

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 September 5, 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. Among the most popular tools are Bayesian approaches, such as
11 BEAST, MrBayes and RevBayes, that use efficient tree sampling routines
12 to create a posterior probability distribution of the phylogenetic tree. A
13 feature of these approaches is the possibility to incorporate known or hy-
14 pothesized structure of the phylogenetic tree through the tree prior. It
15 has been shown that the effect of the prior on the posterior distribution
16 of trees can be substantial.

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

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

21 We compare the inferred tree to the simulated tree, and find that ...

22 **Keywords:** computational biology, evolution, phylogenetics, Bayesian anal-
23 ysis, tree prior

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

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

48 Current phylogenetic tools use tree priors that assume speciation is instan-
49 taneous, whilst we know that, speciation is often a gradual process (Schluter
50 2009). The (constant-rate) birth-death (BD) model is a commonly used tree
51 prior, but it ignores this temporal aspect of speciation. The protracted birth-
52 death (PBD) model, an extension of the BD model, does incorporate the idea
53 that speciation takes time. In this model, a branching event does not give rise
54 to a new species, but to a new species-to-be, called an incipient species. Such an
55 incipient species may go extinct, finish its speciation to become a good species,
56 or give rise to new incipient species. Protracted speciation may explain observed
57 declines in lineage accumulation (Etienne & Rosindell 2012).

58 Unfortunately, a tree prior according to this model, providing the probability
59 of a species tree under the PBD model, is unavailable in current Bayesian phy-
60 logenetic tools. Whilst an approximate formula for this probability has been
61 derived (Lambert *et al.* 2015) and the approximation is very good (Simonet
62 *et al.* 2018), it has not been implemented as tree prior yet. There are various
63 reasons for this. First, the computation of this probability involves solving a set
64 of non-linear differential equations, and while this computation is quite fast, it
65 still takes much more time than the corresponding probability of the BD model
66 which is a simple analytical formula. In a Bayesian MCMC chain, the tree
67 prior probability must be calculated many times, and hence the total compu-
68 tation will take considerably longer with a PBD tree prior. Furthermore, the
69 approximate probability is a probability for the species tree assuming an under-
70 lying incipient species tree. It can be safely used as tree prior when only one
71 individual per species is sampled, but if one has multiple samples per species
72 - which is currently often the case - the methods to account for this such as
73 the multi-species coalescent (Heled & Drummond 2009) may not be compatible
74 with the underlying incipient species tree. More precisely, the phylogeny under

75 the PBD model may contain paraphylies, while the multi-species coalescent was
76 developed exactly to avoid these by explaining them as arising from incomplete
77 lineage sorting. Because of these paraphylies there is no such thing as a true
78 species tree in the PBD model. To get a species-level tree one must sample one
79 incipient species per species. Which incipient species is sampled may therefore
80 have an impact on the species tree.

81 Here we aim to explore the effect of using the BD prior on PBD simulated
82 phylogenies, taking into account possible sampling effects. In brief, we simulate
83 protracted phylogenies using the PBD process, from which we sample a species
84 tree in two very different ways. Given this species tree, we simulate a DNA
85 sequence alignment. Then, we use BEAST2 on these alignments to infer a pos-
86 terior of phylogenies, using a BD prior. We quantify the difference between the
87 (BD) posterior phylogenies and the simulated (PBD) species tree. Furthermore,
88 while we evidently know the clock and site models used in the simulation, us-
89 ing a different clock and/or site model prior in inference may compensate or
90 increase this difference between inferred and simulated tree. To study this, we
91 also explore the effect of a different clock and site model prior in inference.

92 **[RJCB: Start new]** The PBD model has five parameters, depicted in table
93 2. The speciation completion rates λ we use are 0.1, 0.3, 1.0 and 10^9 probability
94 of occurrence per time unit. **[RJCB: Is the unit correct unit now?]** The
95 extinction rates $\mu = \mu_g = \mu_i$ we use are 0.0, 0.1 and 0.2 probability of occurrence
96 per time unit. We use expected mean tree sizes n of 50, 100 and 200 good taxa.
97 From each combination of λ , μ and n , we derive a speciation initiation rate
98 $b = b_i = b_g$, shown in table 3. Our parameters are inspired on existing work
99 Etienne & Rosindell 2012 Etienne *et al.* 2014. We use $\lambda = 10^9 \approx \infty$ to let the
100 PBD model reduce to the BD model.

101 We simulate protracted birth-death trees, using the PBD package (Etienne

2015) in the R programming language (R Core Team 2013). The first tree has a random number generator seed of 1, which is incremented by 1 for each simulated tree. For each combination of λ, μ, b, n , we generate incipient species trees with a crown age of 15 million years. Only trees with the desired number of good taxa are kept.

This research creates two data sets: a general one, to explore parameter space, and one to investigate the effect of sampling incipient species (see below). For the general data set, all the trees with the correct number of good species are kept. For the data set to investigate sampling, only trees with the additional constraint of sampling having an effect are kept. As sampling does not have an effect for $\lambda = \infty$, this parameter value is absent in that data set. [RJCB: End new]

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

See the supplementary information for a visualization of these sampling methods. Based on the sampled species tree, we simulate a DNA alignment that has the same history as this species tree, using the **phangorn** package (Schliep 2011). We set the nucleotides of the DNA alignment to follow a Jukes-Cantor (Jukes *et al.* 1969) nucleotide substitution model, in which all nucleotide-to-nucleotide transitions are equally likely. The DNA sequence of the root ancestor consists of four equally sized single-nucleotide blocks of adenine, cytosine, guanine and thymine respectively. For example, for a DNA sequence length of

12, this would be AAACCCGGGTTT. The order of nucleotides does not matter in this study, because we do not consider several partitions of the sequence with their own parameters. Only the frequency of occurrence matters. In our Bayesian inference (see below) we use the same site model as the (obviously correct) site model prior, but we also explore the effect of assuming a more complex site model prior. We predict with the more complex substitution model, that there will be more noise and hence our inference error will increase. On the other hand, we dare not rule out that the inference error will decrease, due to more flexibility in the more complex prior. We set the mutation rate in such a way to maximize the information contained in the alignment. To do so, we set the mutation rate such that we expect on average one (possibly silent) mutation per nucleotide between crown age and present, which equates to $\frac{1}{15}$ mutations per million years. The DNA sequence length is chosen to provide a resolution of 10^3 years, that is, to have one expected nucleotide change per 10^3 years per lineage on average. As one nucleotide is expected to have on average one (possibly silent) mutation per 15 million years, $15 \cdot 10^3$ nucleotides result in 1 mutation per alignment per 10^3 years (which is coincidentally the same as Möller *et al.* 2018). The simulation of these DNA alignments follows a strict clock model, which we will specify as one of the two clock models assumed in the Bayesian inference (see below).

From an alignment, we run a Bayesian analysis and create a posterior distribution of trees and parameters using the **babette** (Bilderbeek & Etienne 2018) package that sets the input parameters similar to BEAUti 2 and then runs BEAST2. For our site model, we assume either a Jukes-Cantor or GTR nucleotide substitution model. The Jukes-Cantor model is the correct one, as it is used for simulating that alignment, where the GTR model is the site model that is picked as a default by most users. For our clock model, we assume either

156 a strict or relaxed log-normal clock model. Also here, the strict clock model
 157 is the correct one, as it is used for simulating the alignment, but the relaxed
 158 log-normal clock model is the one most commonly used. We set the BD model
 159 as a tree prior, as gauging the effect of this incorrect assumption is the goal of
 160 this study. We assume an MRCA prior with a tight normal distribution around
 161 the crown age, by choosing the crown age as mean, and a standard deviation of
 162 $0.5 \cdot 10^{-3}$ time units, resulting in 95% of the crown ages inferred have the same
 163 resolution (of 10^{-3} time units) as the alignment. We ran the MCMC chain to
 164 generate 1111 states, of which we remove the first 10% (also called the 'burn-
 165 in'). Of the remaining 1000 MCMC states, the effective sample size (ESS) of
 166 the posterior must at least be 200 for a strong enough inference (Drummond &
 167 Bouckaert 2015). An ESS can be increased by increasing the number of samples
 168 or decreasing the autocorrelation between samples. If the ESS is less than 200,
 169 we decrease autocorrelation by doubling the MCMC sampling interval of that
 170 simulation, until the ESS exceeds 200.

171 We compare each posterior phylogeny to the (sampled) species tree using the
 172 nLTT statistic (Janzen *et al.* 2015), from the nLTT package (Janzen 2015). The
 173 nLTT statistic equals the area between the normalized lineages-through-time-
 174 plots of two phylogenies, which has a range from zero (for identical phylogenies)
 175 to one. We use inference error and nLTT statistic interchangeably. Comparing
 176 the simulated species tree with each of the posterior species trees yields a dis-
 177 tribution of nLTT statistics. **[RJC: Start new]** The input trees generated
 178 with a $\lambda = 10^9$ allow us to measure the noise of the experiment. For $\lambda = \infty$,
 179 the PBD model that generates the starting trees reduces to a BD model. In the
 180 following steps, sampling will have no effect, BEAST2 will assume the correct
 181 speciation model, and the difference between inferred tree and true species tree
 182 are explained purely due to this experimental noise. **[RJC: End new]**

183 We produce two data sets as a comma-separated file. The general data set
 184 has 144 [RJC: recal] different combinations of biological parameter combi-
 185 nations, site and clock models. The data set to investigate sampling has 552
 186 [RJC: recal] different combinations of biological parameter combinations,
 187 site models, clock models and sampling methods. The experiment is compu-
 188 tationally intensive: pilot experiments show that the experiment takes roughly
 189 100 days of CPU time and 20 days of wall clock time (which includes the queued
 190 waiting for computational resources) per replicate. Due to this, we choose to
 191 perform ten replicates, so that the complete experiment will take an acceptable
 192 time of roughly seven months.

193 For both data sets, we display the nLTT statistics distribution per biological
 194 parameter combination as a violin plot. [RJC: Start new] We only show
 195 the nLTT distributions that were generated under the (correct) assumptions
 196 of a Jukes-Cantor site model and a strict clock model, separated per sampling
 197 method used. [RJC: End new] We display the nLTT statistic distributions
 198 separated per site or clock model in the supplementary information.

199 2 Results

200 3 Glossary

201 References

- 202 Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A. & Lemey, P. (2012) Ac-
 203 curate model selection of relaxed molecular clocks in bayesian phylogenetics.
 204 *Molecular biology and evolution*, **30**, 239–243.
- 205 Bilderbeek, R.J. & Etienne, R.S. (2018) babette: Beauti 2, beast2 and tracer

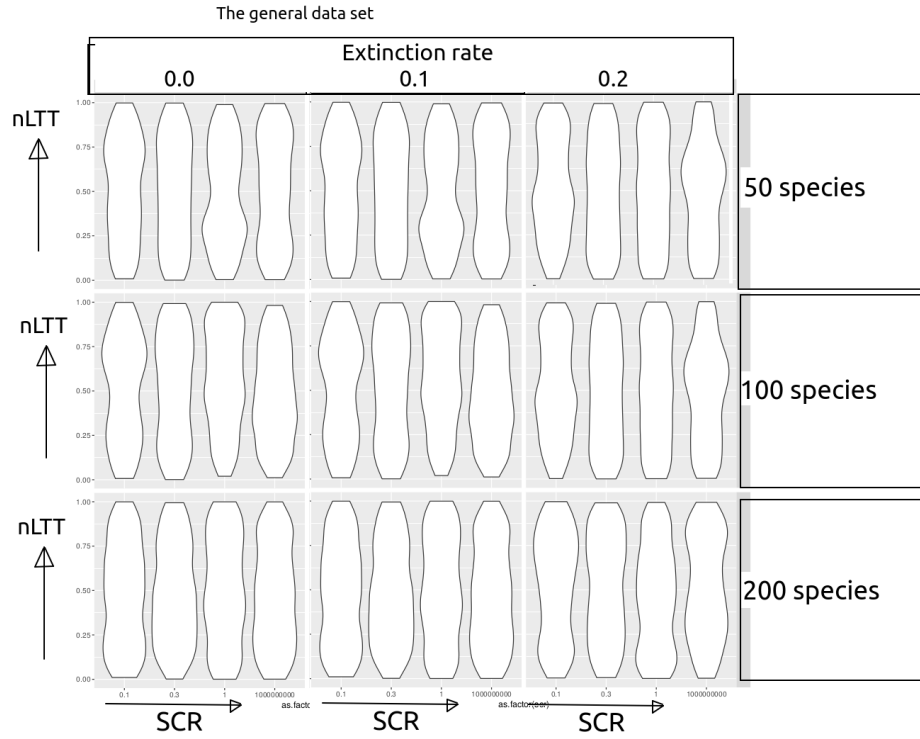


Figure 1: nLTT statistic distribution per biological parameter set, using the general data set, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

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: 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 becomes or is recognized as a good species

Table 1: Glossary

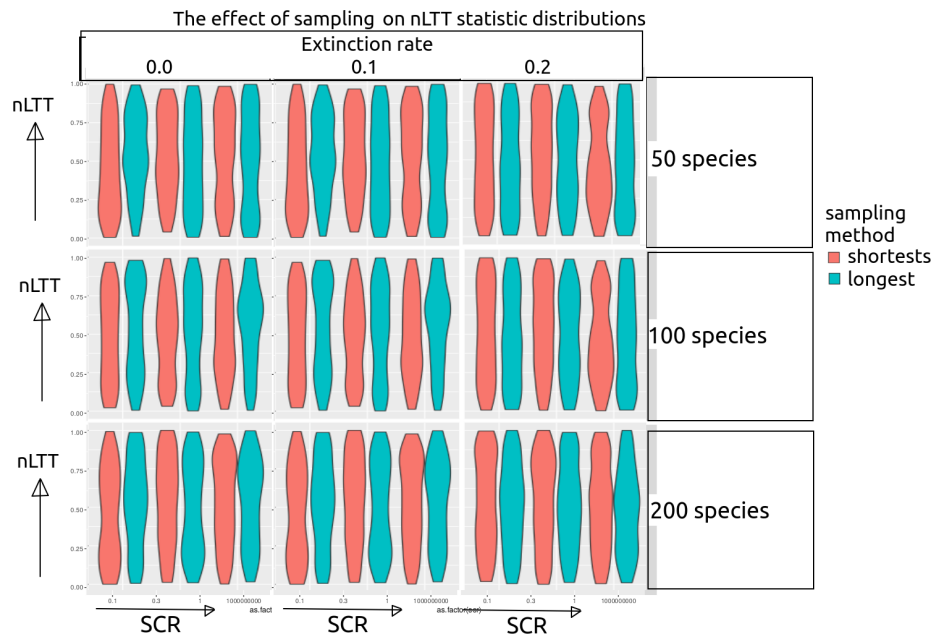


Figure 2: nLTT statistic distribution per biological parameter set per sampling regime, using the data set conditioned on sampling regime having an effect, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

206 for r. *bioRxiv*, p. 271866.

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

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

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

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

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

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

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

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

225 Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R.,
226 Huelsenbeck, J.P. & Ronquist, F. (2016) Revbayes: Bayesian phylogenetic
227 inference using graphical models and an interactive model-specification lan-
228 guage. *Systematic biology*, **65**, 726–736.

229 Huelsenbeck, J.P. & Ronquist, F. (2001) Mrbayes: Bayesian inference of phylo-
230 genetic trees. *Bioinformatics*, **17**, 754–755.

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

232 Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate bayesian compu-
233 tation of diversification rates from molecular phylogenies: introducing a new
234 efficient summary statistic, the nltt. *Methods in Ecology and Evolution*, **6**,
235 566–575.

236 Jukes, T.H., Cantor, C.R. *et al.* (1969) Evolution of protein molecules. *Mam-*
237 *malian protein metabolism*, **3**, 132.

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

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

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

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

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

252 Schliep, K. (2011) phangorn: phylogenetic analysis in r. *Bioinformatics*, **27**,
253 592–593.

- 254 Schluter, D. (2009) Evidence for ecological speciation and its alternative. *Sci-*
 255 *ence*, **323**, 737–741.
- 256 Simonet, C., Scherrer, R., Rego-Costa, A. & Etienne, R. (2018) Robustness of
 257 the approximate likelihood of the protracted speciation model. *Journal of*
 258 *evolutionary biology*, **31**, 469–479.
- 259 Tavaré, S. (1986) Some probabilistic and statistical problems in the analysis of
 260 dna sequences. *Lectures on mathematics in the life sciences*, **17**, 57–86.
- 261 Yang, Z. & Rannala, B. (2005) Branch-length prior influences bayesian posterior
 262 probability of phylogeny. *Systematic Biology*, **54**, 455–470.

263 A Acknowledgements

264 [RJCB: put this section here, as the journal does not request for this]

265 We would like to thank the Center for Information Technology of the University
 266 of Groningen for their support and for providing access to the Peregrine high
 267 performance computing cluster.

268 B Authors' contributions

269 [RJCB: put this section here, as the journal does not request for
 270 this] RSE conceived the idea for this experiment. RJCB created and tested
 271 the experiment, and wrote the first draft of the manuscript. RSE contributed
 272 substantially to revisions.

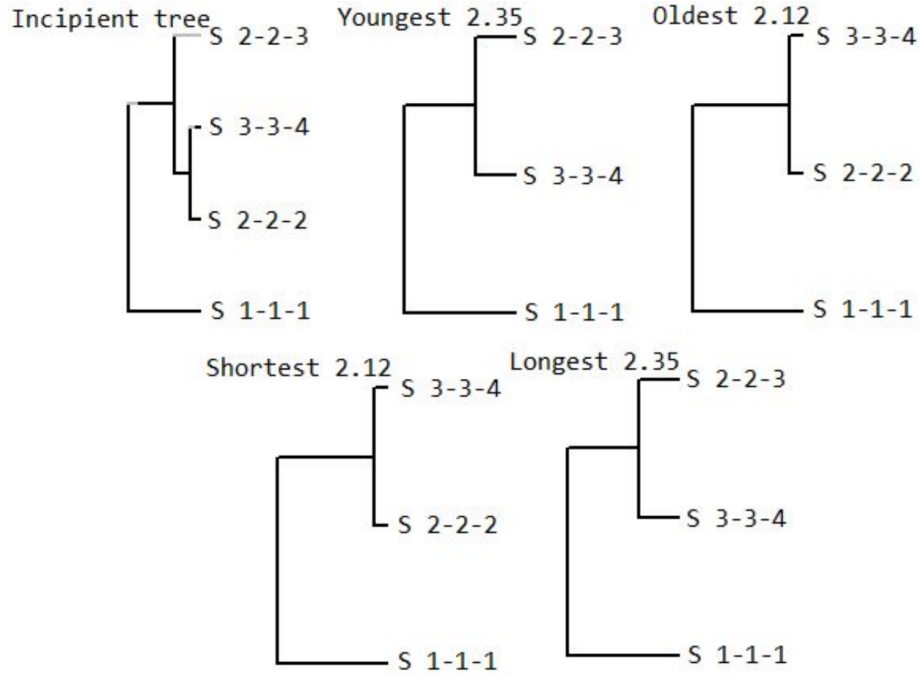


Figure 3: Sampling a species tree from an incipient species tree. At the top left, an incipient species tree is shown, of three different good species (the first and second number in the taxon label) and four different subspecies (the third number in the taxon label). The other four trees are species trees, that use a different sampling method to determine which sub-species is picked to represent a good species. These are: 'Youngest', 'Oldest', 'Shortest' and 'Longest'. With 'Youngest' the youngest sub-species is picked to represent the good species. With 'Oldest' the oldest sub-species is picked to represent the good species. 'Shortest' is the sampling method in which the sub-species are picked to assure the shortest branch lengths. 'Longest' is the sampling method in which the sub-species are picked to assure the longest branch lengths.

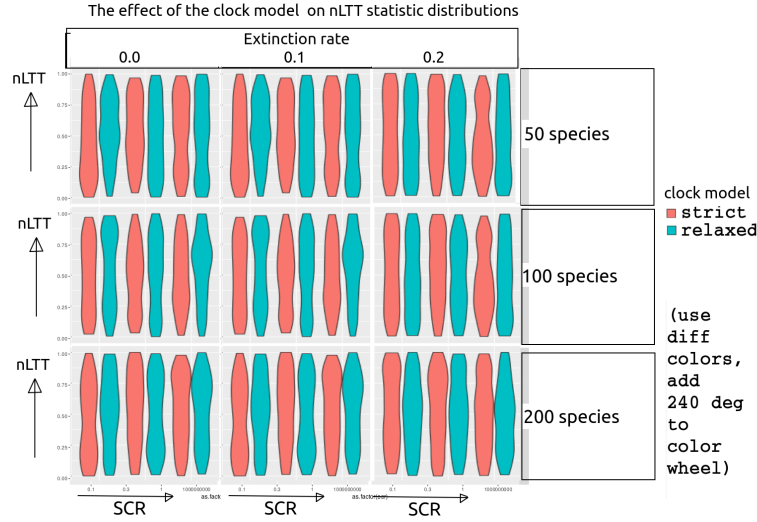


Figure 4: nLTT statistic distribution per biological parameter set per clock model, using the general data set, under the (correct) assumption of a Jukes-Cantor site model.

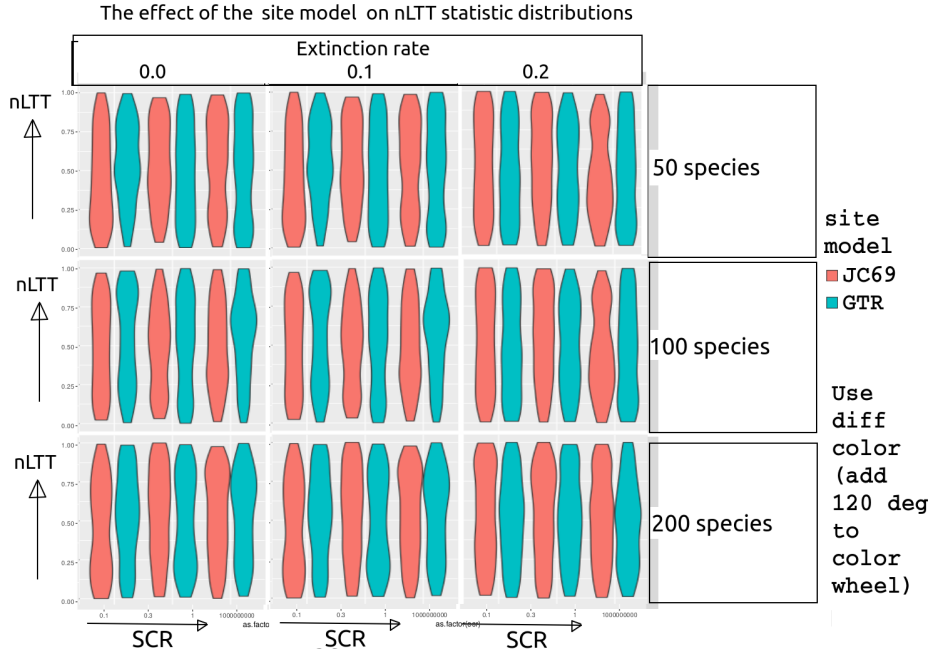


Figure 5: nLTT statistic distribution per biological parameter set per site model, using the general data set, under the (correct) assumption of a strict clock model.

	Description	Values
b_g	Speciation initiation rate of a good species	derived, see 3
b_i	Speciation initiation rate of an incipient species	derived, see 3
λ	Speciation completion rate	0.1, 0.3, 1.0, 10^9
μ_g	Extinction rate of a good species	0.0, 0.1, 0.2
μ_i	Extinction rate of an incipient species	0.0, 0.1, 0.2
n	Number of good taxa	50, 100, 200
t_c	Crown age	15
σ_c	Standard deviation around crown age	0.001
M_s	Sampling method	S, L, R
M_c	Clock model	S, RLN
M_t	Site model	JC69, GTR
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, 2, etc.
R_a	RNG seed alignment simulation	R_i
R_b	RNG seed BEAST2	R_i

Table 2: Overview of the simulation parameters. Above the horizontal line is the biological parameter set. **[RJCB: Start new]** The RNG seed R_i is 1 for the first simulation of the general data set, 2 for the next, and so on. The RNG seeds for the data set investigating the effect of sampling continue from there, but only those RNG seeds are used in which sampling has an effect. **[RJCB: End new]** The sampling methods are abbreviated as such: 'R' denotes random sampling, 'S' is 'shortest' and 'L' is 'longest'. Sampling method M_s is random for the general data set. For the data set exploring the effect of sampling, we use 'shortest' and 'longest' for each value of R_i (which are random seeds in which sampling has an effect). The clock models are abbreviated as 'S' for a strict and 'RLN' for a relaxed log-normal model. The site models are abbreviated as 'JC69' for Jukes-Cantor (Jukes *et al.* 1969) and 'GTR' for the generalized time-reversible model (Tavaré 1986).

	μ	n	λ	b
1	0	50	0.1	0.30944
2	0.1	50	0.1	0.39674
3	0.2	50	0.1	0.48667
4	0	100	0.1	0.36344
5	0.1	100	0.1	0.45283
6	0.2	100	0.1	0.54425
7	0	200	0.1	0.41669
8	0.1	200	0.1	0.50759
9	0.2	200	0.1	0.6001
10	0	50	0.3	0.25717
11	0.1	50	0.3	0.34003
12	0.2	50	0.3	0.42648
13	0	100	0.3	0.30862
14	0.1	100	0.3	0.39455
15	0.2	100	0.3	0.48328
16	0	200	0.3	0.35991
17	0.1	200	0.3	0.44804
18	0.2	200	0.3	0.53841
19	0	50	1	0.2297
20	0.1	50	1	0.30759
21	0.2	50	1	0.38984
22	0	100	1	0.2778
23	0.1	100	1	0.35961
24	0.2	100	1	0.44481
25	0	200	1	0.32617
26	0.1	200	1	0.41078
27	0.2	200	1	0.49818
28	0	50	10^9	0.21589
29	0.1	50	10^9	0.28896
30	0.2	50	10^9	0.36635
31	0	100	10^9	0.26146
32	0.1	100	10^9	0.33872
33	0.2	100	10^9	0.41945
34	0	200	10^9	0.30733
35	0.1	200	10^9	0.38768
36	0.2	200	10^9	0.47099

Table 3: The speciation parameters used. Starting from extinction rate μ ($\mu = \mu_g = \mu_i$), the expected mean number of good species n , speciation completion rate λ , the speciation initiation rate b ($b = b_g = b_i$) follows.

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

Table 4: 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 **[R.JCB: recalc]** columns and 20K rows.

#	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 5: Overview of the 11 parameters estimated by BEAST2