Ornela N. Dehayem, Ryan F. A. Brewer, Luis Valente, Frederic Lens, Rampal S. Etienne. 2025. “Impact of sampling strategy on inference of community assembly processes in phylogenetic island biogeography”. *Methods in Ecology and Evolution*. 16:1507–1520. https://doi.org/10.1111/2041-210X.70058

Class focus area: Simulation

EEB603: Brian O’Meara

All quotes and images from the above paper unless otherwise noted

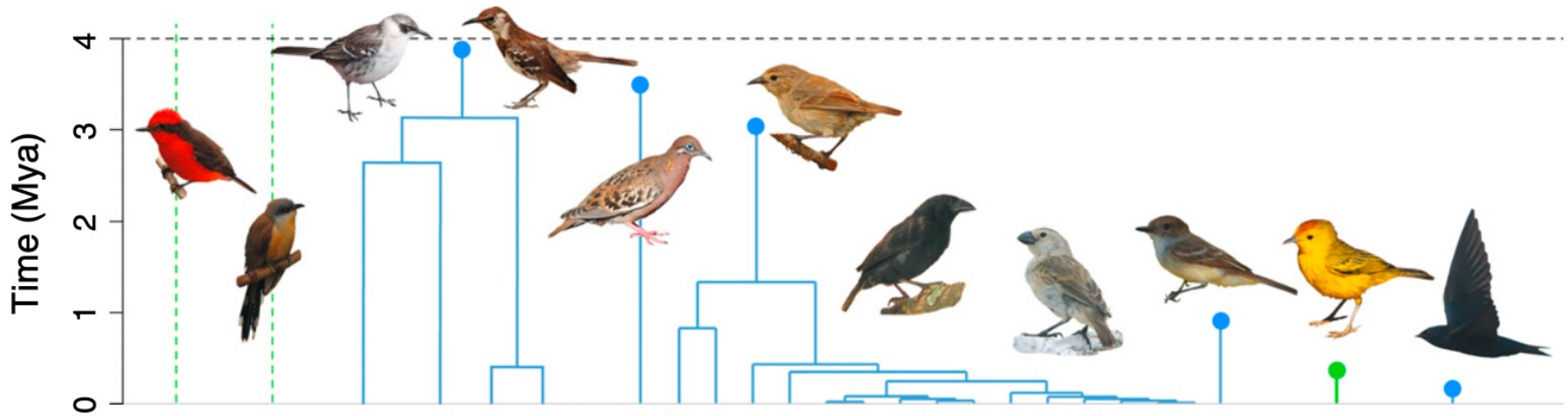

Figure 1 Colonisation and branching times of Galápagos terrestrial birds. Circles represent mean age of colonisation across all phylogenetic data sets. Blue – endemic lineages; green – non-endemic lineages. Dashed lines are shown for the vermilion flycatcher and dark-billed cuckoo as for these species only an upper bound to the age of colonisation is known. The dashed black line shows the approximate age of the oldest currently emerged island in the Galápagos archipelago. The photographs (by Ruben Heleno, Luis Valente and Steve Arlow) show representatives of each of the independent colonisation events. Lineage names are given in Table S1 in the same order as in the figure.

Luis M. Valente, Albert B. Phillimore and Rampal S. Etienne 2015

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

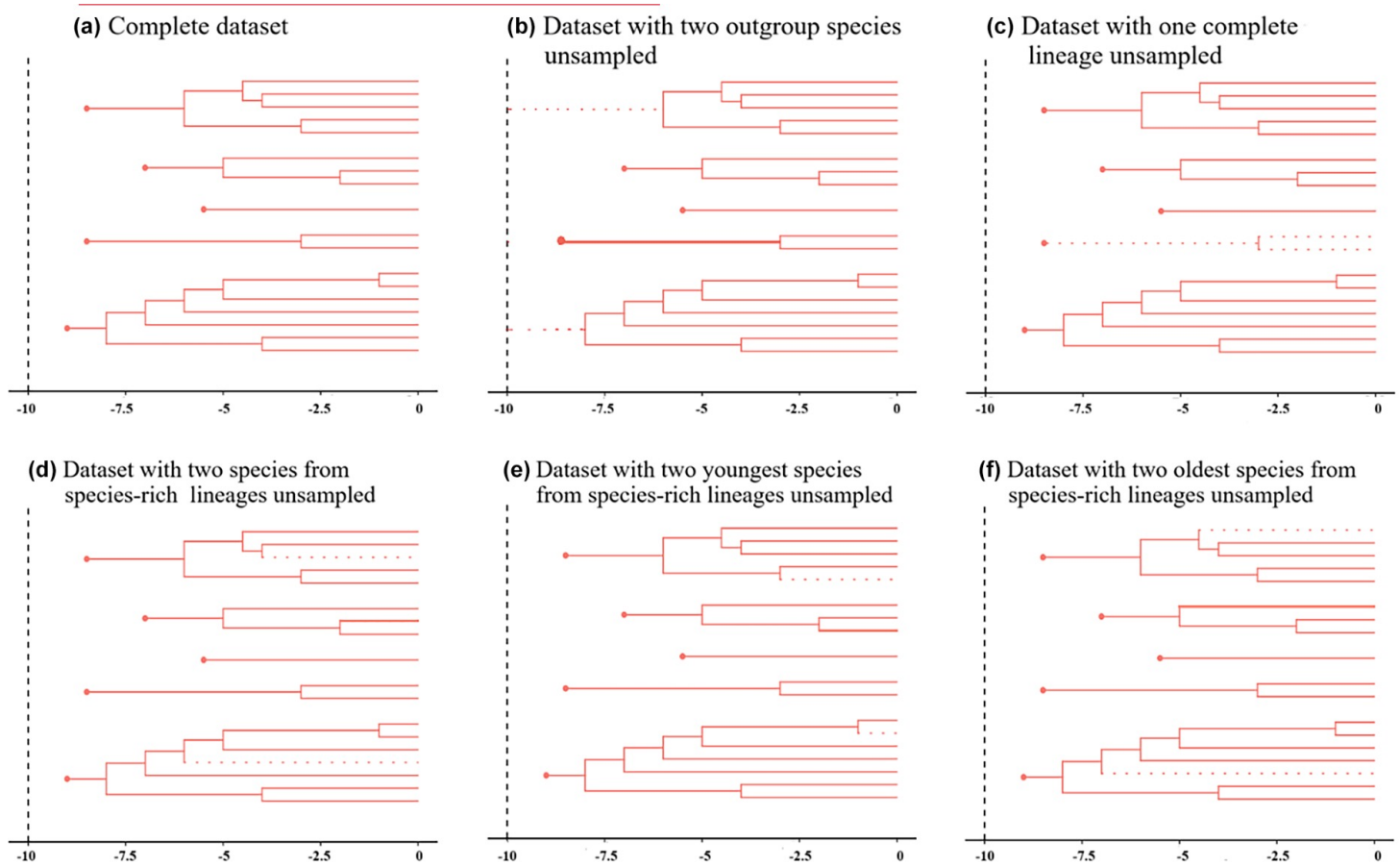

FIGURE 1 Examples of incomplete datasets generated in our study. Each panel represents the phylogeny of an island community. Each tree represents a colonization event that has established one or more descendants on the island and includes one close continental relative. The dotted lines represent non-sampled 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.

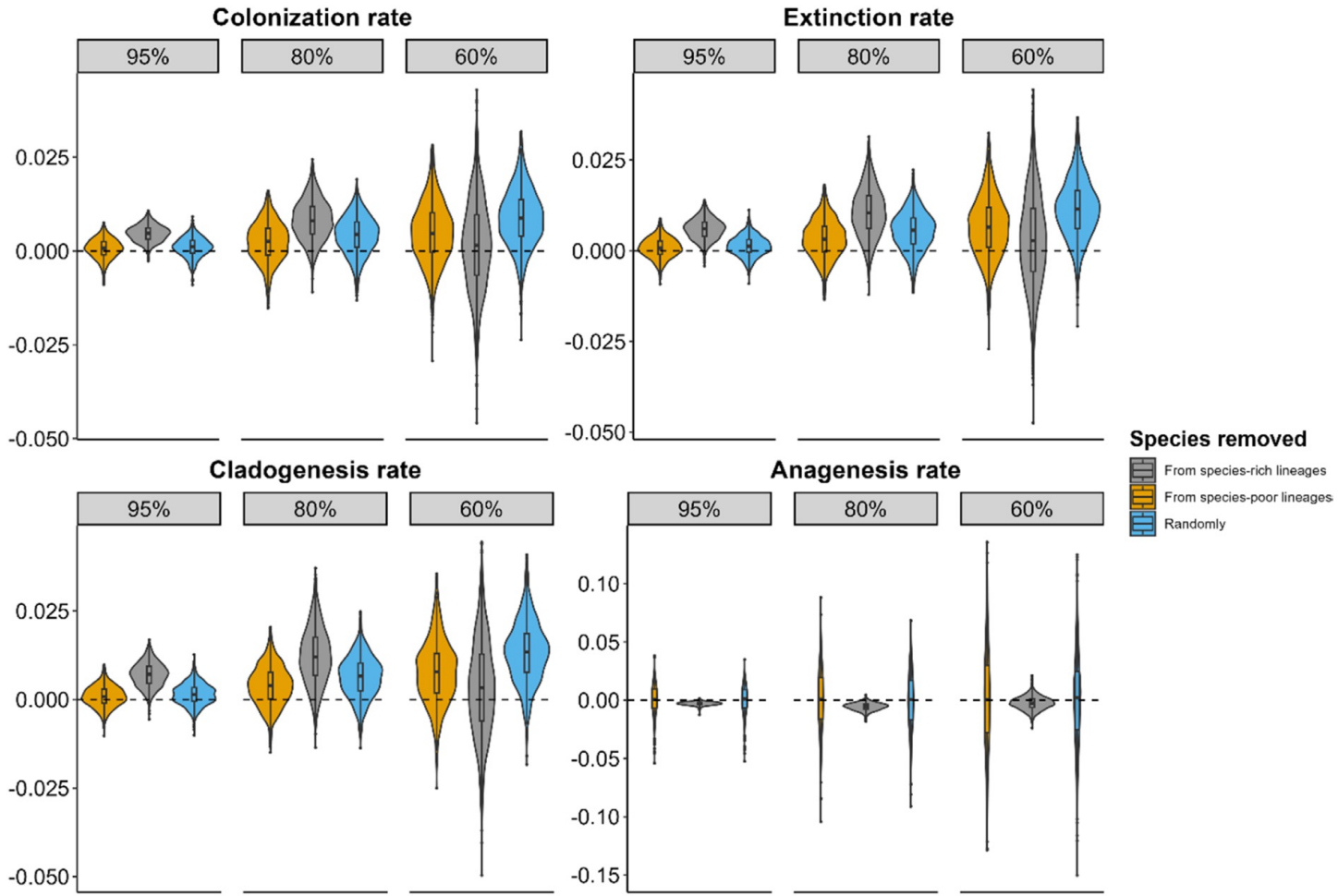

FIGURE 2 Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset. Violin plots show the error distributions for colonization, extinction, cladogenesis and anagenesis rates across simulations from incomplete datasets 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.

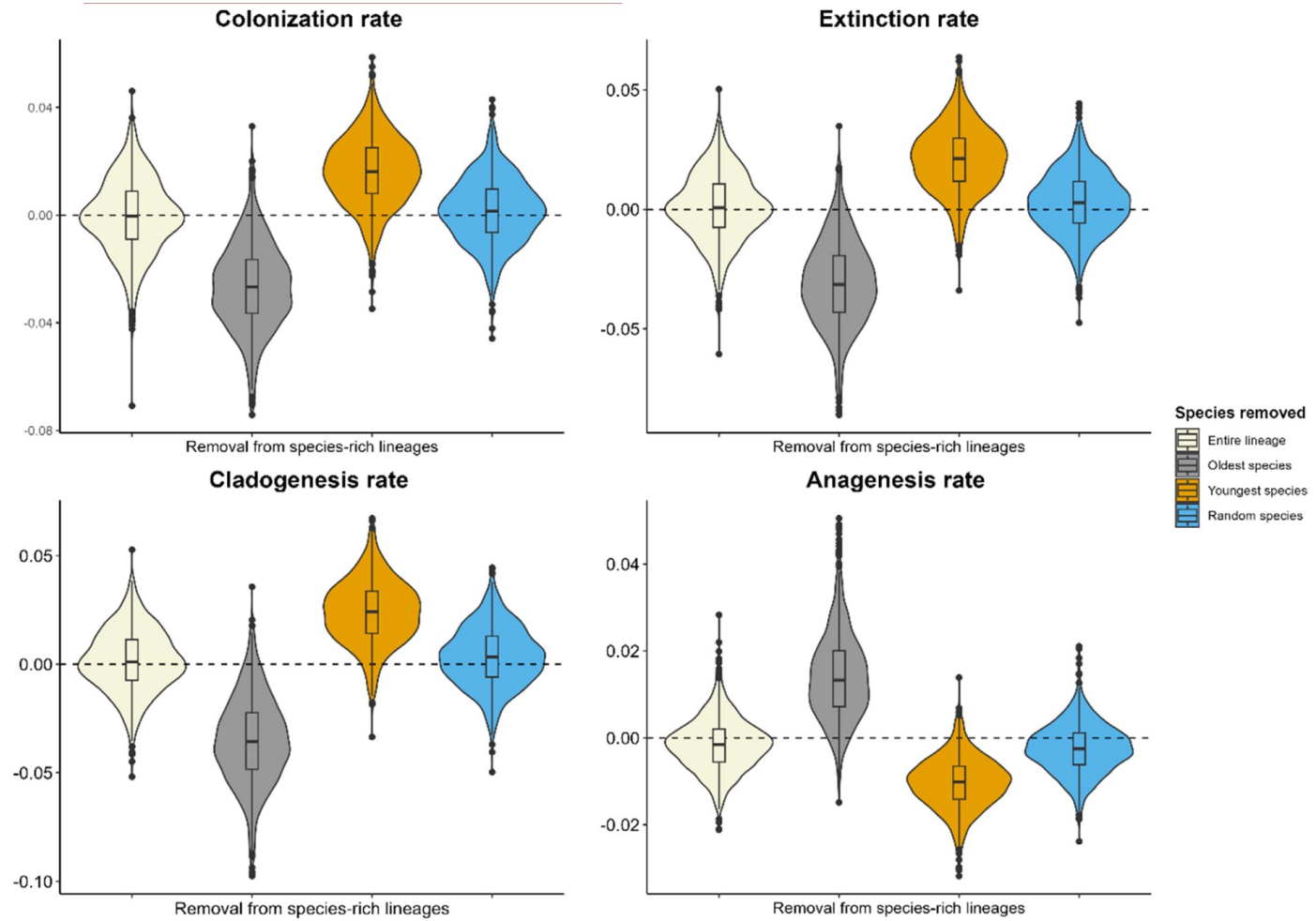

FIGURE 3 Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset. Violin plots show the error distributions for colonization, extinction, cladogenesis and anagenesis rates across simulations from incomplete datasets (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.

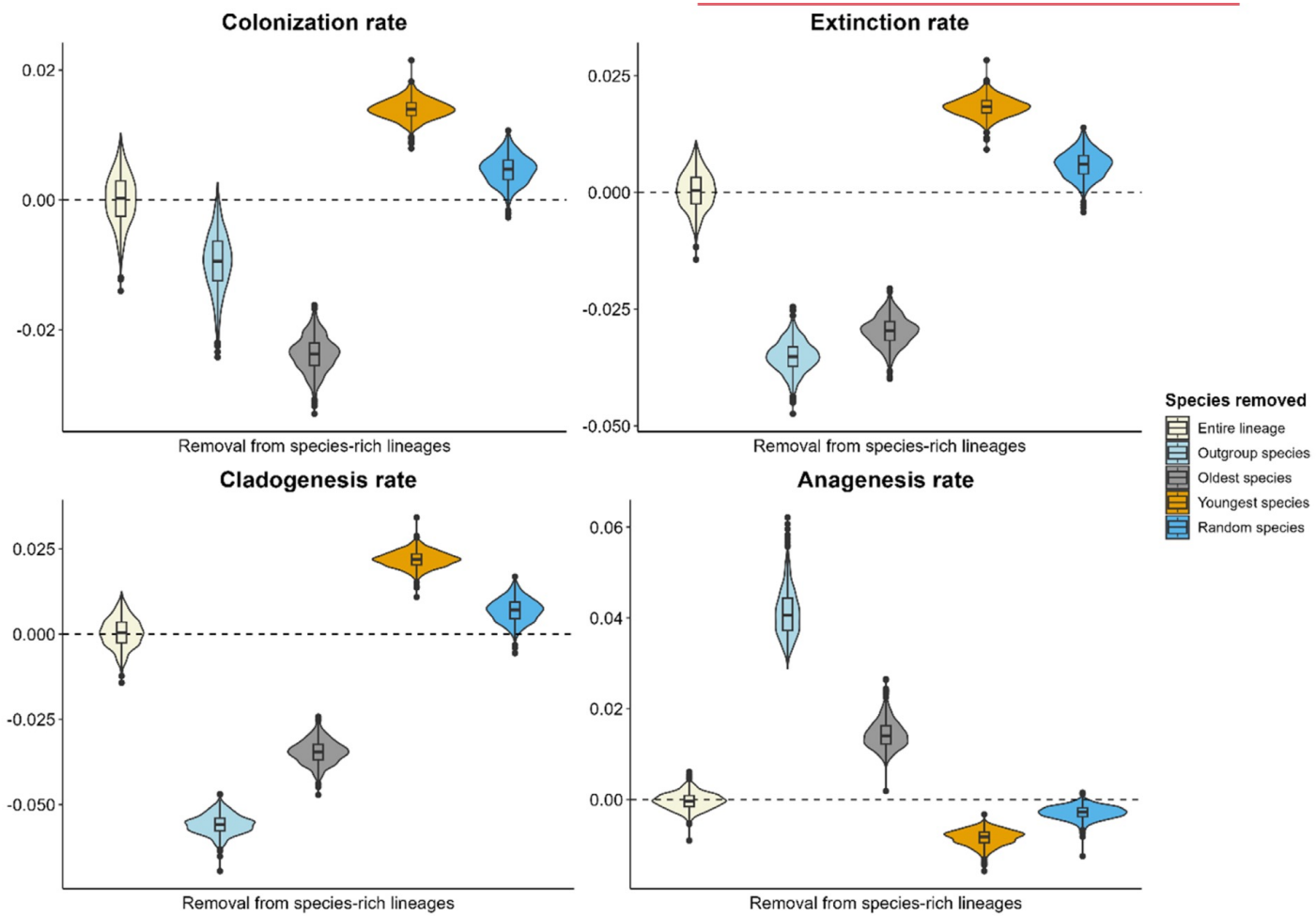

FIGURE 4 Effect of sampling strategy on parameter estimation for simulated datasets generated with the parameters that created the Large dataset. Missing species are removed from species-rich lineages. Violin plots show the error distributions for colonization, extinction, cladogenesis and anagenesis 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.

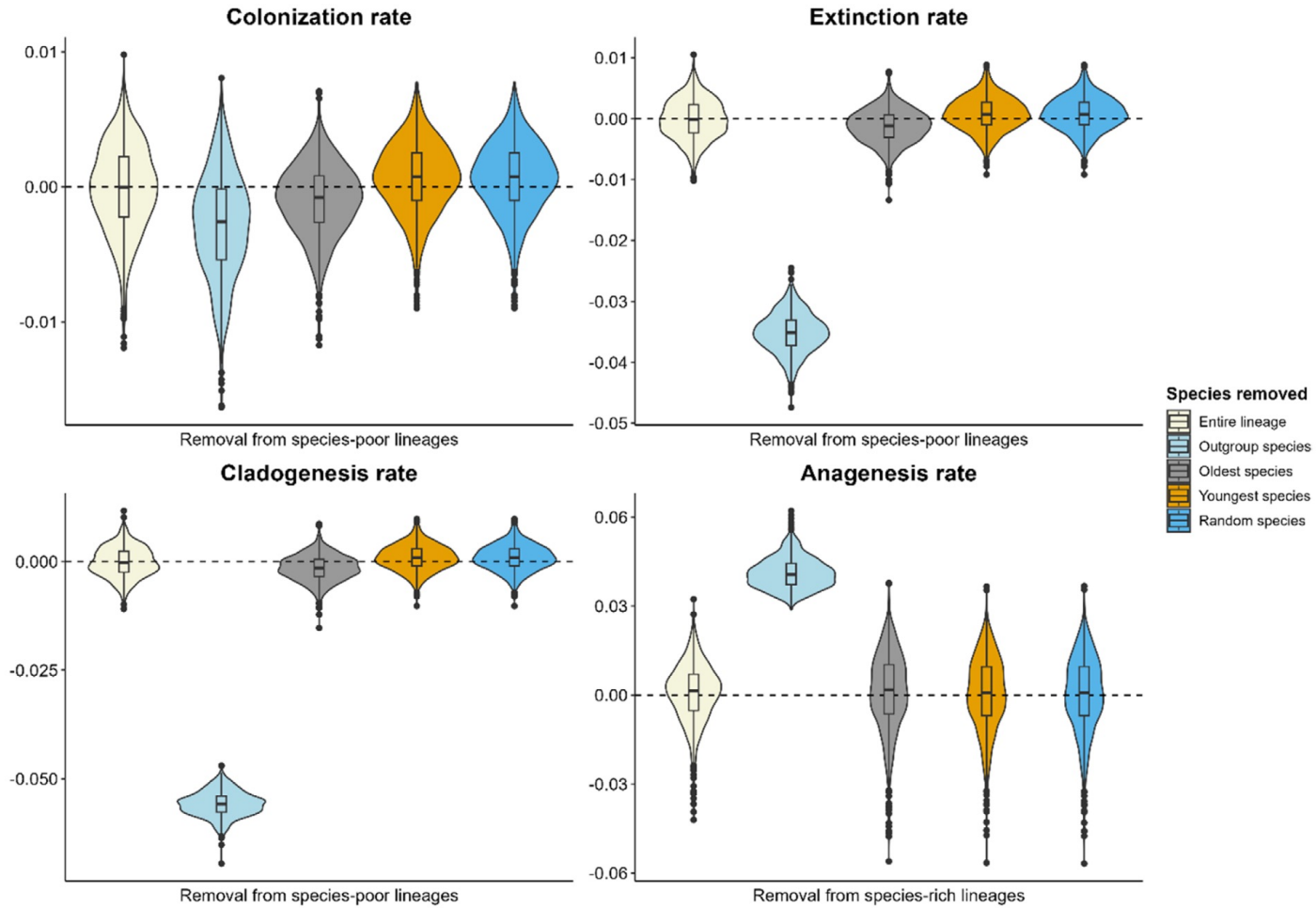

FIGURE 5 Effect of sampling strategy on the estimation of parameters for simulated datasets generated with the parameters that created the Large dataset. Missing species are removed from species-poor lineages. Violin plots show the error distributions for colonization, extinction, cladogenesis, and anagenesis 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.

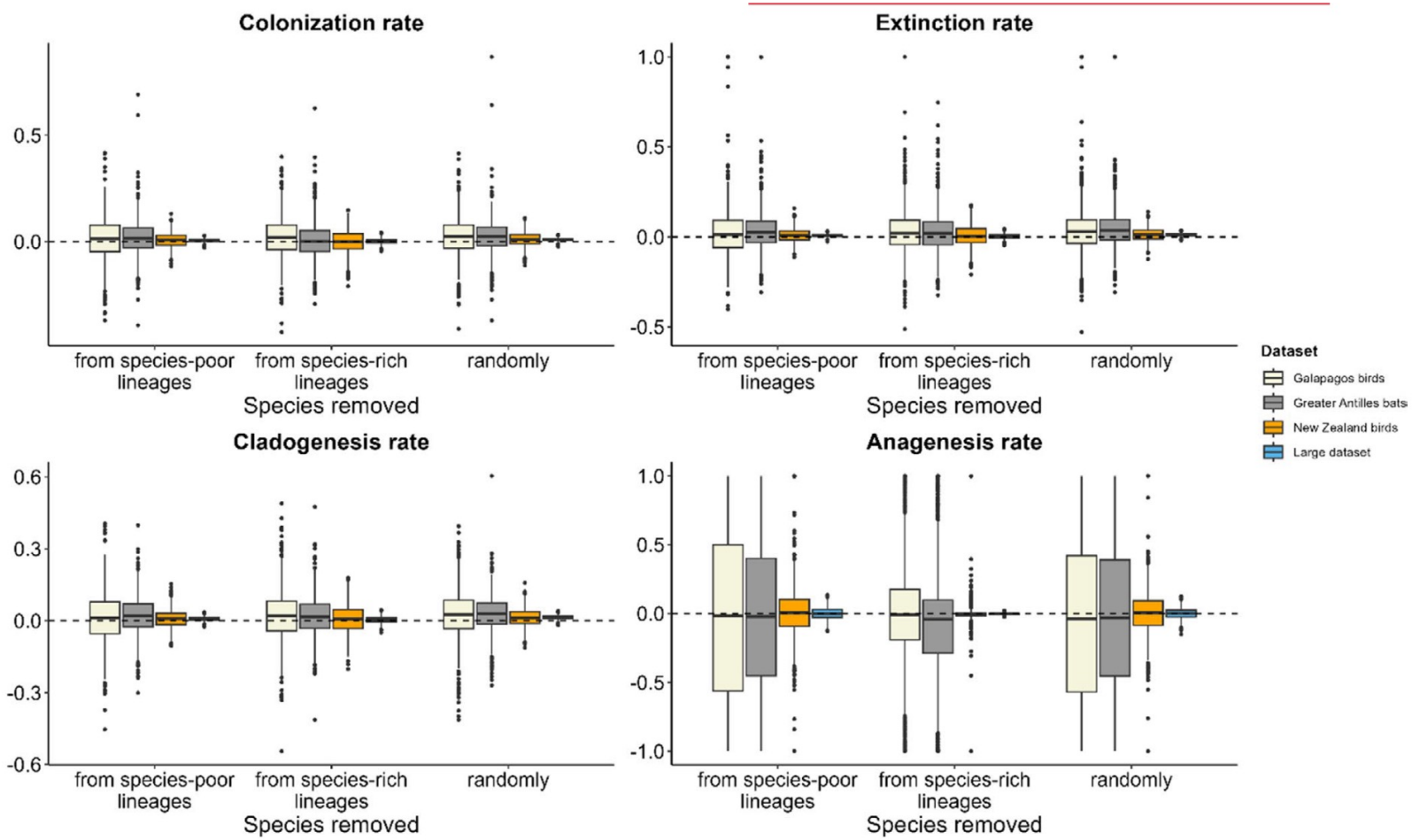

FIGURE 6 Effect of missing data on parameter estimation from datasets of different sizes. Box plots show the distribution of the errors in colonization, extinction, cladogenesis and anagenesis rates in simulations from incomplete datasets (here 40% unsampled species) with different sizes and numbers 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.

- Do species have ages?
- Does it matter that this paper and the DAISIE paper have some of the same authors?
- Could this be done if the generating model were not a DAISIE model?
- Does statistical significance matter for a simulation study?
- How do you simulate a tree?