

1 The error in Bayesian phylogenetic reconstruction  
2 when speciation co-occurs

3 Giovanni Laudanno<sup>1</sup>, Richèl J.C. Bilderbeek<sup>1</sup>, and Rampal S.  
4 Etienne<sup>1</sup>

5 <sup>1</sup>Groningen Institute for Evolutionary Life Sciences, University of  
6 Groningen, Groningen, The Netherlands

7 December 3, 2018

8 **Abstract**

9  
10 **Keywords:** computational biology, evolution, phylogenetics, Bayesian anal-  
11 ysis, tree prior [RJCB: Have you already looked up for a target journal?]  
12 [GL: Honestly I have literally no idea how to select a good journal  
13 for this kind of article.] [RJCB: May I suggest we aim for Molecular  
14 Phylogenetics and Evolution, the same journal as the raket paper? ]

15 **1 Introduction**

- 16 • There are many contemporary tools that provide the possibility to infer a  
17 phylogeny from genetic data (DNA, RNA, proteins). A popular Bayesian

18 phylogenetic tool is called BEAST (Drummond & Rambaut 2007) and its  
19 cousin BEAST2 (Bouckaert *et al.* 2014).

- 20 • BEAST is very flexible, providing the user with the option to set up all  
21 possible phylogenetic priors (e.g. site/clock/speciation model).
- 22 • However, currently available priors can be not suitable to analyze some  
23 specific datasets. With this work we aim to test whether or not the im-  
24 plementation of a new prior model is beneficial to study a specific kind of  
25 diversification process.
- 26 • BEAST2 gives us the possibility to introduce new tree priors to infer  
27 phylogenies based on different assumptions on how the speciation process  
28 takes place.
- 29 • One of such speciation processes is the multiple birth hypothesis, a new  
30 model (described below) and thus currently absent in BEAST.
- 31 • The Multiple birth hypothesis can be useful to explain a phenomenon  
32 that has always puzzled evolutionary biologists: what are the drivers of  
33 the diversification processes for those phylogenies that show an impressive  
34 amount of speciation events in relatively short times? The (constant-rate)  
35 birth-death (BD) model embodies the common assumption that only a  
36 single speciation event can occur at any given time. The multiple-birth-  
37 death (MBD) model relaxes this assumption, allowing events in which  
38 large-scale environmental changes lead to a great number of species in  
39 relatively short time intervals. Such a hypothesis may be a better fit to  
40 describe the burst in systems like cichlid fish diversification in the African  
41 Great Lakes: Malawi, Tanganyika and Victoria (Janzen *et al.* 2016, Janzen  
42 *et al.* 2017).

- 43 • However, it may be that current BD tree priors are good enough at de-  
44 tecting such events, with a (preferred) lower level of complexity. If this is  
45 the case one should always be more keen to adopt the simplest model.
- 46 • Here we present our study with the aim of exploring when using a more  
47 complex MBD tree prior is warranted.
- 48 • We hypothesize that the error made today, using BD tree priors, increases  
49 with an increased number or stronger effect of multiple birth events. This  
50 is straightforward: without multiple birth events or such event having no  
51 effect, the MBD model falls back to a BD model. We expect larger errors  
52 when we deviate more from the BD model's assumptions. Additionally, we  
53 hypothesize MBD having a stronger effect if the normal speciation process  
54 is less pronounced. The more speciations are caused by the BD process,  
55 there are relatively less multiple-birth events. To put this hypothesis,  
56 H1, into an explicit equation, we expect the error made be correlated to  
57 the number of species created by the multiple-birth process over the total  
58 number of species created:

$$< e > = f\left(\frac{n_{taxa}^{MBD}}{n_{taxa}^{BD} + n_{taxa}^{MBD}}\right) \quad (1)$$

59 Where  $< e >$  denotes the expected error,  $f$  is a monotonously increas-  
60 ing function of unknown shape,  $n_{taxa}^{MBD}$  is the number of taxa created in  
61 multiple-birth events and  $n_{taxa}^{BD}$  is the number of taxa created by the stan-  
62 dard BD speciation process.

- 63 • We have the hypothesis, H7, that there is no difference in inference errors  
64 if the MBD process generates many species in many modest MB events  
65 or few intense MB events.

H1: we expect the error made be correlated to the number of species created by the multiple-birth process over the total number of species created

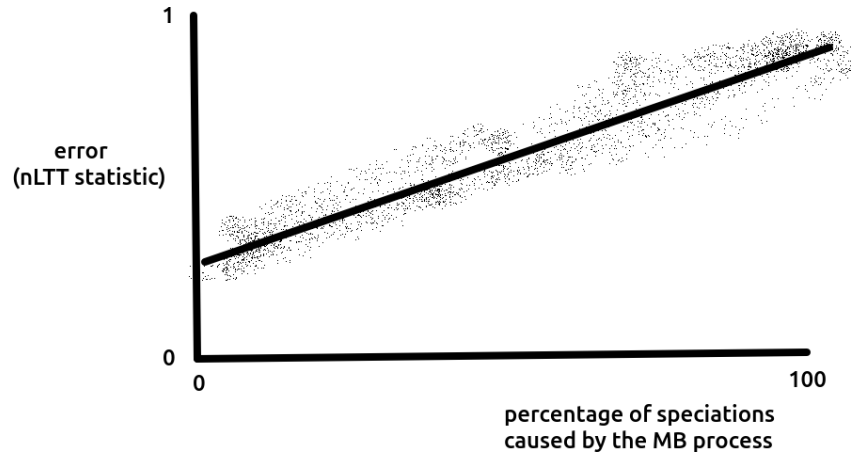


Figure 1: Hypothesis 1: we expect the error made be correlated to the number of species created by the multiple-birth process over the total number of species created

- We have the hypothesis, H2, that the effect of extinction rates is neutral, as extinctions will hit lineages created by both speciation processes equally.

H2: We expect the effect of extinction rates to be neutral

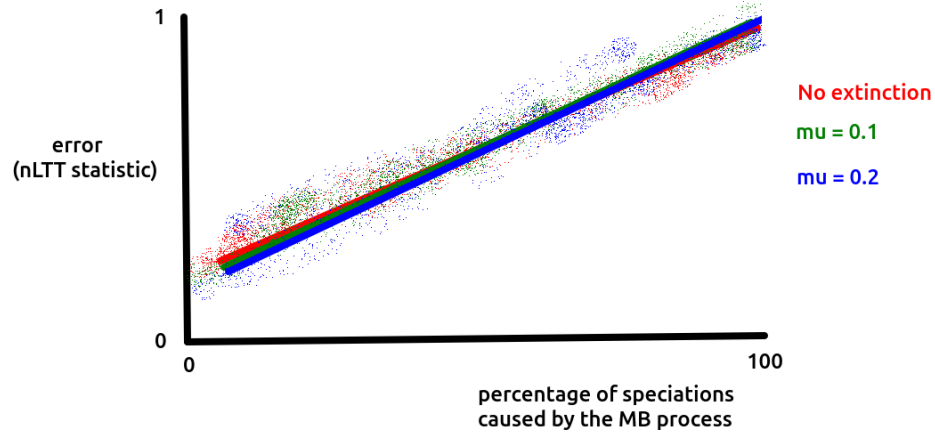


Figure 2: Hypothesis 2: the effect of extinction rates is neutral

- Due to the proportionality of the term within  $f$ , we have the hypothesis,

69 H3, that the timing of a multiple birth event (be it close to the crown age  
70 or close to the present) to have no effect. Compared to a late multiple  
71 birth event, an early multiple birth event may have a longer-lasting effect  
72 (as the next speciation event will be later), but it will create less new  
73 species, as there are still fewer taxa.

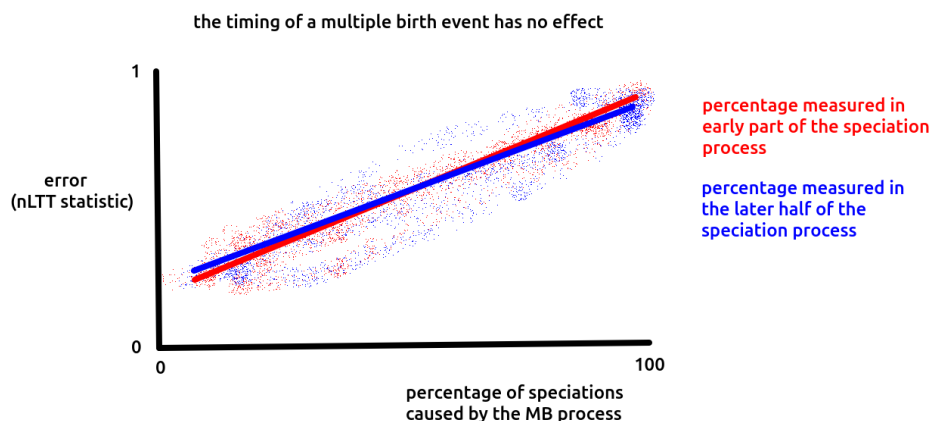


Figure 3: Hypothesis 3: the timing of a multiple birth event has no effect.

## 74 2 Methods

### 75 2.1 Model

- 76 • Current phylogenetic tools assume that only a single speciation event can  
77 occur at any given time. While this assumption is useful to construct  
78 a wide variety of successful models (e.g Maddison *et al.* 2007, Valente  
79 *et al.* 2015, Etienne *et al.* 2012, Etienne *et al.* 2014), they disallow for  
80 environmental changes that trigger speciations in multiple clades at a same  
81 point in time.
- 82 • The (constant-rate) birth-death (BD) model embodies the common as-  
83 sumption that only a single speciation event can occur at any given time.

The multiple-birth-death (MBD) model relaxes this assumption, allowing events in which large-scale environmental changes lead to a great number of species in relatively short time intervals. Such hypothesis can be useful to describe, for example, systems like cichlid fish diversification in the African Great Lakes: Malawi, Tanganyika and Victoria (Janzen *et al.* 2016, Janzen *et al.* 2017).

- In the MBD model, parameters  $\lambda$  and  $\mu$  correspond, respectively, to the common per-species speciation and extinction rates present also in the standard BD model. Additionally, MBD relies on two additional parameters. Parameter  $\nu$  is the rate at which an environmental change is triggered. When such event is triggered, all species present in the phylogeny at that moment have a probability  $q$  to speciate at that time, which is independent on  $\lambda$ . Polytomies are not allowed in such process as each species can speciate only once at the time.
- It is also possible to write down a likelihood function for such processes as in Laudanno 2018.

## 2.2 Simulations

- To prove our hypothesis we simulate two twin datasets. All the simulations are produced in continuous time, using the Doob-Gillespie algorithm.
- We start simulating  $N_S = 1000$  MBD trees, with either 50, 100 and 200 taxa. We have the hypothesis H5, that the number of taxa does not have an effect on the error being made, as there is no diversity dependency in any of the processes. We have the hypothesis H6, that for a higher number of taxa the variance in the error decreases, as more information is present in the simulated phylogenies.

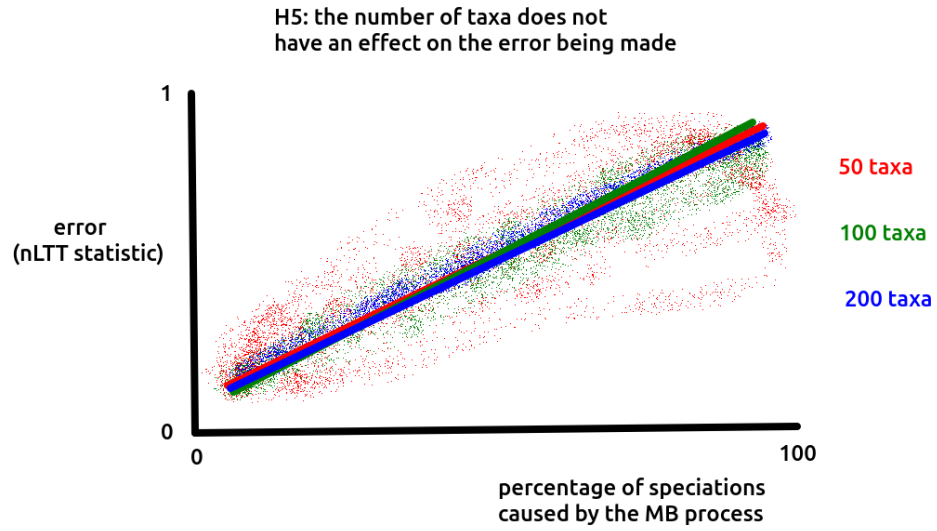


Figure 4: Hypothesis 5: the number of taxa does not have an effect on the error being made

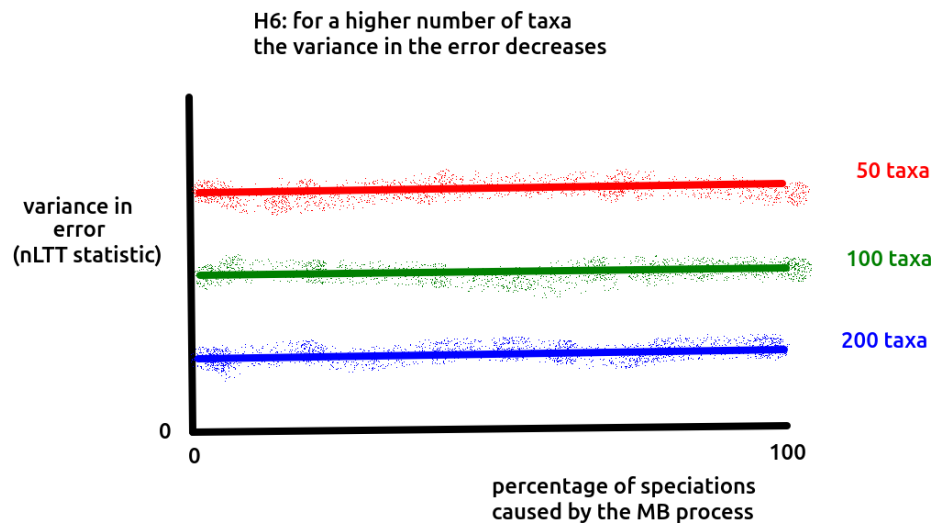


Figure 5: Hypothesis 6: for a higher number of taxa the variance in the error decreases

- 109 • From each MBD tree, a DNA sequence alignment is simulated. For each

110 sequence alignment we then perform a Bayesian analysis to recover a poste-

111 rior distribution of trees, each composed of  $N_P$  phylogenies. Such analysis

112 is performed using the 'pirouette' package (Bilderbeek 2018) to call the

113 BEAST2 tool suite from R. We let the Bayesian analysis assume a BD

114 prior in both cases, to investigate the extent of the error we make under

115 this assumption.
- 116 • For each tree generated under the MBD model we aim to generate a

117 "twin" tree under the BD model. With the word "twin" we denote a

118 tree generated starting from the respective MBD tree, in order to perform

119 a fair comparison with it. This operation has to be done, because we

120 want to compare two trees that are generated by different processes. To

121 do so we infer the parameters  $\lambda_{BD}$  and  $\mu_{BD}$  from the MBD maximizing

122 the likelihood under a BD model. To perform this operation we use the

123 function "bd\_ML" from the package "DDD" (Etienne *et al.* 2012).
- 124 • We then exploit such parameters to generate a BD tree using the func-

125 tion "tess.sim.taxa.age" from the package "TESS" (Hhna 2013). We

126 simulate the tree in such a way the new tree has the same number of tips

127 and the same crown age as the MBD tree. We furthermore require that

128 the BD tree conserve the topology of the MBD tree. We have hypothesis

129 H4 that, compared to the MBD trees, the error will be less in the BD

130 twin tree. The difference between the errors made in MBD and twin BD

131 trees indicates the impact the MBD process has on the error we make in

132 inference using a contemporary BD prior.

133 We want the MBD and twin BD trees to contain the same amount of in-

134 formation, i.e. the same number of DNA mutations and the same number

135 of taxa at the present:



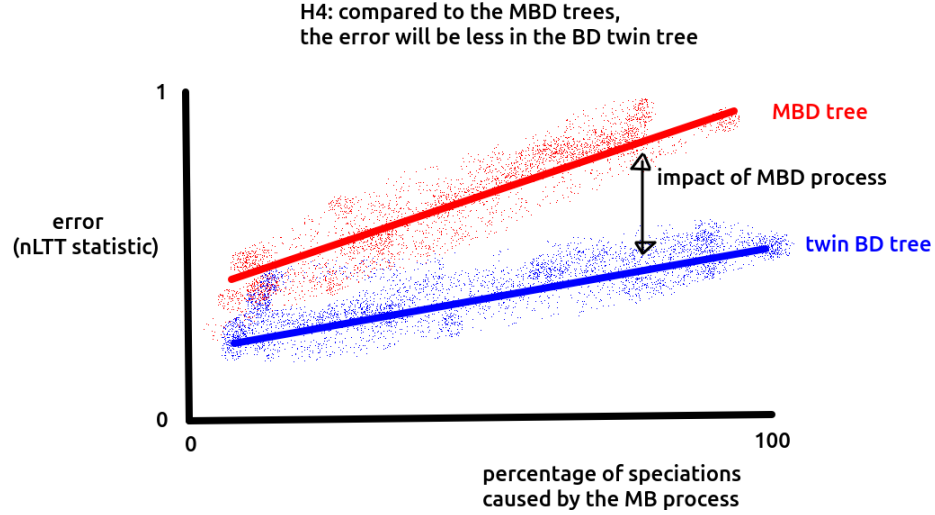


Figure 6: Hypothesis 4: compared to the MBD trees, the error will be less in the BD twin tree

$$m_{MBD} = m_{BD} \quad (2)$$

136 The expected number of mutations  $m$  of a phylogeny with crown age  $-T$   
 137 (with  $T > 0$ ) in fact is given by [RJCB: So one of use likes '-T', the  
 138 other likes 'T'. How to resolve this? ]

$$m = L \cdot \rho \cdot \int_0^T n(t) dt \quad (3)$$

139 where  $L$  is the number of DNA nucleotides,  $\rho$  is the per-site per-species  
 140 mutation rate and  $n(t)$  the number of species at each time.

141 The parameter we'll tune is  $\rho$  ... [RJCB: elaborate here :-)]

142 Since we cannot know  $n_{BD}(t)$  before running simulations we need to re-  
 143 place it with a proxy. For this reason we will use the average number of  
 144 species in time according to the BD model. It's well known that this is  
 145 equal to [GL: insert proper citation]

$$< n_{BD} > (t) = n_0 \cdot e^{(\mu_{BD} - \lambda_{BD})t} \quad (4)$$

where  $n_0 = n_{BD}(-T) = n_{MBD}(-T)$  is the initial number of species at the crown age. From 2, 3 and 4 follows:

$$m_{MBD} = L \cdot \rho \cdot \int_0^T < n_{BD} > (t) dt = L \cdot \rho \cdot n_0 \cdot \left[ \frac{e^{(\mu_{BD} - \lambda_{BD})T} - 1}{\mu_{BD} - \lambda_{BD}} \right] \quad (5)$$

If we set  $\mu_{BD} = \mu_{MBD}$  and reverse this relation we can extrapolate the value of  $\lambda_{BD}$  to use to generate BD trees.

- We explained how we set the parameters for each twin BD tree. Using this rules we generate a BD dataset. We repeat the analysis, producing alignments for each tree and subsequently using BEAST to produce a posterior for each of them.

## 2.3 Measuring the inference error

- So far we have simulated two datasets of trees under the two models:  $\{T_i^{BD}\}_{i=1}^{N_S}$  and  $\{T_i^{MBD}\}_{i=1}^{N_S}$ . We used them to generate a dataset of alignments for each model:  $\{X_i^{BD}\}_{i=1}^{N_S}$  and  $\{X_i^{MBD}\}_{i=1}^{N_S}$ . From each dataset we produced a posterior distribution from a BD prior:  $P_i(\theta|X_i^{BD}, BD)$  and  $P_i(\theta|X_i^{MBD}, BD)$ . **[GL: 1) We might want to rename the models, e.g. BD = (0) and MBD = (1). These names with capital letters are too big and ugly; ] [RJCB: I would strongly prefer MBD and BD, as I feel replacing the big ugly capital letters by short pretty numbers hurts readability even more ]**
- To compare the results for the two models we measure the inference error

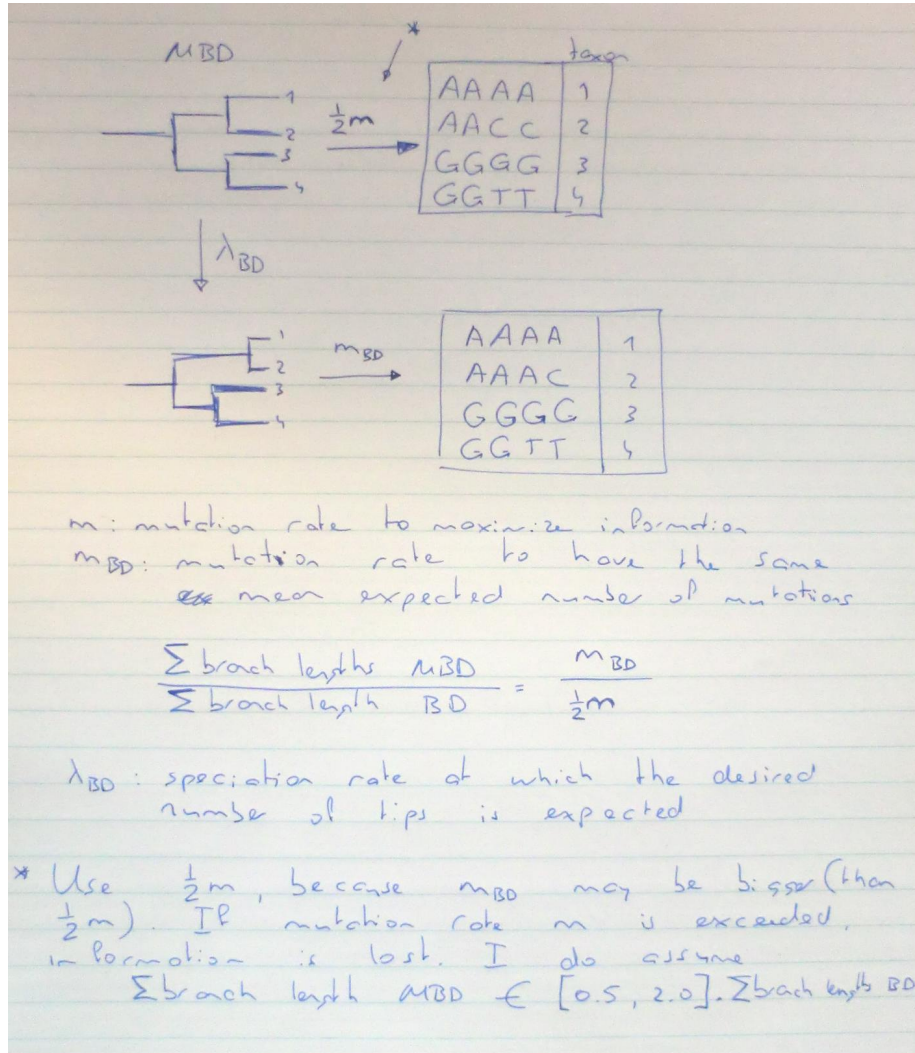


Figure 7: How to create twin trees and alignments. From a focal MBD tree, a twin tree is produced as such: (1) estimate the  $\lambda_{BD}$  to get the same expected number of tips, (2) simulate a BD tree with that amount of tips (discard trees with different number of tips), (3) estimate a mutation rate to get an alignment with the same expected number of mutations, (4) simulate alignments with that amount of mutations (discard those that don't, the picture shows an alignment that should be discarded)

using the nLTT statistic between known/true tree and posterior/inferred trees (Janzen 2015). To obtain such statistics the procedure is the following:

- From each tree  $T_{i,j}^M$  (with  $j = 1, \dots, N_S$ ) belonging to the posterior  $P_i(\theta|X_i^M, BD)$  and relative to the model  $M$ , we extrapolate the lineage-through-time (LTT), in other words we measure the number of species as a function of time  $n_{i,j}(t)$ . To allow a comparison we normalize dividing by the maximum number of species of each tree, i.e. the number of tips at the present  $N_{i,j}(t) = \frac{n_{i,j}(t)}{n_{i,j}^{max}}$ . We then define the nLTT measure as

$$nLTT_{i,j} = \int_0^T |N_{i,j}(t) - N_{T_i}| dt$$

[GL: I am running out of letters :(] [RJCB: Haha! I suggest to use the same equation and symbols as equation 1 in the nLTT article of Janzen, Hoehna and Etienne, 2015: ]

$$\Delta nLTT = \int_0^1 |nLTT_1(t) - nLTT_2(t)| dt$$

## 2.4 Model selection

We simulate alignments using the simplest nucleotide substitution model (JC69), the simplest clock model (strict). It is thus imminent to assume these models in our Bayesian inference. Nevertheless, the phylogeny the alignment was based on, could have followed either an MBD or BD tree model, where we in both cases assume a BD tree model. This will have an unknown effect on our inference: it may theoretically be that an MBD model generates (a tree that generates) an alignment in which a different site and/or clock model is favored.

We investigate this by measuring if the generative model (with the simplest

nucleotide substitution and simplest clock model) is indeed selected to be the best fitting model. To be precise, we look at the model with the highest marginal likelihood (also called evidence MacKay & Mac Kay 2003),  $f(D|M)$ , which is the probability of the data D given model M. In the context of this research, D consists of the DNA alignment, and M is the combination of site, clock and tree models.

To estimate the marginal likelihood, we use an algorithm named nested sampling Skilling *et al.* 2006. Nested sampling is attractive to use in a phylogentic context, as it gives a good estimation, requires little tuning Russel *et al.* 2018. Nested sampling is available as a BEAST2 package and can be used by babette Bilderbeek & Etienne 2018.

The nested sampling algorithm stops its run when the marginal likelihood estimation error reaches below a certain tolerance. Similar to Russel *et al.* 2018, we use a (relative) error tolerance  $\epsilon$  of  $10^{-13}$ , 1 particle to explore the parameter space and 100 active points. To achieve the latter, we use the MCMC chain length  $L_c$  of 1M (as also used in the parameter estimates), and a sub-chain length  $L_{sc}$  of 10K.

The models we use in our model comparison are the four combinations of two site models and two clock models. We use the JC69 site model, which is the (generative and) simplest model and GTR, the site model with most degrees of freedom. For the clock models, we use the strict clock model, which is the (generative and) simplest clock model, and the RLN clock model.

From these four marginal likelihood estimates, we calculate the weight of the generative model and plot this in figure 2. We do this for both the alignments derived from the MBD tree and the BD twin tree. We expect that the generative model has the heighest weight in both the MBD and

215 BD alignments. We expect this weight to be higher in the BD alignments.

## 216 3 Results

217 •

218 •

## 219 References

- 220 Bilderbeek, R.J. (2018) *pirouette: create a posterior from a phylogeny*.
- 221 Bilderbeek, R.J. & Etienne, R.S. (2018) babette: Beauti 2, beast 2 and tracer  
222 for r. *Methods in Ecology and Evolution*.
- 223 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard,  
224 M.A., Rambaut, A. & Drummond, A.J. (2014) Beast 2: a software platform  
225 for bayesian evolutionary analysis. *PLoS computational biology*, **10**, e1003537.
- 226 Drummond, A.J. & Rambaut, A. (2007) Beast: Bayesian evolutionary analysis  
227 by sampling trees. *BMC evolutionary biology*, **7**, 214.
- 228 Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A.  
229 & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies  
230 closer to agreement with the fossil record. *Proc R Soc Lond B: Biol Sci*, **279**,  
231 1300–1309.
- 232 Etienne, R.S., Morlon, H. & Lambert, A. (2014) Estimating the duration of  
233 speciation from phylogenies. *Evolution*, **68**, 2430–2440.
- 234 Hhna, S. (2013) Fast simulation of reconstructed phylogenies under global time-  
235 dependent birth–death processes. *Bioinformatics*, **29**, 1367–1374.

236 Janzen, T. (2015) *nLTT: Calculate the NLTT Statistic*. R package version 1.1.

237 Janzen, T., Alzate, A., Muschick, M., Maan, M.E., van der Plas, F. & Etienne,  
 238 R.S. (2017) Community assembly in lake tanganyika cichlid fish: quantifying  
 239 the contributions of both niche-based and neutral processes. *Ecology and*  
 240 *Evolution*, **7**, 1057–1067.

241 Janzen, T., Alzate, A., Muschick, M., van der Plas, F. & Etienne, R.S. (2016)  
 242 Stochastic processes dominate community assembly in cichlid communities in  
 243 lake tanganyika.

244 Laudanno, G. (2018) *MBD: Multiple Birth Death Diversification*. R package  
 245 version 0.1.

246 MacKay, D.J. & Mac Kay, D.J. (2003) *Information theory, inference and learn-*  
 247 *ing algorithms*. Cambridge university press.

248 Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary char-  
 249 acter’s effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.

250 Russel, P.M., Brewer, B.J., Klaere, S. & Bouckaert, R.R. (2018) Model selection  
 251 and parameter inference in phylogenetics using nested sampling. *Systematic*  
 252 *Biology*, p. syy050.

253 Skilling, J. *et al.* (2006) Nested sampling for general bayesian computation.  
 254 *Bayesian analysis*, **1**, 833–859.

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

	Description	Value(s)
$L_c$	MCMC chain length	$10^6$
$L_{sc}$	MCMC sub-chain length	$10^4$
$\epsilon$	relative error tolerance in marginal likelihood estimation	$10^{-13}$

Table 1: Overview of the simulation parameters.