Received: 18 September 2024 | Accepted: 21 March 2025

DOI: 10.1111/2041-210X.70058

BRITISH ECOLOGICAL SOCIETY Methods in Ecology and Evolution

RESEARCH ARTICLE

Impact of sampling strategy on inference of community assembly processes in phylogenetic island biogeography

Ornela N. Dehayem[1,2] | Ryan F. A. Brewer[1,2] | Luis Valente[1,2] | Frederic Lens[1,3] | Rampal S. Etienne[1]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

[2]Naturalis Biodiversity Center, Leiden, The Netherlands

[3]Institute of Biology Leiden, Leiden University, Leiden, The Netherlands

Correspondence
Ornela N. Dehayem
Email: o.dehayem@gmail.com

Funding information
Dutch Research Council, Grant/Award Number: OCENW.KLEIN.498

Handling Editor: Will Pearse

Abstract

1. Phylogenetic trees are increasingly used to infer the processes that shape biodiversity patterns, such as diversification, dispersal and trait evolution. However, collecting many samples and sequencing them for phylogenetic reconstruction is challenging and costly, and achieving high sampling fractions can be logistically impossible for species-rich groups or regions. When studying isolated environments such as islands, this issue applies to the species that make up the island community, as well as outgroup taxa when sampling from a large pool of mainland relatives is required.
2. In this study, we use simulations of the island biogeography model DAISIE (Dynamic Assembly of Islands through Speciation, Immigration and Extinction) to investigate the best sampling strategy to minimize error in the estimation of the model parameters when there is a constraint on the number of species that can be sampled. We compare three different community-level sampling strategies for islands: (1) prioritizing species-rich island lineages; (2) prioritizing species-poor island lineages; and (3) random selection of island lineages. Furthermore, we explore the effect of the nature of the missing data within each (species-rich or species-poor) lineage by testing the impact of incomplete sampling of the oldest and youngest extant species within each lineage and assess the effect of excluding outgroup species or even entire island lineages.
3. Parameter estimates of speciation, colonization and extinction rates of island lineages show slightly larger errors when the unsampled species belong to species-rich lineages. Within clades, we observed larger errors when the unsampled species were the oldest or outgroup species (i.e. mainland species sampled to determine the stem age of the clade).
4. When sampling is limited by time and/or budget, our study suggests prioritizing sampling of the phylogenetically most distinct species from the most diversified island lineages along with their mainland relatives.

KEYWORDS
biodiversity patterns, colonization, extinction, island biogeography, sampling strategies, simulation, speciation

1 | INTRODUCTION

Understanding the processes that have contributed to the incredible variety of life on Earth has long been a key goal in biogeography, ecology and evolution, prompting research into the underlying mechanisms governing the dynamic evolution of biodiversity. In recent decades, the advent of new molecular techniques has resulted in increasingly detailed and complete phylogenetic trees, which have facilitated the emergence of novel methods that leverage information from these phylogenetic trees to study their past diversification dynamics (Etienne et al., 2012; Etienne & Rosindell, 2012; Hey, 1992; Maddison et al., 2007; Morlon, 2014; Nee et al., 1992, 1994). However, these methods generally require a good level of taxonomic sampling (i.e. including the majority of known species in the phylogeny), which is often difficult to achieve in practice. It has been shown that incomplete sampling can introduce inaccuracies in the inferences from diversification models, which, in turn, may compromise the precision of understanding biogeographical history or diversification dynamics (FitzJohn et al., 2009; Garamszegi & Møller, 2011; Molina-Venegas & Lima, 2021; Mynard et al., 2023; Sun et al., 2020).

The extent to which taxonomic sampling may impact the results of biogeographical and diversification analyses has long been a concern in the field, and its effects have been examined for methods relying on a single lineage (a single phylogenetic tree), generally of large (continental) clades (Chang et al., 2020; Davis et al., 2013; Mynard et al., 2023). However, studies on the effect of taxonomic sampling in methods that use information from communities and/or multiple lineages (including different phylogenetic trees) or that focus on non-continental systems, such as islands, are lacking. Biological communities on islands are among the most effective model systems for investigating biogeographical and diversification mechanisms. With their distinct boundaries, relatively recent origins, and simplified ecosystems compared to continents, islands offer ideal arenas for studying evolution and ecology and have long been regarded as excellent natural laboratories for biodiversity research (Gillespie & Roderick, 2014; Helmus et al., 2014; Losos & Ricklefs, 2009; Warren et al., 2015). Their isolation often leads to the development of unique and endemic species, with oceanic islands harbouring 15%–20% of the world's terrestrial species, despite covering only 3.5% of the Earth's surface (Matthews & Triantis, 2021; Whittaker et al., 2017). Hence, their combination of tractable features and high biodiversity provides ideal settings for studying species diversification (Losos & Ricklefs, 2009; Matthews & Triantis, 2021).

Even though islands are small and generally more species-poor than continents, comprehensive phylogenetic data collection of an entire insular assemblage is often challenging. Collecting all island species and their mainland relatives, especially for very diverse communities such as tropical islands (Ralimanana et al., 2022), requires significant time and budget for fieldwork, collecting permits for all species including those that are rare or protected, and additional budget to generate and analyse the sequences. In addition to time and budget constraints, some islands may be geographically remote and some localities within islands may be difficult to access due to harsh terrains. Geopolitical constraints, the recent extinction of certain species and the sometimes narrow spatial range of endemics also put an extra burden on data collection. Consequently, most phylogenetic datasets used to infer the diversification and biogeography of island communities are incomplete (Valente et al., 2015, 2019; Valente, Illera, et al., 2017).

Understanding the pattern of species assembly on islands involves unravelling the evolutionary processes responsible for their origin, and how they vary over time and across lineages (Florencio et al., 2021). In an island context, the DAISIE (Dynamic Assembly of Islands through Speciation, Immigration and Extinction) framework can be used to infer the rates of the processes that govern the assembly of an entire community based on phylogenetic information, which includes lineage colonization times, branching times (extracted from time-calibrated molecular phylogenies), endemicity status (whether or not species are endemic to the island) and island age (geological age of emergence, Valente et al., 2015). While other diversification models and methods (e.g. state-dependent diversification models and corresponding inference methods) have been studied regarding their sensitivity to incomplete sampling (e.g., state-dependent diversification methods) (Davis et al., 2013; Mynard et al., 2023), DAISIE uses an unconventional data format and its sensitivity to incomplete sampling is not yet known.

Recent studies emphasize that accurately reconstructing species diversity dynamics is crucial for understanding equilibrium states in island communities (Valente, Illera, et al., 2017). Ignoring incomplete data can introduce significant biases in parameter estimation, potentially distorting conclusions about key processes such as colonization and extinction. Missing data on island taxa may lead to incorrect estimation of key clade attributes, such as stem and crown ages, and the overall tree topology. Similarly, missing data on the mainland relatives of island lineages can result in the overestimation of divergence times between these mainland relatives and island lineages, affecting the inferred timing and tempo of diversification events (García-Verdugo et al., 2019; Lambert et al., 2022). Fortunately, as with many contemporary diversification methods, the DAISIE can account for incomplete sampling of phylogenetic data, by integrating over the possible placements of missing taxa within the clade from which they originated. However, although this can certainly mitigate some issues associated with missing data, its accuracy does not match the ideal situation when all species are included. Moreover, the DAISIE assumes random sampling, which may not always be the case. Therefore, it is important to determine to what extent incomplete datasets can provide information for reliable inference of biogeographical and diversification parameters.

Here, we used process-based simulations of the DAISIE model to identify which type of incomplete datasets minimize errors in parameter estimates in cases where not all phylogenetic information can be included. To this end, we simulated different incomplete sampling scenarios and fitted the DAISIE inference model to evaluate the error in the rates of colonization, anagenesis, cladogenesis, extinction, and the clade-level carrying capacity for each of the incomplete datasets. We first explored three island-level sampling strategies, one with random sampling and two that sample species based on the diversity of the clade from which they originate, either only sampling from the most diverse island lineages (containing four or more species) or only from the least diverse island lineages (containing less than four species). With this analysis, our objective was to identify whether prioritizing taxonomic sampling based on the size of the island lineage could be beneficial or detrimental.

For each of the three community-wide sampling strategies, we further explored the impact of failing to sample certain types of species, by investigating scenarios with incomplete sampling of (i) the oldest species (extant species that have the longest individual tip branches) versus young species (extant species that have the shortest individual tip branches), (ii) outgroup species (continental species closely related to each island lineage, which are used to estimate the stem age of the clade and thereby the upper limit of its colonization time), and (iii) entire lineages (that is, failing to sample all species from a given island lineage). Subsequently, for all the above scenarios of missing data, we assessed the robustness and sensitivity of diversification and island biogeography parameters and identified which specific regions of the phylogeny, in terms of unsampled species, have the greatest impact on the precision and accuracy of parameter estimation. We repeated analyses under different fractions of unsampled species to assess how total sampling percentage affects findings. We discuss these results in the context of islands, but also in the context of biogeography and diversification analyses in general.

2 | MATERIALS AND METHODS

Our pipeline to assess the impact of incomplete sampling is as follows: first, we fitted the DAISIE model to four existing, but quite different, empirical island datasets to estimate parameters (colonization, speciation, extinction, and eventually the carrying capacity). This provided us with various realistic parameter sets. Second, we simulated "complete" island datasets using these parameters. Third, we created "incomplete" datasets from these simulated complete datasets by removing species or lineages under a variety of sampling scenarios. Fourth, we fitted the DAISIE to the complete and incomplete datasets to estimate parameters and compare them to the generating parameters and to the estimates under other sampling scenarios. Below we describe each of these steps in detail.

2.1 | Model description

For our analysis, we used the DAISIE, a framework that leverages the phylogenetic information of an island community to infer the rates of processes that have shaped its diversity and composition (Valente et al., 2015). The model assumes that the assembly of species on an island is primarily driven by three key processes: colonization from a fixed mainland (where no diversification or extinction occurs within the mainland source community over time), speciation through anagenesis (evolution into a species distinct from the mainland ancestor due to reduced gene flow between island and mainland populations) or cladogenesis (the process where an ancestral species gives rise to two or more species), and extinction. Together, these processes result in a community that can be captured in one or multiple phylogenetic trees, each representing a colonization event from the mainland that has led to an island lineage comprising one or more species (Figure 1a). To estimate per-lineage rates of colonization, speciation through anagenesis and cladogenesis, and extinction, the DAISIE uses information on island age (geological age of emergence of the island), lineage colonization times (estimated from genetic divergence of these island species and their mainland counterparts), lineage endemicity status and lineage branching times (extracted from time-calibrated molecular phylogenies) (Valente et al., 2015). Additionally, the DAISIE allows colonization and cladogenetic speciation to be diversity-dependent, meaning that the rates of colonization and cladogenesis decline as the number of species on the island increases (Etienne et al., 2012). This allows estimation of the diversity limits of clades and helps assess whether a community is close to this limit, thereby determining whether diversity dependence has influenced community dynamics. In contrast, the rates of anagenesis and extinction are assumed to be uniform across lineages and remain constant over time. These estimates can be used to simulate the dynamics of the island community through time, providing insights into the processes shaping island biodiversity and species richness evolution (Hauffe et al., 2020; Jiménez-Ortega et al., 2023; Neves et al., 2021; Valente et al., 2020; Valente, Illera, et al., 2017; Xie et al., 2023). The DAISIE accounts for missing species in clades, whether due to extinction before the present or incomplete sampling, by integrating across all possible placements weighted by their probability according to the model. The model can also address uncertainty in colonization times in the following way. If the closest continental relatives for some lineages are unknown or unsampled, the DAISIE allows the maximum colonization time to be entered as input; subsequently, it will integrate over all possible ages between the maximum colonization time and the present or the crown age of the radiation. When only the upper limit of the divergence time is known, the model integrates over all possible colonization times, that is from that upper limit to the crown age for endemic clades, or from the upper limit to the present for endemic singleton lineages. If the upper limit is not known, the model integrates from the island's origin (island age) to the present.

2.2 | Fitting DAISIE to empirical datasets

To use realistic parameter sets in our simulations, we first used maximum likelihood (ML) to fit the diversity-independent DAISIE model to four empirical datasets with varying sizes in terms of total species numbers and numbers of colonization events, allowing us to investigate the impact of different sampling scenarios on datasets with different sizes (Table 1). The four empirical datasets have different levels of phylogenetic sampling, but our aim of fitting the DAISIE to these datasets is not to investigate the impact of sampling on these specific datasets, but rather to obtain a set of empirically informed

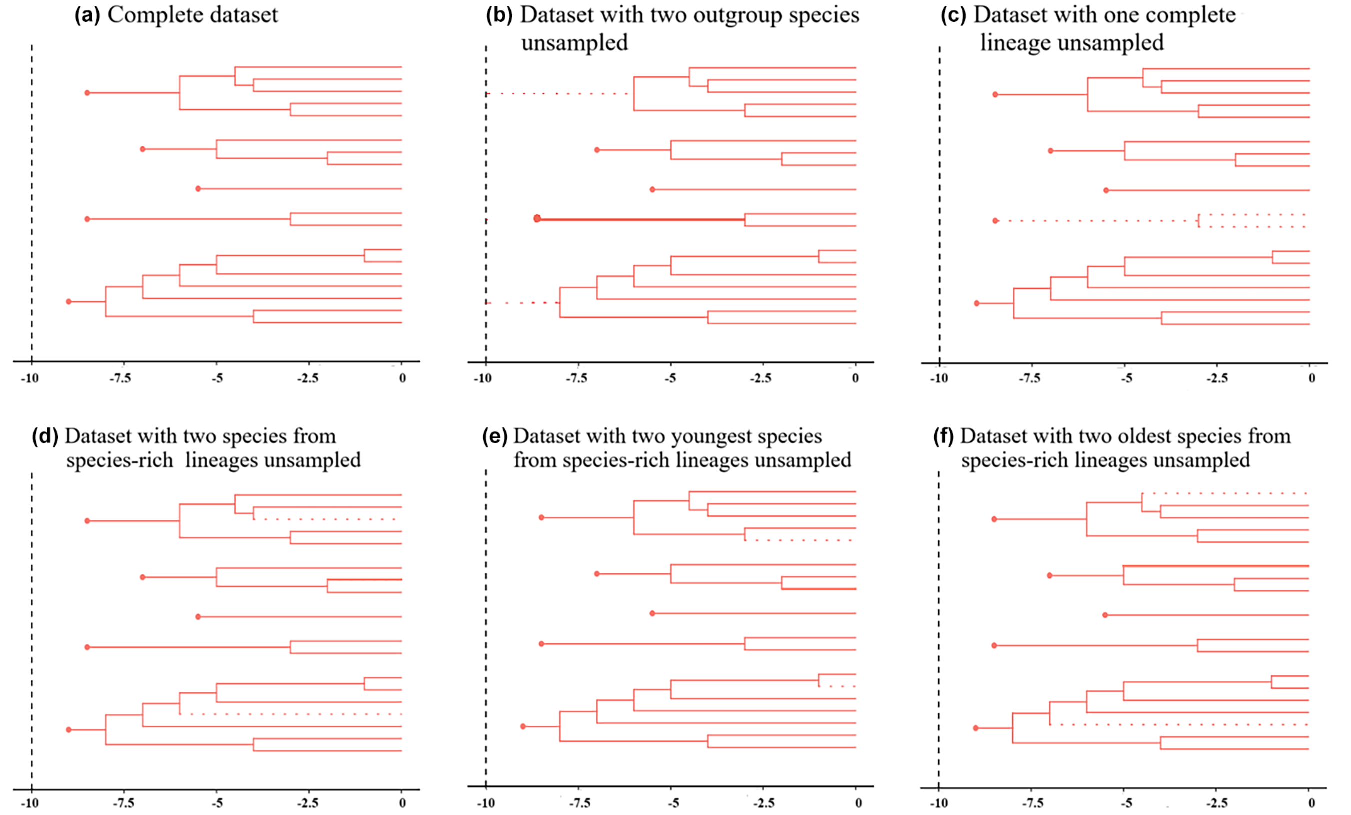
(a) Complete dataset
(b) Dataset with two outgroup species unsampled
(c) Dataset with one complete lineage unsampled
(d) Dataset with two species from species-rich lineages unsampled
(e) Dataset with two youngest species from species-rich lineages unsampled
(f) Dataset with two oldest species from species-rich lineages unsampled
-10
-7.5
-5
-2.5
0

FIGURE 1 Examples of incomplete datasets generated in our study. Each panel represents the phylogeny of an island community. Each tree represents a colonization event that has established one or more descendants on the island and includes one close continental relative. The dotted lines represent non-sampled species. Dots at the stem of a lineage indicate the upper limit of colonization time of each lineage. (a) Complete dataset, (b) dataset with two outgroup species unsampled (thus the time of colonization is not known for these two island clades, only island species richness and within-island branching times [if any] are known), (c) dataset with one complete lineage unsampled (colonization and branching times are not known, only species richness is known), (d) dataset with two species from species-rich lineages unsampled, (e) dataset with two youngest species from species-rich lineages unsampled and (f) dataset with two oldest species from species-rich lineages unsampled.

TABLE 1 Summary of the four empirical datasets and the simulated dataset.

Dataset	Mean total species	Mean number of colonization events	Data source	Island age (million years)
Galapagos birds	25	8	Valente et al. (2015)	4
Greater Antilles bats	37	16	Valente, Etienne, et al. (2017)	20
Hispaniola frogs	65	5	Etienne et al. (2023)	30
New Zealand birds	72	39	Valente et al. (2019)	52
Large dataset (simulated)	1268	581	This study	21

parameters to generate simulated datasets that are similar to these empirical datasets in terms of the number of species, endemicity status and clade size distribution. We note that there might be other models that fit the data better, but our purpose is not to find the best-fitting model but to generate a parameter set that would generate datasets resembling empirical datasets.

2.3 | Simulating complete island data sets

We used a standard algorithm that allows stochastic simulations in continuous time (Gillespie, 1976) and that is implemented in the DAISIE package version 4.6.0, to simulate island phylogenetic datasets for the ML parameter sets obtained from fitting the model to each empirical dataset. It is worth noting that in a few cases, some of these simulated datasets consisted of exclusively species-rich or species-poor clades (according to our definition of rich or poor, see below). This occurrence can be attributed to the stochastic nature of the model. We excluded these specific datasets from our analysis, as we aimed to examine the impact of unsampled species from both rich and poor clades (see below). This means that the comparison of the parameter estimated from the complete dataset and the true generating parameters becomes biased. However, our primary interest lies in comparing the parameters estimated from the complete dataset with those estimated from the generated incomplete dataset. We initially generated enough datasets to ensure that, after discarding those with only species-rich or species-poor clades, we still have a total of 1000 datasets. To investigate the impact of unsampled species on larger and more complex datasets (e.g. comparable to the number of plants of a diverse oceanic archipelago, such as the Canary Islands), we also created a set of 1000 datasets with a parameter set that simulated an island community with a higher number of species (1000 – 1200) species and colonization events (400 – 600) (Table 1). We will refer to these as the 'Large' dataset in the following sections.

2.4 | Simulating incomplete island datasets

To account for common limitations in data collection, we assumed that only a subsample of the total number of known species in the community could be sampled. Therefore, we trimmed each of the simulated datasets to replicate various scenarios of incomplete sampling, using different fractions of unsampled species (5%, 20%, 40% and 60%) with two different thresholds for considering a lineage species-rich (four and seven). For each complete simulated dataset, we created three different incomplete datasets using the following three primary sampling strategies:

- Primary sampling strategy 1: In this strategy, a given fraction of unsampled species is randomly pruned from clades within the focal community to create the incomplete dataset.
- Primary sampling strategy 2: This strategy prioritizes sampling species from the most diversified lineages ('species-rich' lineages). We pruned the given fraction of unsampled species from the species-poor lineages (≤four species) in the phylogeny to create the incomplete dataset.
- Primary sampling strategy 3: This strategy prioritizes sampling species from the least diversified clades ('species-poor' lineages). We pruned the given fraction of unsampled species from the species-rich lineages (>four species) in the phylogeny to create the incomplete dataset.

For each of the three primary sampling strategies listed above, we created five distinct types of incomplete datasets, which we referred to as 'secondary sampling strategies'. Each secondary sampling strategy represents a scenario of incomplete sampling within clades:

- Secondary sampling strategy 1: This dataset was produced by removing (i.e. not sampling) species randomly within each clade.
- Secondary sampling strategy 2: This dataset was produced by removing the oldest species from clades. We define the oldest species as the island species that have the longest individual branch in the tree. This allowed us to assess the effect of unsampled old species on parameter estimation.
- Secondary sampling strategy 3: This dataset was produced by removing the youngest species from clades. We define the youngest species as the island species that have the shortest individual branch in the tree. This scenario aimed to examine the impact of unsampled young species on parameter estimation.
- Secondary sampling strategy 4: This dataset was produced by removing outgroup species. Outgroup species are continental ('mainland') species that are closely related to the island clade. They can provide a reference point for estimating the divergence time of the island clades from their mainland ancestors, and hence the colonization time (Martín-Hernanz et al., 2023). Within the DAISIE framework, excluding the outgroup species of a clade corresponds to removing its stem age. This incomplete dataset was only created for the Large dataset because other datasets have too few colonization events, which prevents comparing the effect of excluding outgroup species from rich lineages to excluding outgroup species from poor lineages. But even for this dataset, we could only consider a scenario with 5% of unsampled species, as some datasets contained very few species-rich or species-poor clades.
- Secondary sampling strategy 5: This dataset was produced by removing some clades entirely (regardless of clade size), including the stem age.

It should be emphasized that in our analyses, pruning away a specific species, or an entire clade, means removing the corresponding phylogenetic data (stem age or branching times). However, for all removal scenarios, the number of species removed from each clade ('unsampled species') and their endemicity status are known (e.g. mimicking the empirical case where this is known based on checklists of island species) and thus they are included in the analyses by specifying this when preparing the data to be run in the DAISIE R package.

In total, we examined 15 scenarios of incomplete sampling (5 secondary sampling strategies for each of the three primary sampling strategies), each with four degrees of incompleteness (5%, 20%, 40% and 60%).

We note that the sampling effort for the incomplete data (b), where some outgroup species are missing, differs from the other datasets. This is because, typically, multiple mainland relatives are sampled to identify the closest mainland relative. To ensure that the comparison between the different sampling methods is not biased by differences in sampling effort (number of lineages samples and sequenced), we assumed that a single outgroup species is sampled for each clade on the island.

2.5 | Fitting DAISIE to complete and incomplete simulated datasets

To determine the best sampling method, we fitted the DAISIE model that was used to simulate the data to each complete and incomplete dataset generated by each sampling method. We estimated the rates of speciation (anagenesis and cladogenesis), extinction, colonization and the carrying capacity for each sampling method. Comparing the estimates from the incomplete datasets and the corresponding complete dataset allows us to determine the effectiveness of each incomplete dataset generated by different sampling methods in estimating rates of processes that generated the community.

Importantly, fitting the DAISIE to the complete simulated datasets does not guarantee that the true generating parameters (those parameters that were used to simulate the complete simulated datasets) will be correctly estimated, as the DAISIE has some inherent baseline error (Neves et al., 2021; Xie et al., 2023). By fitting the model to different incomplete datasets, the error in parameter estimates can stem from both the incompleteness and the DAISIE model itself. Thus, we compared the parameter estimates from the incomplete dataset to (i) those obtained using the complete simulated dataset and (ii) the true generating values. This helps to isolate the sources of error and provides a clearer understanding of how the model performs when working with complete data before evaluating its performance with incomplete datasets.

3 | RESULTS

The results presented below are based on the analysis of the Large dataset. The results for the four empirical datasets (Galapagos birds, New Zealand birds, Hispaniola frogs and Greater Antilles bats) can be found in the Supporting Information. We selected a threshold of four species to distinguish between species-poor and species-rich clades, defining clades with more than five species on the island as species-rich. The aim was to investigate the impact of missing species across clades with varying degrees of species richness. We found that the results remain consistent even when a different species-richness threshold is applied (Figure S15).

We began our analysis by assessing the model's accuracy in estimating the true generating values (i.e. the initial parameters used to simulate the complete simulated datasets) from the complete datasets (beige colour violin plots in Figures S4 and S5). The results show that the model performs well in recovering the correct colonization, cladogenesis, anagenesis and extinction rates (Figures S4 and S5). Notably, the median of the relative error (relative to the real values used to simulate the datasets) in parameter estimates across the 1000 complete simulated datasets was very low, close to zero (Figures S4 and S5). This was also true for the incomplete datasets.

Next, we compared the parameter estimates obtained from incomplete datasets to those derived from complete datasets (e.g. Figure 2). Generally, except for the rate of anagenesis, the width of the distribution of the error in parameter estimates did not show significant differences across incomplete datasets with the same proportion of missing species (e.g. Figure 2). However, we observed that the median errors across the 1000 simulated incomplete datasets were slightly lower when species were removed from species-poor clades compared to when species were unsampled from species-rich or randomly selected clades (e.g. Figure 2). For the anagenesis rate, the width of the error distribution tended to decrease when the unsampled species were from the most species-rich clades (e.g. Figure 2; Figures S1–S3).

Furthermore, incomplete sampling had distinct effects on parameter estimates when different groups of species were not sampled within clades (i.e. depending on the secondary sampling strategy). The width of the distribution of the error in parameter estimates did not show significant differences across all incomplete datasets with the same proportion of missing species, but the median errors across the 1000 incomplete datasets were higher when the proportion of unsampled species within clades were the oldest species and even higher when they were outgroup species (Figures 3–5; Figures S6–S14). For datasets with many colonization events (e.g. the Large dataset and the New Zealand birds), median errors tended to be lower when unsampled species were from entirely unsampled lineages (Figures 3–5; Figures S11 and S13). The difference in the impact of not sampling these groups of species on parameter estimates was especially noticeable when species were removed (i.e. not sampled) from species-rich lineages (Figures 3–5; Figures S6–S14).

We also assessed how the size of the dataset impacts the consequences of incomplete sampling. Specifically, we investigated how different-sized communities, characterized by varying levels of species diversity and colonization events (Table 1), responded to incomplete sampling. In general, when considering the same proportion of unsampled species (5%, 20%, 40% and 60%), our study demonstrates that larger incomplete datasets tend to yield more precise parameter estimates compared to smaller incomplete datasets, as expected (Figure 6). Moreover, across all datasets, the width of the distribution of the error in parameter estimates consistently decreases as the sampling fraction increases (Figure 2). However, at intermediate sampling (sampling fraction = 60%), when missing species are from species-rich lineages, the median errors in parameter estimates are lower than for higher sampling (sampling fraction = 80%) (e.g. Figure 2).

4 | DISCUSSION

Incomplete taxonomic sampling is inherent in almost all phylogenetic research, and this also applies to small, 'simpler' systems such as islands. Although island taxa typically occur in a more confined area compared to the often more widely distributed continental taxa that can cover multiple continents, sampling all species of island clades is often not possible (Bank et al., 2021; Ralimanana et al., 2022). Under the realistic assumption that most empirical phylogenetic studies will not achieve complete sampling, we aimed to identify how the level of sampling can affect biogeographical and diversification analyses of island species, and to identify whether certain taxonomic sampling strategies may be more efficient to mitigate limitations resulting from incomplete sampling.

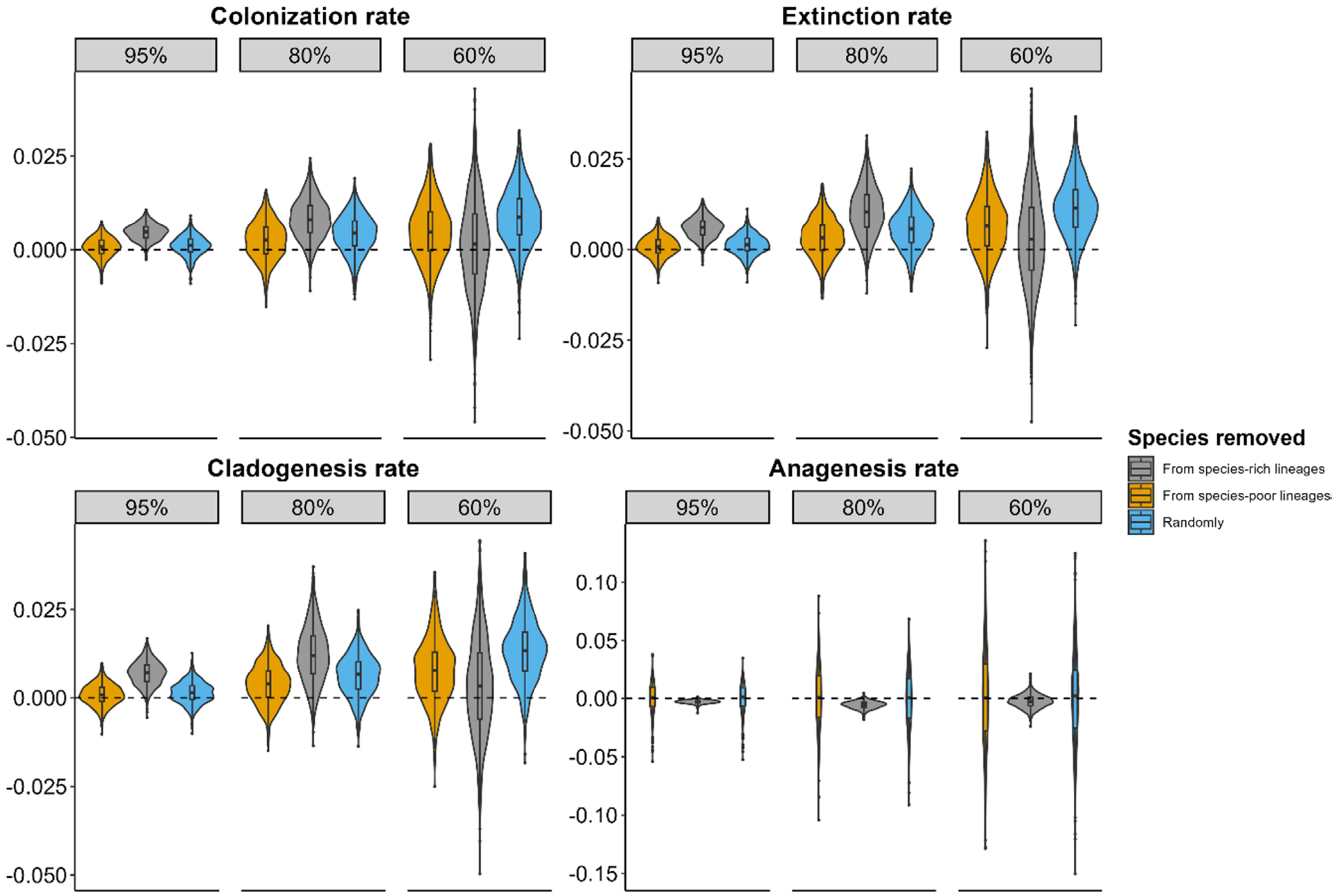

FIGURE 2 Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset. Violin plots show the error distributions for colonization, extinction, cladogenesis and anagenesis rates across simulations from incomplete datasets with varying sampling fractions. These datasets were generated using the three primary sampling strategies, with species in clades sampled randomly. Removing (i.e. not sampling) species from species-rich lineages results in less accurate parameter estimates. All errors are calculated relative to the parameter estimates for the complete simulated dataset.

To address the issue of incomplete taxonomic sampling, we used the DAISIE model as a case study to identify sampling strategies that minimize errors in parameter estimation as well as quantify the errors for a given existing sampling strategy. We ran simulations of different island datasets with variable levels of complexity (e.g. number of species and colonizations) where mimicked incomplete sampling by removing species according to the size of their clade (species-rich clade vs. species-poor), their phylogenetic position (outgroup species, oldest species and youngest species) or by excluding some lineages entirely.

Our study indicates that the impact of unsampled species on parameter estimation varies depending on both clade size (primary sampling strategies) and the specific species that are unsampled (secondary sampling strategies). We found that missing species from the most species-rich clades resulted in a slightly larger impact on parameter estimates than pruning species from species-poor clades or randomly selected clades (Figure 2). We argue that this is because species-rich clades have multiple potential positions for the missing species within the phylogeny, which increases the likelihood of misplacement and consequently introduces greater errors in parameter estimates. This finding leads to the clear recommendation to prioritize sampling of species-rich island lineages in the DAISIE context. However, it may also have implications for other types of models. While the distinction between species-rich and species-poor clades is not relevant for single-lineage diversification analyses (e.g. analyses of clade-wide homogenous rates in Bayesian Analysis of Macroevolutionary Mixtures (BAMM)—Rabosky et al., 2014 or R: Phylogenetic ANalyses of DiversificAtion (RPANDA)—Morlon et al., 2016), it may be important for methods that test for diversification rate shifts at specific nodes or branches (e.g. shift models in BAMM—Rabosky et al., 2014, Diversity-Dependent Diversification (DDD) key innovation—Etienne et al., 2012) or trait-dependent diversification (e.g. state-dependent speciation and extinction [SSE] models) (Beaulieu & O'Meara, 2016; Herrera-Alsina et al., 2019; Maddison et al., 2007). In such methods, there are effectively subclades with different diversities that are governed by different diversification regimes. We hypothesize that failing to sample species from the more diverse subclades may have a greater impact on the accuracy of parameter estimates, but this will need to be tested for each specific model.

Regarding sampling strategies within targeted clades (secondary sampling strategies), the most significant effect on parameter

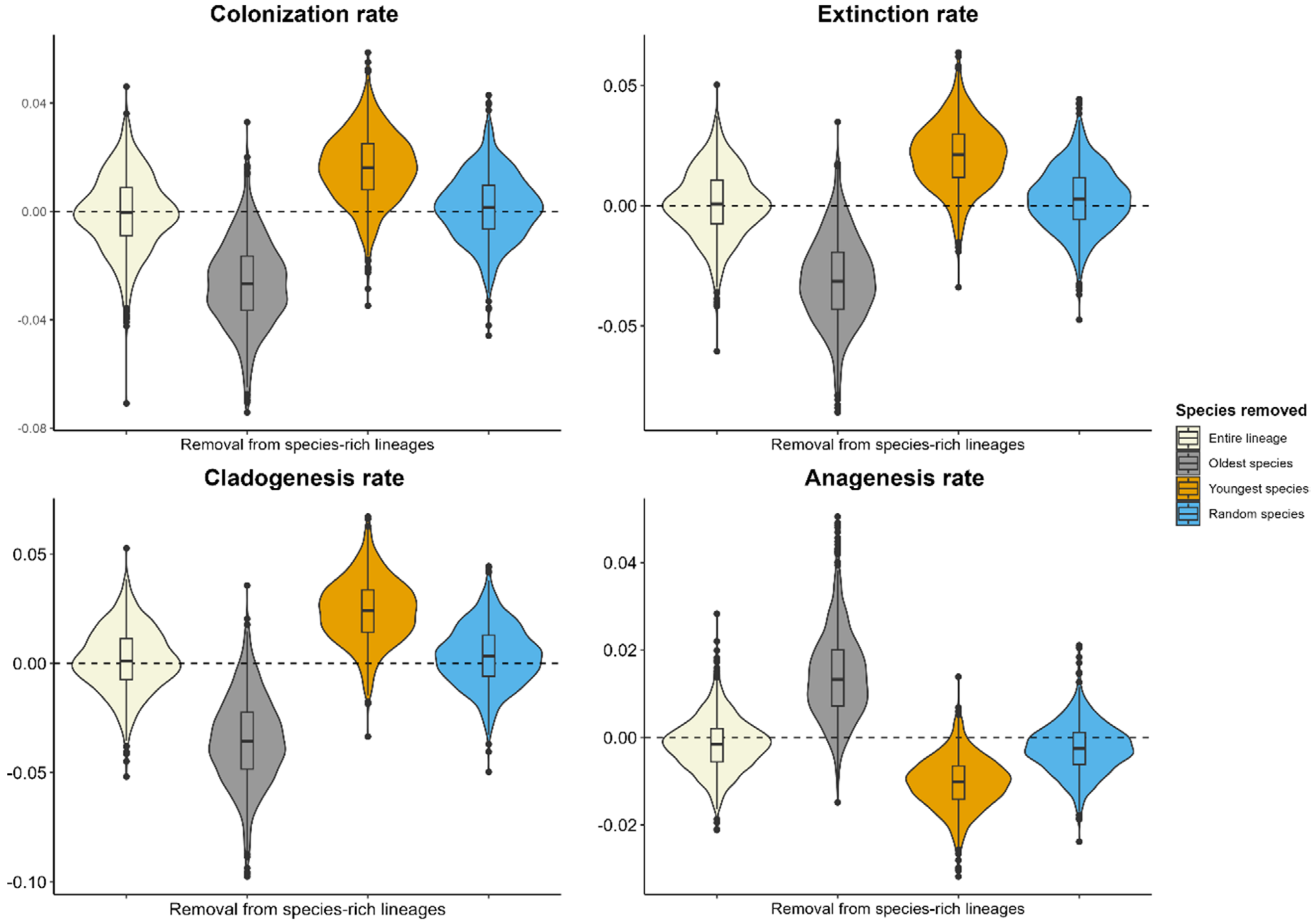

FIGURE 3 Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset. Violin plots show the error distributions for colonization, extinction, cladogenesis and anagenesis rates across simulations from incomplete datasets (here 40% missing species) generated under the secondary sampling strategies. Removing (i.e. not sampling) the oldest species results in less accurate parameter estimates. All errors are calculated relative to the parameter estimates for the complete simulated dataset.

estimates was observed in datasets lacking complete information on clade colonization times (Figures 4 and 5). This corresponds to secondary sampling strategies where phylogenetic data for outgroup species is unavailable for certain clades. Our findings align with those of Valente et al. (2018), who observed that including clade colonization times in phylogenetic information leads to more precise and accurate parameter estimates, compared to relying solely on species richness and endemicity status. This suggests that information about mainland relatives is essential for understanding how communities have assembled on the islands. Outgroup species provide essential data on the divergence time between extant island lineages and their closest mainland relatives, which is necessary for estimating clade colonization times (Martín-Hernanz et al., 2023). When this outgroup data are absent or incomplete, divergence times between mainland relatives and island lineages may be overestimated (García-Verdugo et al., 2019; Lambert et al., 2022), potentially affecting the inferred timing and tempo of diversification events. Beyond the DAISIE, for single-clade diversification models and models that are mostly relevant to continental clades (e.g. studying diversification of continental genus), this suggests that having a good estimate of the stem age of the clade in question may be important, and we therefore recommend paying more attention to outgroup sampling for diversification and biogeographical approaches in general (rather than focussing exclusively on the ingroup).

We also observed a higher impact on parameter estimates when the unsampled species in incompletely sampled clades were the oldest species (species with the longest individual branches): The error was higher when the unsampled species were the oldest than when the unsampled species were randomly chosen or were the youngest species (species with the shortest individual branches), and was also higher than the error in the scenario where entire lineages were left out (Figures 3–5). The higher error found when failing to sample older species may be due to the fact that later stages of an island radiation have more branches and more species that can speciate, making it more likely that species are missing from late-diverging branches. Hence, removing these is then more or less in line with the model, whereas removing the oldest species is unlikely according to

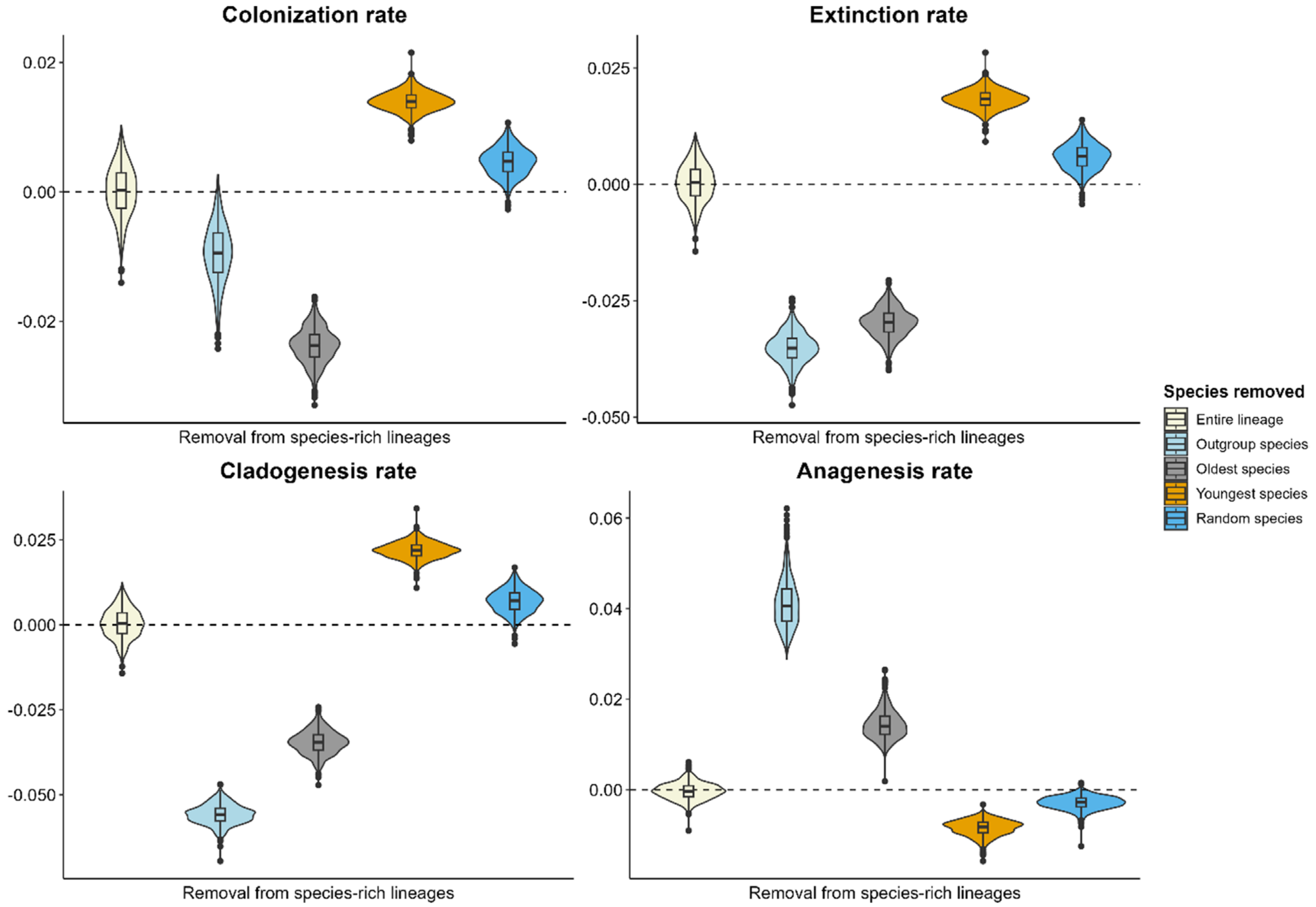

FIGURE 4 Effect of sampling strategy on parameter estimation for simulated datasets generated with the parameters that created the Large dataset. Missing species are removed from species-rich lineages. Violin plots show the error distributions for colonization, extinction, cladogenesis and anagenesis rates across simulations from incomplete datasets (here 5% missing species) generated under the secondary sampling strategies. Removing (i.e. not sampling) outgroup species or the oldest species results in less accurate parameter estimates. All errors are calculated relative to the parameter estimates for the complete simulated dataset.

the model and hence it is less well able to account for these missing data. Random sampling, on the other hand, aligns more closely with the way missing species are distributed within a clade, resulting in a smaller effect on parameter estimates compared to the targeted sampling of the oldest or youngest species. However, when the proportion of missing species was small (5% and 20%), datasets with a larger number of colonization events (e.g. the Large dataset and the New Zealand birds) where some species-rich lineages were completely unsampled also led to lower median error in parameter estimates (Figures 3–5; Figures S11 and S13). We argue that this happens because, at small proportions, the missing species typically represent one or very few clades relative to the dataset's size. Consequently, the dataset retains much of its overall structure, reducing the effect of the missing species on the parameter estimates. While this result is based on the DAISIE model, it is likely that this may also affect other models. For instance, when studying large continental clades to test for speciation, extinction or carrying-capacity shifts (e.g. BAMM or DDD), we anticipate that failing to sample older species of subclades for which a diversification shift has been detected may cause similar issues.

Except for datasets with incomplete information on the colonization times (i.e. outgroup species unsampled), the difference in the impact of missing species on parameter estimates was more noticeable when unsampled species were removed (i.e. unsampled) from species-rich lineages (Figures 4 and 5). For lineages with few species, removing the oldest, youngest or randomly selected species often results in similar tree structures due to the limited complexity of such lineages (Figure 5; Figures S6–S14). This explains why the sampling strategy may not matter as much for small clades, compared to species-rich lineages.

Missing species have a relatively minor impact on parameter estimates for large datasets (the Large dataset). For instance, as shown in Figure S16, the errors in parameter estimates for the Large dataset with a 60% sampling fraction are equal to or lower than those observed for smaller datasets at a 95% sampling fraction. Similar findings were reported by Mynard et al. (2023) in state-dependent speciation and extinction models, who observed that larger, incomplete phylogenies can be more effective than smaller, more complete sub-clades to study patterns of trait-dependent diversification. This supports the idea that larger datasets are generally more robust and

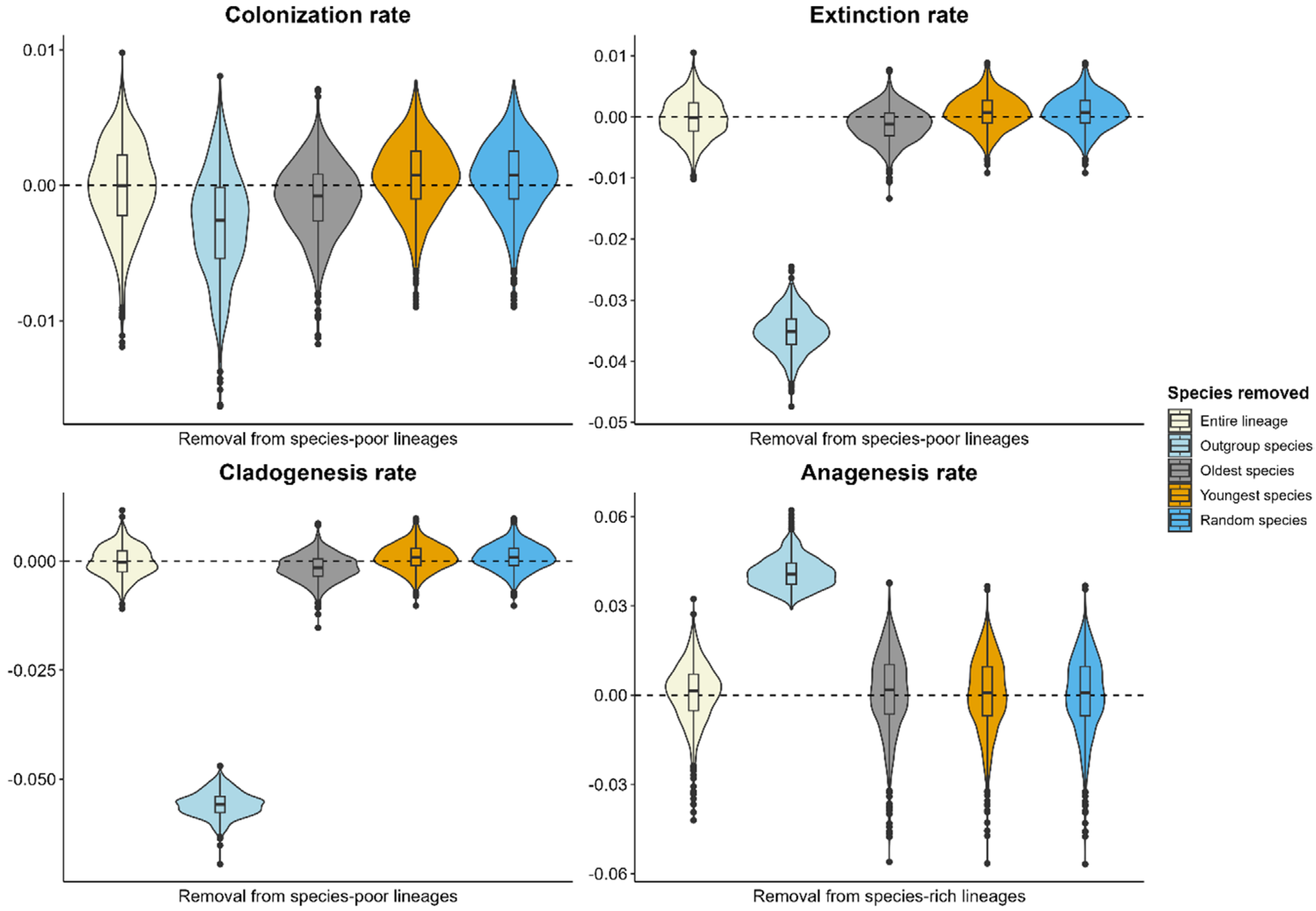

FIGURE 5 Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset. Missing species are removed from species-poor lineages. Violin plots show the error distributions for colonization, extinction, cladogenesis, and anagenesis rates across simulations from incomplete datasets (here 5% missing species) generated under the secondary sampling strategies. Removing (i.e. not sampling) outgroup species or the oldest species results in less accurate parameter estimates. All errors are calculated relative to the parameter estimates for the complete simulated dataset.

provide greater statistical power to detect underlying evolutionary patterns (Davis et al., 2013; Mynard et al., 2023). For smaller communities with fewer species, however, achieving complete sampling should be of higher priority. We do note that larger clades are often older and the assumption that the same rates and processes operate across these extended time scales may be violated more often than for smaller, younger clades. That is, uncertainty is then of a more fundamental nature.

As expected, we also observed that a higher sampling fraction (low proportions of missing species) results in more accurate and precise parameter estimates (Figure 2; Figures S1–S14), emphasizing the importance of comprehensive data collection in accurately characterizing the diversification dynamics of a biological community. Previous studies have also shown that incorporating more phylogenetic data further enhances the accuracy and precision of parameter estimates (Valente et al., 2018) as well as model selection (Davis et al., 2013; Mynard et al., 2023). Interestingly, at intermediate sampling (e.g. sampling fraction = 80%), the median errors are higher than for lower sampling (e.g. sampling fraction = 60%). We argue that this happens because, with a lower proportion of sampled species, the sampling process resembles random sampling better and consequently may in some cases have a lower impact on the parameter estimates. However, the distribution width error across the 1000 simulated datasets remains larger than when the sampling fraction is lower. As the sampling fraction increases further (e.g., 95%), the error decreases as the dataset becomes closer to the complete dataset.

Given our findings, the results of previous studies that have used the DAISIE appear to be qualitatively robust, with all analysed datasets having a sampling fraction of at least 85% at species level. This suggests that the overall qualitative insights from these studies remain reliable. However, some of the former studies (e.g. Hauffe et al., 2020; Valente, Illera, et al., 2017) have a high proportion of unknown colonization times, at least for some of the islands, which would be equivalent to the scenarios in our simulations where the outgroup or an entire species-poor lineage is removed. Therefore, the latter studies may benefit from increased sampling in future.

The DAISIE model relies on phylogenetic data to estimate rates of evolutionary processes that have shaped island communities, which in turn are used to simulate and understand island dynamics.

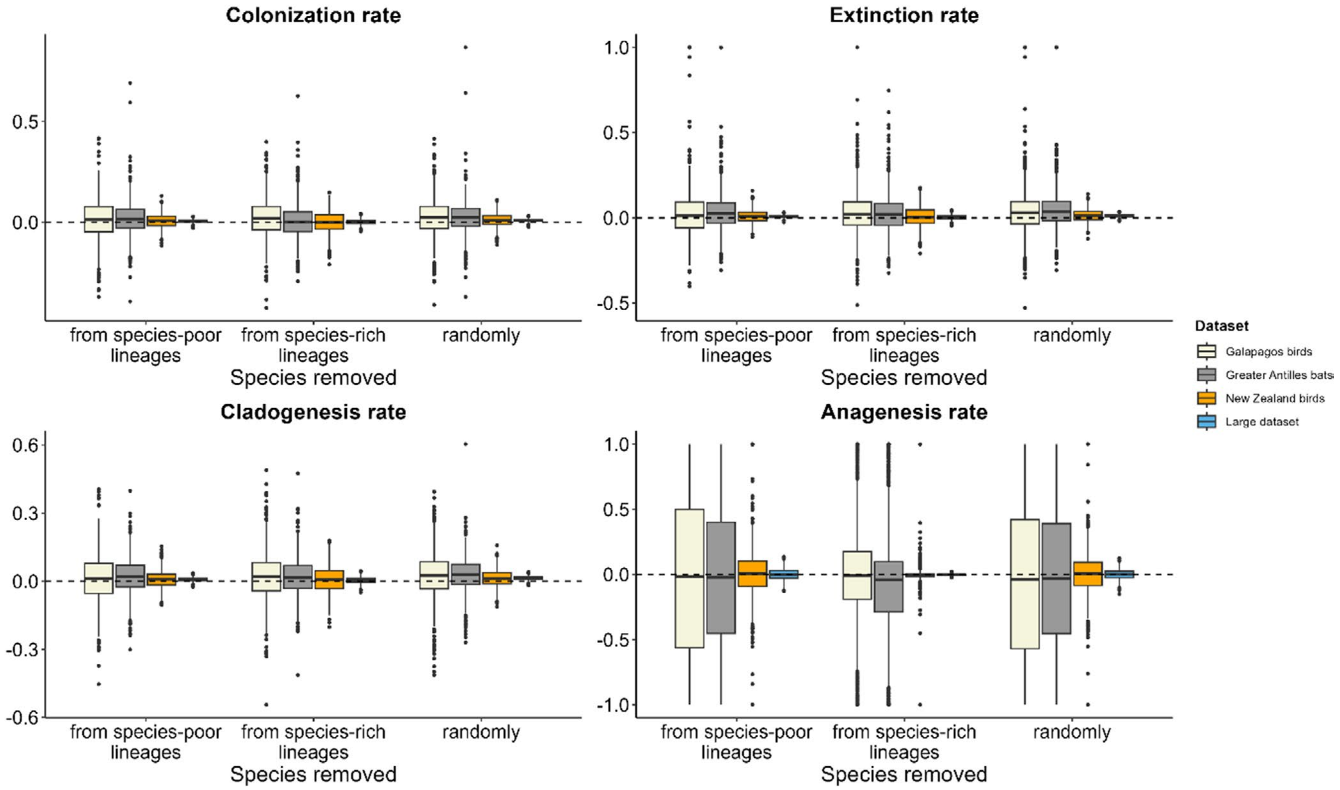

FIGURE 6 Effect of missing data on parameter estimation from datasets of different sizes. Box plots show the distribution of the errors in colonization, extinction, cladogenesis and anagenesis rates in simulations from incomplete datasets (here 40% unsampled species) with different sizes and numbers of colonizations generated by the three primary sampling strategies (removal from species-poor, removal (i.e. not sampled) from species-rich lineages, and random removal). All errors are calculated relative to the parameter estimates for the complete simulated dataset.

By studying the impact of different types of incomplete phylogenetic datasets on parameter estimates, our study offers practical guidance on selecting the optimal sampling strategy that minimizes errors in parameter estimates under the realistic scenario that not all species can be sampled. It is important to emphasize that our approach to determining the optimal species sampling strategy assumes some prior knowledge of the community of interest—for example, which clades are species-rich or poor, or which species are the oldest and the youngest.

Regarding prior information on lineage sizes, there are several comprehensive databases and research studies on island biodiversity that can provide prior knowledge about certain communities. For example, the list of wild species in the Canary Islands published in Acebes et al. (2009) offers detailed information on the region's flora and fauna, and Beierkuhnlein et al. (2021) list provides an extensive checklist of Canary Islands' plant species. Additionally, documentation such as the flora of Hawaii (Wagner et al., 2023) or checklists of Galapagos flora and fauna (McMullen, 1999; Parent et al., 2008) show the availability of prior knowledge for certain regions.

Regarding young versus old species, our secondary sampling strategies 2 and 3 require identifying the youngest and oldest species within a lineage, and therefore require a prior phylogeny with branch lengths. Because we are conducting a simulation study, we can readily obtain this information from our simulated phylogenies. However, in a real empirical study, identifying young and old species is much more challenging, because this information cannot be obtained in the absence of a phylogeny (which is exactly what the researchers would be aiming to achieve). While the identification of young and old species cannot be done for lineages without an existing phylogeny, this approach can be applied to groups for which a previous phylogeny already exists. There are many examples of such a situation—for instance, the many ongoing international initiatives to produce phylogenies using the same molecular markers, such as the Plant and Fungal Trees of Life Project (PAFTOL), which aims to reconstruct phylogenies using a standardized set of reduced-representation genomic sequences Zuntini et al. (2024), or the 'Bird 10,000 Genomes' (B10K) project (Stiller et al., 2024) project, which aims to reconstruct the avian tree of life with full genome data. In these cases, previously published phylogenies using a single or a few molecular markers often already exist, but researchers want to update those phylogenies using standard loci or an increased number of loci. When such prior phylogenies exist, researchers could take advantage of them to identify young and old species. Likewise, a community-wide phylogeny approach, such as the one we are discussing in the manuscript, requires a high-quality, standardized dataset from as many lineages as possible on a given island. This may involve resampling clades with existing phylogenies to ensure consistency in genetic markers, sequencing methods and protocols, as well as increasing the number of markers used to enhance the robustness of the phylogenetic relationship estimates. Again, information on young or old lineages could be obtained from the already existing phylogenies.

Even in cases where nearly complete sampling has been achieved but time or funding does not allow sequencing all species, our results can help researchers set priorities on which species to sequence: (1) focus on as many species as possible in species-rich clades, (2) if previous phylogenetic information is available, target the oldest species within clades for the new phylogeny and (3) include outgroup species (closely related relatives). Although our analysis utilizes specific software tools, we believe that it represents a general model, as any macroevolutionary approach inherently involves the fundamental processes of colonization, speciation, and extinction, with generalizable results to guide field sampling for future biogeographic studies.

AUTHOR CONTRIBUTIONS

Ornela N. Dehayem, Ryan F. A. Brewer, Luis Valente, Frederic Lens and Rampal S. Etienne conceived the ideas and designed the methodology. Ornela N. Dehayem conducted the analysis and wrote the first draft. All authors contributed to developing ideas and revising the manuscript.

ACKNOWLEDGEMENTS

We would like to thank Pedro Neves, Raphaël Scherrer, Lizzie Roeble and Vicente Castillo for their input. The Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine and Hábrók high-performance computing clusters. The Dutch Research Council for financial support through an (OCENW.KLEIN.498) grant awarded to FL and RSE.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70058.

DATA AVAILABILITY STATEMENT

The empirical datasets used in this study are available in the R package 'DAISIE' available on CRAN at https://rdrr.io/cran/DAISIE/. The code used to generate the complete and incomplete datasets has been archived on Zenodo and is available at https://doi.org/10.5281/zenodo.15319719 (Dehayem et al., 2025).

ORCID

Ornela N. Dehayem https://orcid.org/0009-0009-0507-1215
Rampal S. Etienne https://orcid.org/0000-0003-2142-7612

REFERENCES

Acebes, J., Leon Arencibia, M. C., Rodríguez Navarro, M. L., Del Arco Aguilar, M., Gallo, G., Paz, P., Delgado, O., Osorio, V. E., & Wildpret, W. (2009). *Listas de Especies Silvestres de Canarias. Hongos, Plantas y Animales Terrestres 2009*. Gobierno de Canarias.

Bank, S., Cumming, R. T., Li, Y., Henze, K., Le Tirant, S., & Bradler, S. (2021). A tree of leaves: Phylogeny and historical biogeography of the leaf insects (Phasmatodea: Phylliidae). *Communications Biology*, *4*(1), 932. https://doi.org/10.1038/s42003-021-02436-z

Beaulieu, J. M., & O'Meara, B. C. (2016). Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Systematic Biology*, *65*(4), 583–601.

Beierkuhnlein, C., Walentowitz, A., & Welss, W. (2021). Flocan—A revised checklist for the flora of the canary islands. *Diversity*, *13*(10), 480.

Chang, J., Rabosky, D. L., & Alfaro, M. E. (2020). Estimating diversification rates on incompletely sampled phylogenies: Theoretical concerns and practical solutions. *Systematic Biology*, *69*(3), 602–611.

Davis, M. P., Midford, P. E., & Maddison, W. (2013). Exploring power and parameter estimation of the bisse method for analyzing species diversification. *BMC Evolutionary Biology*, *13*, 1–11.

Dehayem, O. N., Brewer, R. F. A., Valente, L., Lens, F., & Etienne, R. S. (2025). *Code and data for "impact of sampling strategy on inference of community assembly processes in phylogenetic island biogeography"*. https://doi.org/10.5281/zenodo.15319719

Etienne, R. S., Haegeman, B., Dugo-Cota, Á., Vilà, C., Gonzalez-Voyer, A., & Valente, L. (2023). The phylogenetic limits to diversity-dependent diversification. *Systematic Biology*, *72*(2), 433–445.

Etienne, R. S., Haegeman, B., Stadler, T., Aze, T., Pearson, P. N., Purvis, A., & Phillimore, A. B. (2012). Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1732), 1300–1309.

Etienne, R. S., & Rosindell, J. (2012). Prolonging the past counteracts the pull of the present: Protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, *61*(2), 204–213.

FitzJohn, R. G., Maddison, W. P., & Otto, S. P. (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology*, *58*(6), 595–611.

Florencio, M., Patiño, J., Nogué, S., Traveset, A., Borges, P. A., Schaefer, H., Amorim, I. R., Arnedo, M., Ávila, S. P., Cardoso, P., de Nascimento, L., Fernández-Palacios, J. M., Gabriel, S. I., Gil, A., Gonçalves, V., Haroun, R., Illera, J. C., López-Darias, M., Martínez, A., … Santos, A. M. C. (2021). Macaronesia as a fruitful arena for ecology, evolution, and conservation biology. *Frontiers in Ecology and Evolution*, *9*, 718169.

Garamszegi, L. Z., & Møller, A. P. (2011). Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. *Systematic Biology*, *60*(6), 876–880.

García-Verdugo, C., Caujapé-Castells, J., & Sanmartín, I. (2019). Colonization time on island settings: Lessons from the hawaiian and canary island floras. *Botanical Journal of the Linnean Society*, *191*(2), 155–163.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, *22*(4), 403–434.

Gillespie, R. G., & Roderick, G. K. (2014). Geology and climate drive diversification. *Nature*, *509*(7500), 297–298.

Hauffe, T., Delicado, D., Etienne, R. S., & Valente, L. (2020). Lake expansion elevates equilibrium diversity via increasing colonization. *Journal of Biogeography*, *47*(9), 1849–1860.

Helmus, M. R., Mahler, D. L., & Losos, J. B. (2014). Island biogeography of the anthropocene. *Nature*, *513*(7519), 543–546.

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*(2), 317–328.

Hey, J. (1992). Using phylogenetic trees to study speciation and extinction. *Evolution*, *46*(3), 627–640.

Jiménez-Ortega, D., Valente, L., Dugo-Cota, Á., Rabosky, D. L., Vilà, C., & Gonzalez-Voyer, A. (2023). Diversification dynamics in caribbean rain frogs (*Eleutherodactylus*) are uncoupled from the anuran community and consistent with adaptive radiation. *Proceedings of the Royal Society B*, *290*(1990), 20222171.

Lambert, J., Neves, P., Bilderbeek, R., Valente, L., & Etienne, R. (2022). The effect of mainland dynamics on data and parameter estimates in island biogeography. *bioRxiv*, 2022.01.13.476210.

Losos, J. B., & Ricklefs, R. E. (2009). Adaptation and diversification on islands. *Nature*, *457*(7231), 830–836.

Maddison, W. P., Midford, P. E., & Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, *56*(5), 701–710.

Martín-Hernanz, S., Nogales, M., Valente, L., Fernández-Mazuecos, M., Pomeda-Gutiérrez, F., Cano, E., Marrero, P., Olesen, J. M., Heleno, R., & Vargas, P. (2023). Time-calibrated phylogenies reveal mediterranean and pre-mediterranean origin of the thermophilous vegetation of the canary islands. *Annals of Botany*, *131*(4), 667–684.

Matthews, T. J., & Triantis, K. (2021). Island biogeography. *Current Biology*, *31*(19), R1201–R1207.

McMullen, C. K. (1999). *Flowering plants of the Galápagos*. Cornell University Press.

Molina-Venegas, R., & Lima, H. (2021). Should we be concerned about incomplete taxon sampling when assessing the evolutionary history of regional biotas? *Journal of Biogeography*, *48*(9), 2387–2390.

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, *17*(4), 508–525.

Morlon, H., Lewitus, E., Condamine, F. L., Manceau, M., Clavel, J., & Drury, J. (2016). Rpanda: An r package for macroevolutionary analyses on phylogenetic trees. *Methods in Ecology and Evolution*, *7*(5), 589–597.

Mynard, P., Algar, A. C., Lancaster, L. T., Bocedi, G., Fahri, F., Gubry-Rangin, C., Lupiyaningdyah, P., Nangoy, M., Osborne, O. G., Papadopulos, A. S., Sudiana, M., Juliandi, B., Travis, J. M. J., & Herrera-Alsina, L. (2023). Impact of phylogenetic tree completeness and mis-specification of sampling fractions on trait dependent diversification models. *Systematic Biology*, *72*(1), 106–119.

Nee, S., May, R. M., & Harvey, P. H. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *344*(1309), 305–311.

Nee, S., Mooers, A. O., & Harvey, P. H. (1992). Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(17), 8322–8326.

Neves, P. S., Lambert, J. W., Valente, L., & Etienne, R. S. (2021). The robustness of a simple dynamic model of island biodiversity to geological and sea-level change. *Journal of Biogeography*, *49*(11), 2091–2104.

Parent, C. E., Caccone, A., & Petren, K. (2008). Colonization and diversification of galápagos terrestrial fauna: A phylogenetic and biogeographical synthesis. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *363*(1508), 3347–3361.

Rabosky, D. L., Grundler, M., Anderson, C., Title, P., Shi, J. J., Brown, J. W., Huang, H., & Larson, J. G. (2014). Bamm tools: An r package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, *5*(7), 701–707. https://doi.org/10.1111/2041-210X.12199

Ralimanana, H., Perrigo, A. L., Smith, R. J., Borrell, J. S., Faurby, S., Rajaonah, M. T., Randriamboavonjy, T., Vorontsova, M. S., Cooke, R. S., Phelps, L. N., Sayol, F., Andela, N., Andermann, T., Andriamanohera, A. M., Andriambololonera, S., Bachman, S. P., Bacon, C. D., Baker, W. J., Belluardo, F., … Antonelli, A. (2022). Madagascar's extraordinary biodiversity: Threats and opportunities. *Science*, *378*(6623), eadf1466.

Stiller, J., Feng, S., Chowdhury, A.-A., Rivas-González, I., Duchêne, D. A., Fang, Q., Deng, Y., Kozlov, A., Stamatakis, A., Claramunt, S., Nguyen, J. M. T., Ho, S. Y. W., Faircloth, B. C., Haag, J., Houde, P., Cracraft, J., Balaban, M., Mai, U., Chen, G., … Zhang, G. (2024). Complexity of avian evolution revealed by family-level genomes. *Nature*, *629*(8013), 851–860.

Sun, M., Folk, R. A., Gitzendanner, M. A., Soltis, P. S., Chen, Z., Soltis, D. E., & Guralnick, R. P. (2020). Estimating rates and patterns of diversification with incomplete sampling: A case study in the rosids. *American Journal of Botany*, *107*(6), 895–909.

Valente, L., Etienne, R. S., & Davalos, L. M. (2017). Recent extinctions disturb path to equilibrium diversity in caribbean bats. *Nature Ecology & Evolution*, *1*(2), 1–7.

Valente, L., Etienne, R. S., & Garcia-R, J. C. (2019). Deep macroevolutionary impact of humans on New Zealand's unique avifauna. *Current Biology*, *29*(15), 2563–2569.

Valente, L., Illera, J. C., Havenstein, K., Pallien, T., Etienne, R. S., & Tiedemann, R. (2017). Equilibrium bird species diversity in atlantic islands. *Current Biology*, *27*(11), 1660–1666.

Valente, L., Phillimore, A. B., & Etienne, R. (2018). Using molecular phylogenies in island biogeography: It's about time. *Ecography*, *41*(10), 1684–1686.

Valente, L., Phillimore, A. B., Melo, M., Warren, B. H., Clegg, S. M., Havenstein, K., Tiedemann, R., Illera, J. C., Thébaud, C., Aschenbach, T., & Etienne, R. S. (2020). A simple dynamic model explains the diversity of island birds worldwide. *Nature*, *579*(7797), 92–96.

Valente, L. M., Phillimore, A. B., & Etienne, R. S. (2015). Equilibrium and non-equilibrium dynamics simultaneously operate in the galápagos islands. *Ecology Letters*, *18*(8), 844–852.

Wagner, W. L., Khan, N. R., & Lorence, D. H. (2023). *Flora of the Hawaiian islands website*. https://naturalhistory2.si.edu/botany/hawaiianflora/

Warren, B. H., Simberloff, D., Ricklefs, R. E., Aguilée, R., Condamine, F. L., Gravel, D., Morlon, H., Mouquet, N., Rosindell, J., Casquet, J., Conti, E., Cornuault, J., Fernández-Palacios, J. M., Hengl, T., Norder, S. J., Rijsdijk, K. F., Sanmartín, I., Strasberg, D., Triantis, K. A., … Thébaud, C. (2015). Islands as model systems in ecology and evolution: Prospects fifty years after macarthur-wilson. *Ecology Letters*, *18*(2), 200–217.

Whittaker, R. J., Fernández-Palacios, J. M., Matthews, T. J., Borregaard, M. K., & Triantis, K. A. (2017). Island biogeography: Taking the long view of nature's laboratories. *Science*, *357*(6354), eaam8326.

Xie, S., Valente, L., & Etienne, R. S. (2023). Can we ignore trait-dependent colonization and diversification in island biogeography? *Evolution*, *77*(3), 670–681.

Zuntini, A. R., Carruthers, T., Maurin, O., Bailey, P. C., Leempoel, K., Brewer, G. E., Epitawalage, N., Françoso, E., Gallego-Paramo, B., McGinnie, C., Negrão, R., Roy, S. R., Simpson, L., Romero, E. T., Barber, V. M. A., Botigué, L., Clarkson, J. J., Cowan, R. S., Dodsworth, S., … Baker, W. J. (2024). Phylogenomics and the rise of the angiosperms. *Nature*, *629*, 843–850.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Effect of sampling strategy on parameter estimation for simulated datasets generated with the parameters estimated from the Galapagos birds.

Figure S2. Effect of sampling strategy on parameter estimation for simulated datasets generated with the parameters estimated from the Greater Antilles bats.

Figure S3. Effect of sampling strategy on parameter estimation for simulated datasets generated with the parameters estimated from the New Zealand birds.

Figure S4. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset.

Figure S5. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset.

Figure S6. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters estimated from the Galapagos birds.

Figure S7. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters estimated from the Galapagos birds.

Figure S8. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters estimated from the Greater Antilles bats.

Figure S9. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters estimated from the Greater Antilles bats.

Figure S10. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters estimated from the Hispaniola frogs.

Figure S11. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters estimated from the New Zealand birds.

Figure S12. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters estimated from the New Zealand birds.

Figure S13. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset.

Figure S14. Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset.

Figure S15. Effect of sampling strategy on the estimation of parameters for simulated data sets generated with the parameters that created the Large dataset.

Figure S16. Effect of the effect of missing data on datasets of different size.

How to cite this article: Dehayem, O. N., Brewer, R. F. A., Valente, L., Lens, F., & Etienne, R. S. (2025). Impact of sampling strategy on inference of community assembly processes in phylogenetic island biogeography. *Methods in Ecology and Evolution*, *00*, 1–14. https://doi.org/10.1111/2041-210X.70058