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

June 6, 2018

Abstract

The tools for reconstructing phylogenetic relationships between taxonomic units (e.g. species) have become very advanced in the last three decades. **[RSE: test comment, so Rampal can see this in action]**

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, prior choice

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 successor 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. From these priors and an alignment of DNA, RNA or protein sequences, they create a posterior distribution of parameter estimates (of the models used as a prior) and phylogenies, 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 a wrong 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. To gain more insight of incorrect prior choice, we also explore the effect of two clock and site priors.

The PBD model has five biological parameters, depicted in table 2, which we explore in a factorial fashion, excluding some combinations. We simulate a PBD process for those combinations in which the 95% quantile of the expected number of good species is less than 1250. The quantile is calculated with a recently added function to the PBD package, based on equation 6 of Etienne *et al.* 2014. This calculation assumes $b = b_g = b_i$, we used $b = \max(b_g, b_i)$. 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 0.3 and 0.5 speciation initiation events per good/incipient species per time unit. The speciation completion rates we use are 0.1, 0.3, 1.0 and 10^9 speciation completion events per (incipient species) species per time unit. We use $10^9 \approx \infty$ to mimic the BD model, because the PBD model reduces to the BD model for $\lambda = \infty$. This allows us to measure the baseline error, which is the difference between inferred tree and true species tree that arises purely due to noise because the generating model and the model used in inference are identical in this case. The extinction rates of good and incipient species, μ_g and μ_i respectively, that we use are 0.0, 0.1 and 0.2 extinction events per good/incipient species per time unit.

From each biological parameter set, we simulate a protracted birth-death tree, using the PBD package (Etienne 2015) in the R programming language (R Core Team 2013), all with a crown age of 15 million years. Each protracted birth-death tree uses a different random number generatior seed, which makes all runs independent, resulting in a balanced data set. **[RJCB: Rampal assumed runs with close seeds were related. I assume I have convinced him otherwise]]**

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 (Cantor & Jukes 1969) nucleotide substitution model, in which all nucleotide-to-nucleotide transitions are equally likely. 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. 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.

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 that alignment, but the relaxed log-normal clock model is the 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 **[RJCB: I chose the ESS of the posterior over the ESS of the tree likelihood (and the others displayed in table 4). For both something can be said. Agree on this choice?]** 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 by the nLTT statistic (Janzen *et al.* 2015), using 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.

We produce two data sets as a comma-separated file. The general data set has 348 different combinations of biological parameter combinations, site and clock models. The data set to investigate sampling has 496 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 per replicate. Due to this, we choose to perform ten replicates,

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

so that the complete experiment will take an acceptable time of roughly seven months.

For both data sets, we display the nLTT statistics distribution per biological parameter combination as a violin plot. We only show combinations for which $\lambda_g = \lambda_i$ and $\mu_g = \mu_i$, to simplify the interpretation of these results. Additionally, we only show the nLTT distributions that were generated under the (correct) assumptions of a Jukes-Cantor site model and a strict clock model. We display the nLTT statistic distributions for both site and clock models in the supplementary information. We show the effect of sampling, by seperating the nLTT statistics distribution per sampling method used.

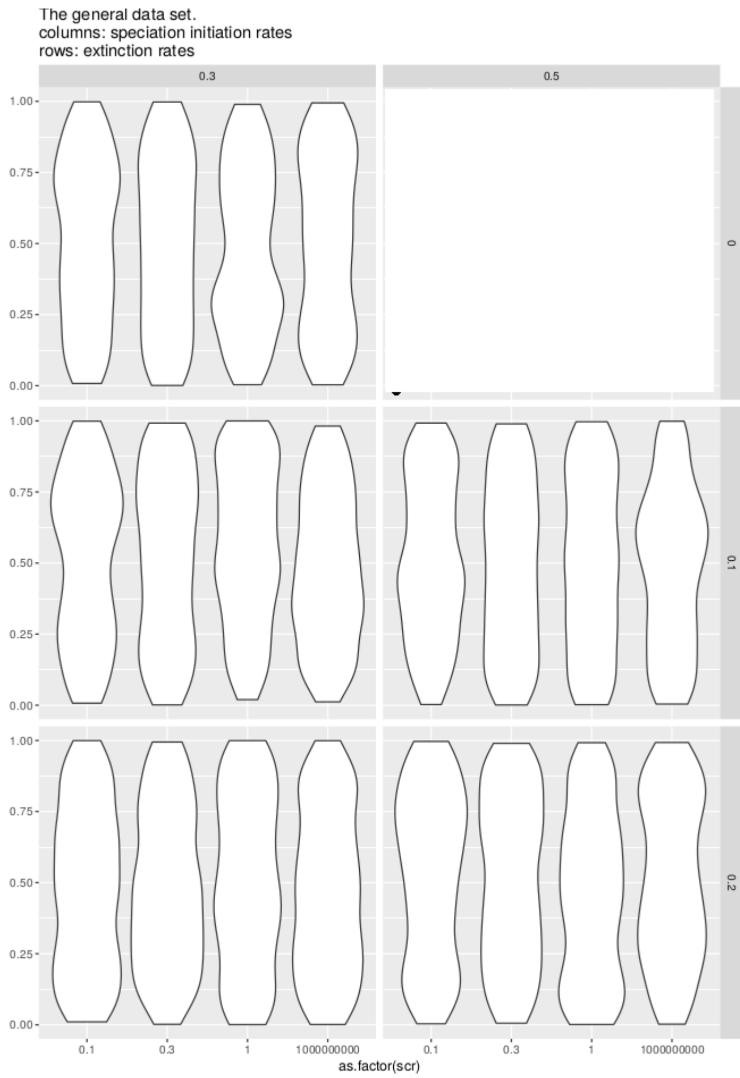

Figure 1: nLTT statistic distribution per biological parameter set, using the balanced data set, for the subset of combinations in which $\lambda_g = \lambda_i$, $\mu_g = \mu_i$, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

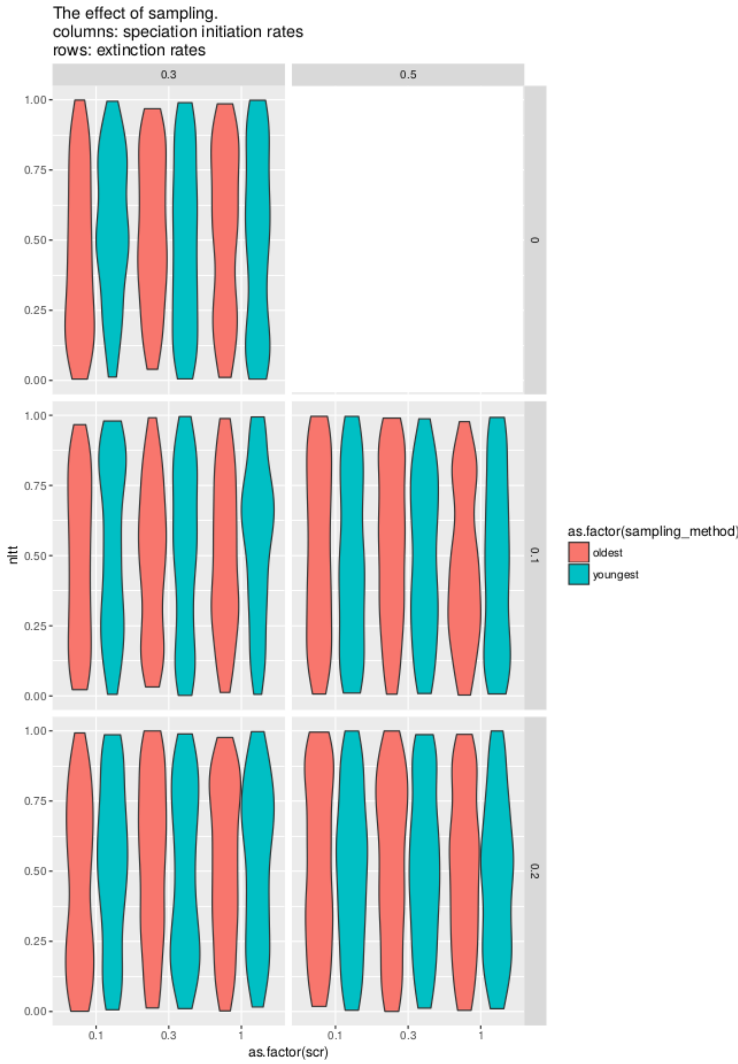

Figure 2: nLTT statistic distribution per biological parameter set per sampling regime, using the data set conditioned on sampling regime having an effect, for the subset of combinations in which $\lambda_g = \lambda_i$, $\mu_g = \mu_i$, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

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.

Bilderbeek, R.J. & Etienne, R.S. (2018) babette: Beauti 2, beast2 and tracer for r. *bioRxiv*, p. 271866.

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.

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

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

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.

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.

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.

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

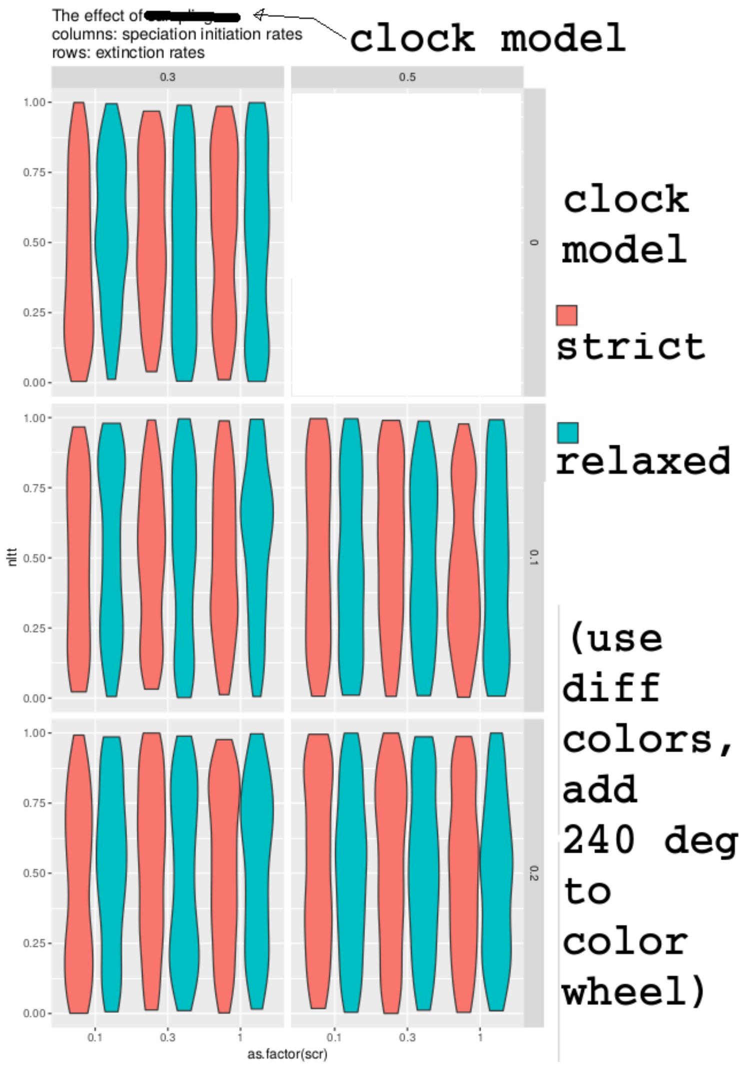

Figure 3: nLTT statistic distribution per biological parameter set per clock model, using the balanced data set, for the subset of combinations in which $\lambda_g = \lambda_i$, $\mu_g = \mu_i$, under the (correct) assumption of a Jukes-Cantor site model.

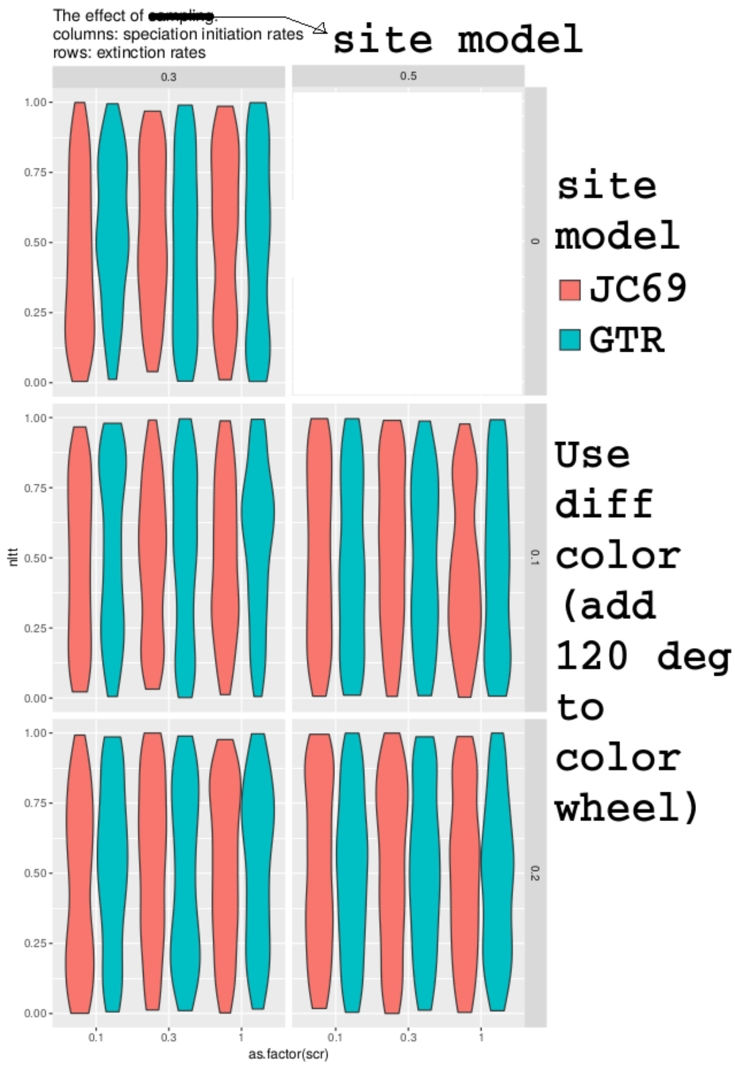

Figure 4: nLTT statistic distribution per biological parameter set per site model, using the balanced data set, for the subset of combinations in which $\lambda_g = \lambda_i$, $\mu_g = \mu_i$, under the (correct) assumption of a strict clock model.

	Description	Values
b_g	Speciation initiation rate of a good species	0.3, 0.5
b_i	Speciation initiation rate of an incipient species	0.3, 0.5
λ	Speciation completion rate	0.1, 0.3, 1.0, ∞
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, 0.4
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 to 20K
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. The RNG seed R_i is 1 for the first simulation, 2 for the next, etc. 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' for odd values of R_i, and 'longest' for even values of R_i. The clock models are abbreviated as 'S' is a strict and 'RLN' is a relaxed log-normal model. The site models are abbreviated as 'JC69' for Jukes-Cantor and 'GTR' for the generalized time-reversible model.

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.

#	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