

1 The error in Bayesian phylogenetic reconstruction
2 when speciation is not instantaneous

3 Richèl J.C. Bilderbeek¹ and Rampal S. Etienne¹

4 ¹Groningen Institute for Evolutionary Life Sciences, University of
5 Groningen, Groningen, The Netherlands

6 May 23, 2018

7 **Abstract**

8 The tools for reconstructing phylogenetic relationships between taxo-
9 nomic units (e.g. species) have become very advanced in the last three
10 decades.

11 Among the most popular tools are Bayesian approaches, such as BEAST,
12 MrBayes and RevBayes, that use efficient tree sampling routines to create
13 a posterior probability distribution of the phylogenetic tree. A feature of
14 these approaches is the possibility to incorporate known or hypothesized
15 structure of the phylogenetic tree through the tree prior. It has been
16 shown that the effect of the prior on the posterior distribution of trees
17 can be substantial.

18 Currently implemented tree priors assume that speciation is instantane-
19 ous, where we know that speciation can be a gradual process.

20 Here we explore the effects of ignoring the protractedness of the spe-
21 ciation process with an extensive simulation study.

22 We compare the inferred tree to the simulated tree, and find that

23 **Keywords:** computational biology, evolution, phylogenetics, prior choice

24 1 **Introduction**

25 The computational tools that are currently available to the phylogeneticists
26 go beyond the wildest imagination of those living four decades ago. Advances
27 in computational power allowed the first cladograms to be inferred from DNA
28 alignments in 1981 (Felsenstein 1981), and the first Bayesian tools emerged in
29 1996 (Rannala & Yang 1996), providing unprecedented flexibility in the setup
30 of a phylogenetic model.

31 Currently, the most popular Bayesian phylogenetics tools are BEAST (Drum-
32 mond & Rambaut 2007) and its successor BEAST2 (Bouckaert *et al.* 2014),
33 MrBayes (Huelsenbeck & Ronquist 2001) and RevBayes (Höhna *et al.* 2016).
34 They allow to incorporate known or hypothesized structure of a phylogenetic
35 tree-to-be-inferred through model priors. From these priors and an alignment
36 of DNA, RNA or protein sequences, they create a posterior distribution of pa-
37 rameter estimates (of the models used as a prior) and phylogenies, in which
38 more probable combinations are represented more often. Each of these tools
39 use efficient tree sampling routines to rapidly create an informative posterior.

40 The model priors in Bayesian phylogenetic reconstruction can be grouped
41 into three categories: (1) site model, specifying nucleotide substitutions, (2)
42 clock model, specifying the rate of mutation per lineage in time, and (3) tree
43 model, constituting the speciation model underlying branching events (specia-
44 tion) and branch termination (extinction). The choice of a wrong site model
45 (Posada & Buckley 2004), clock model (Baele *et al.* 2012) or tree prior (Möller
46 *et al.* 2018; Yang & Rannala 2005) is known to affect the posterior.

47 Current phylogenetic tools use tree priors that assume speciation is instan-
48 taneous, whilst we know that, speciation is often a gradual process (Schluter

2009). The (constant-rate) birth-death (BD) model is a commonly used tree prior, but it ignores this temporal aspect of speciation. The protracted birth-death (PBD) model, an extension of the BD model, does incorporate the idea that speciation takes time. In this model, a branching event does not give rise to a new species, but to a new species-to-be, called an incipient species. Such an incipient species may go extinct, finish its speciation to become a good species, or give rise to new incipient species. Protracted speciation may explain observed declines in lineage accumulation (Etienne & Rosindell 2012).

Unfortunately, a tree prior according to this model, providing the probability of a species tree under the PBD model, is unavailable in current Bayesian phylogenetic tools. Whilst an approximate formula for this probability has been derived (Lambert *et al.* 2015) and the approximation is very good (Simonet *et al.* 2018), it has not been implemented as tree prior yet. There are various reasons for this. First, the computation of this probability involves solving a set of non-linear differential equations, and while this computation is quite fast, it still takes much more time than the corresponding probability of the BD model which is a simple analytical formula. In a Bayesian MCMC chain, the tree prior probability must be calculated many times, and hence the total computation will take considerably longer with a PBD tree prior. Furthermore, the approximate probability is a probability for the species tree assuming an underlying incipient species tree. It can be safely used as tree prior when only one individual per species is sampled, but if one has multiple samples per species -which is currently often the case- the methods to account for this such as the multi-species coalescent (Heled & Drummond 2009) may not be compatible with the underlying incipient species tree. More precisely, the phylogeny under the PBD model may contain paraphylies, while the multi-species coalescent was developed exactly to avoid these by explaining them as arising from incomplete

lineage sorting. Because of these paraphylies there is no such thing as a true species tree in the PBD model. To get a species-level tree one must sample one incipient species per species. Which incipient species is sampled may therefore have an impact on the species tree.

Here we aim to explore the effect of using the BD prior on PBD simulated phylogenies, taking into account possible sampling effects. In brief, we simulate protracted phylogenies using the PBD process, from which we sample a species tree in two very different ways. Given this species tree, we simulate a DNA sequence alignment. Then, we use BEAST2 on these alignments to infer a posterior of phylogenies, using a BD prior. We quantify the difference between the (BD) posterior phylogenies and the simulated (PBD) species tree.

2 Methods (but we are not allowed to keep this header)

The PBD model has five biological parameters (see 2), which we explore in a factorial fashion, excluding some combinations. We only simulate a PBD process for phylogenies in which speciation initiation exceeds extinction rate ($b_i > \mu_i$ and $b_g > \mu_g$), and in which the expected number of extant good species is less than 1000. **[NOTE: accoring to Rampal this has been solved analytically (for the Protracted Birth and Death model). Where?]**. We use 1000 good species as a threshold, to prevent overly taxon-poor and taxon-rich phylogenies respectively. The parameter values chosen are based on the parameter sets used by Etienne *et al.* 2014, as these parameters were shown to result in reasonably sized phylogenies and using the same set allows us to compare results. For the speciation initiation rates of good and incipient species, b_g and b_i respectively, we use 0.1, 0.5 and 1.0 speciation initiation events per

101 good/incipient species per time unit. The speciation completion rates we use
 102 are 0.1, 0.3, 1.0 and 10^9 speciation completion events per (incipient species)
 103 species per time unit. We use $10^9 \approx \infty$ to mimic the BD model, because the
 104 PBD model reduces to the BD model for $\lambda = \infty$. This allows us to measure
 105 the baseline error, which is the difference between inferred tree and true species
 106 tree that arises purely due to noise because the generating model and the model
 107 used in inference are identical in this case. The extinction rates of good and
 108 incipient species, μ_g and μ_i respectively, that we use are 0.0, 0.1, 0.2 and 0.4
 109 extinction events per good/incipient species per time unit.

110 From each biological parameter set, we simulate a protracted birth-death
 111 tree, using the PBD package (Etienne 2015) in the R programming language
 112 (R Core Team 2013), all with a crown age of 15 million years. Each protracted
 113 birth-death tree uses a different random number generator seed, which makes all
 114 runs independent, resulting in a balanced data set. **[NOTE: maybe discuss**
 115 **with Rampal to remove confusion. Rampal requests a histogram of**
 116 **branch lengths to follow a geometric distribution]**

117 From each incipient species tree, we construct a species tree, by sampling one
 118 incipient/good species per good species. For example, when an incipient species
 119 branched off from its mother lineage, both of these subspecies are recognized
 120 as representing the species, and hence both can be picked as an (equally good)
 121 representative of the species. Here, we use three sampling scenarios, in which
 122 we pick the representative randomly or in such a way that this results in either
 123 the shortest or longest branch lengths. See the supplementary information for
 124 a visualization of these sampling methods.

125 Based on the sampled species tree, we simulate a DNA alignment that has
 126 the same history as this species tree, using the **phangorn** package (Schliep 2011).
 127 We assume that the nucleotides of the DNA alignment follow a Jukes-Cantor

128 (Cantor & Jukes 1969) nucleotide substitution model, in which all nucleotide-to-
 129 nucleotide transitions are equally likely. In our Bayesian inference (see below)
 130 we use the same site model as the (obviously correct) site model prior. One
 131 could explore other substitution models in the simulations and in the Bayesian
 132 inference, but we chose this simple model because we are primarily interested
 133 in the effect of the choice of tree prior. If anything, our results are conservative:
 134 with a more complex substitution model, there will be more noise and hence our
 135 inference error will increase. We set the mutation rate in such a way to maximize
 136 the information contained in the alignment. To do so, we set the mutation rate
 137 such that we expect on average one (possibly silent) mutation per nucleotide
 138 between crown age and present, which equates to $\frac{1}{15}$ mutations per million years.
 139 The DNA sequence length is chosen to provide a resolution of 10^3 years, that is,
 140 to have one expected nucleotide change per 10^3 years per lineage on average. As
 141 one nucleotide is expected to have on average one (possibly silent) mutation per
 142 15 million years, $15 \cdot 10^3$ nucleotides result in 1 mutation per alignment per 10^3
 143 years (which is coincidentally the same as Möller *et al.* 2018). The simulation
 144 of these DNA alignments follows a strict clock model, which we will specify as
 145 the known clock model prior in the Bayesian inference.

146 From an alignment, we run a Bayesian analysis and create a posterior dis-
 147 tribution of trees and parameters using the **babette** (?) package that sets the
 148 input parameters similar to BEAUti 2 and then runs BEAST2. For our site
 149 and clock model, we assume a Jukes-Cantor nucleotide substitution model and
 150 strict clock model, as those are also used for simulating that alignment. We set
 151 the BD model as a tree prior, as gauging the effect of this incorrect assump-
 152 tion is the goal of this study. We assume an MRCA prior with a tight normal
 153 distribution around the crown age, by choosing the crown age as mean, and a
 154 standard deviation of $0.5 \cdot 10^{-3}$ time units, resulting in 95% of the crown ages

155 inferred have the same resolution (of 10^{-3} time units) as the alignment. We
 156 ran the MCMC chain to generate 1111 states, of which we remove the first 10%
 157 (also called the 'burn-in'). Of the remaining 1000 MCMC states, the effective
 158 sample size (ESS) of the posterior [**NOTE: there is a parameter estimate**
 159 **called 'posterior'. I choose to pick that one, and I assume it is the**
 160 **wisser choice over 'prior' and parameter estimates. Must discuss]** must
 161 at least be 200 for a strong enough inference (Drummond & Bouckaert 2015).
 162 An ESS can be increased by increasing the number of samples or decreasing
 163 the autocorrelation between samples. If the ESS is less than 200, we decrease
 164 autocorrelation by doubling the MCMC sampling interval of that simulation,
 165 until the ESS exceeds 200.

166 We compare each posterior phylogeny to the (sampled) species tree by the
 167 nLTT statistic (Janzen *et al.* 2015), using the nLTT package (Janzen 2015). The
 168 nLTT statistic equals the area between the normalized lineages-through-time-
 169 plots of two phylogenies, which has a range from zero (for identical phylogenies)
 170 to one. We use inference error and nLTT statistic interchangeably. Compar-
 171 ing the simulated species tree with each of the posterior species trees yields a
 172 distribution of nLTT statistics.

173 We produce two data sets as a comma-seperated file. We set the number
 174 of replicates for each parameter combination such, that this file and a possible
 175 copy can be handled in R's memory. Each row will then contain a parameter set
 176 and the generated nLTT statistics (see 3 for the exact data specification). The
 177 abovementioned memory constraints allows for $2 \cdot 10^3$ rows. With 48 [**NOTE:**
 178 **recalculate]** combinations of biological parameter, there will be 168 [**NOTE:**
 179 **recalculate]** replicates per parameter set.

180 For both data sets, we plot the nLTT statistics distribution per parameter set
 181 using a violin plot, as such a plot maintains information about the distribution.

182 To simplify the interpretation of these plots, only nLTT statistics distribution
 183 are shown for $\lambda_g = \lambda_i$ and $\mu_g = \mu_i$.

184 3 Results

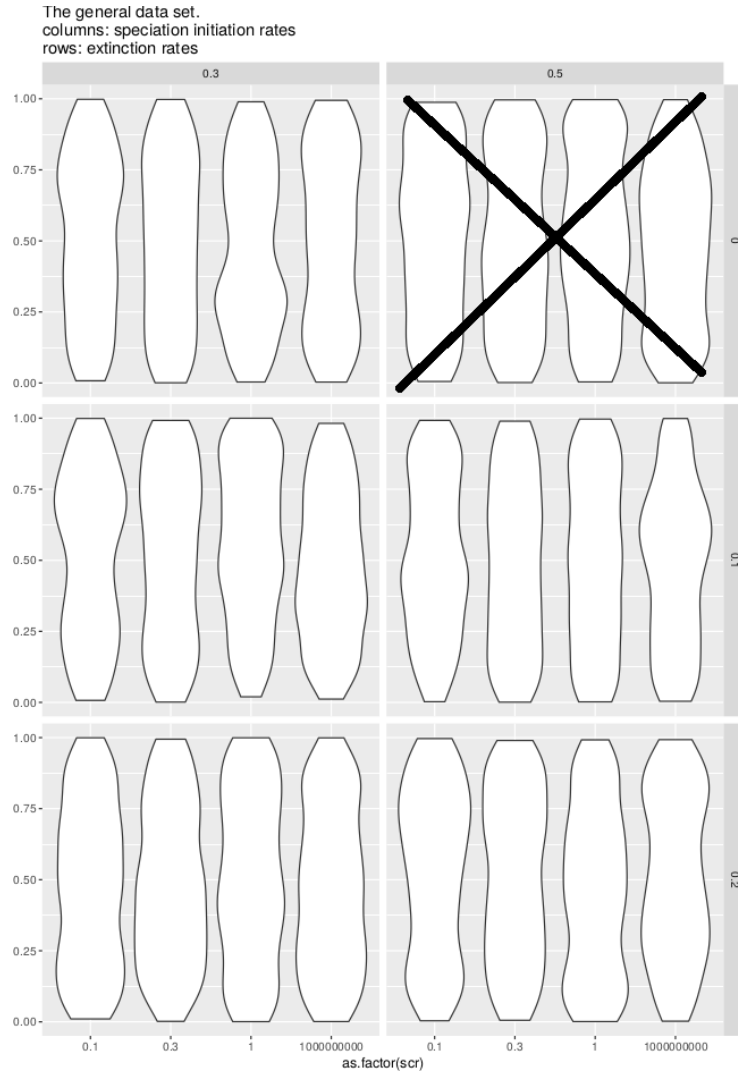


Figure 1: nLTT statistic distribution per biological parameter set, using the balanced data set

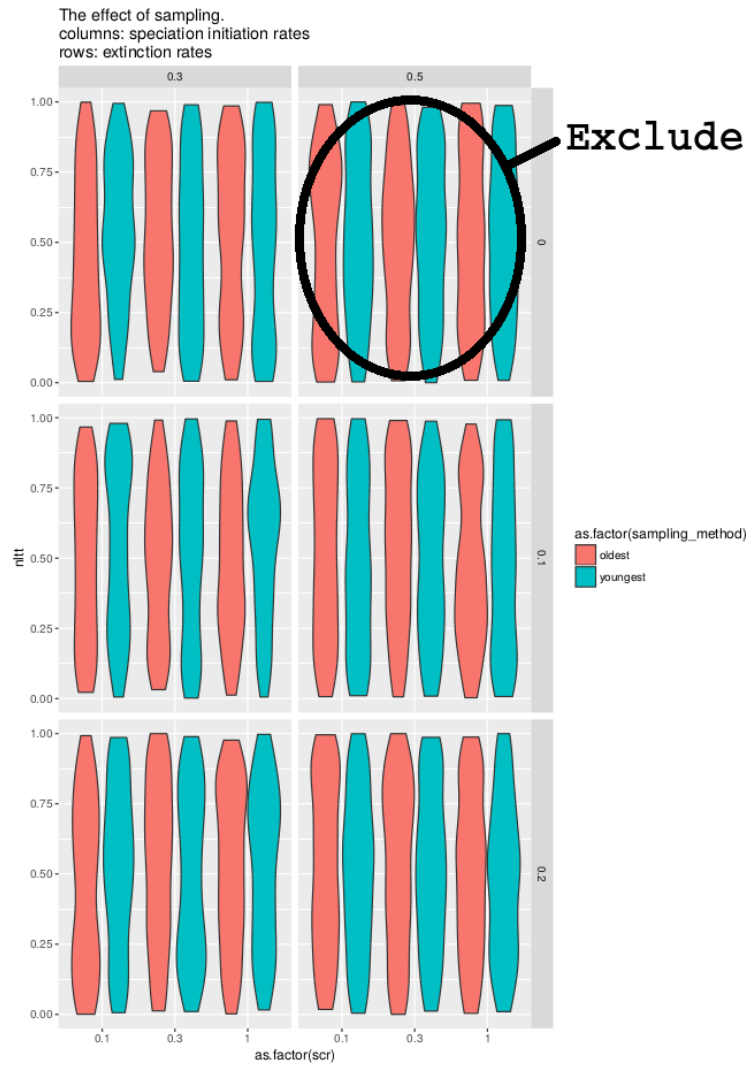


Figure 2: nLTT statistic distribution per biological parameter set per sampling regime, using the data set conditioned on sampling regime having an effect

Term	Definition
Phylogenetics	The inference of evolutionary relationships of groups of organisms using genetics
Model prior	Knowledge or assumptions about the ontogeny of evolutionary histories
Posterior	A collection of phylogenies and parameter estimates, in which more probable combinations (determined by the data and the model prior) are presented more frequently
Protracted speciation	The process in which speciation takes two events to complete: a speciation-initiation event and a speciation-completion event
Speciation initiation	The start of a speciation event creating an incipient species
Speciation completion	The end of a speciation event, in which an incipient species is recognized as a good species

Table 1: Glossary [NOTE: this is requested by the journal]

185 4 Glossary

186 5 Acknowledgements

187 [NOTE: journal does not request for this. Suggest to remove, but how
188 to acknowledge Peregrine otherwise?] We would like to thank the Center
189 for Information Technology of the University of Groningen for their support and
190 for providing access to the Peregrine high performance computing cluster.

191 6 Authors' contributions

192 [NOTE: journal does not request for this] RSE conceived the idea for this
193 experiment. RJCB created and tested the experiment, and wrote the first draft
194 of the manuscript. RSE contributed substantially to revisions.

195 References

196 Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A. & Lemey, P. (2012) Ac-
197 curate model selection of relaxed molecular clocks in bayesian phylogenetics.

198 *Molecular biology and evolution*, **30**, 239–243.

199 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard,
200 M.A., Rambaut, A. & Drummond, A.J. (2014) Beast 2: a software platform
201 for bayesian evolutionary analysis. *PLoS computational biology*, **10**, e1003537.

202 Cantor, J. & Jukes, T. (1969) Mammalian protein metabolism. *Evolution of*
203 *protein molecules Academic Press, New York, NY*, pp. 21–132.

204 Drummond, A.J. & Bouckaert, R.R. (2015) *Bayesian evolutionary analysis with*
205 *BEAST*. Cambridge University Press.

206 Drummond, A.J. & Rambaut, A. (2007) Beast: Bayesian evolutionary analysis
207 by sampling trees. *BMC evolutionary biology*, **7**, 214.

208 Etienne, R.S. (2015) *PBD: Protracted Birth-Death Model of Diversification*. R
209 package version 1.1.

210 Etienne, R.S., Morlon, H. & Lambert, A. (2014) Estimating the duration of
211 speciation from phylogenies. *Evolution*, **68**, 2430–2440.

212 Etienne, R.S. & Rosindell, J. (2012) Prolonging the past counteracts the pull of
213 the present: protracted speciation can explain observed slowdowns in diver-
214 sification. *Systematic Biology*, **61**, 204–213.

215 Felsenstein, J. (1981) Evolutionary trees from dna sequences: a maximum like-
216 lihood approach. *Journal of molecular evolution*, **17**, 368–376.

217 Heled, J. & Drummond, A.J. (2009) Bayesian inference of species trees from
218 multilocus data. *Molecular biology and evolution*, **27**, 570–580.

219 Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R.,
220 Huelsenbeck, J.P. & Ronquist, F. (2016) Revbayes: Bayesian phylogenetic

inference using graphical models and an interactive model-specification language. *Systematic biology*, **65**, 726–736.

Huelsenbeck, J.P. & Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

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

Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nltt. *Methods in Ecology and Evolution*, **6**, 566–575.

Lambert, A., Morlon, H. & Etienne, R.S. (2015) The reconstructed tree in the lineage-based model of protracted speciation. *Journal of mathematical biology*, **70**, 367–397.

Möller, S., du Plessis, L. & Stadler, T. (2018) Impact of the tree prior on estimating clock rates during epidemic outbreaks. *Proceedings of the National Academy of Sciences*, p. 201713314.

Posada, D. & Buckley, T.R. (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, **53**, 793–808.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rannala, B. & Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, **43**, 304–311.

Schliep, K. (2011) phangorn: phylogenetic analysis in r. *Bioinformatics*, **27**, 592–593.

Parameter	Description	Values
b_g	Speciation initiation rate of a good species	0.1, 0.5, 1.0
b_i	Speciation initiation rate of an incipient species	0.1, 0.5, 1.0
λ	Speciation completion rate	0.1, 0.3, 1.0, ∞
μ_g	Extinction rate of a good species	0.0, 0.1, 0.2, 0.4
μ_i	Extinction rate of an incipient species	0.0, 0.1, 0.2, 0.4
t_c	Crown age	15
σ_c	Standard deviation around crown age	0.001
M	Sampling method	'shortest', 'longest' or random
r	Mutation rate	$\frac{1}{15}$
l_a	DNA alignment length	15K
f_i	MCMC sampling interval	1K or more
R_i	RNG seed incipient tree and randomly sampled species tree	1 to 20K
R_a	RNG seed alignment simulation	R_i
R_b	RNG seed BEAST2	R_i

Table 2: Overview of the 12 simulation parameters. Above the horizontal line is the biological parameter set. Sampling method M is random for the general data set. For the data set exploring the effect of sampling, we use 'shortest' for odd values of R_i , and 'longest' for even values of R_i . R_i is 1 for the first simulation, 2 for the next, etcetera.

n	Description
12	simulation parameters, see table 2
1000	nLTT statistic values
11	ESSes of all parameters estimated by BEAST2 (see specs below)

Table 3: Specification of the data sets. Each row will contain one experiment, where the columns contain parameters, measurements and diagnostics. This table displays the content of the columns. n denotes the number of columns a certain item will occupy, resulting in a table of 1023 columns and 20K rows.

- 246 Schluter, D. (2009) Evidence for ecological speciation and its alternative. *Sci-*
247 *ence*, **323**, 737–741.
- 248 Simonet, C., Scherrer, R., Rego-Costa, A. & Etienne, R. (2018) Robustness of
249 the approximate likelihood of the protracted speciation model. *Journal of*
250 *evolutionary biology*, **31**, 469–479.
- 251 Yang, Z. & Rannala, B. (2005) Branch-length prior influences bayesian posterior
252 probability of phylogeny. *Systematic Biology*, **54**, 455–470.

#	Description
1	posterior
2	likelihood
3	prior
4	treeLikelihood
5	TreeHeight
6	BirthDeath
7	BDBirthRate
8	BDDeathRate
9	logP.mrca
10	mrcatime
11	clockRate

Table 4: Overview of the 11 BEAST2 estimated parameters