Realistic scenarios of missing taxa in phylogenetic comparative methods and their effects on model selection and parameter estimation

Rafael S Marcondes [Corresp. 1]

[1] Museum of Natural Science and Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, United States

Corresponding Author: Rafael S Marcondes
Email address: rmarco3@lsu.edu

Model-based analyses of continuous trait evolution enable rich evolutionary insight. These analyses require a phylogenetic tree and a vector of trait values for the tree's terminal taxa, but rarely do a tree and dataset include all taxa within a clade. Because the probability that a taxon is included in a dataset depends on ecological traits that have phylogenetic signal, missing taxa in real datasets should be expected to be phylogenetically clumped or correlated to the modelled trait. I examined whether those types of missing taxa represent a problem for model selection and parameter estimation. I simulated univariate traits under a suite of Brownian Motion and Ornstein-Uhlenbeck models, and assessed the performance of model selection and parameter estimation under absent, random, clumped or correlated missing taxa. I found that those analyses perform well under almost all scenarios, including situations with very sparsely sampled phylogenies. The only notable biases I detected were in parameter estimation under a very high percentage (90%) of correlated missing taxa. My results offer a degree of reassurance for studies of continuous trait evolution with missing taxa, but the problem of missing taxa in phylogenetic comparative methods still demands much further investigation. The framework I have described here might provide a starting point for future work.

1 **Realistic scenarios of missing taxa in phylogenetic comparative methods and their**

2 **effects on model selection and parameter estimation**

4 Rafael S. Marcondes

6 Louisiana State University Museum of Natural Science and Department of Biological Sciences

7 Baton Rouge, LA 70802, USA

9 rmarco3@lsu.edu

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

33   **Abstract**

34   Model-based analyses of continuous trait evolution enable rich evolutionary

35   insight. These analyses require a phylogenetic tree and a vector of trait values for the

36   tree's terminal taxa, but rarely do a tree and dataset include all taxa within a clade.

37   Because the probability that a taxon is included in a dataset depends on ecological traits

38   that have phylogenetic signal, missing taxa in real datasets should be expected to be

39   phylogenetically clumped or correlated to the modelled trait. I examined whether those

40   types of missing taxa represent a problem for model selection and parameter estimation.

41   I simulated univariate traits under a suite of Brownian Motion and Ornstein-Uhlenbeck

42   models, and assessed the performance of model selection and parameter estimation

43   under absent, random, clumped or correlated missing taxa. I found that those analyses

44   perform well under almost all scenarios, including situations with very sparsely sampled

45   phylogenies. The only notable biases I detected were in parameter estimation under a

46   very high percentage (90%) of correlated missing taxa. My results offer a degree of

47   reassurance for studies of continuous trait evolution with missing taxa, but the problem of

48   missing taxa in phylogenetic comparative methods still demands much further

49   investigation. The framework I have described here might provide a starting point for

50   future work.

51

52

53

54

55

Introduction

56

57   Phylogenetic comparative biology is a thriving field that uses phylogenies towards

58   the goal of elucidating the historical mechanisms that have given rise to species and their

59   traits. At the core of phylogenetic comparative methods are statistical models that

60   translate ideas about evolutionary processes into the language of mathematics, thus

61   allowing biological explanations to be quantitatively weighed against one another, and

62   their implications to be explored in detail (Butler and King 2004, Cressler et al. 2015,

63   Brown and Thomson 2018, Zuk and Travisano 2018). Model-based phylogenetic studies

64   of univariate continuous trait evolution require at least two inputs: a phylogenetic tree of

65   the species under study, and a vector of trait values for those species. One or more

66   models are then typically fitted to the trait vector, conditioned on the phylogeny. The fits

67   of these models can then be compared (often using Akaike information criteria) to assess

68   which one offers the best explanation of the data. In addition, parameter estimates from

69   fitted models can give insights into features of the evolutionary process that generated

70   the data, such as adaptive peaks, rates of evolution and strength of selection (Beaulieu

71   et al. 2012, O'Meara and Beaulieu 2014).

72   Phylogenies used in comparative studies are typically estimated from molecular

73   sequences and trait data are typically generated from specimens deposited in natural

74   history collections or, less often, from observations of live organisms. Because of

75   differences in ecological characters such as range size, habitat preference, life history

76   and behavior, not all taxa are equally likely to be available for inclusion in a molecular tree

77   or trait dataset (Garamszegi and Møller 2011). Therefore, comparative studies often have

78   missing taxa, that is, taxa that are members of the clade under study, but which

79 researchers have been unable to include in their analyses because they were either not

80 included in the phylogeny or not accessible for measurement of traits (Thomson and

81 Shaffer 2010, Slater at al. 2012, Reddy 2014, Rabosky 2015). This usually results in the

82 missing taxa being excluded from the analyses, which has the potential to introduce

83 biases in model selection and parameter estimation (Garamszegi and Møller 2011,

84 Pennell et al. 2016).

85   Missing data have received significant attention in the context of nucleotide

86 sequences used in phylogeny estimation (e.g., Wiens et al 2011, Jiang et al 2014, Eaton

87 et al 2017). In stark contrast, missing data have received little attention in phylogenetic

88 comparative biology (but see Garamszegi and Møller 2011), and, in the rare occasions

89 when researchers assessed impacts of missing taxa in comparative models, they were

90 simulated in a random fashion with respect both to trait values and to phylogeny (e.g.,

91 Ingram and Mahler 2013). In the next few paragraphs, I argue that missing taxa are

92 unlikely to occur randomly and describe more realistic scenarios where missing taxa are

93 phylogenetically clumped and/or correlated to the trait of interest. Next, I describe a set

94 of simulations where my aims were two-fold: first, to conduct an initial, exploratory

95 investigation of the impacts of realistic missing taxa on a limited set of models of univariate

96 continuous trait evolution; and, second, by presenting a basic framework that can be

97 taken up by other investigators, to instigate more attention and future research on this

98 neglected issue in phylogenetic comparative biology.

99

100   **Realistic scenarios of missing taxa in comparative datasets**

101   The multifarious ecological traits that influence the probability that a taxon is

102   sampled have phylogenetic signal. Consequently, missing taxa should be expected to be

103   phylogenetically clumped. The most important ecological trait influencing sampling

104   probability is rarity, broadly understood as the character of a taxon that has a low

105   abundance and/or a small geographical distribution (Gaston 1994). It is intuitive that rarer

106   taxa are more difficult to detect, observe, capture and collect, and thus more likely to be

107   missing from datasets than common taxa. Both components of rarity, abundance and

108   range size, have been shown to have phylogenetic signal, *i.e.*, they are heritable at a

109   macroevolutionary level. For example, 41% of the variation in population density of North

110   American birds can be attributed to their taxonomic family (Maurer et al. 1991), and

111   taxonomic affiliation also explains variation in abundance among Neotropical rainforest

112   mammals (Arita et al 1990). As for range size, it has been shown to be heritable in studies

113   of mammals (DeSantis et al. 2012), birds (Waldron 2007, Herrera-Alsina and Villegas-

114   Patraca 2014), mollusks (Jablonski 1987) and herbaceous plants (Qian and Ricklefs

115   2004). That rarity has phylogenetic signal was also indicated by Fritz and Purvis' (2010)

116   finding that threatened species of British birds and of the world's mammals tended to be

117   phylogenetically clumped.

118   Beyond rarity, various other axes of an organism's ecological niche can affect its

119   probability of being sampled in comparative studies, and the pervasiveness of

120   phylogenetic niche conservatism (Losos 2008, Wiens et al. 2010, Crisp and Cook 2012)

121   is evidence of the high phylogenetic signal of ecological niches. Because taxa inhabiting

122   climates that are hostile to humans (for example, boreal organisms; Malaney and Cook

123   2018) are less likely to be sampled, climate is a particularly important niche axis in this

124   context, and it has been repeatedly shown to display phylogenetic signal (reviewed by

125   Wiens and Graham 2005). Another important niche axis, habitat preference, also has

126   phylogenetic signal (Barr and Scott 2014) and is likely to affect sampling probability,

127   because forest-based organisms are more difficult to detect and collect than nonforest

128   organisms.

129   These taxonomic biases in rarity presumably translate into taxonomic biases in

130   availability of specimens and genetic samples. For example, Malaney and Cook (2018)

131   reported strong taxonomic imbalance in collections of North American mammals, with

132   Rodentia being comparatively overrepresented in relation to all other mammal orders,

133   and Reddy (2014) described similar imbalances for the availability of genetic data for the

134   world's birds.

135   In addition to being phylogenetically clumped, missing taxa might sometimes be

136   directly correlated to the trait under study, when taxa with a higher (or lower) trait value

137   have a lower probability of being sampled. For example, because taxa with small range

138   sizes are more difficult to sample, they will be missing from molecular datasets more often

139   than taxa with large ranges (Reddy 2014). This means that unsampled taxa will have a

140   smaller average range size than sampled taxa. Therefore, if we were undertaking a

141   comparative study of range size with incomplete taxon sampling, the distribution of trait

142   values in our sample will differ from the real distribution, potentially biasing model-based

143   analyses. This should also be the case for other ecological traits tightly linked to rarity

144   and sampling probabilities, such as abundance and some niche axes.

145   In sum, there is ample evidence of phylogenetic signal in ecological traits likely to

146   influence sampling probability for comparative studies. Consequently, sampling

147 probability itself should display phylogenetic signal, and missing taxa are likely to be

148 phylogenetically clumped. There is also reason to expect some types of traits to be

149 directly correlated to sampling probabilities. Here, I present simulation-based analyses

150 examining how those realistic scenarios of missing taxa might affect the performance of

151 models of continuous trait evolution.

152

153 **Methods**

154 **Overview**

155   My simulations were designed to examine how realistic scenarios of missing taxa

156 affect model selection and parameter estimates for a single continuous trait in whose

157 evolution we are interested, hereafter referred to as **T** (Table 1). I simulated **T** under a

158 number of different models and then pruned 10%, 50% or 90% of terminal taxa from the

159 tree under three different schemes (Figure 1): (1) randomly (rMT); (2) phylogenetically

160 clumped missing taxa (cluMT); and (3) correlated to **T** (corMT). For comparison, I also

161 ran simulations with no missing taxa (nMT). The sampling status of a taxon can be thought

162 of as a binary character, hereafter referred to as **S** (Table 1), with states 0 (missing) and

163 1 (sampled). Once **T** and **S** were simulated, I pruned tips from the tree and dataset based

164 on **S** and, finally, I fitted all models to **T** and assessed support for the generating model,

165 as well as precision and bias of parameter estimates. I repeated the simulation 1000 times

166 for each combination of type of missing taxa, percentage of missing taxa, and model of **T**

167 trait evolution.

168   Before presenting the details of my simulations, I reiterate that I did not seek to

169 exhaust every possible scenario of missing taxa in phylogenetic comparative biology.

170   Rather, I sought to explore the implications of two very specific scenarios that I argue are

171   likely in real datasets. There undoubtedly exist other possible, if less probable,

172   configurations of missing taxa, and they may have different impacts on model

173   performance, but addressing those was beyond the scope of my study. I also

174   acknowledge that, even within the scenarios I studied, some of my simulation settings

175   necessarily entailed some degree of arbitrariness, for example in the size of the trees or

176   in the parameter values I used to simulate various traits. Different choices might have led

177   to different results, but while exploring those choices would certainly be productive, it was

178   not my present objective in this initial, exploratory study.

179

180   **Details**

181   I started each simulation by generating a random phylogenetic tree under a pure-

182   birth model using the function *rphylo* in the R package *ape* (Paradis and Schliep 2019)

183   and rescaling the tree to unit height using the function *rescale* in the R package *geiger*

184   (Pennell et al. 2014). The size of each initial tree was set so that the number of tips *after*

185   dropping missing taxa was always equal to 300. For example, when simulating 10%

186   missing taxa, the initial number of tips was 333, and when simulating 50% missing taxa it

187   was 600. This ensured that my results were affected only by missing taxa *per se*, and not

188   by tree size, which is known to affect the performance of models of trait evolution

189   (Beaulieu et al. 2012, Boettiger et al. 2012).

190   To represent regimes underlying variation in BMS and OUM parameters (see

191   below), I simulated the evolution of a binary trait **R** under an equal rates MK model with

192   a transition rate $q=0.5$ (Table 1, Figure 1d), and I ensured that the smallest of the two

193   regimes always included no fewer than 25% and no more than 45% of tips *after* dropping

194   the tips representing missing taxa. To simulate this trait, I used the function

195   *simulate_mk_model* in the R package *castor* (Louca and Doubeli 2017).

196   After simulating the tree and **R**, I simulated **T** (Table 1), the continuous trait of

197   interest, under one of four models using the function *OUwie.sim* in the R package *OUwie*

198   (Beaulieu et al. 2012): single-rate Brownian Motion (BM), single-optimum Ornstein-

199   Uhlenbeck (OU), multiple-rate Brownian Motion (BMS), and multiple-optimum Ornstein-

200   Uhlenbeck (OUM). Under BM, changes in trait value are purely nondirectional and

201   governed by a single parameter, σ^2 (sigma-square), that determines the rate of evolution

202   (Felsenstein 1985). Under OU, trait evolution is controlled by a nondirectional component

203   represented by σ^2 as well as by a directional component under which trait values change

204   preferentially towards an optimum (θ, theta) with strength of attraction α (alpha) (Hansen

205   1997). BMS and OUM represent variations of BM and OU in which σ^2 and θ, respectively,

206   are allowed to assume different values depending on regimes (trait **R** in my simulations)

207   reconstructed *a priori* on the phylogeny (Butler and King 2004, O'Meara et al 2006,

208   Beaulieu et al 2012).

209   For each model, I set the rate parameter σ^2 at the root (σ^2_0) to a value of 0.5, and

210   only for the BMS model it shifted to $\sigma^2_1=1$ in the derived regime. For the OU and OUM

211   models, I set the optimum parameter θ to 10 at the root (θ_0), with a shift to $\theta_1=11$ in the

212   derived regime under OUM. The α parameter of OU and OUM models, representing the

213   strength of attraction to the optimum, was always constant at 1.5. This α value

214   corresponds to a phylogenetic half-life (the time, as a proportion of the tree height, that a

215   trait takes to evolve halfway towards the adaptive peak θ) of 0.46, thus representing an

216   OU process of moderate strength.

217   For the cluMT scenario (Figure 1B), I simulated **S**, the trait determining the

218   sampling status of each tip (Table 1), under a threshold model (Felsenstein 2005, Fritz

219   and Purvis 2010), where **S** is underlain by a continuous liability trait **L** (Table 1). The state

220   of **S** for each tip depends exclusively on whether **L** is above or below a certain threshold

221   value. Because sampling status is likely to be a highly complex trait determined by

222   innumerable neutral and adaptive evolutionary forces, its evolution should resemble a

223   purely nondirectional process over macroevolutionary time and be best described by a

224   simple Brownian Motion model (O'Meara et al. 2006), which I thus chose to simulate **L**. I

225   used a σ^2 value of 1 in the function *OUwie.sim*. The threshold for **S** was set based on the

226   desired percentage of missing taxa, so that, for instance, when simulating 10% missing

227   taxa, the tips with the 10% lowest values of **L** were assigned state 0 and the tips with the

228   90% highest values were assigned state 1.

229   For corMT (Figure 1C), I simulated **S** non-phylogenetically, using the R native

230   function *sample.int* to sample tips to be dropped. I provided that function with an integer

231   corresponding to the total number of tree tips (argument *n*), the desired number of tips to

232   be dropped (*size*), and a vector of weights for obtaining the elements of the vector being

233   sampled (*prob*). Those weights were calculated, for each tip, as:

234

235
$$w = \frac{t}{\mathrm{sum(T)}} - \frac{\mathrm{min\,(T)}}{\mathrm{sum(T)}}$$

236

237   ,where t is the value of the trait of interest for that tip, and T is the vector of t values for all

238   tips. *Sample.int* then used those weights to sample *size* integers from the interval 1:*n*.

239   The sampled integers correspond to tips to drop. This procedure results in the sampling

240   probability of each tip being linearly proportional to its **T** value. The tip with the lowest **T**

241   value had a 100% chance of being sampled, and tips with a higher **T** value were

242   probabilistically more likely, but not certain, to be missing (Figure 1C).

243   Due to the different ways in which missing tips were sampled, in the cluMT and

244   corMT scenarios the number of missing tips always corresponded exactly to the desired

245   proportion, but in the rMT case, that number varied slightly due to the probabilistic

246   sampling. For example, under 50% rMT the actual proportion of missing tips varied from

247   0.43 to 0.57. However, this does not bias my inferences because over 1000 simulations

248   the number of missing tips will average to the desired proportion, and the results will

249   reflect that.

250   For each simulation in each scenario of missing taxa, I quantified the phylogenetic

251   signal in **S** using Fritz and Purvis' (2010) D statistic, calculated with the function *phylo.d*

252   in *caper* (Orme et al. 2017). D equals 0 when a binary trait has evolved under a threshold

253   model with a Brownian liability as described above, and 1 when it has a phylogenetically

254   random distribution at the tips of the tree.

255   Once **T** had been simulated and tips had been pruned based on **S**, I used *OUwie*

256   to fit each of the four models to **T** and assess their support using sample size-corrected

257   Akaike information criteria (AICc; Burham and Anderson 1998), under the expectation

258   that the generating model should have the lowest AICc score. For each combination of

259   generating model, scenario of missing taxa, and percentage of missing taxa, I also

260 calculated the median delta AICc of the generating model in the set of simulations in

261 which it was not the top model. Delta AICc equals the AICc score of the focus model

262 minus the AICc score of the model with the most support (lowest AICc), and is thus a

263 measure of relative support of a model compared to the top model.

264   Finally, I computed the bias and precision of parameter estimates under each

265 scenario to assess how they were affected by missing taxa. I calculated bias as the mean

266 parameter estimate minus the generating parameter value. I calculated precision as the

267 median absolute deviation (median deviation from the median) of parameter estimates

268 among simulations. I used this statistic in lieu of the simple variance because it is more

269 robust to outliers. I normalized both bias and precision by dividing them by the generating

270 parameter values.

271

272   **Results**

273   The mean value of Fritz and Purvis' D statistic for **S**, the trait determining the

274 sampling status of each tip, was 1.004 and -0.019 across all simulations in the rMT and

275 cluMT scenarios respectively, indicating, as intended, low phylogenetic signal of missing

276 taxa in the former, and high phylogenetic signal of missing taxa in the latter. For corMT,

277 because **S** was correlated to **T**, its phylogenetic signal was computed separately

278 depending on **T**'s model of evolution. The mean value of Fritz and Purvis' D statistic in

279 that scenario was 0.921, 0.931, 0.944, and 0.932 for **T** simulated under BM, BMS, OU

280 and OUM, respectively.

281   Clumped and correlated missing taxa resulted at most in a slight increase in the

282 model selection error rate compared to no missing taxa or to random missing taxa, as

283   indicated by the number of simulations in which the generating model did not have the

284   lowest AICc among the four models (Table 2, Figure 2). For example, when the

285   generating model was OUM, it failed to receive the most support in 22 out of 1000

286   simulations under a no missing taxa scenario, 4 out of 1000 under 90% rMT, 24 out of

287   1000 under 90% cluMT, and 11 out of 1000 under 90% corMT. The model selection error

288   rate did not consistently increase with the percentage of missing taxa under any scenario,

289   but under all scenarios the error rate was higher (always ~30%) when BM was the

290   generating model, usually followed by OU, then by BMS and OUM (Table 2, Figure 2).

291   BM was most often confused for BMS and OU, and more rarely confused for OUM (Figure

292   2). Likewise, BMS was confused more often for BM and OU than for OUM (Figure 2). In

293   contrast, OU was often confused for OUM. Finally, OUM was only very rarely not selected

294   as the top model when it was the generating model (Table 2, Figure 2), indicating that it

295   leaves the strongest signature on the data among all models examined here.

296   Further revealing that missing taxa have little impact on model selection, the

297   median delta AICc of the generating model in the set of simulations where it was not the

298   top model was almost always ≤2 (Table 3). This means, following Burham and Anderson's

299   (1998) rule of thumb that "models having delta AICc≤2 have substantial support", that the

300   generating model still had relatively high support even in the cases where it was not the

301   top model.

302   I also examined the bias (Table 4) and precision (Table 5) of parameter estimates

303   across simulations in each scenario. Because I normalized these metrics, they are directly

304   comparable across parameters. The absolute bias (Table 4) of σ^2_0 was always low

305   (<0.0374) for all models and almost always remained so for σ^2_1 in BMS, even though σ^2_1

306   covered much smaller proportions of my simulated trees. The absolute bias of θ_0 also

307   remained fairly low (<0.01), but the absolute bias of θ_1 was noticeably higher across most

308   situations (often >0.2). Finally, α had by far the greatest bias of all parameters (almost

309   always >0.1).

310   Notably, although biases were generally low, they often tended to be systematic,

311   meaning they were always in the same direction. σ^2 was consistently underestimated in

312   BM and BMS models and overestimated in OU and OUM. α was always overestimated,

313   and θ_1 always underestimated. Only for θ_0 was bias not systematic, as it was sometimes

314   positive and sometimes negative. Regarding precision of parameter estimates, across all

315   scenarios it tended to be to be better (meaning parameter estimates were more

316   concentrated around the generating parameter value) for θ, followed by σ^2, and worst for

317   α (Table 5).

318   Interestingly, looking at each parameter individually, both bias and precision were

319   largely similar across almost all scenarios, revealing that missing taxa had little effect on

320   parameter estimation. Only at 90% corMT did the bias of most parameters worsen relative

321   to other scenarios and percentages (Table 4, Figure 3). For example, the bias of σ^2_0 in

322   the BM model was -0.0084 under nMT, -0.0067 under 90% rMT, -0.0012 under 90%

323   cluMT, and jumped to -0.0374 under 90% corMT. Similarly, the bias of θ_0 in the OUM

324   model was -0.0033 at nMT, -0.0032 at 90% rMT, -0.0036 at 90% cluMT, and much greater

325   at -0.0142 at 90% corMT. In contrast, the precision of parameter estimates did not

326   deteriorate consistently even under 90% corMT (Table 5). For example, for σ^2_0 in the BM

327   model, the precision was 0.0834 under 90% corMT compared to 0.0800 under nMT and

328   0.0862 under 10% corMT.

329

330   **Discussion**

331   **Model performance in the presence of missing taxa**

332   Random and phylogenetically clumped missing taxa had little effect on model

333   selection and parameter estimation for univariate continuous traits evolving along

334   phylogenetic trees, even when 90% of taxa in a tree were missing. This result can be

335   intuitively understood by appreciating that neither of these two scenarios change the

336   distribution of trait values at the tips of the tree. Under random missing taxa, each tree tip

337   has the same chance of being missing, independent of its trait value or phylogenetic

338   position. Under phylogenetically clumped missing taxa, the chances of a tip being missing

339   are independent of trait values, meaning that if a taxon is missing, it is likely that its close

340   relatives will also be missing. These close relatives probably have similar trait values, but

341   the distribution of trait values at the tips will not change because clumps of missing taxa

342   will be evenly distributed across the tree, and trait values are not correlated across these

343   clumps (Figure 1). The only notable effect of missing taxa was when 90% of taxa were

344   missing under a scenario in which sampling probability was correlated to the trait of

345   interest (corMT). In that situation, there was a noticeable deterioration in the bias, but not

346   the precision, of all parameters. This result can be understood by realizing that this is the

347   only type of missing taxa that alters the distribution of trait values at the tips, thus

348   misleading estimates.

349

350   **Model performance without missing taxa**

351   Even though my main interest was exploring the implications of missing taxa, and

352   those proved to be minor, it is also valuable to discuss performance in a scenario without

353   missing taxa, seeing as model performance, in particular of OU models, has received

354   significant recent attention in the form of simulation studies (Beaulieu et al. 2012, Ho and

355   Ané 2014, Cooper et al. 2015, Cressler et al. 2015).

356   BM was by far the model that most often failed to be selected by AICc when it was

357   in fact the generating model, with an error rate of ~30% across all scenarios (Table 2,

358   Figure 2). This suggests that BM is quite prone to mimicking patterns expected under

359   other models, echoing Slater et al.'s (2012b) finding, in a simulated birth-death phylogeny

360   with 100 extant tips, that statistical power is low to favor BM over other candidate models

361   even when it is the generating model. In particular, in my simulations with no missing taxa,

362   a single-peak OU model was mistakenly selected over BM in 12.2% of the simulations

363   (Figure 2) when BM was the generating model. This result can be compared to Cooper

364   et al.'s (2015), who similarly simulated BM datasets and computed how often an OU

365   model was mistakenly selected as the best-fitting model. However, instead of using

366   Akaike information criteria for model selection, they used likelihood-ratio tests and Bayes

367   factors. The error rate they found was considerably lower than what I found using AICc,

368   for example 5.5% and 0.2% in pure-birth trees with 200 tips using likelihood-ratio tests

369   and Bayes factors respectively.

370   To explore the discrepancy between mine and Cooper et al.'s (2015) results, I ran

371   likelihood ratio tests (LRTs) on the results of my nMT simulations under BM (details

372   presented as supplementary material). I found that LRTs incorrectly preferred OU over

373   the true BM model in 6.7% of the simulations, a number much lower than the AICc error

374   rate of 12.2% and more in line with Cooper et al.'s (2015) 5.5%. This result suggests that

375   LRTs may perform better than AICc as a tool to discriminate between BM and OU models

376   in comparative datasets, an issue that, while beyond the purview of the present paper,

377   deserves further investigation.

378   Regarding parameter estimates, α was the hardest parameter to estimate in every

379   scenario, seeing as it had the largest biases and poorest precisions, sometimes by orders

380   of magnitude, compared to all other parameters (Tables 4 and 5). This result is consistent

381   with Cressler et al.'s (2015), who found the related parameter η (selection opportunity,

382   equaling the product of α and the height of the phylogeny) the hardest parameter to

383   estimate, and θ the easiest in OUM models. In my results α and σ^2 estimates tended to

384   have a consistent upward bias in OUM models, whereas θ_1 was biased downwards and

385   θ_0 did not have a consistent upwards or downwards bias (Figure 3). These results are

386   similar to Beaulieu et al.'s (2012), except, interestingly, for θ_0, which Beaulieu et al. (2012)

387   found to be underestimated as well, although in that case simulations were based on

388   more complex models.

389

390   **Broader implications for models of trait evolution in the presence of missing**

391   **taxa**

392   My results bear on the issue of birth-death polytomy resolver (BDPR) algorithms,

393   such as those of Kuhn et al. (2011) and Thomas et al. (2013), that impute the position of

394   missing taxa onto a phylogeny based solely on taxonomic information and a

395   diversification model. Jetz et al. (2012) used a BDPR to add unsampled taxa to a species-

396   level phylogeny of the world's birds, amounting to 33% of all species in the final tree. This

397   approach was shown by Rabosky (2015) to systematically and severely bias downstream

398   analyses of trait macroevolution. For example, when using a BDPR algorithm, Rabosky

399   (2015) found the Brownian rate parameter σ^2 to be overestimated by an average factor

400   of 3.51 times, compared to a bias of no more than 0.02 in any of my scenarios with missing

401   taxa (Table 4). Therefore, my results suggest that the reliability of BDPR algorithms may

402   be a moot point for studies concerned mainly with continuous traits, because model

403   selection and parameter estimation might perform better on phylogenies with a very high

404   proportion of realistic missing taxa than on BDPR-imputed phylogenies. This is in

405   accordance with Rabosky's (2015) "genes only" scenario, where he ran trait evolution

406   analyses with minimal bias after pruning BDPR-imputed species from the Jetz et al.

407   (2012) tree. However, this is not to say that BDPR algorithms are never useful. We need

408   to remember that BDPRs do demonstrably improve the performance of comparative

409   studies that do not include traits, such as speciation-extinction analyses (Rabosky 2015).

410   Even though my and Rabosky's (2015) findings offer some reassurance for the

411   use of incompletely sampled phylogenies for studies of continuous trait evolution, caution

412   must be taken with ecological traits that are correlated to sampling probabilities, a

413   scenario that caused biased estimates in my simulation. This includes traits tightly linked

414   to rarity, such as abundance and range size (Gaston 1994), as well as niche axes such

415   as habitat type. Rarity is an emergent property of species and not a property of individuals.

416   Traits with potential to bias model-based analyses will likely also be similarly emergent

417   species-level traits. However, it is difficult to generalize about the types of traits that will

418   negatively affect parameter estimates, because they will vary depending on the ecology

419   and life-history of the clade under study. Because ecology and life-history are most well-

420   known by the researchers investigating specific clades, it is advisable that these

421   researchers use their biological intuition to stay mindful of traits that they believe may be

422   linked to rarity and sampling probabilities in their study organisms, especially in situations

423   with extremely high proportions of missing taxa.

424   The statistical power of a tree and comparative dataset to correctly identify the

425   model that generated the data and to estimate model parameters scales with tree size

426   (Beaulieu et al. 2012, Boetigger et al. 2012). For example, trees with up to 128 tips

427   sometimes do not allow the correct selection of an OUM model over OU (Beaulieu et al.

428   2012). The problem of low statistical power in small trees is related to, but distinct from,

429   the problem of missing taxa. Phylogenetic trees may have small sizes and consequent

430   low statistical power for three reasons: (1) because taxon sampling is complete but the

431   clade under study is in fact a small clade; (2) because the clade under study is known to

432   be larger but researchers were unable to sample many known taxa, or (3) because the

433   clade is larger but contains many undiscovered, undescribed or recently extinct species,

434   so that researchers may not even be aware that their study contains missing taxa. In my

435   simulations, I isolated the missing taxa problem from the tree size problem by keeping

436   the number of tips constant *after* pruning missing taxa, but this is not realistic. In practice

437   the two problems cannot be addressed separately, as under reason 2 above, and may

438   not even be distinguishable, as under reason 3 above. Assessment of statistical power

439   should be standard in comparative analyses. A simulation-based method to accomplish

440   this was described by Boetigger et al. (2012), but it implicitly assumes complete taxon

441   sampling (reason 1 above) and it is unclear if it is adequate in situations with missing

442   taxa. Slater et al. (2012a) described a method to account for missing taxa in comparative

443   studies, but it assumes that missing taxa are phylogenetically random, a scenario that I

444   argue is unlikely. As far as I am aware, there are no available methods to account for

445   clumped, correlated or unknown (reason 3 above) missing or recently extinct taxa.

446   In fact, the scenarios produced by my simulations, where I pruned terminal

447   branches from a tree to represent missing taxa, are indistinguishable from scenarios that

448   would be produced by recent extinctions (*i.e.*, extinction at the tips of the phylogeny only).

449   Therefore, my findings might also be interpreted in that context as meaning that random,

450   clumped or correlated recent extinction appear to have little effect on model performance.

451   My results should not, however, be extended to scenarios with extinction in internal

452   branches, which are likely to be prevalent in real datasets (Stadler 2010, Slater et al.

453   2012b, Stadler et al. 2018). As a case in point, Slater et al. (2012b) found the performance

454   of model selection to significantly deteriorate when randomly-simulated fossil (extinct)

455   taxa were pruned from trees. This deterioration is likely to be even greater if extinction is

456   clumped (Vamosi and Wilson 2008, Rabosky 2009) or correlated (FitzJohn 2010, Harvey

457   and Rabosky 2018) to trait values.

458   The roster of phylogenetic models of trait evolution is large and ever-growing, and

459   here I examined the effects of missing taxa on only a very small part of that universe.

460   Rabosky (2015) found that results for discrete traits were generally similar to those for

461   continuous traits, but before further analyses it is difficult to speculate on how my findings

462   might apply to other models. Furthermore, the scenarios I used to simulate missing taxa

463   were quite limited. Whereas it is clear that missing taxa will often be phylogenetically

464   clumped and/or correlated to the investigated traits, I cannot be sure that the ways in

465   which I simulated those scenarios are optimal. In particular, my non-phylogenetic

466 simulation of corMT is perhaps less than ideal. A more realistic way to simulate that

467 scenario would be under a multivariate phylogenetic model where the value of a liability

468 **L** is governed both by a Brownian Motion process and by a correlation with a trait of

469 interest **T**, which may be evolving the same or a different model than **L**. However, to the

470 best of my knowledge, that type of multivariate model is not currently available.

471   Another caveat is that, my simulations notwithstanding, scenarios of missing taxa

472 indisputably exist that do severely mislead phylogenetic comparative methods. For

473 example, without delving into details, a situation in which all taxa at one extreme of the

474 trait value distribution are deterministically missing from a dataset (i.e., all tips with the

475 X% greatest values of **T** are always unsampled) would truncate the distribution of trait

476 values, altering its mean and variance and likely leading to selection of incorrect models

477 and poor performance of parameter estimates. However, that scenario is unrealistic,

478 because even the rarest taxa will always have a non-zero chance of being discovered,

479 collected and sampled.

480   As a final cautionary note, my study, as all simulation studies, is only valuable to

481 the extent that our models are accurate. In reality, evolution is likely to proceed in a much

482 more heterogenous manner than reflected in any of our current, simplistic models of trait

483 evolution. We do not know, and might never know, the real generating "models" of

484 empirical datasets. Moreover, actual data, unlike the data I simulated here, are not ideal

485 (*e.g.*, they have measurement error). Therefore, I recommend that my results be not

486 uncritically taken to be applicable whenever any of the models I tested here is fit to real

487 data. The processes that generated a given real dataset may or may not be as robust to

488 missing taxa as are our simple models. Rather, I expect my results to prompt empirical

489   phylogenetic comparative biologists to think critically about missing taxa in their datasets

490   and how they might be distributed. The better the fit of a dataset to a model is, the more

491   likely it is that my findings on the effects of missing taxa will be applicable.

492

493   **Conclusion**

494   My results demonstrate that realistic scenarios of missing taxa do not affect the

495   performance of selected simple models of univariate continuous trait evolution, except for

496   parameter estimates in situations with a very high proportion of correlated missing taxa.

497   Understanding the impact of missing taxa in comparative methods is a multifaceted

498   problem that still demands much additional work. Topics that are ripe for further study,

499   considering my results, include the effect of missing taxa on models of discrete trait

500   evolution, exploration of better ways to simulate correlated missing taxa, examinations of

501   the level of phylogenetic signal in missing taxa in real datasets, and disentangling the

502   problem of small tree size from the problem of missing taxa. The framework I have

503   provided here will be a starting point for such studies. To facilitate that, R scripts to

504   replicate my simulations are available on GitHub. Proponents of existing and novel

505   models of trait evolution will benefit by using this framework to more explicitly and

506   thoroughly investigate model performance under realistic scenarios of missing taxa.

507

508   **Acknowledgments**

509   This study benefitted from feedback from Jeremy Brown, Robb Brumfield, Jake

510   Esselstyn and all participants in the Phyleaux discussion group at LSU. All simulations

511   were run in the Odyssey computing cluster at Harvard University, access to which is

512 thanks to Scott Edwards. This research was supported in part by NSF grant DEB-

513 1146265, and by a "Science Without Borders" doctoral fellowship to RSM from Brazil's

514 National Council for Scientific and Technological Development (CNPq; 201234/2014-9).

515

516 **Data and code archival statement**

517   R scripts to replicate the analyses and results of the simulations are available at

518 https://github.com/rafmarcondes/Missing_taxa.

519

520 **References cited**

521

522 Arita, H. T., J. G. Robinson, K. H. Redford, H. T. Arita, J. G. Robinson, and K. H.

523   Redford (1990). Rarity in Neotropical Forest Mammals and Its Ecological

524   Correlates. Conservation Biology 4:181–192.

525 Barr, W. A., and R. S. Scott (2014). Phylogenetic comparative methods complement

526   discriminant function analysis in ecomorphology. American Journal of Physical

527   Anthropology 153:663–674. doi: 10.1002/ajpa.22462

528 Beaulieu, J. M., D. C. Jhwueng, C. Boettiger, and B. C. O'Meara (2012). Modeling

529   stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive

530   evolution. Evolution 66:2369–2383. doi: 10.1111/j.1558-5646.2012.01619.x

531 Boettiger, C., G. Coop, and P. Ralph (2012). Is your phylogeny informative? Measuring

532   the power of comparative methods. Evolution 66:2240–2251. doi: 10.1111/j.1558-

533   5646.2011.01574.x

534 Brown, J. M., and R. C. Thomson (2018). Evaluating Model Performance in

535   Evolutionary Biology. Annual Review of Ecology, Evolution, and Systematics

536   49:annurev-ecolsys-110617-062249. doi: 10.1146/annurev-ecolsys-110617-062249

537   Butler, M. A., and A. A. King (2004). Phylogenetic Comparative Analysis: A Modeling

538   Approach for Adaptive Evolution. The American Naturalist 164:683–695. doi:

539   10.1086/426002

540   Burham, K. P., and D. R. Anderson. (1998). Model Selection and Inference: A Practical

541   Information-Theoretuc Approach. Springer Verlag, New York.

542   Cressler, C. E., M. A. Butler, and A. A. King. (2015). Detecting adaptive evolution in

543   phylogenetic comparative analysis using the Ornstein–Uhlenbeck model.

544   Systematic Biology, 64(6), 953-968.

545   Crisp, M. D., and L. G. Cook (2012). Phylogenetic niche conservatism: What are the

546   underlying evolutionary and ecological causes? New Phytologist 196:681–694. doi:

547   10.1111/j.1440-1827.1997.tb04503.x

548   DeSantis, L. R. G., R. A. Tracy, C. S. Koontz, J. C. Roseberry, and M. C. Velasco

549   (2012). Mammalian niche conservation through deep time. PLoS ONE 7. doi:

550   10.1371/journal.pone.0035624

551   Eaton, D. A. R., E. L. Spriggs, B. Park, and M. J. Donoghue (2017). Misconceptions on

552   Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering

553   Plants. Systematic Biology 66(3):399-412.

554   Felsenstein, J. (1985). Phylogenies and the comparative method. The American

555   Naturalist 125(1):1-15.

556   Felsenstein, J. (2005). Using the quantitative genetic threshold model for inferences

557   between and within species. Philosophical Transactions of the Royal Society B

558   360:1427-1434.

559   FitzJohn, R. G. (2010). Quantitative traits and diversification. *Systematic biology*, *59*(6),

560   619-633.

561   Freckleton, R. P., P. H. Harvey, and M. Pagel (2013). Phylogenetic Analysis and

562   Comparative Data. The American Naturalist 160:712–726. doi: 10.1086/343873

563   Fritz, S. A., and A. Purvis (2010). Selectivity in mammalian extinction risk and threat

564   types: A new measure of phylogenetic signal strength in binary traits. Conservation

565   Biology 24:1042–1051. doi: 10.1111/j.1523-1739.2010.01455.x

566   Garamszegi, L. Z., and A. P. Møller (2011). Nonrandom variation in within-species

567   sample size and missing data in phylogenetic comparative studies. Systematic

568   Biology 60:876–880. doi: 10.1093/sysbio/syr060

569   Gaston, K. J. (1994). Rarity. Chapman & Hall, London.

570   Hansen, T. F. (1997). Stabilizing selection and the comparative analysis of adaptation.

571   Evolution 51:1341–1351.

572   Harvey, M. G., and D. L. Rabosky. (2018). Continuous traits and speciation rates:

573   Alternatives to state-dependent diversification models. *Methods in Ecology and*

574   *Evolution*, *9*(4), 984-993.

575   Herrera-Alsina, L., and R. Villegas-Patraca (2014). Biologic interactions determining

576   geographic range size: A one species response to phylogenetic community

577   structure. Ecology and Evolution 4:968–976. doi: 10.1002/ece3.959

578   Ingram, T., and D. L. Mahler (2013). SURFACE: Detecting convergent evolution from

579   comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike

580   Information Criterion. Methods in Ecology and Evolution 4:416–425. doi:

581   10.1111/2041-210X.12034

582 Jablonski, D. (1987). Heritability at the Species Level: Analysis of Geographic Ranges

583   of Cretaceous Mollusks. Science 238:360-363.

584 Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers (2012). The global

585   diversity of birds in space and time. Nature 491:444–8. doi: 10.1038/nature11631

586 Jiang, W., S-Y Chen, H. Wang, D-Z. Li, and J. J. Wiens (2014). Should genes with

587   missing data be excluded from phylogenetic analyses? Molecular Phylogenetics

588   and Evolution 80:308-318.

589 Kuhn, T. S., A. Mooers, and G. H. Thomas (2011). A simple polytomy resolver for dated

590   phylogenies. Methods in Ecology and Evolution 2:427–436. doi: 10.1111/j.2041-

591   210X.2011.00103.x

592 Losos, J. B. (2008). Phylogenetic niche conservatism, phylogenetic signal and the

593   relationship between phylogenetic relatedness and ecological similarity among

594   species. Ecology Letters 11:995–1003. doi: 10.1111/j.1461-0248.2008.01229.x

595 Louca, S., and M. Doebeli. (2017). Efficient comparative phylogenetics on large trees.

596   *Bioinformatics*, *34*(6), 1053-1055.

597 Malaney, J. L., and J. A. Cook (2018). A perfect storm for mammalogy: declining

598   sample availability in a period of rapid environmental degradation. Journal of

599   Mammalogy 99:773–788. doi: 10.1093/jmammal/gyy082

600 Maurer, B. A. (1991). Concluding remarks: Birds, body size and evolution. In: Acta XX

601   Congressus Internationalis Ornithologici. New Zealand Ornithological Congress

602   Board, Wellington.

603 O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright (2006). Testing for

604   different rates of continuous trait evolution using likelihood. Evolution; international

605   journal of organic evolution 60:922–33. doi: 10.1111/j.0014-3820.2006.tb01171.x

606 O'Meara, B. C., and J. M. Beaulieu (2014). Modelling Stabilizing Selection: The

607   Attraction of Ornstein–Uhlenbeck Models. In Modern Phylogenetic Comparative

608   Methods and Their Application in Evolutionary Biology (L. Z. Garamszegi, Editor).

609   Springer, Berlin, Heidelberg.

610 Orme, D., R. Freckleton, G. Thomas, T. Petzoldt, S. Fritz, N. Isaac and W.Pearse.

611   (2017). The caper package: comparative analysis of phylogenetics and evolution in

612   R, version 1.0.1.

613 Paradis, E. and K. Schliep. (2019). ape 5.0: an environment for modern phylogenetics

614   and evolutionary analyses in R. Bioinformatics 35(3):526-528.

615 Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn,

616   M. E. Alfaro, and L. J. Harmon. (2014). geiger v2. 0: an expanded suite of methods

617   for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*(15),

618   2216-2218.

619 Pennell, M. W., R. G. FitzJohn, and W. K. Cornwell (2016). A simple approach for

620   maximizing the overlap of phylogenetic and comparative data. Methods in Ecology

621   and Evolution 7:751–758. doi: 10.1111/2041-210X.12517

622 Qian, H., and R. E. Ricklefs (2004). Geographical distribution and ecological

623   conservatism of disjunct genera of vascular plants in eastern Asia and eastern

624   North America. Journal of Ecology 92:253–265. doi: 10.1111/j.0022-

625   0477.2004.00868.x

626 Rabosky, D. L. (2009). Heritability of extinction rates links diversification patterns in

627   molecular phylogenies and fossils. Systematic Biology, 58(6), 629-640.

628   Rabosky, D. L. (2015). No substitute for real data: A cautionary note on the use of

629   phylogenies from birth-death polytomy resolvers for downstream comparative

630   analyses. Evolution 69:3207–3216. doi: 10.1111/evo.12817

631   Reddy, S. (2014). What's missing from avian global diversification analyses? Molecular

632   Phylogenetics and Evolution 77:159–165. doi: 10.1016/j.ympev.2014.04.023

633   Slater, G. J., L. J. Harmon, D. Wegmann, P. Joyce, L. J. Revell,and M. E. Alfaro.

634   (2012a). Fitting models of continuous trait evolution to incompletely sampled

635   comparative data using approximate Bayesian computation. Evolution66-3: 752-

636   762.

637   Slater, G. J., L. J. Harmon, and M. E. Alfaro. (2012b). Integrating fossils with molecular

638   phylogenies improves inference of trait evolution. Evolution 66(12), 3931-3944.

639   Stadler, T. (2010). Sampling-through-time in birth-death trees. Journal of Theoretical

640   Biology 267:396-404.

641   Stadler, T., A. Gavryushkina, R. C. Warnock, A. J. Drummond, and T. A. Heath. (2018).

642   The fossilized birth-death model for the analysis of stratigraphic range data under

643   different speciation modes. Journal of Theoretical Biology, *447*, 41-55.

644   Thomas, G. H., K. Hartmann, W. Jetz, J. B. Joy, A. Mimoto, and A. O. Mooers (2013).

645   PASTIS: An R package to facilitate phylogenetic assembly with soft taxonomic

646   inferences. Methods in Ecology and Evolution 4:1011–1017. doi: 10.1111/2041-

647   210X.12117

648   Thomson, R. C., and H. B. Shaffer (2010). Rapid progress on the vertebrate tree of life.

649   BMC Evolutionary Biology 8:19.

650   Vamosi, J. C., and J. R. Wilson. (2008). Nonrandom extinction leads to elevated loss of

651   angiosperm evolutionary history. Ecology Letters, *11*(10), 1047-1053.

652   Waldron, A. (2007). Null Models of Geographic Range Size Evolution Reaffirm Its

653   Heritability. The American Naturalist 170:221–231. doi: 10.1086/518963

654   Wiens, J. J., D. D. Ackerly, A. P. Allen, B. L. Anacker, L. B. Buckley, H. V. Cornell, E. I.

655   Damschen, T. Jonathan Davies, J. A. Grytnes, S. P. Harrison, B. A. Hawkins,R. D.

656   Holt, C. M McCain, and P. R. Stephens. (2010). Niche conservatism as an

657   emerging principle in ecology and conservation biology. Ecology Letters 13:1310–

658   1324. doi: 10.1111/j.1461-0248.2010.01515.x

659   Wiens, J. J., and C. H. Graham (2005). Niche Conservatism: Integrating Evolution,

660   Ecology, and Conservation Biology. Annual Review of Ecology, Evolution, and

661   Systematics 36:519–539. doi: 10.1146/annurev.ecolsys.36.102803.095431

662   Wiens, J. J., and M. C. Morrill (2011). Missing Data in Phylogenetic Analysis: Reconciling

663   Results   from Simulations and Empirical Data. Systematic Biology 60(5):719-731.

664   Zuk, M., and M. Travisano (2018). Models on the Runway: How Do We Make Replicas

665   of the World? The American Naturalist 192:1–9. doi: 10.1086/697508

666

667

Summary of simulated traits and their generating models.

1   Table 1. Summary of simulated traits and their generating models.

Abbreviation of trait	Description	Model
R	Binary trait representing regimes required for BMS and OUM models for **T**	Equal-rates MK model
T	Continuous trait about whose evolution we are interested in making inferences	BM, BMS, OU or OUM
S	Binary trait determining the sampling status of each tip (0=missing, 1=sampled)	Threshold model for cluMT; non-phylogenetic for corMT
L	Continuous trait representing the liability underlying **S** in the threshold model in the cluMT scenario	BM

Model selection error rate.

Number of simulations, for each combination of type of missing taxa, percentage of missing taxa, and generating model for the trait of interest T, in which the model with the greatest support (lowest AICc) was not the generating model. The numbers are always out of 1000 replicated simulations.

1 Table 2. Model selection error rates. Number of simulations, for each combination of

2 type of missing taxa, percentage of missing taxa, and generating model for the trait of

3 interest T, in which the model with the greatest support (lowest AICc) was not the

4 generating model. The numbers are always out of 1000 replicated simulations.

		BM	BMS	OU	OUM
No missing taxa		306	61	181	22
Random missing taxa	10%	296	53	176	13
	50%	298	70	161	11
	90%	266	83	196	4
Clumped missing taxa	10%	279	63	167	18
	50%	291	63	195	22
	90%	318	60	194	24
Correlated missing taxa	10%	300	47	178	25
	50%	305	73	177	16
	90%	381	99	169	11

Delta AICc of the generating model when it was not selected as the top model.

Median delta AICc of the generating model for the trait of interest **T** in the set of simulations, under each combination of type of missing taxa, percentage of missing taxa and generating model, in which the generating model was not the model with the greatest support.

1 Table 3. Median delta AICc of the generating model for the trait of interest T in the set of

2 simulations, under each combination of type of missing taxa, percentage of missing taxa

3 and generating model, in which the trait of interest was not the model with the greatest

4 support (lowest AICc).

		BM	BMS	OU	OUM
No missing taxa		1.592	2.127	1.422	1.791
Random missing taxa	10%	1.676	1.471	1.215	1.826
	50%	1.759	1.927	1.459	25.669
	90%	1.577	1.508	1.391	4.261
Clumped missing taxa	10%	1.966	1.784	1.047	1.370
	50%	1.866	2.174	1.196	2.325
	90%	1.811	2.079	1.210	1.515
Correlated missing taxa	10%	1.865	2.364	1.414	1.364
	50%	1.621	1.442	1.243	3.016
	90%	2.252	2.353	1.441	3.875

Bias of parameter estimates

Bias of parameter estimates from models fitted to data simulated under various scenarios and proportions of missing taxa. Bias was calculated as the generating parameter value, minus the mean estimated parameter across 1000 simulations, divided by the generating parameter value. Types of missing taxa: nMT: no missing taxa; rMT: random missing taxa; cluMT: phylogenetically clumped missing taxa; corrMT missing taxa correlated to the trait of interest).

				Normalized bias per scenario							
Parameter	Model	nMT	10% rMT	50% rMT	90% rMT	10% cluMT	50% cluMT	90% cluMT	10% corMT	50% corMT	90% corMT
σ^2_0	BM	-0.0084	-0.0060	-0.0006	-0.0062	-0.0002	-0.0014	-0.0012	-0.0032	-0.0090	-0.0374
σ^2_0	BMS	-0.0100	-0.0040	-0.0080	-0.0060	-0.0060	-0.0120	-0.0080	-0.0080	-0.0120	-0.0300
σ^2_1	BMS	-0.0090	-0.0180	-0.0080	-0.0150	-0.0090	-0.0080	0.0000	-0.0070	-0.0240	-0.0550
σ^2_0	OU	0.0280	0.0280	0.0240	0.0400	0.0120	0.0240	0.0240	0.0240	0.0220	0.0280
θ_0	OU	0.0002	-0.0004	0.0001	0.0002	-0.0003	0.0001	0.0004	0.0000	-0.0056	-0.0217
α	OU	0.1073	0.0907	0.0807	0.0853	0.0680	0.0873	0.0887	0.1107	0.0887	0.1460
σ^2_0	OUM	0.0460	0.0440	0.0560	0.1040	0.0520	0.0420	0.0340	0.0500	0.0560	0.0720
θ_0	OUM	0.0033	0.0034	0.0029	0.0032	0.0034	0.0036	0.0036	0.0025	-0.0011	-0.0142
θ_1	OUM	-0.0220	-0.0215	-0.0183	-0.0132	-0.0215	-0.0222	-0.0200	-0.0231	-0.0261	-0.0413
α	OUM	0.1347	0.1207	0.1207	0.1347	0.1400	0.1280	0.1040	0.1473	0.1280	0.1600

Precision of parameter estimates

Precision of parameter estimates from models fitted to data simulated under various scenarios and proportions of missing taxa. Precision was calculated as the median deviation from the median estimated parameter across 1000 simulations, divided by the generating parameter value. Types of missing taxa: nMT: no missing taxa; rMT: random missing taxa; cluMT: phylogenetically clumped missing taxa; corrMT missing taxa correlated to the trait of interest).

		Normalized precision per scenario									
Parameter	Model	nMT	10% rMT	50% rMT	90% rMT	10% cluMT	50% cluMT	90% cluMT	10% corMT	50% corMT	90% corMT
σ^2_0	BM	0.0800	0.0768	0.0838	0.0782	0.0792	0.0816	0.0810	0.0862	0.0756	0.0834
σ^2_0	BMS	0.1080	0.1000	0.1040	0.1040	0.1080	0.1000	0.1120	0.1020	0.1020	0.1020
σ^2_1	BMS	0.1340	0.1380	0.1420	0.1440	0.1430	0.1420	0.1400	0.1350	0.1460	0.1460
σ^2_0	OU	0.1180	0.1180	0.1160	0.1500	0.1240	0.1140	0.1140	0.1220	0.1220	0.1460
θ_0	OU	0.0090	0.0084	0.0079	0.0068	0.0085	0.0090	0.0088	0.0082	0.0082	0.0071
α	OU	0.2893	0.2587	0.2727	0.2520	0.2840	0.2673	0.2580	0.2673	0.2567	0.2547
σ^2_0	OUM	0.1260	0.1220	0.1360	0.1660	0.1160	0.1200	0.1180	0.1300	0.1260	0.1580
θ_0	OUM	0.0106	0.0090	0.0089	0.0077	0.0102	0.0098	0.0094	0.0090	0.0089	0.0078
θ_1	OUM	0.0141	0.0145	0.0147	0.0144	0.0141	0.0149	0.0160	0.0135	0.0134	0.0108
α	OUM	0.2893	0.2760	0.2727	0.2847	0.2813	0.2787	0.3027	0.3147	0.2953	0.2873

Figure 1

Illustration of the various simulated configurations of missing taxa.

(A-C) Trees depicting a trait **T** simulated under an OUM model, and random (A), phylogenetically clumped (B), or correlated (C) missing taxa. (D) Selective regimes, simulated under an MK model, underlying variation in the theta parameter in the OUM model used to simulate **T** in A-C. The tree in this figure contains 333 tips, 33 of which were simulated to be missing in each panel, thus corresponding to a 10% missing taxa scenario.

- Missing tip

Trait value

A

B

C

D

OFF
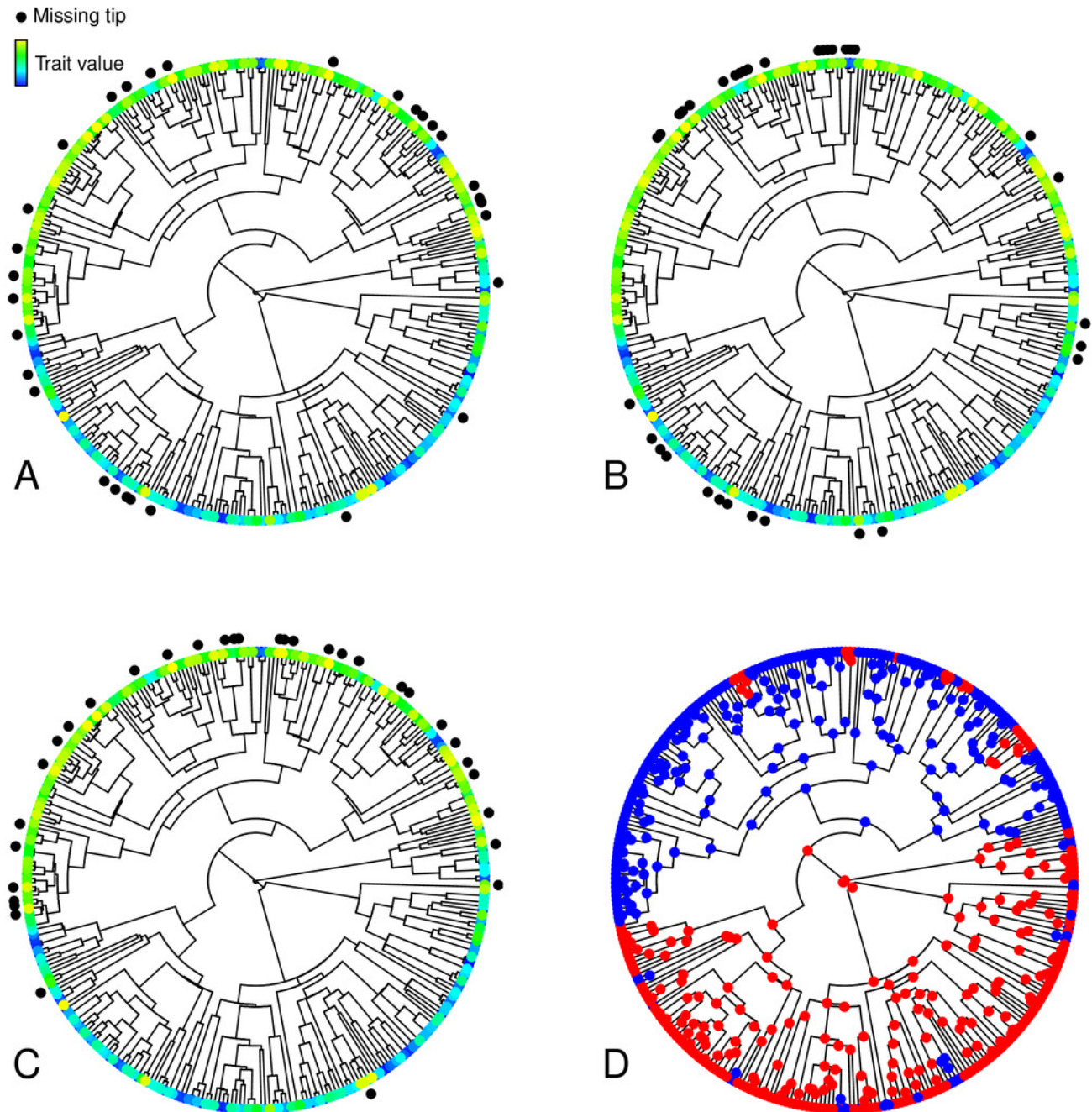
Figure 2

Best-fitting models selected in each set of simulations

Bar plots depicting the number of simulations in which each model was selected by AICc as the best fitting model, under each combination of type of missing taxa, percentage of missing taxa, and trait of interest generating model. Types of missing taxa: nMT: no missing taxa; rMT: random missing taxa; cluMT: phylogenetically clumped missing taxa; corrMT missing taxa correlated to the trait of interest)
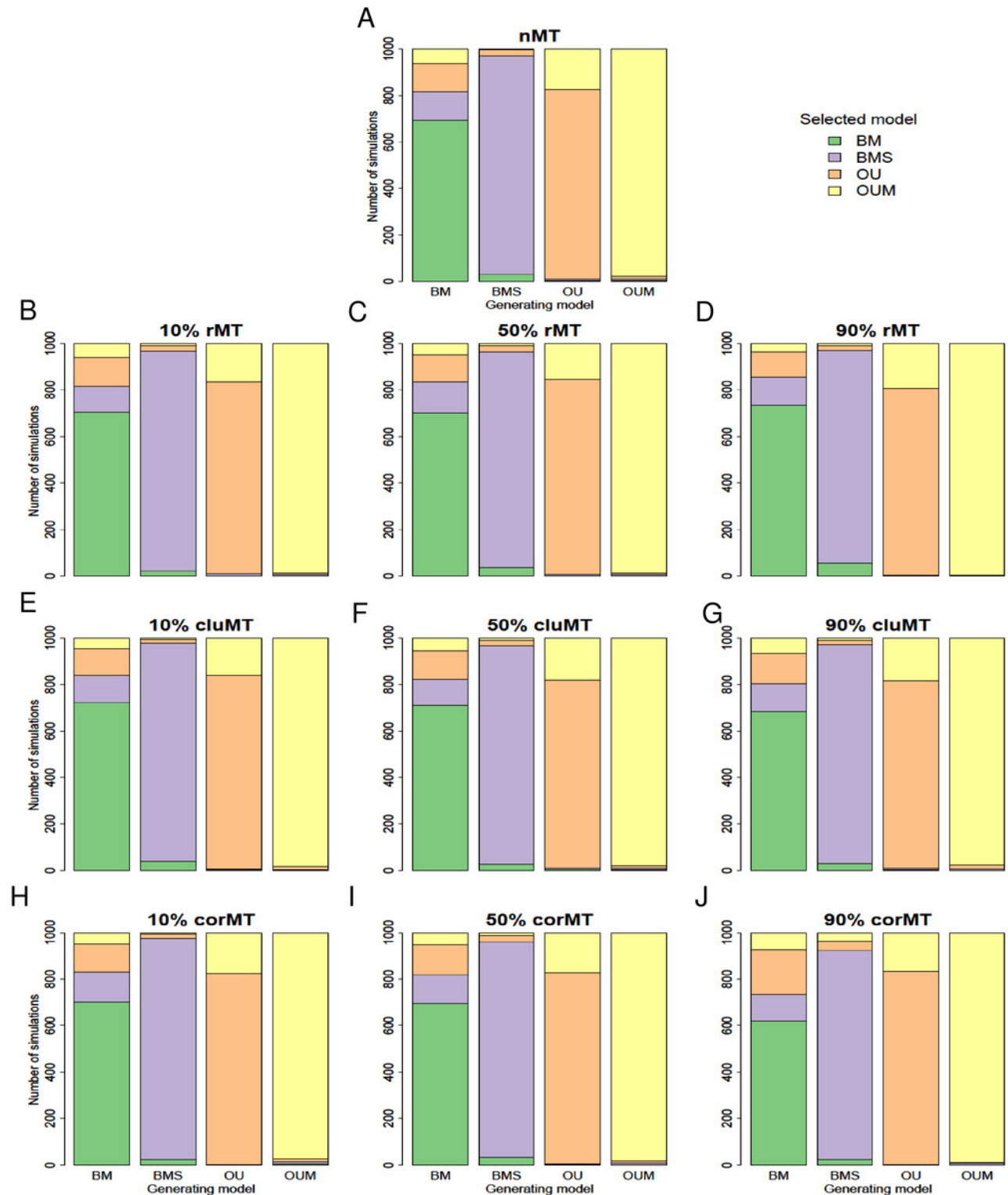
Figure 3

Distribution of parameter estimates

Distribution of estimated parameter values for each parameter in each combination of type of missing taxa, percentage of missing taxa and trait of interest generating model. The dotted lines represent the real, generating parameter values. The type and percentage of missing taxa are indicated at the bottom of the figure, whereas the model and parameters are shown at the right-hand edge of the figure.Types of missing taxa: nMT: no missing taxa; rMT: random missing taxa; cluMT: phylogenetically clumped missing taxa; corrMT missing taxa correlated to the trait of interest)