The error in Bayesian phylogenetic reconstruction when speciation is not instantaneous

Richèl J.C. Bilderbeek[1] and Rampal S. Etienne[1]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

October 11, 2018

Abstract

The tools for reconstructing phylogenetic relationships between taxonomic units (e.g. species) have become very advanced in the last three decades. Among the most popular tools are Bayesian approaches, such as BEAST, MrBayes and RevBayes, that use efficient tree sampling routines to create a posterior probability distribution of the phylogenetic tree. A feature of these approaches is the possibility to incorporate known or hypothesized structure of the phylogenetic tree through the tree prior. It has been shown that the effect of the prior on the posterior distribution of trees can be substantial.

Currently implemented tree priors assume that speciation is instantaneous, where we know that speciation can be a gradual process.

Here we explore the effects of ignoring the protractedness of the speciation process with an extensive simulation study.

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

Keywords: computational biology, evolution, phylogenetics, Bayesian analysis, tree prior

1 Introduction

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

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

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

Current phylogenetic tools use tree priors that assume speciation is instantaneous, 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. Furthermore, while we evidently know the clock and site models used in the simulation, using a different clock and/or site model prior in inference may compensate or increase this difference between inferred and simulated tree. To study this, we also explore the effect of a different clock and site model prior in inference.

[RJCB: Start new] The PBD model has five parameters, depicted in table 2. The per species speciation completion rates λ we use are 0.1, 0.3, 1.0 and 10^9. This means that there is a $0.15\,dt$ ($0.35\,dt$, $1.0\,dt$, $10^9\,dt$) probability of speciation completion occurring in an infinitesimal time dt. We use per species extinction rates of $\mu = \mu_g = \mu_i$ 0.0, 0.1 and 0.2. We use tree sizes of n of 50, 100 and 200 good taxa. For each combination of λ, μ and n, we use a speciation initiation rate $b = b_i = b_g$ so that the expected mean number of species $\mathbf{E}(\overline{n}; b, \lambda, \mu)$, given a b, λ and μ, equals the desired number of species n. b is calculated using the PBD R package (Etienne 2015) for each parameter combination and shown in table 3. We use $\lambda = 10^9 \approx \infty$ as our control for

which the PBD model reduces to the BD model.

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.

We create 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. There is an additional selection criterion, for the data set to investigate sampling: to generate that data set, only incipient species trees are kept, on which the two sampling methods (see below) result in different species trees. As sampling will never 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 aligments 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 a strict or relaxed log-normal clock model. Also here, the strict clock model is the correct one, as it is used for simulating the alignment, but the relaxed log-normal clock model is the one most commonly used. We set the BD model as a tree prior, as gauging the effect of this incorrect assumption is the goal of this study. We assume an MRCA prior with a tight normal distribution around the crown age, by choosing the crown age as mean, and a standard deviation of $0.5 \cdot 10^{-3}$ time units, resulting in 95% of the crown ages inferred have the same resolution (of 10^{-3} time units) as the alignment. We ran the MCMC chain to generate 1111 states, of which we remove the first 10% (also called the 'burn-in'). Of the remaining 1000 MCMC states, the effective sample size (ESS) of the posterior must at least be 200 for a strong enough inference (Drummond & Bouckaert 2015). An ESS can be increased by increasing the number of samples or decreasing the autocorrelation between samples. If the ESS is less than 200, we decrease autocorrelation by doubling the MCMC sampling interval of that simulation, until the ESS exceeds 200.

We compare each posterior phylogeny to the (sampled) species tree using the nLTT statistic (Janzen *et al.* 2015), from the `nLTT` package (Janzen 2015). The nLTT statistic equals the area between the normalized lineages-through-time-plots of two phylogenies, which has a range from zero (for identical phylogenies) to one. We use inference error and nLTT statistic interchangeably. Comparing the simulated species tree with each of the posterior species trees yields a distribution of nLTT statistics. The input trees generated with a $\lambda = 10^9$ allow us to measure the noise of the experiment. For $\lambda = \infty$, the PBD model that generates the starting trees reduces to a BD model. In the subsequent steps,

sampling will have no effect in this case, because BEAST2 will assume the correct speciation model, so the difference between inferred tree and true species tree are explained purely due to this experimental noise.

[RJCB: Note there is no twinning, as in the razzo package. Already sent an email to discuss this with RSE.]

[RJCB: Start new] As described above, per alignment, we do four different Bayian analyses, as we use two different site models and two different clock models. We know the generative site model (which is JC69) and clock model (a strict clock), but want to explore how often the correct model is indeed preferred. Per alignment, we measure the estimated marginal likelihood (which is the probability of the data given the model) for each of the four models and measure their relative propertions. To estimate the marginal likelihood, we use the novel Nested Sampling approach (Russel *et al.* 2018), which we configure to have an effective sample size (in estimating the marginal likelihood) of at least 200, by using a single-particle sub-chain with a length equal to the MCMC interval. **[RJCB: There are no rules described in the setup of the NS (for example, what should the relation between MCMC chain length and NS sub-chain length be?), but one can experiment with a sub-chain length and number of particles. I will need to do some pilot studies to verify if this is the correct way to do and describe this.]** **[RJCB: End new]**

We produce two data sets as a comma-separated file. The general data set has 144 **[RJCB: recalc]** different combinations of biological parameter combinations, site and clock models. The data set to investigate sampling has ?552 **[RJCB: recalc]** different combinations of biological parameter combinations, site models, clock models and sampling methods. The experiment is computationally intensive: pilot experiments show that the experiment takes roughly

100 days of CPU time and 20 days of wall clock time (which includes the queued waiting for computational resources) per replicate. Due to this, we choose to perform ten replicates, so that the complete experiment will take an acceptable time of roughly seven months.

[RJCB: Start new] In this article, we showcase the effect of sampling and the certainty when selecting a (site and clock) model. We show the effect sampling has on the inference error, for the nLTT distributions generated from assuming the (correct) Jukes-Cantor site model and a strict clock model. We do so per parameter combination, displaying the distribution as a violin plot. In the supplementary information, we display these error distributions of the general data set all combined, or separated per site or clock model. The certainty in the model selection is again shown per parameter setting, as a stacked bar. **[RJCB: I am unsure how to depict that stacked bar and its uncertainty. Sure, the stacked bars (without uncertainty) will work, but information about uncertainty is lost. See figure 'Showing a stacked bar with uncertainty' for three equivalent ways to display a stacked bar with uncertainty.]**

[RJCB: End new]

2 Results

3 Glossary

References

Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A. & Lemey, P. (2012) Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular biology and evolution*, **30**, 239–243.

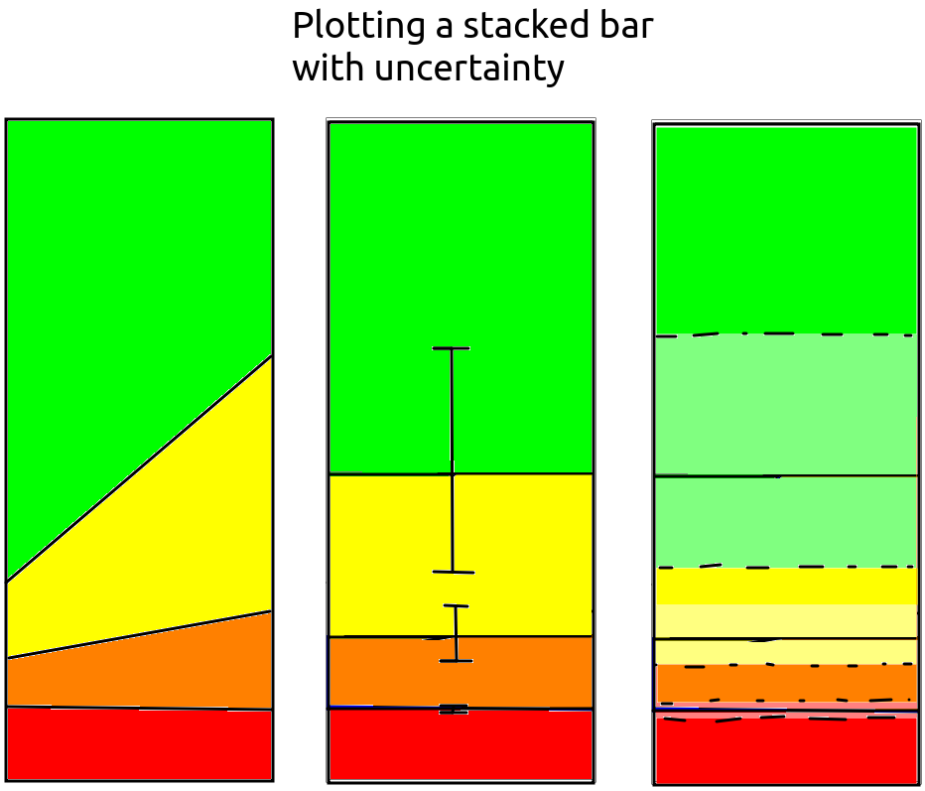

Figure 1: Showing a stacked bar with uncertainty. These three different ways display the same underlying data.

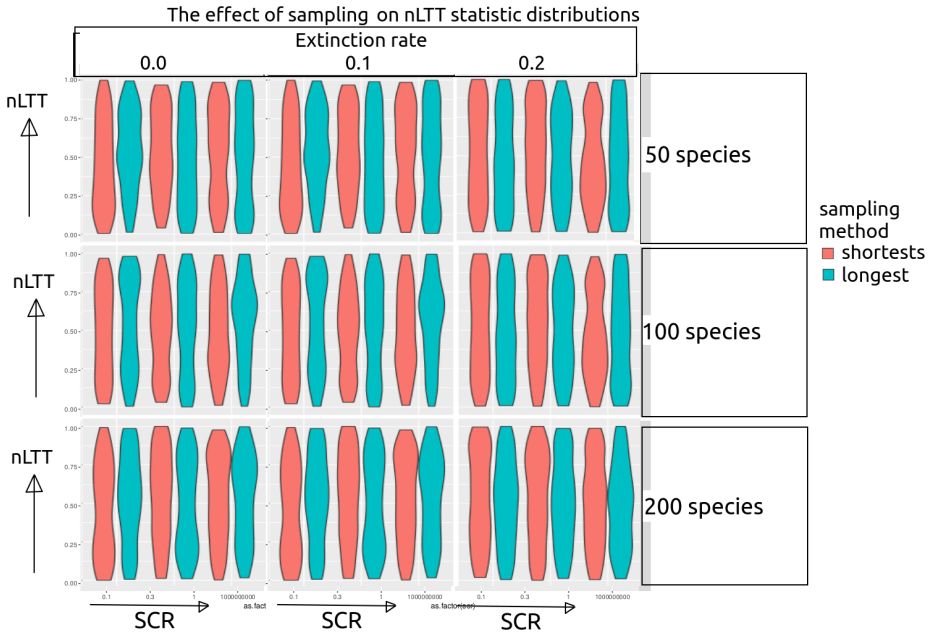

Figure 2: nLTT statistic distribution per biological parameter set per sampling regime, using the data set conditioned on sampling regime having an effect, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

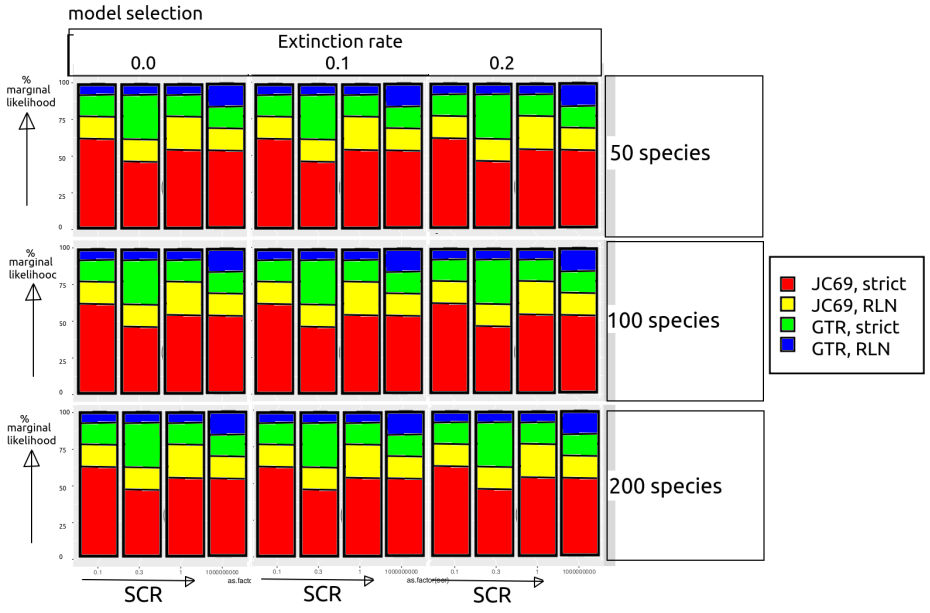

Figure 3: Model preference on the general data set.

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

Bilderbeek, R.J. & Etienne, R.S. (2018) babette: Beau ti 2, beast 2 and tracer for r. *Methods in Ecology and Evolution*.

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

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

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

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

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

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

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

Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R., 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.

Jukes, T.H., Cantor, C.R. *et al.* (1969) Evolution of protein molecules. *Mammalian protein metabolism*, **3**, 132.

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.

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

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

Schluter, D. (2009) Evidence for ecological speciation and its alternative. *Science*, **323**, 737–741.

Simonet, C., Scherrer, R., Rego-Costa, A. & Etienne, R. (2018) Robustness of the approximate likelihood of the protracted speciation model. *Journal of evolutionary biology*, **31**, 469–479.

Tavaré, S. (1986) Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, **17**, 57–86.

Yang, Z. & Rannala, B. (2005) Branch-length prior influences bayesian posterior probability of phylogeny. *Systematic Biology*, **54**, 455–470.

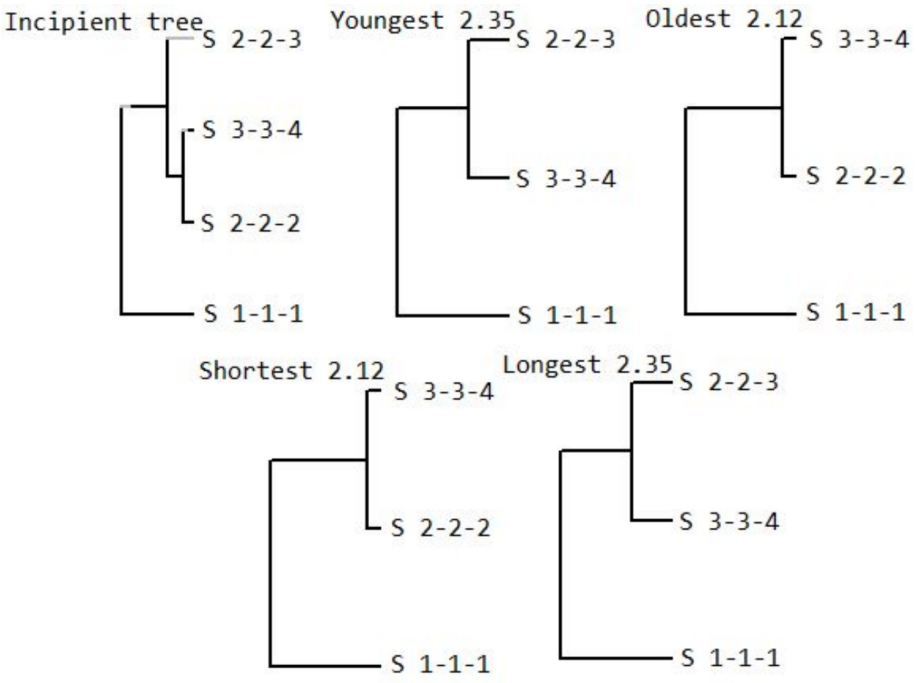

Figure 4: Sampling a species tree from an incipient species tree. At the top left, an incipient species tree is shown, of three different good species (the first and second number in the taxon label) and four different subspecies (the third number in the taxon tabel). The other four trees are species 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.

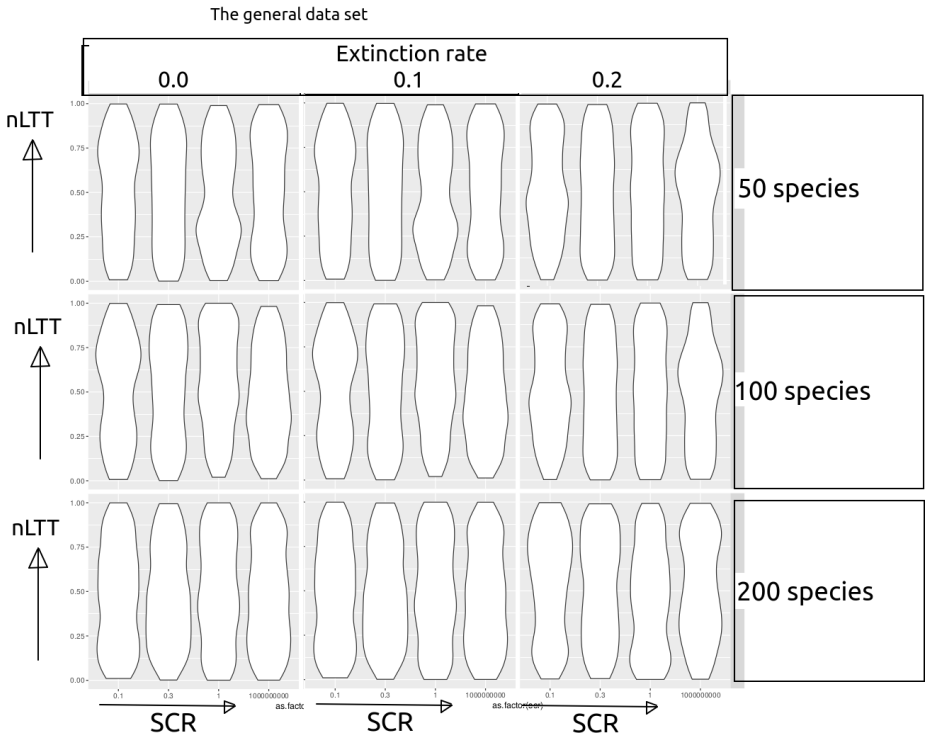

Figure 5: nLTT statistic distribution per biological parameter set, using the general data set, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

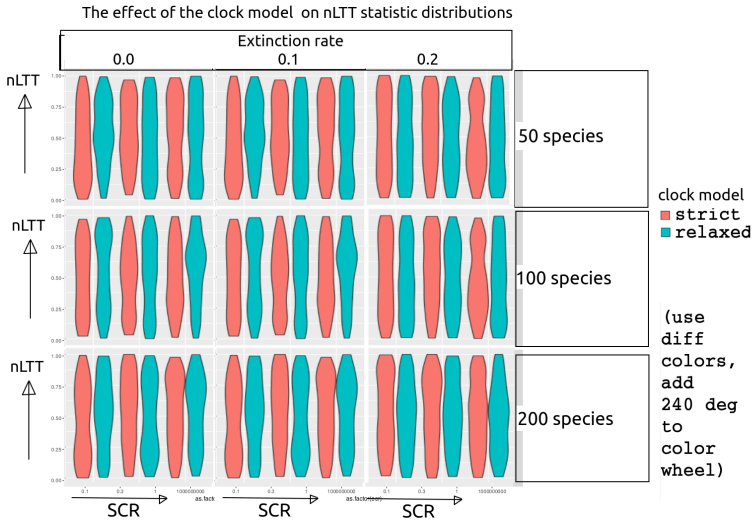

Figure 6: nLTT statistic distribution per biological parameter set per clock model, using the general data set, under the (correct) assumption of a Jukes-Cantor site model.

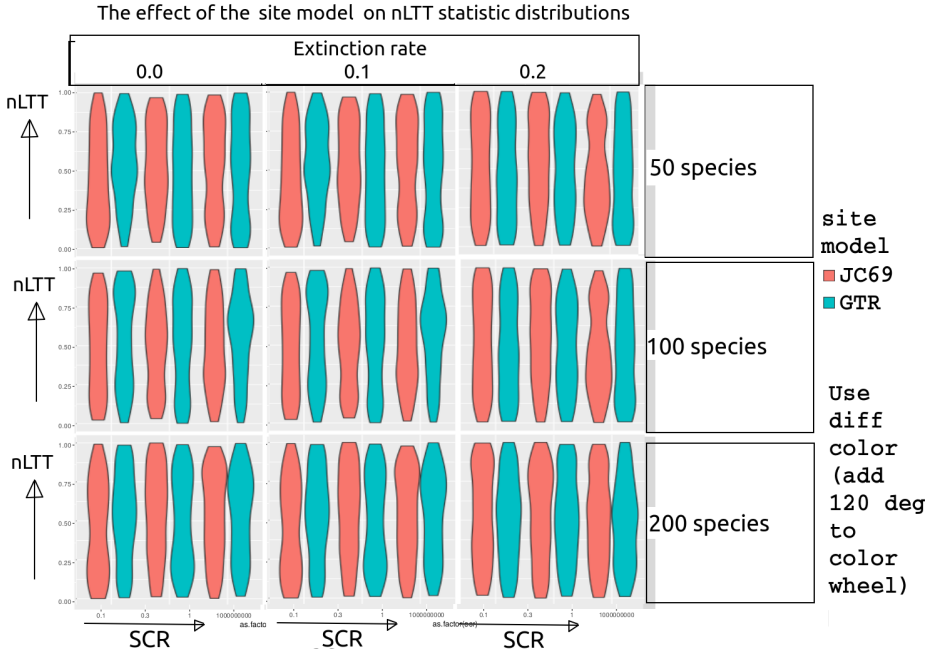

Figure 7: nLTT statistic distribution per biological parameter set per site model, using the general data set, under the (correct) assumption of a strict clock model.

	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 initation 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)
1	Marginal likelihood estimate
1	Marginal likelihood estimation uncertainty
1	Marginal likelihood ESS

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 **[RJCB: 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

A Acknowledgements

[RJCB: put this section here, as the journal does not request for this] We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

B Authors' contributions

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