thus, when applying SSE methods to empirical datasets, there are increased risks of false inferences of trait dependent diversification when some sub-clades are heavily under-sampled. Mis-specifying the sampling fraction severely affected the accuracy of parameter estimates: parameter values were over-estimated when the sampling fraction was specified as lower than its true value, and under-estimated when the sampling fraction was specified as higher than its true value. Our results suggest that it is better to cautiously

under-estimate sampling efforts, as false positives increased when the sampling fraction was over-estimated. We encourage SSE studies where the sampling fraction can be reasonably estimated and provide recommended best practices for SSE modelling.

Keywords: Trait dependent diversification, SSE models, phylogenetic tree completeness, sampling fraction

Understanding the origins of diversity and why some clades are more diverse than others are fundamental questions in evolutionary biology. Estimating the diversification rate of lineages, and how they vary across phylogenetic trees, is essential to developing this understanding. The per-lineage rates of speciation and extinction in a clade can be affected by environmental and geographical factors (Ricklefs 2007), time (or clade age; Henao Diaz et al. 2019), and species' traits (Jablonski 2008; Rabosky and McCune 2010). Possessing a certain trait state can promote speciation by increasing fitness or reproductive output (Liem 1973; Weber and Agrawal 2014; Helmstetter et al. 2016; Laenen et al. 2016; Igea et al. 2017). Conversely, some trait states may hinder speciation or increase extinction rates; for example, specialisation has been shown to reduce speciation rates in some clades (Day et al. 2016), but promotes it in others (Resl et al. 2018; Tonini et al. 2020). Although differences in diversification rates across trait states occur frequently (Jablonski 2008), it does not necessarily mean that the trait itself drives alone the clades' diversification dynamics.

The State-dependent Speciation and Extinction (SSE) framework was developed to determine the impact that the evolution of a trait has on subsequent patterns of lineage diversification through time, by linking the presence/absence (or value) of trait states to diversification rates. To implement this approach and determine the most likely mode of diversification, Examined Trait Dependent (ETD) models, which include the trait hypothesised to affect diversification, are compared to models with Constant Diversification Rates (CR models) and Concealed Trait Dependent (CTD) models (also known as Character-Independent (CID) models) (Beaulieu and O'Meara 2016). CTD models account for the possibility that diversification rates do not vary in relation to the focal trait, but rather with some unmeasured trait(s). The CTD model is always as complex (in terms of the number of character states) as the ETD model (in CID notation, the number of character states has to be specified e.g., CID-2 has two character states). The use of CTD models successfully reduces

false inferences of trait dependent diversification that were problematic in earlier SSE tools (Rabosky and Goldberg 2015; Beaulieu and O'Meara 2016) such as BiSSE (Maddison et al. 2007a) and QuaSSE (Fitzjohn 2010). The improved suit of SSE models includes HiSSE (hidden-state-dependent speciation and extinction) (Beaulieu and O'Meara 2016), GeoHiSSE (a biogeographical version) (Caetano et al. 2018), MuHiSSE (a multi-trait state version) (Nakovet al. 2019; Zenil-Ferguson et al. 2019; Román-Palacios et al. 2020), MiSSE (Vasconcelos et al. 2022), and SecSSE (several examined and concealed state dependant speciation and extinction) (Herrera-Alsina et al. 2019).

The accuracy of model selection (i.e., comparing ETD, CTD and CR models) and parameter estimation (speciation, extinction, and transition rate estimates) in state-dependent diversification analyses is dependent on several factors, including: the similarity of true speciation rates across trait states, the number of transitions among trait states, completeness of trait information, and accuracy and completeness of the phylogenetic tree. It is harder to find evidence for trait dependent diversification when speciation rates are similar to each other (Beaulieu and O'Meara 2016) and when extinction rates are high (Herrera-Alsina et al. 2019). Incomplete trait information adds uncertainty to models: a missing tip in the phylogenetic tree or lack of trait information for an extant species will render the contribution of that branch to the analysis null. Some tools, such as SecSSE, can account for partial trait state data which reduces the negative effect of missing trait information (Herrera-Alsina et al. 2019). Missing tips can result in loss of trait state transitions; this is a crucial piece of information for SSE modelling and lost transitions will likely negatively impact the analysis. Moreover, phylogenetic trees are sources of uncertainty themselves, with potential inaccuracies both in the topology of the tree and the branch lengths (Felsenstein 1985; Donoghue and Ackerly 1996).

Diversification analyses are normally intended for complete clades, but phylogenetic trees may be incompletely sampled, or only a (potentially polyphyletic) subset of taxa may be chosen for analysis. For example, the number of species/phylotypes in the entire clade may be unknown (e.g., in microbes) or the taxonomic scope of a study could be geographically limited. In SSE models, the sampling fraction is the percentage of taxa in the clade included in the phylogenetic tree (Nee et al. 1994; Fitzjohn et al. 2009; Chang et al. 2020), is specified separately in each trait state. The sampling fraction setting typically assumes that taxa are missing from the clade at random; however, this is rarely true and does not reflect realistic sampling associated with empirical phylogenies. Older, more abundant lineages are less likely to be excluded than younger lineages in incompletely sampled trees (Davies et al. 2011). In some studies, certain sub-clades, for example those from the tropics in contrast to temperate counterparts, may be highly under sampled or completely excluded, potentially affecting the accuracy of SSE analysis (Titley et al. 2017). In studies that focus on a specific geographic region, species from outside that region would be removed. Plant collection biases have been well documented (Rich and Woodruff 1992; Moerman and Estabrook 2006; Corlett 2016; Daru et al. 2018) and some families (e.g., Asteraceae, Cyperaceae, and Poaceae) have phylogenetic biases in collection frequency (Daru et al. 2018), so sampling is neither uniform nor random across these phylogenies. Animals that are particularly prone to taxonomic bias include invertebrates such as Insecta, Arachnida, and Gastropoda (Troudet et al. 2017). Larger animals are not exempt from biases; for example, there is a lack of genetic data for bird species from tropical regions (Reddy 2014). Indeed, there is a general pattern of under-sampling of tropical species compared to those in temperate areas (Chek et al. 2003; Collen et al. 2008; Titley et al. 2017). Microbial phylogenetic reconstructions are likely to almost always be under-sampled and there may often be biases related to geography or

environmental conditions, for example with higher sampling occurring under certain pH conditions (Gubry-Rangin et al. 2015).

In cases where the total number of species in a clade is unknown, the sampling fraction could be mis-specified, and this too may affect diversification analyses. Issues with clade specific sampling factions have previously been documented (Beaulieu 2020) but little work has been done on the effects of sampling extent, tree imbalance, and mis-specification on SSE model accuracy (Nee et al. 1994; Fitzjohn et al. 2009; Chang et al. 2020). However, understanding the consequences of different sampling regimes on model comparison and parameter estimation is paramount to give confidence to SSE studies with incomplete phylogenetic trees. Conducting SSE analyses in a Bayesian framework may be a useful method to reduce uncertainty around parameter estimation and improve the ability to detect signals of trait-dependence diversification. For instance, with Bayesian analysis is possible to account for sampling fraction uncertainty, by providing a range of possible sampling fractions as a prior, although this is not something that has been implemented in studies yet.

Here, we provide the first in-depth evaluation of how incomplete phylogenetic information affects the performance of SSE models that incorporate concealed/hidden traits. We simulated data sets (phylogenetic trees along with trait information) and fitted SSE models in order to evaluate model selection under a number of different scenarios. Specifically, we evaluate the ability to select the correct type of relationship between trait evolution and branching patterns, and the accuracy in estimating parameter values (speciation, extinction, and transition rates). We focus on four key variables: (1) sampling fraction; (2) phylogenetic tree size; (3) sampling regime, in which phylogenies were sampled randomly or with taxonomic bias; and (4) misspecification of sampling fraction (where the true clade size is either under- or over-specified). We also used Bayesian analysis with a prior on the sampling fraction to deal with unknown sampling fractions.

MATERIALS & METHODS

Trait and Phylogeny Simulation

We simulated phylogenetic trees and accompanying trait data under three types of trait dependent diversification – Examined Trait Dependent (ETD), Concealed Trait Dependent (CTD), and Constant Rate (CR). We applied the same simulation procedure described in (Herrera-Alsina et al. 2019). In a nutshell, the simulations are run in continuous time, where first the waiting times are taken from an exponential distribution whose parameter is the sum of the per-lineage rates of events (i.e., speciation, extinction, and trait evolution). Then, a species is randomly taken to undergo one of the events and a record of it is kept. Specifically, we simulated the evolution of two traits that were independent from one another. Each trait had three states (examined trait with states 1, 2 and 3; concealed trait with states A, B, and C, see below) and all transitions across the three states were possible at a single rate (q). This means that the shift from state 1 to state 2 had the same rate than shifting from state 2 to state 1. Notice that both traits are combined to yield a nine-state system (1A, 2A, 3A, 1B, 2B, 3B, 1C, 2C, 3C) where double transitions (e.g., the change from 1A to 2C) are not possible. Although more complex systems (traits states ≥ 4) are possible in SecSSE (Herrera-Alsina et al. 2019), we chose three to minimise complexity. One of the traits, the examined trait, influenced diversification such that every time a lineage switched to a different trait state, its speciation/extinction rate was adjusted accordingly. This trait was the only factor affecting speciation and extinction dynamics and its' three states were numerically coded (i.e., 1, 2, 3). In contrast, the other trait, the concealed trait, did not influence diversification, but rather evolved neutrally over time. This trait's states were denoted alphabetically (A, B, C). Because the concealed trait had the same number of states as the examined trait (Fig. 1), the CTD model used here is equivalent to the character independent (CID-3) model in HiSSE (Beaulieu and O'Meara 2016; Herrera-Alsina et al. 2019). We kept track of the evolution of

both traits and at the end of the simulation we retained the species' trait states of either the examined trait (ETD-generating model) or the concealed trait (CTD-generating model). Note that the CTD model accounts for trait dependent diversification of an unmeasured trait (the concealed trait), and it is equal in complexity to the observed trait model. We also ran a simulation where both traits were neutral, and the rate of diversification did not change across time or lineages (CR-generating model). In the ETD-generating model, speciation rates differed only across examined trait states $\lambda 1A = \lambda 1B = \lambda 1C \neq \lambda 2A = \lambda 2B = \lambda 2C \neq \lambda 3A = \lambda 3B = \lambda 3C$, while in the CTD-generating model, speciation rates differed only across concealed trait states $\lambda 1A = \lambda 2A = \lambda 3A \neq \lambda 1B = \lambda 2B = \lambda 3B \neq \lambda 1C = \lambda 2C = \lambda 3C$ (Fig. 1). In the CR-generating model, all speciation rates were the same regardless of trait state $\lambda 1A = \lambda 1B = \lambda 1C = \lambda 2A = \lambda 2B = \lambda 2C = \lambda 3A = \lambda 3B = \lambda 3C$. The resulting simulated datasets included phylogenetic trees as well as the trait states associated with them.

We set the speciation and extinction rates within the ranges previously used to test the performance of SecSSE analyses (Herrera-Alsina et al. 2019). For ETD and CTD generated phylogenies, the speciation rates (λ) were: 0.1, 0.3, 0.5; for CR generated phylogenies, all trait states had the same speciation rate of 0.3. The extinction rate (μ) was set either low (0.001) or regular (0.05). We did not include models with variable extinction rates (and/or different transition rates) which adds to the model complexity and may lead to confounding effects (Davis et al. 2013). The transition rate (q) was set to 0.4 for all transitions in all phylogenies, and all transitions between the three states were possible (Fig. 1). The transition rate is somewhat high, although not unrealistic: while some empirical SSE studies have found transition rates similar to the rate we chose (e.g., pollination in hummingbirds, Wessinger et al. 2019), other studies have found lower rates (e.g., habitat preference in Diatoms Nakov et al. 2019; host-plant association in dragonflies Letsch et al. 2016; body shape and habitat preference in marine fishes Rincon-Sandoval et al. 2020), or much higher rates (mycorrhizal

association in fungi Looney et al. 2016). This also enabled testing the performance of SecSSE under new conditions, as its performance has already been documented with transition rates of 0.05 and 0.1 (Herrera-Alsina et al. 2019). Phylogenetic trees were simulated in three size groups; as younger trees have fewer tips, this was achieved by altering the crown age of trees: large (1000 – 5000 tips; age = 23 MY), medium (450 – 650 tips; age = 19 MY) and small (100 – 250 tips; age = 13.4 MY). The range in number of species within each category of tree size is due to stochasticity and extinction rate (even with the same clade age). For each tree size, we simulated 100 phylogenies of each diversification mode (ETD, CTD, CR) with low extinction, giving a total of 300 trees per size group. For small and medium tree sizes, we also simulated 100 phylogenies of each diversification mode with regular extinction.

Phylogenetic Tree Trimming

Phylogenetic trees and accompanying trait data were trimmed (removal of extant tips), either randomly or with a taxonomic bias (SI Fig. S1), to generate five sampling fraction (SF) levels in 20% intervals, that is SF: 100% (the full tree), 80%, 60%, 40%, and 20%. For random trimming, tips were randomly removed from across the phylogenetic tree (which is how SSE models treat the sampling fraction of a phylogeny, with the sampling fraction specified for each trait state (Nee et al. 1994; Fitzjohn et al. 2009; Chang et al. 2020)). To generate taxonomically biased sampling, we selected one or two subclades (containing 20 – 30% of the clade's size) to be heavily trimmed (removal of 80 – 90 % of tips) (SI Fig. S1). This resulted in slight variation in the final sampling fraction (± 2%) in each SF level, but this was less than the variation in trait state percentages (see below).

Randomly trimmed sets were created for all three phylogenetic tree sizes while taxonomic bias trimming was performed only on medium sized phylogenetic trees. Note that we did not explicitly evaluate the effect of a *trait* bias: tip loss was done agnostically with

respect to underlying trait state distributions. High tip ratio bias (e.g., one trait state accounting for < 10% of tips) can reduce model power and accuracy of parameter estimates (Davis et al. 2013). We therefore checked for tip ratio bias and found that all trait states were trimmed to a similar percentage – each of the three trait states accounted for ~33% of tips (SI, Fig. S2) and there were no differences in trait state percentages across tree sizes or trimming methods. Although our transition rate (q = 0.4) guarantees that transitions events are distributed throughout time and lineages, unsampling some tree tips might lead or not to the loss of trait state transitions events. The type of trait state transitions lost during trimming may affect model power. Therefore, during simulation of the regular extinction rate phylogenetic trees, the trait state transitions were recorded; for those datasets where the inference fails to select the right model, we explored whether they feature asymmetric transition lost (e.g., more 1 -> 2 lost than 2 -> 1) or dissimilar prevalence of concealed/examined transitions (e.g., more A -> B lost than 1 -> 2).

Sampling fraction settings in Maximum Likelihood (ML) framework

Using the above sets of simulated trees and trait states at the tips, SSE model analyses were performed under two different scenarios: (1) with the sampling fraction correctly specified, and (2) with the sampling fraction incorrectly specified (see Table 1 for details). The sampling fraction was specified per trait state: for example, for SF 60%, the sampling fraction was specified as 0.6 for each of the three trait states (notice that in hisse there is only global sampling fraction that accounts for phylogenetic incompleteness). Additionally, we provided a narrow and a wide prior for sampling fraction in a Bayesian context (see below).

Evaluation of SSE Models in ML

To test the ability of SSE models to select the correct (generating) model of trait dependent diversification under the above scenarios, we ran SecSSE analyses under the three models (ETD, CTD and CR) and compared Akaike information criterion values (AICc) (Bekara et al. 2005) and Akaike Weights (Wagenmakers et al. 2004). The percentage of false positives (i.e., erroneous model selection of ETD diversification) is calculated from the number of cases where ETD was selected as the best model in CTD and CR generated datasets, divided by the total number of CTD and CR generated datasets. False negatives (i.e., erroneous rejection of examined trait dependent diversification) are the percentage of ETD generated datasets that had CTD or CR selected as the best model. For medium sized, regular extinction trees, that were ETD generated, we also tested an ECTD model, which is a combination of the ETD and CTD, equivalent to the MuHiSSE. If ECTD was selected as the best model, it would indicate that the trait of interest has some effect on diversification dynamics but is not solely responsible. We explored the robustness of the results in regard to tree characteristics (tree size and imbalance). The Sackin Index (Sackin 1972) was used as a measure of tree imbalance: it is the average path length from tree root to tip (Blum and François 2005), and the less balanced the tree, the larger its Sackin Index value.

Evaluation of SSE Models in Bayesian framework

We also explored model selection in the presence of incomplete tree sampling under a Bayesian framework. Bayesian implementations of SSE models are available in RevBayes (Höhna et al. 2014), mcmc-diversitree (Silvestro et al. 2014), and BEAST (the "SSE" package; Mendes et al. 2018 in Bouckaert et al. 2019). RevBayes is particularly flexible, enabling various priors to be set, for example on the root frequencies, and allows for uncertainty in the tree topology and branch lengths to be marginalized over. The scripts provided in Freyman and Höhna (2018) are intended to compare the likelihood of state-

dependent and state-independent models using RevBayes (Höhna et al. 2016), which fits our purpose. We simulated 50 ETD datasets using the same procedure as in our main analysis (but considering only two examined and two concealed states, tree size ranged from 100 to 250 species; see note on computing time below), where 60% of the species were kept under both trimming methods: random and biased. We ran RevBayes's routine for state-dependent and state-independent models on each simulated dataset with two different setups for the sampling fraction. In one setup, we used a uniform prior with lower bound of 0.3 and upper bound of 0.9 (i.e., wide prior), whereas the other setup featured lower bound of 0.5 and upper bound of 0.7 (i.e., narrow prior). We used a stepping-stone approach to compute the marginal likelihood and bayes factor (Kass and Raftery 1995) to find the model with the highest statistical support. The major computational resources necessary to conduct this analysis prevented us from testing other scenarios under Bayesian framework (for the RevBayes analysis we used 192000 hours of computing time: two sampling methods x two priors x two dependence modes x 50 trees, 20 days each).

RESULTS

Effects of Phylogenetic Tree Size and Sampling Fraction, when the Sampling Fraction is Known and Taxonomically Unbiased

To test the effect of phylogenetic tree size, we compared randomly trimmed phylogenies of large (1000 – 5000 tips), medium (450 – 650 tips), and small (150 – 250 tips) sizes. As expected, SecSSE performed best with larger trees and those with more complete sampling (Fig. 2; SI Table S1). Correct model selection was reduced in smaller phylogenetic trees, and under low sampling fractions across all phylogeny sizes. For large trees, the false positive rate was ~ 6.5% when the Sampling Fraction (SF) ≥ 60, and increased to 17% false positives

rate at SF 40 (SI Table S1, set 10; Fig 2). Medium sized trees had false positive rates of ~ 14.5% when SF ≥ 80, this rate became > 22% at SF 40. Small trees had a higher false positive rate of 20% at SF 100, this decreased to 16% at SF 40, however, the rates of false negatives increased dramatically: from 9% at SF 100, to 55% at SF 40 (SI Table S1; Fig 2). In contrast, false negatives were negligible (< 5%) for large and medium sized trees when the sampling fraction was ≥ 60% (Fig. 2 and Table S1). Results were very similar for trees with higher extinction rates (Fig.2 and Table S3).

Akaike weights were higher in correctly selected models compared to incorrectly selected models (Fig. 3); the difference was most pronounced in large trees, and less so in small trees. Mean Akaike weight values for correctly and incorrectly selected models were more similar at lower sampling fractions (Fig. 3). The ECTD model was heavily penalised in the AICc analysis due to too many free parameters (11 in ECTD compared to 5 in ETD) and as such was never selected as the best model, even with ECTD simulated phylogenetic trees (SI Table S4). With so many variables in the ECTD model, the parameter space becomes too wide, resulting in some model runs being incomplete even after 10 optimization cycles.

Tree imbalance, as determined via Sackin index, did not affect correct model selection for complete or randomly trimmed phylogenies The random sampling procedure removed tree tips randomly and consequently transition events were also removed randomly as we found that number of transitions out of a given state were not different than transitions going into that state (e.g., 1 -> 2 = 2 -> 1). Moreover, the number of transitions lost across the examined trait were similar than in the concealed trait (SI Fig. S3).

Parameter estimation was more accurate and precise at higher sampling fractions, whereas variation in parameter estimates increased with decreasing sampling fraction (Fig. 4). Small trees had the largest variation in their parameter estimates, while large trees had the smallest variation (Fig. 4). The only exception was the net diversification rates of CTD

generated trees, which showed little difference in parameter estimate variation across tree size (Fig. 4). In most cases, large trees at SF ≥ 40 had less variation in parameter estimates compared to medium sized trees at SF 100 (Fig. 4). Extinction rates for phylogenies generated with low extinction (μ = 0.001) were generally over-estimated (e.g., for medium trees at SF 100, mean = 0.0284 ± 0.0371; median = 0.0111), whereas in the regular extinction sets (μ = 0.05), extinction rates were marginally under-estimated (e.g., for medium trees at SF 100, mean = 0.0442 ± 0.0418; median = 0.0384). There were occasional high outliers for the transition rate estimates; this occurred with ETD, CTD and CR generated trees, but only affected medium sized trees at SF 20 and small trees at SF 60 or lower (Fig. 4).

Effects of Sampling Bias when the Sampling Fraction is Correctly Specified

Medium sized phylogenetic trees that were trimmed randomly, or with taxonomic bias, were compared to test the effects of sampling bias. Random sampling led to better model performance than biased sampling for both model selection and parameter estimation (SI Table S1, Fig. 2, and Fig. 4). Randomly trimmed trees had lower rates of false negatives than biased trimmed trees (SI Table S1). Biased trimmed trees had a lower rate of false positives compared to randomly trimmed trees at SF 80, but at SF 60 this was reversed (SI Table S1). At SF 20, the percentage of false positives was considerably higher for biased trimmed trees compared to randomly trimmed trees (SI Table S1). In biased trimmed sets, most false positives came from an erroneous ETD selection of a CR generated tree (SI Table S1). In contrast, for randomly trimmed trees, most false positives came from an erroneous ETD selection of a CTD generated tree (SI Table S1). Net diversification rate estimates were considerably more accurate in randomly sampled trees (Fig. 5). For biased trimmed trees, as sampling fraction decreased, net diversification rate estimates became more inaccurate (Fig. 5). Figure 5 shows that, even though the rate estimates are not accurate, the model is able to

detect differences across the three states and accommodates the rates of net diversification to maximize this difference.

Effects of Mis-specifying the Sampling Fraction

Mis-specifying the sampling fraction by ±20% (of the total number of tips) reduced the accuracy of model selection (Fig. 6) and parameter estimates (Fig. 5). Random and biased sampled phylogenetic trees were affected by misspecification of the sampling fraction in a similar manner. Specifying the sampling fraction as higher than its true value often caused an increase in false positives (Fig. 6), while under-specifying the sampling faction gave similar or slightly lower rates of false positives compared to correctly specified SF sets (Fig. 4; SI Table S2). Similar to correctly specified SF sets, false negatives were negligible (≤ 5%) when the true SF was ≥ 60%, irrespective of the degree of misspecification; the only exception to this was Set 6b, which had a much higher false negative rate (SI Table S2).

Mis-specifying the sampling fraction affected net diversification rate estimates in some sets, most noticeably 'bias ETD', 'random ETD' and 'random CR' (Fig 5., SI Fig. S4). Values were over-estimated when sampling fraction was specified as lower than its true value (i.e., the clade was larger than thought), and under-estimated when sampling fraction was specified as higher than its true value (i.e., the clade was smaller than thought; SI Fig. S4). These effects were most noticeable at lower sampling fractions. Transition rate estimates were similar in correctly specified and mis-specified sets (SI Fig. S4). Top left panel in Figure 4 shows that even though net diversification rate estimates were inaccurate when the sampling fraction was mis-specified (i.e., boxplots with wide ranges), particularly in small trees, there was still little overlap between rate estimates for each trait state. This suggests that the model was able to correctly detect which trait state had a comparatively higher net

diversification rate and which trait state had a lower rate (i.e., the median for estimates are roughly at the true value).

Specifying the sampling fraction as a range

When using RevBayes, we found that the signal of state-dependent diversification is correctly recovered in all cases under both sampling methods when the prior distribution of sampling fraction was narrow (Table 2). However, when the uncertainty around the true completeness of the dataset is higher and the prior distribution of sampling fraction is wider, in 10% of cases the state-independent model was wrongly selected as being the best performing. Even though bayes factor did not point to ETD as the best performing model in those datasets, we note that the rate estimates taken from the MCMC posterior distribution did include the true generating rates. These distributions have an important overlap which is related to the failure to detect state-dependent diversification (SI Fig. S5).

DISCUSSION

In this study, we have explored how incomplete sampling of phylogenetic trees, and misspecification of the sampling fraction, can affect the ability of SSE models to detect trait dependent diversification and estimate diversification rates. We found that both taxonomic biased sampling and mis-specifying the sampling fraction can severely decrease the accuracy of parameter estimation. Sampling fraction misspecification had more minor effects on model selection, with false positive rates only increasing when the sampling fraction was over-specified. Taxonomic biased sampling reduced the accuracy of parameter estimates, more so at lower sampling fractions, and sometimes increased the rates of false positives and false negatives compared to random sampling. When using biased sampled phylogenetic trees, there is a greater risk of obtaining a false positive from a neutrally diverging phylogeny. Smaller phylogenetic trees and any sized phylogenetic trees under low true sampling

fractions (< 60%) have higher rates of false positives, rendering them less suitable for current SSE analyses. Although SecSSE was the main method used in this study, our results are relevant to other SSE methods that incorporate hidden traits. The CTD model used is equivalent to the CID-3 model in HISSE (Beaulieu and O'Meara 2016; Herrera-Alsina et al. 2019), meaning our general findings are applicable to HISSE and its relatives.

Effects of Tree Size

Model selection accuracy was severely reduced in small phylogenies (150 – 250 tips) even at high sampling fractions (i.e., nearly complete phylogenetic trees). This concurs with previous work using BiSSE, which also showed that small trees (fewer than 300 tips) are less suitable for SSE modelling (Davis et al. 2013). This is because the statistical power of SSE models partially depends on the number of taxa in the phylogenetic tree (Davis et al. 2013). Across all phylogenetic tree sizes, rates of false positives and false negatives were elevated at lower sampling fractions. This is due to decreased sizes of phylogenetic trees and increased uncertainty in the models, both of which come as a consequence of lower sampling fractions.

When interpreting parameter estimates from SSE models, it is therefore important to consider phylogenetic tree size. Interestingly, large phylogenetic trees (1000 – 5000 tips) under low sampling fraction (40%) had similar or better parameter estimates than medium sized phylogenetic trees (450 – 650 tips) at high sampling fractions (Fig. 4). This suggests that it may be better to use a larger, but incomplete phylogeny, rather than a smaller more complete subclade, to study the patterns of trait dependent diversification. The larger (and more complete) the phylogenetic tree, the more accurate the speciation and transition rate estimates.

As in other SSE performance studies (Maddison et al. 2007b; Beaulieu and O'Meara 2016), extinction rate estimates were imprecise, due to the lack of information about

extinction in phylogenetic trees (Rabosky 2006). In contrast to some studies (e.g., Höhna et al., 2011), but in agreement with others (e.g., Beaulieu and O'Meara, 2016), we found that some extinction rate estimates tended to be over-estimated. Specifically, the low extinction sets ($\mu = 0.001$) tended to have elevated extinction rate estimates. We believe this is due to the extinction rate parameter being set so low, because in the regular extinction rate sets ($\mu = 0.05$), extinction rate estimates were generally under-estimated. However, there was a large amount of variation in all parameter estimates. Previous work using BiSSE and simulated phylogenies with 500 tips, showed that speciation rate estimates remain accurate down to ~ 50% sampling fraction (Fitzjohn et al. 2009). We concur with this finding and add that speciation rate estimates for larger phylogenetic trees (≥ 1000 tips in the complete in tree) could remain accurate at slightly lower sampling fractions (~ 40%), although accuracy is improved with greater sampling. As suggested by Beaulieu and O'Meara (2016), more accurate net diversification rate estimates can be obtained from larger phylogenetic trees, making it possible to distinguish between trait-states with smaller rate differences. However, while larger trees are better suited to SSE analyses in terms of model selection and parameter estimation, the time and computational power required for these analyses is high.

Tip ratio bias and loss of trait state transitions

We found that datasets where the inference analysis failed to select the right model have very similar structure in terms of transition type lost than those datasets where the analysis recovered the right generating underlying process. However, it was often the same phylogenetic trees that had incorrect model selection at different sampling fractions, suggesting that there may be something inherent within these trees that made it more difficult for the model to detect the correct diversification type. The number of trait state transitions lost may have a different impact on simulations with different parameter settings.

Further studies will need to use scenarios with asymmetrical transition rates. Differences in transition rates may be more important when there are trait biases, that is when some trait states are more likely to be sampled than others. Equally, trait biases are more prevalent when transition rates are asymmetrical (Davis et al. 2013). For example, in the scenario where it is easy to transition into a specialist state, but harder to transition out of the specialised state, this asymmetry could lead to trait state biases with more tips in the specialised state, unless this specialised state also had a lower speciation or higher extinction rate. It would be interesting to test how loss of trait state transitions affects SSE models when there are trait biases. Other future work that could be beneficial to further understanding how loss of trait state transitions from incomplete phylogenetic trees affect SSE models include: (1) exploring transitions lost under different transition rate scenarios such as low, medium, and high transition rates and (2) asymmetrical transition rates and different speciation/extinction rates.

Effects of Sampling Regime

Phylogenetic trees may suffer sampling bias due to certain sub-clades containing greater numbers of rare or undescribed species. In other cases, some species may be deliberately removed from the clade. Overall, if phylogenetic trees are incomplete, our results show that is better for them to be randomly sampled rather than sampled with taxonomic bias. At high sampling fraction (80%), biased sampling represents a minor source of inaccuracy: parameter estimates were similar to those from randomly sampled phylogenies, and rates of false positives were lower than in randomly sampled phylogenies. However, when sampling is less complete (sampling fraction $\leq$ 60%), parameter values became over-estimated in biased sampled sets but remained accurate in randomly sampled sets. As sampling fraction decreased, rates of false negatives became higher in biased sampled sets compared to

randomly sampled sets. This means there is a greater risk of erroneously rejecting trait dependent diversification when a phylogeny is sampled with taxonomic bias.

Similarly to Fitzjohn et al. (2009), we found that randomly trimmed CR generated trees maintained ~ 85% correct model selection across all sampling fractions. In contrast, we found that biased trimmed CR generated trees led to more false positives and to considerable reductions in correct model selection at lower sampling fractions. This means that there is a greater risk of erroneously finding evidence for trait dependent diversification from a biased sampled phylogenetic tree. This may be due to sampling method, as some sub-clades were heavily trimmed (to simulate under sampling), leading to longer branch lengths, and in the CR model branching patterns are the only information available to estimate diversification.

Currently, the only way to specify the sampling fraction in most SSE methods is by trait state, so it is not always possible to account for alternative sampling methods or taxonomic biases. Clade specific sampling fractions were previously enabled in HiSSE but were removed as they caused mathematical errors (Beaulieu 2020); however, clade specific sampling fractions still exist in diversitree (FitzJohn 2012). RevBayes can also account for different sampling methods: uniform (random), diversified, and empirical (clade specific) sampling strategies can be accommodated (https://revbayes.github.io/tutorials/divrate/sampling.html). When it is not possible to sample clades completely, it is recommended to assess the degree of bias in sampling in SSE modelling. When clades are biased sampled and ≤ 60 % complete, extra caution is advised when interpreting the results of SSE models. It is also important to bear in mind that parameter estimates will likely be higher than their true value when trees are biased sampled and have a low sampling fraction.

Effects of Mis-specifying the Sampling Fraction

It can be difficult to accurately set the sampling fraction of empirical phylogenetic trees as the actual number of extant species is often unknown. Organisms which may be particularly problematic in this regard include bacteria, archaea, and fungi, as these groups contain many undescribed species (Barns et al. 1994; Lambais et al. 2006; Mueller et al. 2007; Brock et al. 2009; Öpik et al. 2013; Looney et al. 2016). Our findings indicate that it is not acceptable to guess the sampling fraction if it is completely unknown: inaccurate sampling fraction estimates have a high risk of false positives and inaccurate parameter estimates. Therefore, SSE modelling is most suitable for incomplete reconstructions when the sampling fraction is known with some degree of accuracy. Sensitivity analysis to sampling fraction specification should be performed to provide confidence to results when the sampling fraction has been estimated.

Sampling fraction specifications of ±20% inaccuracy led to inaccurate parameter estimates. In incompletely sampled phylogenetic trees, the apparent number of speciation and character change events is reduced because tips/branches are missing from the tree. If incomplete sampling is not accounted for, this can lead to likelihoods favouring lower diversification and transition rate estimates (Fitzjohn et al. 2009). When the sampling fraction is thought to be higher than it truly is (over-specified; e.g., there is a 60% complete phylogenetic tree, but it is specified as 90% complete), not all tips and speciation events will be accounted for, leading to lower speciation rate estimates. Conversely, when the sampling fraction is thought to be lower than it truly is (under-specified; e.g., there is a 90% complete phylogenetic tree, but it is specified as only 60% complete), the model will account for more tips and speciation events than there actually were, leading to higher diversification rate estimates.

We suggest that parameter estimates are interpreted cautiously when there is uncertainty around the sampling fraction approximation. When considering net diversification, the differences in rate estimates with sampling fraction misspecification were only present with biased sampled phylogenetic trees. Even with low levels of sampling, the model correctly detects differences in diversification rates across states so that it tries to maximize the difference between trait 1 and 3 rates. Even though the estimated difference in rates is quite close to the true one (0.4), the overall estimates are inaccurate. This is more evident under bias sampling (Figure 5, bottom row) where net diversification rate for trait 1 is underestimated and at the same time, for trait 3 is overestimated. Interestingly, specifying the correct sampling fraction does not lead to better rate estimates under biased sampling. This is likely to be result of 1) confounding extinction with missing branches due to sampling, and 2) SSE models always assume that non-sampled branches are randomly distributed. In the presence of bias sampling, overspecifying sampling fraction yields to better estimates. When diversification rate is separated into rates of speciation and extinction, estimates are highly affected by sampling fraction misspecification. However, the relative speciation rate estimates (i.e., which trait states have the lowest and highest speciation rates) is robust to changes in sampling fraction specification.

Due to unknown numbers of undescribed species and taxonomic uncertainties, it may be more common for researchers to over-specify the sampling fraction, thinking that they have a greater proportion of the phylogeny sampled than they actually do (Vieites et al. 2009; Pimm et al. 2014; Chan et al. 2018; Dickens et al. 2019). Our results show that, in contrast, cautiously under-specifying the sampling fraction may not be as bad as over-specifying it: false positive rates were elevated when the sampling fraction was over-specified but remained similar to (or even slightly lower than) correctly specified sets when the sampling fraction was under-specified. With an 80% complete (randomly trimmed, medium sized)

phylogenetic tree, when the sampling fraction was correctly specified (as 80%), the rate of false positives was 14.5%; when the sampling fraction was under-specified as 60% complete, the false positive rate was slightly lower (13.5%), but when the sampling fraction was over-specified as 100% complete, false positives were higher (18%). As sampling fraction decreased, rates of false negatives became higher in biased sampled sets compared to randomly sampled sets. This means there is a greater risk of erroneously rejecting trait dependent diversification when a phylogeny is sampled with taxonomic bias. In SSE analyses, we recommend carrying out sensitivity analysis of sampling fraction specification that spans, at a minimum ±20% of the estimated fraction, to determine if the results are robust to variation in the sampling fraction specification. The range of sampling fractions used should reflect how much uncertainty there is in the completeness of the tree.

Bayesian analysis

One alternative method for dealing with an uncertain sampling fraction utilises Bayesian analyses with a hyperprior placed on the sampling fraction. No empirical studies have thus far used this method for dealing with sampling fraction uncertainty. More commonly used methods to specify the sampling fraction in Bayesian SSE studies are to specify the probability of sampling species within the clade, based on the total number of known species (tips) and the number of species sampled (e.g., Wessinger et al. 2019; Varga et al. 2021), or assuming that all extant species have been included in the phylogeny (e.g., Condamine et al. 2018). Studies using RevBayes and HiSSE that may have benefitted from incorporating uncertain sampling include a study on weevils (Letsch et al. 2018) and a study on basidiomycete fungi (Varga et al. 2021), because these taxa likely have undescribed species. The use of RevBayes in SSE analysis is promising, especially when sampling fraction is not

well known. However, it is computationally slow which compromises its applicability in large datasets.

Conclusions and Best Practices

Much progress has been made from the early days of SSE models but work still needs to be done to make SSE methods more robust to the sampling issues explored here. Areas which require further attention include dealing with different types of sampling, uncertainty in tree topology, exploring how loss of trait state transitions affect SSE models, and developing robust methods for confident analyses of smaller phylogenetic trees. Additionally, a more thorough exploration of the efficacy of using a sampling fraction prior within a Bayesian framework is needed. Most empirical studies will to some extent violate the assumption that sampling is uniform and random across the phylogeny. It would be highly desirable for SSE methods to be able to account for taxonomic bias. One possibility which could provide more information to the model and decrease uncertainty around the sampling fraction, could be to allow for sampling fraction specification both by trait state and by clades at the same time; however, this would be challenging to develop (for birth-death process see Höhna et al. 2011).

This work has helped to inform how much error in sampling fraction estimates is acceptable, enabling confident SSE modelling of phylogenies where the sampling fraction can be reasonably estimated. To conclude, we provide suggestions for best practices when using SSE methods on incompletely sampled phylogenetic trees.

- It may be better to use a larger but somewhat incomplete phylogeny, rather than a smaller but more complete subclade. Larger (≥ 450 tips) and more complete (≥ 60% SF) phylogenetic trees are most suitable for SSE analyses, but tree size is generally more important than completeness.

- Taxonomic biases in sampling can be problematic when phylogenetic trees are < 80% complete. We recommend assessing the degree of bias in sampling, as there are greater risks of false positives and inaccurate parameter estimates when trees are ≤ 60% complete and have been sampled with taxonomic bias. If possible, additional sampling of missing tips is advised, to reduce sampling bias and increase the sampling faction of the phylogeny: increased taxon sampling remains one of the best methods to increase accuracy of inferences drawn from phylogenetic trees (Heath, Hedtke and Hillis 2008). Inclusion of tips with uncertain trait states is possible in some packages (e.g., SecSSE), and is one possible way to increase the sampling fraction.
- For SSE methods, we do not recommend excluding tips from phylogenetic trees, for example because of regional sampling, as this lowers the sampling fraction, increases uncertainty in the model, and may increase sampling bias.
- Misspecification of sampling fraction can reduce correct model selection and leads to inaccuracies in parameter estimates. We advise that SSE modelling is most suitable for incomplete phylogenies when the number of extant species in the clade is known with some accuracy. It is worth conducting a thorough examination to estimate the sampling fraction as precisely as possible. We suggest two methods for dealing with uncertainty in the sampling fraction: 1) using the ML approach, sensitivity analyses should be performed across an appropriate range of sampling fractions (at least ± 20% of the estimated sampling fraction), in order to confirm that results are robust to variation in sampling fraction specification; 2) using Bayesian analyses of SSE models in order to specify a range of possible sampling fractions as a prior. These methods are not mutually exclusive, and the most confident results may be obtained by implementing both approaches.

Acknowledgements

All simulations and analyses were performed on the University of Aberdeen HPC, Maxwell.

Funding

This work was funded by Newton Fund (UK)/NERC (UK)/RISTEKDIKTI (Indonesia) grants awarded to JT, BJ, ACA, ASTP, CG-R, GB and LTL (Grant numbers: NE/S006923/1; NE/S006893/1; 2488/IT3.L1/PN/2020; and 3982/IT3.L1/PN/2020). GB is funded by a Royal Society University Research Fellowship (UF160614).

Data Availability

The data underlying this article is available on dryad repository:

https://doi.org/10.5061/dryad.wwpzgmsjp

References

Barns S.M., Fundyga R.E., Jeffries M.W., Pace N.R. 1994. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment (archaebacteria/phylogeny/thermophfly/molecular ecology). Proc. Nati. Acad. Sci. USA. 91.

Beaulieu J.M. 2020. The Problem with Clade-specific Sampling Fractions. .

Beaulieu J.M., O'Meara B.C. 2016. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. Systematic Biology. 65:583–601.

Bekara M., Hafidi B., Fleury G. 2005. Smoothing parameter selection in nonparametric regression using an improved kullback information criterion. Proceedings - 8th International Symposium on Signal Processing and its Applications, ISSPA 2005. 2:887–890.

Blum M.G.B., François O. 2005. On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. Mathematical Biosciences. 195:141–153.

Brock P.M., Döring H., Bidartondo M.I. 2009. How to know unknown fungi: The role of a herbarium. New Phytologist. 181:719–724.

Caetano D.S., O'Meara B.C., Beaulieu J.M. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution (N Y). 72:2308–2324.

Chan K.O., Grismer L.L., Brown R.M. 2018. Comprehensive multi-locus phylogeny of Old World tree frogs (Anura: Rhacophoridae) reveals taxonomic uncertainties and potential cases of over- and underestimation of species diversity. Molecular Phylogenetics and Evolution. 127:1010–1019.

Chang J., Rabosky D.L., Alfaro M.E. 2020. Estimating Diversification Rates on Incompletely Sampled Phylogenies: Theoretical Concerns and Practical Solutions. Systematic Biology. 69:602–611.

Chek A.A., Austin J.D., Lougheed S.C. 2003. Why is there a tropical-temperate disparity in the genetic diversity and taxonomy of species? .

Collen B., Ram M., Zamin T., Mcrae L. 2008. The tropical biodiversity data gap: addressing disparity in global monitoring. Mongabay.com Open Access Journal-Tropical Conservation Science. 1.

Corlett R.T. 2016. Plant diversity in a changing world: Status, trends, and conservation needs. Plant Diversity. 38:10–16.

Daru B.H., Park D.S., Primack R.B., Willis C.G., Barrington D.S., Whitfeld T.J.S., Seidler T.G., Sweeney P.W., Foster D.R., Ellison A.M., Davis C.C. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytologist. 217:939–955.

Davies T.J., Allen A.P., Borda-de-Água L., Regetz J., Melián C.J. 2011. Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification. Evolution (N Y). 65:1841–1850.

Davis M.P., Midford P.E., Maddison W. 2013. Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. .

Day E.H., Hua X., Bromham L. 2016. Is specialization an evolutionary dead end? Testing for differences in speciation, extinction and trait transition rates across diverse phylogenies of specialists and generalists. J Evol Biol. 29:1257–1267.

Dickens J.K., McMahon L., Binnie S.E. 2019. The butterflies of a Cerrado--Atlantic Forest ecotone at Laguna Blanca reveal underestimation of Paraguayan butterfly diversity and need for conservation. Journal of Insect Conservation. 23:707–728.

Donoghue M.J., Ackerly D.D. 1996. Phylogenetic uncertainties and sensitivity analyses in comparative biology. Philosophical Transactions of the Royal Society B: Biological Sciences. 351:1241–1249.

Felsenstein J. 1985. Phylogenies and the Comparative Method Author (s): Joseph Felsenstein Source : The American Naturalist , Vol . 125 , No . 1 (Jan ., 1985), pp . 1-15 Published by : The University of Chicago Press for The American Society of Naturalists Stable URL : ht. American Society of Naturalists. 125:1–15.

Fitzjohn R.G. 2010. Quantitative traits and diversification. Systematic Biology. 59:619–633.

Fitzjohn R.G., Maddison W.P., Otto S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology. 58:595–611.

Francisco Henao Diaz L., Harmon L.J., Sugawara M.T.C., Miller E.T., Pennell M.W. 2019. Macroevolutionary diversification rates show time dependency. Proc Natl Acad Sci U S A. 116:7403–7408.

Gubry-Rangin C., Kratsch C., Williams T.A., McHardy A.C., Embley T.M., Prosser J.I., Macqueen D.J. 2015. Coupling of diversification and pH adaptation during the evolution of terrestrial Thaumarchaeota. Proc Natl Acad Sci U S A. 112:9370–9375.

Helmstetter A.J., Papadopulos A.S.T., Igea J., Van Dooren T.J.M., Leroi A.M., Savolainen V. 2016. Viviparity stimulates diversification in an order of fish. Nature Communications. 7.

Herrera-Alsina L., Van Els P., Etienne R.S. 2019. Detecting the Dependence of Diversification on Multiple Traits from Phylogenetic Trees and Trait Data. Systematic Biology. 68:317–328.

Igea J., Miller E.F., Papadopulos A.S.T., Tanentzap A.J. 2017. Seed size and its rate of evolution correlate with species diversification across angiosperms. PLoS Biology. 15.

Jablonski D. 2008. Species selection: Theory and data. Annual Review of Ecology, Evolution, and Systematics. 39:501–524.

Laenen B., Machac A., Gradstein S.R., Shaw B., Patiño J., Désamoré A., Goffinet B., Cox C.J., Shaw A.J., Vanderpoorten A. 2016. Increased diversification rates follow shifts to bisexuality in liverworts. New Phytologist. 210:1121–1129.

Lambais M.R., Crowley D.E., Cury J.C., Büll R.C., Rodrigues R.R. 2006. Bacterial diversity in tree canopies of the Atlantic Forest. Science (1979). 312:1917.

Liem K.F. 1973. Evolutionary Strategies and Morphological Innovations: Cichlid Pharyngeal Jaws. Systematic Biology. 22:425–441.

Looney B.P., Ryberg M., Hampe F., Sánchez-García M., Matheny P.B. 2016. Into and out of the tropics: global diversification patterns in a hyperdiverse clade of ectomycorrhizal fungi. Molecular Ecology. 25:630–647.

Maddison W.P., Midford P.E., Otto S.P. 2007a. Estimating a binary character's effect on speciation and extinction. Systematic Biology. 56:701–710.

Maddison W.P., Midford P.E., Otto S.P. 2007b. Estimating a binary character's effect on speciation and extinction. Systematic Biology. 56:701–710.

Moerman D.E., Estabrook G.F. 2006. The botanist effect: Counties with maximal species richness tend to be home to universities and botanists. Journal of Biogeography. 33:1969–1974.

Mueller G.M., Schmit J.P., Leacock P.R., Buyck B., Cifuentes J., Desjardin D.E., Halling R.E., Hjortstam K., Iturriaga T., Larsson K.H., Lodge D.J., May T.W., Minter D., Rajchenberg M., Redhead S.A., Ryvarden L., Trappe J.M., Watling R., Wu Q. 2007. Global diversity and distribution of macrofungi. Biodiversity and Conservation. 16:37–48.

Nee S., May R.M., Harvey P.H. 1994. The reconstructed evolutionary process. .

Öpik M., Zobel M., Cantero J.J., Davison J., Facelli J.M., Hiiesalu I., Jairus T., Kalwij J.M., Koorem K., Leal M.E., Liira J., Metsis M., Neshataeva V., Paal J., Phosri C., Põlme S., Reier Ü., Saks Ü., Schimann H., Thiéry O., Vasar M., Moora M. 2013. Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi. Mycorrhiza. 23:411–430.

Pimm S.L., Jenkins C.N., Abell R., Brooks T.M., Gittleman J.L., Joppa L.N., Raven P.H., Roberts C.M., Sexton J.O. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. Science (1979). 344.

Rabosky D.L. 2006. LIKELIHOOD METHODS FOR DETECTING TEMPORAL SHIFTS IN DIVERSIFICATION RATES. Evolution (N Y). 60:1152–1164.

Rabosky D.L., Goldberg E.E. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Systematic Biology. 64:340–355.

Rabosky D.L., McCune A.R. 2010. Reinventing species selection with molecular phylogenies. Trends in Ecology and Evolution. 25:68–74.

Reddy S. 2014. What's missing from avian global diversification analyses? Molecular Phylogenetics and Evolution. 77:159–165.

Resl P., Fernández-Mendoza F., Mayrhofer H., Spribille T. 2018. The evolution of fungal substrate specificity in a widespread group of crustose lichens. Proceedings of the Royal Society B: Biological Sciences. 285.

Rich T.C.G., Woodruff E.R. 1992. Recording bias in botanical surveys. 19.

Ricklefs R.E. 2007. Estimating diversification rates from phylogenetic information. Trends in Ecology and Evolution. 22:601–610.

Sackin M.J. 1972. "Good" and "Bad" Phenograms. Systematic Biology. 21:225–226.

Titley M.A., Snaddon J.L., Turner E.C. 2017. Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. PLOS ONE. 12:e0189577-.

Tonini J.F.R., Ferreira R.B., Pyron R.A. 2020. Specialized breeding in plants affects diversification trajectories in Neotropical frogs. Evolution (N Y). 74:1815–1825.

Troudet J., Grandcolas P., Blin A., Vignes-Lebbe R., Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. Scientific Reports. 7:1–14.

Vasconcelos T., O'Meara B.C., Beaulieu J. M. 2022. A flexible method for estimating tip diversification rates across a range of speciation and extinction scenarios. Evolution doi:10.1111/evo.14517

Vieites D.R., Wollenberg K.C., Andreone F., Köhler J., Glaw F., Vences M. 2009. Vast underestimation of Madagascar's biodiversity evidenced by an integrative amphibian inventory. Proc Natl Acad Sci U S A. 106:8267–8272.

Wagenmakers E.-J., Farrell S., Wagenmakers -J. 2004. AIC model selection using Akaike weights. .

Weber M.G., Agrawal A.A. 2014. Defense mutualisms enhance plant diversification. 111.

Figure Captions

Figure 1. Schematic model diagram. Each circle indicates a trait state combination with one examined trait state (1, 2, or 3) and one concealed trait state (A, B, or C). The speciation rate (λ) for each trait state combination differed either with the examined trait (ETD model), or the concealed trait (CTD model). The extinction rate (μ) was the same in every trait state combination. The transition rates (q), shown in grey for the examined trait and red for the concealed trait, was symmetrical, i.e., the same rate was used for every transition, and all transitions between trait states were possible.

Figure 2. False positives (A) and false negatives (B) from all correctly specified sampling fraction sets, including randomly trimmed large, medium, and small trees, as well as bias trimmed trees (medium B), and trees with different extinction rates (small HX and medium HX).

Figure 3. Model selection. Akaike weights of correctly (blue) and incorrectly (red) selected models of each size and trimming method (R = randomly trimmed; B = biased trimmed), at each sampling fraction. All three generating models for each sampling fraction are grouped into the violin plot; for each pair of violin plots n = 300.

Figure 4. Parameter estimates for correctly specified sampling fraction sets of randomly trimmed ETD (examined trait dependent), CTD (concealed trait dependent) and CR (constant rate) generated trees. Colours indicate tree sizes. The horizontal line indicates the true parameter value. Note that these plots include all generated trees, regardless of whether the generating model was selected as the best model or not.

Figure 5. Net diversification rate estimates of the medium size ETD generated phylogenetic trees, that were randomly (R; top row) or biased (B; bottom row) trimmed. ND1, ND2 and ND3 indicate the net diversification rate for trait state 1, 2 and 3 respectively. Colours indicate if the sampling fraction was correctly specified or mis-specified. The horizontal line indicates the true parameter value. Plots include all generated trees for each set, regardless of whether the generating model was selected as the best model or not.

Figure 6. Percentages of false positives and false negatives in Random and Bias sampled sets. Colours indicated the difference from the true sampling fraction, that is if the true sampling fraction is 80, at -20 the sampling fraction was specified as 60, and at +20 the sampling fraction was specified as 100.

Table 1. Mis-specified sampling fraction sets. 100% is a complete phylogenetic tree; 20% is a phylogenetic tree containing only 20% of tips from the complete tree. In mis-specified SF, columns indicate what the True SF was mis-specified as; for example, True SF 80%, was mis-specified as SF 100% and SF 60%. Mis-specified sets were done on random and bias trimmed medium sized trees.

	Sampling Fraction (%)				
True SF	100	80	60	40	20
Mis-specified SF	80	100 60	100 80 40	100 80	-

Table 2. Number of datasets that were incorrectly (column CTD) and correctly selected (column ETD) during RevBayes analysis using a narrow and wide priors for sampling fraction. The table also includes two different methods of tree tip sampling.

Prior	Sampling	CTD	ETD
Narrow	random	0	50
	bias	0	47
Wide	random	8	42
	bias	1	49

Figure 1

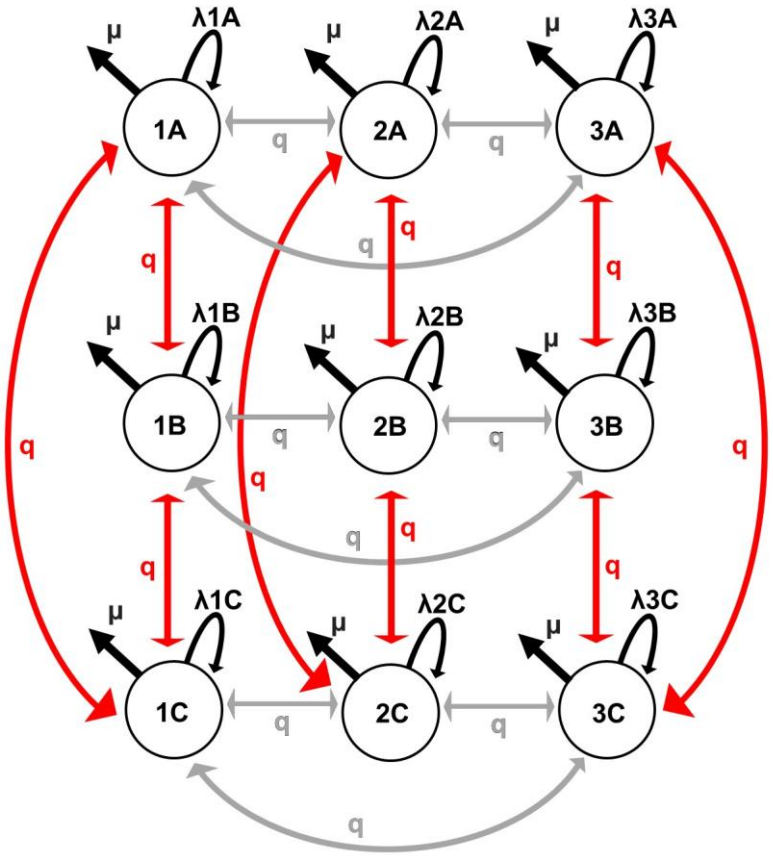

Figure 2

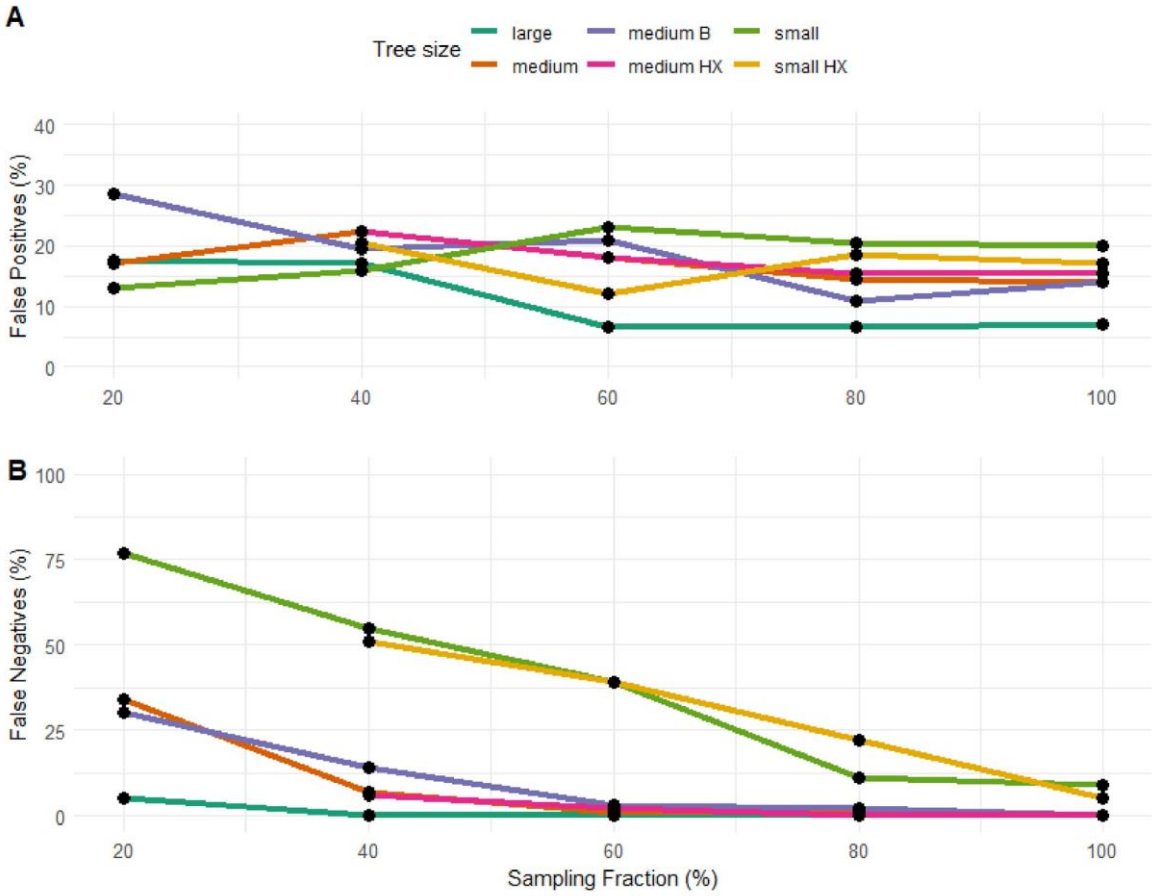

Figure 3

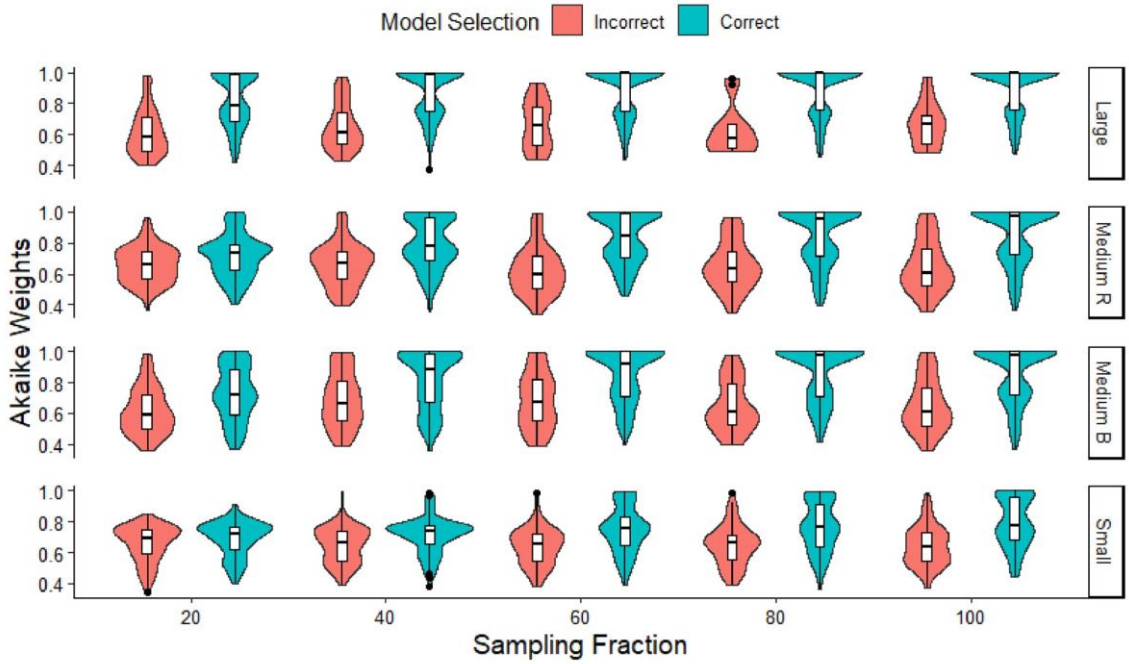

Figure 4

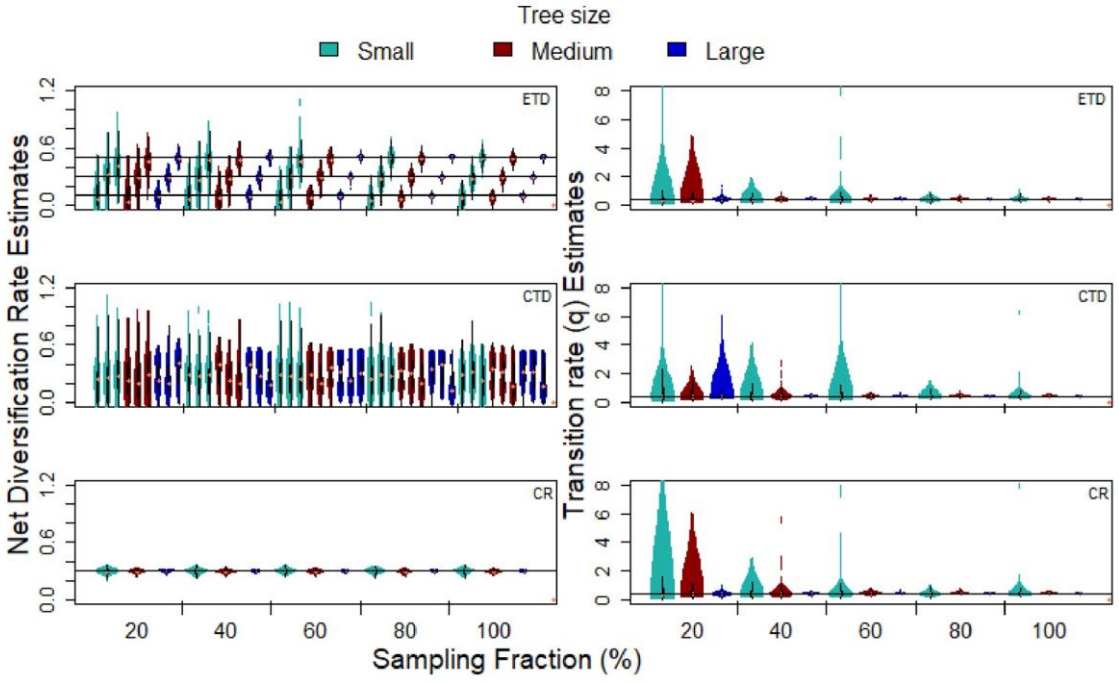

Figure 5

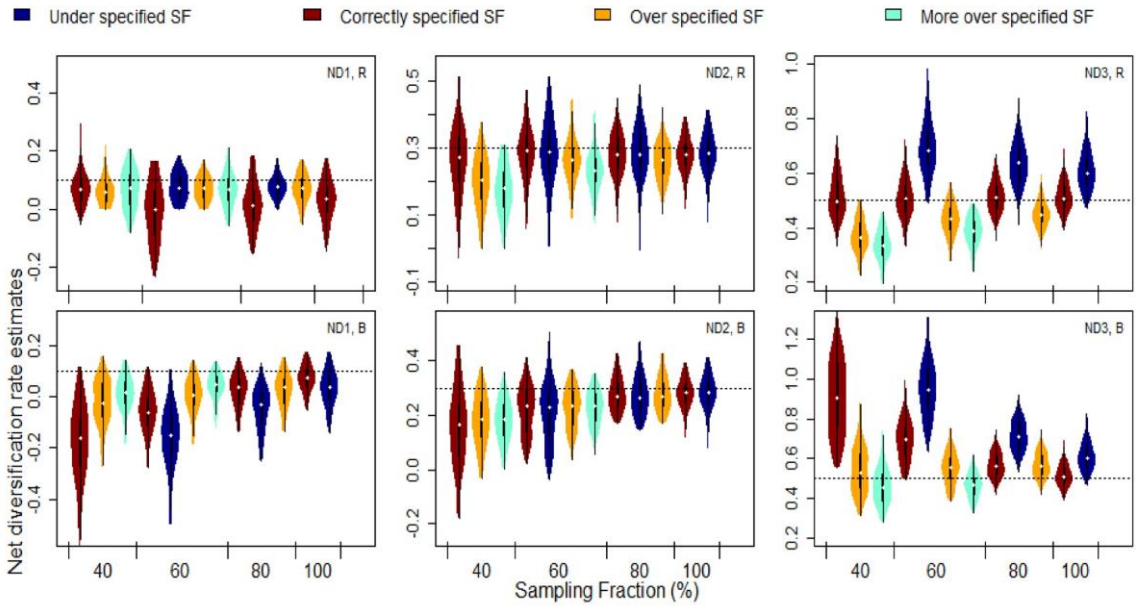

Figure 6